The emerging landscape of Social Media Data Collection: anticipating trends and addressing future challenges

Laura Sáez Ortuño

2023

PhD in Business | Laura Sáez Ortuño

UNIVERSITAT DE
BARCELONA

PhD in Business

The emerging landscape of Social Media Data Collection: anticipating trends and addressing future challenges.

Laura Sáez Ortuño

PhD in Business

Thesis title:

The emerging landscape of Social Media Data Collection: anticipating trends and addressing future challenges.

PhD student:

Laura Sáez Ortuño

Advisors:

Santiago Forgas Coll
Rubén Huertas García

Date:

May 2023

LIBERTAS PERFVNDET · OMNIA LVCE
UNIVERSITAT DE BARCELONA

To my family, in the broadest sense that the term can have.

It takes a village to raise a child.

African proverb

There is a driving force more powerful than steam, electricity and nuclear power: the will.

Albert Einstein

ACKNOWLEDGEMENTS

It is often said that a true dream should never be forgotten, and for me, as a scientist, pursuing a PhD has always been my dream. After several years away from academia, during which I focused more intensively on the business world and raising my children, I returned to academia a few years ago. I started by teaching classes at the Faculty of Computer Science and Mathematics at UB, which I continue to enjoy, and later embarked on the difficult but passionate journey of pursuing a PhD. After facing a first rejection and working on it for several years, in January 2023, my first article was published in the prestigious Journal of Business Research with an Impact Factor of 11.0. In February 2023, after facing another rejection on an article based on many years of working on a research project, my second article was published in the world well recognized Journal of Innovation & Knowledge, Impact Factor of 11.2.

It hasn't been an easy journey. There were moments when I thought I couldn't handle it all and wanted to give up. But fortunately, I had a huge team around me that helped me not to falter. So, my first gratitude is for them: my two children, Pol and Aina, for being understanding during my busy times working on my PhD; my husband, Xavier Carreras, for being my partner at home and in business since I was 19; my mum, Pilar Ortuño that has been always a referent for me and my parent, Santiago Sáez, especially for his first computing lessons on my early age; my in-laws, Lola Meix and Ramón Carreras for creating a perfect environment in Ivorra for my research; and my brothers, Hector Sáez and Alex Espinosa for their enthusiast behaviour. Without their support, it would not have been possible. I feel fortunate to have the best family that I have ever dreamed of.

I also want to express my gratitude to CoRegistros, S.L.U., and all its team, for generously sharing the data and allowing me to have a flexible schedule.

I would like to express my deepest gratitude to Dra. Paloma Miravitlles, Dr. Jordi Roca, Dr. Eloi Puertas, Dr. Jaume Argerich and Dr. Martí Sagarra for their unwavering support since the beginning of this thesis.

I am also immensely grateful to the members of Universitat de Barcelona who have provided their valuable support at different stages of this arduous process. Special thanks go to Dra. Fariza Achcaoucaou, Dra. Ana Nuñez Prats, Dr. Karim Lekadir, Dr. Sergio Escalera, Dr. Francisco Tarragó, Dr. Jaume Valls, Dra. Mercè Bernardo, Dra. Mel Solé, Dra. Esther Hormiga, Dra. Ana Lauroba, Dr. Oriol Pujol, Dr. Ramón Alemany, Dra. Laura Guitart, Dra. Mercè Mach, Dr. Eugeni Graugés, and Dr. Claudio Cruz, as well as all members of Universitat de Barcelona's PAS and PDI, particularly those affiliated with the Universitat de

Barcelona Business School.

I want to express my gratitude to Dra. Marta Riera, Dra. Sònia Fernández-Vidal, Dr. Ignasi Salaet, Dr. Fèlix Casanova, Jordi Puig, Eric Gracia, and Juan Busquets for their support and fore giving me different points of view of the research phenomenon.

My heartfelt thanks go to my best friends, Anna Figueras and Salomé Nogués, for their life support and for always being there for me in the best and not so best moments.

I also want to thank my PhD colleagues for their support and for always accompanying me: Sergi Pons and Irene Sánchez. Anna Roca, Aïda Bravo, Yanqu Zheng, and Feng Xu also thank you so much. I would like to give a special mention to Roger Font, without whose help in the first stage, this would have never become a reality. Thank you to all of you for training with me and helping me every time.

I extend my sincere thanks to the reviewers and editors of the journals for their constructive comments and timely review of my articles.

And above all, I am extremely grateful to my advisors, Dr. Ruben Huertas García and Dr. Javier Sánchez García, and especially to my director and tutor, Dr. Santiago Forgas Coll, for working tirelessly on all the articles that we are honoured to co-author. Santiago has been the best thesis director that I could have ever imagined, both professionally and personally. I am convinced that we will be researching together all our life. I could not have dreamed of a better team to support me through this long and challenging journey that just has started. From the bottom of my heart, thank you for everything!

I extend my heartfelt gratitude to all the rest of family and friends and to everyone who has contributed to making this girl's dream of becoming an academic researcher a today reality.

Thank you all!

SUMMARY

Social media has become a powerful tool for creating and sharing user-generated content across the internet. The widespread use of social media has led to a massive amount of information being generated, presenting a vast opportunity for digital marketing. Through social media, businesses can reach millions of potential consumers and capture valuable consumer data, which can be used to optimise marketing strategies and actions.

The potential benefits and challenges of using social media for digital marketing are also growing in interest among the academic community. While social media offers businesses the opportunity to reach a vast audience and gather valuable consumer data, the volume of information generated can lead to unfocused marketing and negative consequences such as social overload. To make the most of social media marketing, companies need to collect reliable data for specific purposes such as selling products, raising brand awareness, or fostering engagement and for predicting future consumer behaviours. The availability of quality data can help build brand loyalty, but consumers' willingness to share information depends on their level of trust in the company or brand requesting it.

Therefore, this thesis aims to contribute to the research gap through bibliometric analysis of the field, mixed analysis of profiles and motivations of users who provide their data on social media, and a comparison of supervised and unsupervised algorithms for clustering consumers. This research has used a database of more than 5.5 million data collections over a period of 10 years.

Advancements in technology now allow for sophisticated analysis and reliable predictions to be made based on the captured data, which is particularly useful for digital marketing. Several studies have explored digital marketing via social media, with some focusing on a specific field, while others adopt a more multidisciplinary approach. However, due to the rapidly evolving nature of the discipline, a bibliometric approach is required to capture and synthesise the most up-to-date information and add more value to studies in the field.

Thus, the contributions of this thesis are as follows. Firstly, it provides a comprehensive review of the literature on the methods for collecting personal data of consumers from social networks for digital marketing and establishes the most relevant trends through analysis of significant articles, keywords, authors, institutions, and countries. Secondly, this thesis identifies user profiles that lie the most and why. More specifically, this research demonstrates that some user profiles are more inclined to make mistakes, while others intentionally provide false information. The study also shows that the main motivations behind providing false information include amusement and a lack of trust in data privacy and security measures. Finally, this thesis aims to fill the gap in the literature on which algorithm, supervised or non-supervised, can cluster consumers better, who provide their data on social media to predict their future behaviour.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

1.1. The data collection in social media within the marketing discipline

The use of social media in marketing has grown significantly in recent years. Companies and researchers have recognized the potential of social media data in understanding consumer behavior, preferences, and sentiments. Social media platforms such as Facebook, Twitter, Instagram and others provide a vast amount of data that can be analyzed to gain insights into consumer behavior. Social media has become an increasingly popular source of data for businesses, researchers, and individuals (Gruzd *et al.,* 2011). According to Gruzd *et al.* (2011), social media data can provide valuable insights for understanding consumer behavior, conducting market research, and analyzing sentiment. However, to effectively collect data from social media, there are several key steps that should be followed.

First, it is important to identify the social media platforms that are most relevant to the research goals. Different platforms have different audiences and types of content, so it is important to select those that are most appropriate for the study (De Choudhury *et al.,* 2013). For example, Twitter might be useful for analyzing public opinion about a particular topic, while Instagram might be useful for analyzing user-generated content related to a particular brand (Roma & Aloini, 2019).

Once the relevant social media platforms have been identified, the researcher should decide on the type of data to collect. This can include text data, such as tweets or comments, or visual data, such as images or videos. Additionally, researchers should consider whether they want to collect data in real-time or if they are interested in analyzing historical data (Kramer & Guillory, 2016).

To collect data from social media, a variety of tools and techniques can be used, ranging from web scraping tools to APIs and specialized software (Zhu *et al.*, 2014). However, it is important to keep in mind that collecting data from social media comes with ethical considerations. Researchers should respect user privacy and ensure that they are only collecting data that is relevant to their research goals (Boyd & Crawford, 2012). Additionally, researchers should be transparent about their data collection methods and how the data will be used (Bruns & Burgess, 2011).

Thus, social media data can provide valuable insights for a wide range of applications. By following best practices and ethical considerations, researchers can collect and analyze this data in a way that is both effective and responsible (Epstein & Buhovac, 2014).

1.2. Social Media Data Collection

In recent years, the use of social media platforms has become increasingly popular in digital marketing. Companies are using social media data to understand their customers better and to create more targeted marketing campaigns. However, data collection in social media presents unique challenges (Bala & Verma., 2018). This thesis will explore the methods used for collecting social media data in digital marketing, the advantages and disadvantages of these methods, and how companies can utilize social media data to enhance their marketing efforts.

There are several methods used for collecting social media data in digital marketing. One is web scraping, where data is collected from websites through automated means (Diouf *et al.,* 2019). Another method is social media monitoring, where companies use software to monitor social media platforms for mentions of their brand or products (Zhang & Gupta, 2018). Furthermore, companies can use surveys, polls, and sweepstakes to collect data on social media platforms (Boer *et al.*, 2021).

Web scraping is an effective method for collecting large amounts of data quickly. However, it can be challenging to ensure the accuracy of the data collected, and it may not be legal to scrape certain websites. Additionally, social media platforms may block web scraping efforts, making it difficult to collect data. Social media monitoring is a useful method for collecting data on customer sentiment towards a brand or product (Zhang & Gupta, 2018). This data can be used to improve marketing campaigns and customer service efforts. However, social media monitoring can be time-consuming, and it may be challenging to analyze a large amount of data collected. Surveys, polls, and sweepstakes on social media platforms are an effective method for collecting data on customer preferences and opinions (Boer *et al.*, 2021). These methods allow companies to collect data directly from their customers, providing valuable insights into their needs and preferences. However, surveys, polls, and sweepstakes may not provide a representative sample of the customer base, and respondents may not answer honestly (Sue & Ritter, 2012). Table I shows the advantages and disadvantages of the different methods.

Despite the challenges of collecting social media data, companies can still utilize this data to enhance their marketing efforts. By analyzing social media data, companies can gain insights into customer behavior and preferences, allowing them to create more targeted marketing campaigns (Batrinca & Treleaven, 2015). For example, social media data can be used to identify trending topics and hashtags, which can be incorporated into marketing campaigns to increase visibility and engagement. Additionally, social media data can be used to identify key influencers in a particular industry, allowing companies to partner with them to reach a wider audience (Boyd & Ellison, 2007).

Table 1. Advantages and Disadvantages of Social Media Data Collection Methods

Data Collection Method	Advantages	Disadvantages
Web Scraping	Quick	Accuracy Legality Blocked
Social Media Monitoring	Customer sentiment insights	Time-consuming Data analysis
Surveys, Polls and Sweepstakes	Direct customer insights	Representative sample Honesty

Source*:* Own elaboration adapted from Nadaraja, R., & Yazdanifard, R. (2013)

According to Batrinca & Treleaven (2015), social media data, once is collected, can be used by companies to enhance their marketing efforts through the creation of more targeted marketing campaigns, improvement of customer service, development of new products or services, and identification of influencers and partnerships with them. This is supported by a study conducted by De Vries & Carlson (2014), which found that companies that use social media analytics are more likely to achieve higher levels of customer engagement and loyalty. Moreover, social media data can also be used to improve customer service. By monitoring social media conversations and identifying customer complaints or concerns, companies can respond quickly and effectively to resolve issues, improving overall customer satisfaction (Hanna *et al*., 2011). In addition, social media data can aid in the development of new products or services. By analyzing customer feedback and preferences, companies can gain insights into what their target audience wants and needs, allowing them to develop products or services that are more likely to be successful in the marketplace (Chen *et al*., 2011). Social media data can also be used to identify influencers and establish partnerships with them. By analyzing social media activity and engagement levels, companies can identify individuals who have a significant impact on their target audience and develop partnerships with them to promote their products or services (Godes & Mayzlin, 2004) (see table 2).

Table 2. Social Media Data Utilization Process in digital marketing

Collect Social Media Data	This step involves collecting data from social media platforms using various methods such as web scraping, social media monitoring, surveys, and polls.
Analyze Social Media Data	In this step, the collected data is analyzed using various tools and techniques such as sentiment analysis, text mining, and data visualization.
Identify Insights	Based on the analysis of social media data, insights are identified that can be used to inform marketing strategies and tactics.
Develop Marketing Campaigns	The insights obtained from social media data are used to develop targeted marketing campaigns that resonate with the target audience.
Improve Customer Service	Social media data can also be used to improve customer service by identifying customer needs, preferences, and pain points.
Develop New Products or Services	Social media data can provide valuable insights into customer needs and preferences, which can be used to develop new products or services.
Identify Influencers	Social media data can also be used to identify influencers who can help promote the brand and build partnerships with them.

Source: Own elaboration adapted from Sponder (2018)

1.2.1. Data Collection

When collecting social media data, it is essential to consider several important factors. First, it is crucial to select the most relevant social media platforms based on the research goals (Batrinca & Treleaven, 2015). For instance, Twitter may be more suitable for understanding customer sentiments about a specific brand or product, while Instagram may be more useful for gaining insights into user-generated content related to lifestyle or fashion (Arora *et al*., 2019). Second, researchers and companies should decide whether to collect data in real-time or to analyze historical data. According to Stieglitz *et al*. (2018), real-time data collection can provide valuable information on current trends and events, whereas historical data analysis may be more suitable for identifying long-term patterns and trends. Third, various tools can be used for collecting social media data. Web scraping tools, APIs, and specialized software are among the most common tools used for this purpose (Lomborg & Bechmann, 2014). Web scraping tools can be used to automatically collect data from social media platforms, while APIs provide access to real-time data streams. Specialized software can store and analyze data more efficiently. Finally, it is essential to consider ethical implications when collecting social media data. Researchers and companies must respect user privacy and be transparent about their data collection methods and use (Liu *et al*., 2021). Informed consent must be obtained from individuals whose data is being collected, and data must be stored and used in a secure and ethical manner.

Thus, collecting social media data can be an invaluable tool for gaining insights into customer behavior and preferences. However, researchers and companies must carefully consider relevant social media platforms, the type and timing of data collection, tools for data collection, and ethical considerations to ensure that data collection is carried out responsibly and ethically (Stieglitz *et al.*, 2018). (See table 3).

Table 3. Considerations for Collecting Social Media Data

Consideration	Description
Relevant Social Media Platforms	Choose the social media platforms that are most relevant to the research goals.
Type and Timing of Data Collection	Determine the type of data to be collected and whether to collect data in real-time or to analyze historical data.

Tools for Data Collection	Use web scraping tools, APIs, and specialized software to collect and store data.
Ethical Considerations	Respect user privacy and be transparent about data collection methods and use.

Source: Own elaboration adapted from Appel *et al.,* (2020)

1.2.2. Cluster consumers

Cluster analysis has become a popular technique among marketers for segmenting consumers based on shared characteristics and behaviors (Jain *et al.*,1999). Social media platforms have provided marketers with an opportunity to collect and analyze data that can help identify clusters of consumers who have similar interests, preferences, and buying habits (Kaplan & Haenlein, 2010). Boyd and Crawford (2012) suggested that social media data can include user profiles, posts, comments, likes, and other interactions. Machine learning algorithms can then be applied to the data to identify patterns and group users into clusters based on similarities in their behavior (Hastie *et al.*, 2009) (See Table 4).

Punj and Stewart (1983) highlighted the potential use of cluster analysis for identifying groups of consumers interested in a particular product or service. By targeting marketing efforts specifically towards these groups, marketers can tailor their messaging and advertising to appeal to their interests and preferences (Smith, 1956). Cluster analysis can also be used to identify consumer trends and predict future behavior (Gandomi & Haider, 2015). By analyzing social media data, marketers can gain insights into emerging trends and identify opportunities to innovate and improve their products or services (Brynjolfsson & McAfee, 2014).

However, it is important to acknowledge that social media data may not be representative of the entire population (Tufekci, 2014). Moreover, ethical considerations such as user privacy and informed consent must be considered when collecting and analyzing data from social media (Zimmer, 2010). Despite these limitations, cluster analysis can be a powerful tool for marketers to identify and target specific groups of consumers based on their shared characteristics and behaviors (Everitt, 1979). By collecting and analyzing data from social media platforms, marketers can gain valuable insights into consumer preferences and trends, helping them to make informed business decisions and drive growth (Van Dijck, 2013).

Table 4. Success factors for customer clustering in digital marketing

Success Factor	Description	Authors (Year)
Customer Segmentation	Dividing customers into groups based on shared characteristics, allowing for more targeted marketing efforts.	Smith (1956)
Personalization	Tailoring marketing messages and experiences to individual customers based on their preferences and behaviors.	Peppers and Rogers (1993)
Customer Lifetime Value (CLV)	Estimating the total value a customer brings to a company over the course of their relationship to prioritize high-value customer segments.	Reinartz and Kumar (2000)
Data-driven Marketing	Utilizing data to inform marketing decisions, enabling more effective targeting of customers and optimization of marketing strategies.	Kumar *et al.* (2010)
Customer Feedback and Sentiment Analysis	Gathering and analyzing customer feedback and sentiment to determine customer needs and preferences, which can inform customer clustering and marketing strategies.	Cambria and White (2014)
Omni-channel Marketing	Integrating multiple marketing channels to provide a seamless and consistent customer experience, which can help in better understanding and clustering customers.	Verhoef *et al.* (2015)
Customer Journey Mapping	Visualizing the steps customers go through when engaging with a company, which can help in identifying key touchpoints and opportunities for targeted marketing.	Lemon and Verhoef (2016)
Social Media Engagement	Leveraging social media channels to interact with customers and foster brand loyalty, which can enhance customer segmentation and targeting.	Felix *et al.* (2017)

Behavioral Analytics	Analyzing customer behaviors to identify patterns and trends, which can help in refining customer segments and improving marketing strategies.	Homburg *et al.* (2017)
Machine Learning Algorithms	Implementing advanced algorithms and statistical techniques to analyze customer data and discover hidden patterns, contributing to more effective customer clustering.	Bughin *et al.* (2018)

Source: Own elaboration.

1.2.3. Ethical Considerations

The collection of data from social media also requires ethical considerations. It is important to respect user privacy and only collect data that is relevant to the research goals. Additionally, researchers must be transparent about their data collection methods and how the data will be used. When collecting data from social media, researchers must consider the ethical implications of their actions. While social media can provide valuable insights into consumer behavior and preferences, it is important to respect the privacy of the individuals whose data is being collected (Golder *et al.*, 2017).

One key consideration is to only collect data that is relevant to the research goals (Ess & Jones., 2004). Researchers should avoid collecting any unnecessary or sensitive information that could potentially harm users or violate their privacy. For example, collecting location data or personal messages without explicit consent could be seen as invasive and unethical (Wilson *et al*., 2012).

Transparency is also crucial in data collection from social media Researchers should clearly communicate their data collection methods and explain how the data will be used. This can help to build trust with users and ensure that their data is being used responsibly (Zimmer, 2010).

Informed consent is another important aspect of ethical data collection (British Psychological Society, 2009). Researchers should obtain explicit consent from users before collecting any data. This can involve providing clear information about what data will be collected (Kaufman & Rousseeuw, 2009), how it will be used, and any potential risks or benefits associated with participating in the study (Buchanan & Hvizdak, 2009).

The collection of data from social media requires careful consideration of ethical issues (Cronin *et al.,* 2021). By respecting user privacy, being transparent about data collection methods, and obtaining informed consent (See Table 5), researchers can collect data in a responsible and ethical manner.

Table 5. Ethical Principles for Social Media Data Collection

Principle	Description
Relevance	Collect only data that is relevant to the research goals and avoid collecting unnecessary or sensitive information.
Transparency	Clearly communicate data collection methods and explain how the data will be used to build trust with users.
Informed Consent	Obtain explicit consent from users before collecting any data, providing clear information about data collection and usage.
Privacy	Respect user privacy by protecting their personal information and ensuring data is stored and processed securely.
Anonymization	Anonymize data to remove any personally identifiable information (PII) and maintain user anonymity.

Source: Own elaboration adapted from Cronin *et al*., (2021)

1.3. Data collection on Social Media impact

Data collection is the process of gathering information from various sources, including social media platforms, to make informed decisions (Stieglitz *et al.,* 2014). According to Kumar *et al.,* (2016), data collection on social media impact involves tracking social media metrics, such as likes, shares, comments, and followers. These metrics provide insights into how social media is impacting a business's marketing efforts. For instance, if a business notices that its social media followers are declining, it may indicate that its content is not fitting with its target audience (Zhu & Chen, 2015).

Another way of collecting data on social media impact is through social listening. As explained by Stewart & Arnold, (2018), social listening involves monitoring social media platforms for mentions of a brand, product, or service. This helps businesses understand how their target audience perceives their brand and identify areas where they need to improve to enhance customer satisfaction. Data collection on social media impact also involves analyzing social media trends. According to Hootsuite (2021), monitoring social media trends can help businesses identify emerging topics, hashtags, and conversations that they can leverage to enhance brand visibility.

As mentioned by Bullas (2021), some of the critical social media metrics to track include engagement rate, reach, impressions, and click-through rate. These metrics help businesses measure the effectiveness of their social media campaigns and identify areas where they need to optimize their content. In fact, data collection on social media impact is essential for businesses to make informed decisions and optimize their marketing efforts. Businesses can collect data on social media impact through tracking social media metrics, social listening, analyzing social media trends, and identifying the right social media metrics to track. By doing so, businesses can improve their brand visibility, customer engagement, and overall marketing effectiveness (Huang & Sarigöllü, 2012) (see table 6).

Table 6. Relevant Authors and sources on Data Collection on Social Media Impact in Digital Marketing

Author(s)	**Publication**	**Key Points**
Kumar *et al.*, (2016)	Journal of Marketing	Data collection on social media impact involves tracking social media metrics such as likes, shares, comments, and followers
Stewart & Arnold, (2018)	International Journal of Listening	Social listening helps businesses understand how their target audience perceives their brand
Hootsuite (2021)	Hootsuite Blog	Monitoring social media trends helps businesses identify emerging topics, hashtags, and conversations
Bullas (2021)	JeffBullas.com	Critical social media metrics to track include engagement rate, reach, impressions, and click-through rate

Source: Own elaboration.

On the same hand, collecting data on social media impact is a complex process. Researchers must navigate numerous challenges, such as privacy concerns, biases in the data, and the dynamic nature of online interactions. A variety of methods have been employed to collect and analyze social media data, including surveys, interviews, content analysis, and computational methods (Olteanu *et al.*, 2019). Table 7 provides an overview of some key studies in this field, along with the methods and data sources used by the authors.

Table 7. Methods and data sources to collect and analyze social media data.

Author(s)	***Title***	***Method***	***Data Source***	***Key Findings***
Kross et al. (2013)	Facebook Use Predicts Declines in Subjective Well-Being in Young Adults	Longitudinal study	Survey	Increased Facebook use was associated with declines in subjective well-being among young adults.
Friggeri *et al*. (2014)	Rumor Cascades	Data analysis	Facebook data	Rumors on Facebook exhibit a "cascade" effect, spreading rapidly and reaching large audiences.
Allcott and Gentzkow (2017)	Social Media and Fake News in the 2016 Election	Observational study	Survey, web browsing data	Social media played a significant role in the spread of fake news during the 2016 U.S. presidential election.
Primack *et al*. (2017)	Social media use and perceived social isolation among young adults in the US	Cross-sectional study	Survey	High levels of social media use were associated with increased perceived social isolation among young adults.
Vosoughi *et al.* (2018)	The Spread of True and False News Online	Computational study	Twitter data	False news stories were 70% more likely to be retweeted than true stories, and false stories spread faster on Twitter.

Pennycook and Rand (2019)	Fighting Misinformation on Social Media Using Crowdsourced Judgments: A Scalability Experiment	Experimental study	Survey, online platform	Using crowdsourced judgments to identify misinformation on social media is a feasible and scalable approach to reduce the spread of false information.
Orben and Przybylski (2019)	The Association Between Adolescent Well-Being and Digital Technology Use: Re-examining the Evidence	Longitudinal study	Survey, data analysis	No consistent evidence was found for a strong link between digital technology use and adolescent well-being, suggesting a more nuanced relationship than previously assumed.
Paul et de Hart. (2020)	Social Media Use, Political Participation, and Civic Engagement in Election 2016	Meta-analysis	Peer-reviewed studies	Social media has a modest positive effect on political participation and civic engagement, but the relationship varies depending on user characteristics and the type of platform.
Bail *et al.* (2022)	Exposure to Opposing Views on Social Media Can Increase Political Polarization: Results from a Large Field Experiment	Field experiment	Twitter	Exposing users to opposing political views on social media can actually increase political polarization, challenging the idea that exposure to diverse perspectives leads to more moderate opinions.

Source: Own elaboration.

These studies highlight the diverse approaches to collecting and analyzing data on social media impact. Kross *et al.* (2013) used a longitudinal design to investigate the relationship between Facebook use and well-being, employing self-report surveys to gather data. In contrast, Vosoughi *et al.* (2018) employed computational methods to analyze the spread of true and false news on Twitter, providing insights into the dynamics of information dissemination online. Allcott and Gentzkow (2017) combined survey data with web browsing data to investigate the role of social media in the spread of fake news during the 2016 U.S. presidential election. Their study underscores the importance of using multiple data sources to gain a comprehensive understanding of social media's impact. Primack *et al.* (2017) utilized a cross-sectional design to examine the association between social media use and perceived social isolation in young adults. Their findings support the growing body of evidence suggesting that excessive social media use may have negative effects on mental health. Friggeri *et al.* (2014) analyzed Facebook data to explore the phenomenon of rumor cascades, shedding light on how rumors can spread rapidly and reach large audiences on social media platforms. Pennycook and Rand (2019) conducted an experimental study using an online platform to test the feasibility and scalability of using crowdsourced judgments to identify misinformation on social media. The study did not analyze social media data but instead relied on crowdsourced judgments from participants. Orben and Przybylski (2019) collected survey data to examine the association between adolescent well-being and digital technology use. The study did not analyze social media data but instead relied on self-reported data from the participants. Paul & de Hart (2020) conducted a meta-analysis of peer-reviewed studies to investigate the effects of social media on political participation and civic engagement. The study analyzed data from previous studies that had collected social media data. Finally, Bail *et al.* (2022) conducted a field experiment on Twitter to investigate the effects of exposure to opposing political views on social media on political polarization. The study analyzed social media data from Twitter to test their hypothesis.

In summary, data collection on social media impact is an essential aspect of understanding how these platforms shape our lives. Researchers utilize a range of methods and data sources to study various aspects of social media, from its effects on consumer behaviour to its role in the spread of information. As social media continues to evolve, it is critical for researchers to adapt their data collection strategies to capture the dynamic nature of online interactions and their potential consequences for individuals and society (Hanelt *et al*.,2021).

1.4. Research overview

This section presents the objectives of the thesis and the structure of the main chapters. The main objective of this thesis is to analyse personal data collection in social media on the scope of Digital Marketing. To achieve this objective, three specific objectives and seven research questions (RQs) are set out.

1.4.1. Objectives

As has been shown in the previous sections, the data collection in social media and the digital marketing sector are very important for society both economically and socially. In today's digital age, social media platforms have become indispensable tools for businesses to connect with their target audience, build brand awareness, and promote their products or services (Kaplan & Haenlein, 2010). As such, the process of collecting data from social media channels has become an essential aspect of digital marketing, allowing companies to gain valuable insights into consumer behavior, preferences, and sentiment. Moreover, it is a well-established fact in the academic literature that the social media and digital marketing are among the most analyzed areas today (Alalwan *et al.,*2017). Researchers have been actively studying various aspects of social media and digital marketing, ranging from the effectiveness of marketing strategies to the impact of social media on consumer behavior. This trend is likely to continue as social media platforms evolve and new digital marketing techniques emerge (Dwivedi *et al.,* 2021).

The increasing focus on data collection in social media and digital marketing reflects the growing recognition of the critical role that these platforms play in society. Businesses that can effectively leverage social media and digital marketing data are better positioned to compete in today's marketplace and create products and services that meet the needs and desires of their customers. As such, it is essential for companies to stay up to date with the latest trends and research findings in social media and digital marketing to remain competitive and relevant in their respective industries (Kumar *et al*, 2016; Lee *et al.,* 2012).

Taking this into consideration, the first specific objective is described below:

(1) Specific objective 1: To conduct a bibliometric analysis of publications related to data management within digital marketing and social media. It is crucial to gaining a comprehensive understanding of the current state of research in this field.

Bibliometric analysis involves the quantitative analysis of publication data, which enables researchers to identify patterns and trends in research output, citation patterns, and collaboration networks (van Raan, 2004). In the context of data management within digital marketing and social media, a bibliometric analysis can provide insights into the most influential researchers, the most cited articles, the most common research topics, and the most active research institutions. By conducting such an analysis, researchers can gain a better understanding of the existing knowledge base and identify gaps in research that need to be addressed. This information can inform future research efforts and help to ensure that research in this area remains relevant and impactful. Bibliometric analysis has been used in various fields, including marketing and social media research, to provide insights into the current state of research and identify emerging trends (Leung *et al.,* 2017; Thelwall, 2009). In the context of data management within digital marketing and social media, recent studies have used bibliometric analysis to identify the most common research topics and research methods (Ali *et al.,* 2022). Conducting a bibliometric analysis of publications related to data management within digital marketing and social media is a valuable research objective that can provide insights into the current state of research in this field and inform future research efforts. The following objectives are derived from the bibliometric study.

Hence, the second specific objective of the thesis is the following:

(2) Specific objective 2: To investigate the prevalence and motivations of intentionally falsifying personal data in online sweepstakes and quizzes, and to determine the weight of factors such as privacy, trust, and amusement in users' decision-making processes.

Understanding why users intentionally falsify personal data in online sweepstakes and quizzes can help businesses and researchers design better strategies for data collection and analysis. Previous studies have examined factors that influence users' decision to falsify personal data, such as privacy concerns (Milne & Culnan, 2004), trust in the website (Zhang & Gupta, 2018), and amusement (Karpińska-Krakowiak & Modliński, 2018). However, very few studies have examined the prevalence and motivations of intentionally falsifying personal data in online sweepstakes and quizzes. A study by Fatima *et al.,* (2019) investigated the prevalence and motivations of intentionally falsifying personal data among internet users. The study found that the prevalence of intentionally falsifying personal data was high among internet users, with convenience and privacy concerns being the primary motivations for falsification. Similarly, a study by Chen and Dibb (2010) found that privacy concerns were the primary reason for internet users to falsify their personal data in online contexts. Understanding the weight of factors such as privacy, trust, and amusement in users' decision-making processes can help businesses and researchers design better strategies for data collection and analysis. For example, businesses can implement privacy-enhancing technologies, such as encryption and data anonymization, to address privacy concerns and increase trust in their websites (Rana *et al.*, 2022). Researchers can also design surveys and quizzes that are more engaging and amusing, which may reduce the motivation for users to falsify their personal data (Wang *et al.*, 2016). By investigating the prevalence and motivations of intentionally falsifying personal data in online sweepstakes and quizzes, businesses can gain insights into the factors that influence users' decision-making processes. This information can be used to design targeted marketing strategies that are more likely to resonate with their target audience among artificial intelligence (AI) algorithms (Davenport & Ronanki, 2018).

The third objective is described as follows:

(3) Specific objective 3: To analyse and identify segments of the market based on large databases of lead generation companies, proposing the use of AI algorithms.

Clustering marketing is the process of segmenting a large market into smaller, more manageable groups or clusters based on shared characteristics such as demographics, psychographics, and buying behavior (Punj & Stewart, 1983). This information can be used to identify clusters of users with similar motivations and preferences (Liao *et al.*, 2021). For example, businesses can use clustering marketing to identify segments of users in a sector that could have more tendence to sales conversion (Morwitz & Schmittlein, 1992). By customizing marketing messages and strategies to these specific segments, businesses can enhance the effectiveness of their marketing campaigns while also mitigating the risk of users deliberately falsifying personal data. Furthermore, the use of AI algorithms in clustering marketing can significantly enhance the accuracy and effectiveness of the segmentation process (You *et al.*, 2017). By leveraging AI algorithms to analyze large datasets, businesses can identify hidden patterns and trends that may not be easily discernable through traditional methods, enabling them to create more accurate and effective clusters (Tyagi & Chahal, 2022). The research objective is highly relevant to clustering marketing, enabling businesses to create more accurate and effective clusters and design targeted marketing strategies that are more likely to resonate with their target audience (Sarstedt *et al.*, 2013).

The research questions derived from the specific objectives are the following:

RQ1. Are errors mainly produced accidentally, generating misinformation, or intentionally, generating disinformation, when filling in personal data online?

RQ2. Are there generational differences when entering incorrect personal data?

RQ3. Are there any differences with regard to declared sex when entering incorrect personal data?

RQ4. What are the main motivations for intentionally entering incorrect online data, i.e. to generate disinformation?

RQ5. Are unsupervised algorithms efficient algorithms for clustering in marketing using data from online sweepstakes and tests?

RQ6. Are supervised algorithms efficient algorithm for clustering in marketing with data from online sweepstakes and tests?

RQ7. Which of the two types of algorithms are more efficient for clustering in marketing using data from online sweepstakes and tests?

1.4.2. Structure of the main-body chapters

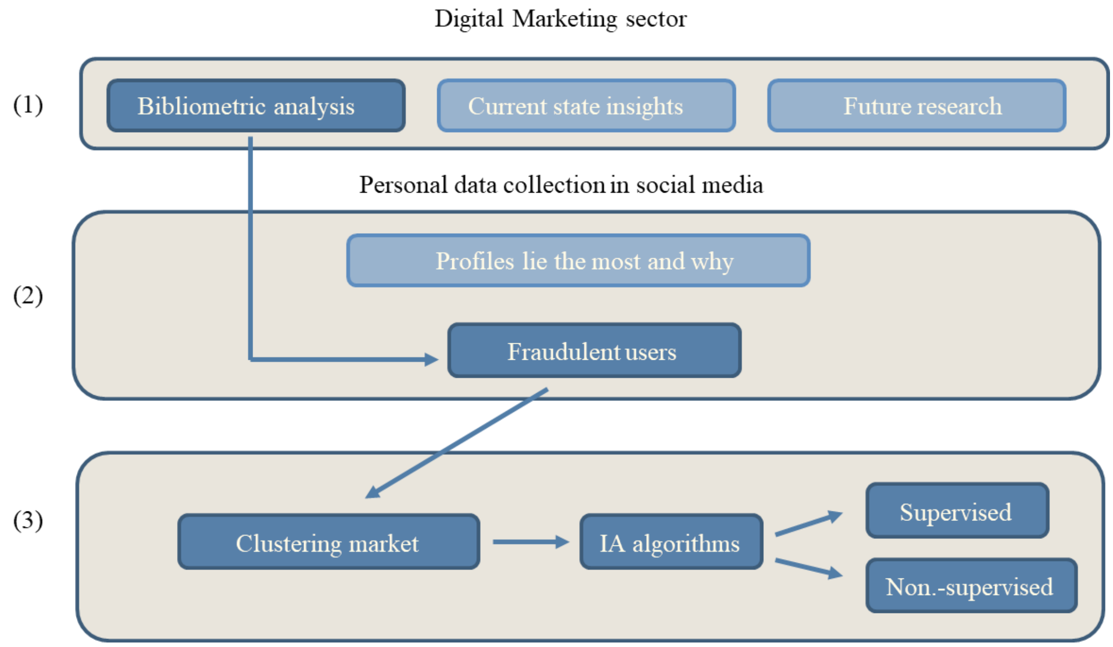
The main chapters (2-4) of the thesis are structured as three distinct yet interconnected academic papers, all of which are aimed at achieving the primary objective. Each chapter corresponds to one of the specific objectives outlined in the study. Table 8 illustrates how each chapter aligns with the specific objective and research questions. Additionally, Figure 1 portrays the research model, which provides an overview of all the objectives of the thesis and simplifies the comprehension of the associations between them.

All three chapters are empirical research studies. Due to the importance of the topic in our era, it was decided to analyse the topic trough a Bibliometric, chapter 2, analysis of publications on the capture of consumers' personal data from social networks in the field of digital marketing. This analysis aims to identify the most relevant trends through analysis of the most significant articles, keywords, authors, institutions, and countries. The study also uses visualisation software to illustrate the relationships established through bibliographic coupling, keyword co-occurrence, authors, and co-citation and discusses the progress of research and suggests future research directions. Therefore, Chapter 3 analyses the issue of online cheaters who falsify their personal data on the internet. It has been conducted three studies to estimate the percentage of users who provide false information, determine their motivations, and characterize their profiles by sex and age. The study uses a combination of quantitative and qualitative methods to estimate the volume of cheaters by stated sex and cohort of the database made up of the information provided by volunteer participants in online sweepstakes and tests. The research also explores the motivations for intentionally falsifying data provided to sweepstake sponsors. The findings can help improve methods for capturing information and detecting cheaters on social networks.

Chapter 4 expands the analysis with a cluster consumer analysis, which is a marketing segmentation technique that groups consumers based on their shared characteristics, such as demographics, behaviour, and preferences. By clustering of consumer profiles by sector, cluster consumer analysis, which is a marketing segmentation technique that groups consumers based on their shared characteristics, such as demographics, behaviour, and preferences. By analysing the clustering of consumer profiles, companies and researchers can better understand target audience and develop more effective marketing strategies.

Thus, the structure of the three central chapters of this thesis (Figure 1) starts from a more general topic, the importance of personal data collection and marketing strategies on social networks and the potential issues associated with them. Table 8 reflects the relationship between chapters and objectives.

Figure 1. Thesis research model



Specific objectives (1) – (3) indicated in parentheses.

Source: Own elaboration

Table 8. Relationship between chapters and objectives

Chapter	Title	Specific Objective	Research questions
Chapter 2	What’s on the horizon? A bibliometric analysis of personal data collection methods on social networks	(1) To conduct a bibliometric analysis of publications related to data management within digital marketing and social media.	The paper does not present a specific research question. Instead, it aims to provide an objective and quantifiable assessment of the current state of the literature in this area.
Chapter 3	Online cheaters: Profiles and motivations of internet users who falsify their data online.	(2) To investigate the prevalence and motivations of intentionally falsifying personal data in online sweepstakes and quizzes, and to determine the weight of factors such as privacy, trust, and amusement in users' decision-making processes.	RQ1. Are errors mainly produced accidentally, generating misinformation, or intentionally, generating disinformation, when filling in personal data online? RQ2. Are there generational differences when entering incorrect personal data? RQ3. Are there any differences with regard to declared sex when entering incorrect personal data? RQ4. What are the main motivations for intentionally entering incorrect online data, i.e. to generate disinformation?
Chapter 4	Market Segmentation Methods: A Comparative Analysis between Unsupervised and Supervised Learning.	(3) To analyse and identify segments of the market based on large databases of lead generation companies, proposing the use of AI algorithms.	RQ5. Are unsupervised algorithms efficient algorithms for clustering in marketing using data from online sweepstakes and tests? RQ6. Are supervised algorithms efficient algorithm for clustering in marketing with data from online sweepstakes and tests? RQ7. Which of the two types of algorithms are more efficient for clustering in marketing using data from online sweepstakes and tests?

Source: Own elaboration.

1.5. Contributions derived from this thesis

The contributions derived from the completion of this thesis are listed in Table 9. This table also shows the relationship of the contributions with the chapters of the thesis. As mentioned above, this thesis merges three articles corresponding to its main chapters.

Table 9. Contributions derived from this thesis.

Authors	Title	Type	Status	Publication details	Relationship with this thesis
Sáez-Ortuño, L., Forgas-Coll, S., Huertas-Garcia, R., Sánchez-García, J	What's on the horizon? A bibliometric analysis of personal data collection methods on social networks	Article	Published (Sáez-Ortuño et al., 2023a).	*Journal of Business Research*, Vol.158, pages 113702 DOI:10.1016/j.jbusres.2023.113702	Chapter 2
Sáez-Ortuño, L., Forgas-Coll, S., Huertas-Garcia, R., Sánchez-García, J	Detección de perfiles de usuarios que falsifican sus datos cuando participan en sorteos y test online	Conference	Published	*Proceedings book of XXXI International ACEDE Conference*	Previous version of Chapter 3
Sáez-Ortuño, L., Forgas-Coll, S., Huertas-Garcia, R., Sánchez-García, J	Ready to lie? An approach to the main motivations in online sweepstakes and quizzes	Conference	Published	*Proceedings book of EMAC Annual Conference 2023*	Previous version of Chapter 3
Sáez-Ortuño, L., Forgas-Coll, S., Huertas-Garcia, R., Sánchez-García, J	Online cheaters: Profiles and motivations of internet users who falsify their data online	Article	Published (Sáez-Ortuño et al., 2023b).	*Journal of Innovation & Knowledge*, Vol.8(2), pages 100349. DOI:10.1016/j.jik.2023.100349	Chapter3

Source: Own elaboration.

CHAPTER 2. WHAT'S ON THE HORIZON? A BIBLIOMETRIC ANALYSIS OF THE COLLECTION OF PERSONAL DATA ON SOCIAL NETWORKS [1]

[1] This chapter has been adapted from Sáez-Ortuño *et al*. (2023a).

Abstract

Purpose: The aim of this paper is to study the publications in the Web of Science Core Collection (WoS) on this issue from 1997 to 2022 (n=866) in the field of Digital Marketing.

Design/methodology/approach: A bibliometric approach is used to identify the most relevant trends by analysing the most significant articles, keywords, authors, institutions, and countries. The study also maps the bibliographic material graphically using visualisation software (VOS). It analyses the bibliographic coupling, the co-occurrence of keywords, authors and how articles are connected to each other through co-citation analysis.

Findings: The results indicate that the USA and Australia are the countries that publish the most in this field, while Finland and Australia have the highest number of publications per capita.

Originality/value: Although studies analysing digital marketing and social media have been published, this study is one of the first focusing on capture of consumers' personal data from social networks. Finally, the progress of research is discussed and directions for future research are suggested.

Keywords: bibliographic coupling of authors, bibliometric analysis, consumer studies, VOS Viewer, co-occurrences, network analysis.

2.1. Introduction

Social Media (SM) are defined by Kaplan and Haenlein (2010) as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0 and enable the creation and sharing of user-generated content." Social networks offer a multitude of possibilities for digital marketing because they allow access to a wide variety of information and have generated a new form of consumer behaviour. Millions of users can be reached through SM, who could become present and/or future consumers (Kumar *et al.*, 2016). Consumer data can be captured from social media via digital marketing actions that have increased tremendously in number in recent years (Statista, 2021). Today's technology means they can be analysed using ever more sophisticated analytical techniques (De Luca *et al.*, 2021; Dwivedi *et al.*, 2021; Grewal *et al.*, 2021; Grishikashvili *et al.*, 2021, Kharchenko, 2019; Wang & Wang, 2020). A wide variety of methods are employed to ensure that this data is used to make the most reliable predictions possible. Digital marketing is an area that has benefited from these techniques due to the effectiveness and usefulness of these new approaches to optimise proposed strategies and actions in this area (Choudrie *et al.*, 2021; Lies, 2019).

Several studies from different points of view have reviewed the literature on digital marketing via social media (Ghorbani *et al.*, 2021; Krishen *et al.*, 2021). Although these studies provide valid information, technology and hence the discipline are evolving so exponentially that a bibliometric approach focused on capturing, aggregating and synthesising the most up-to-date information and data available is needed (Paul & Criado, 2020; Randhawa *et al.*, 2016).

The Web of Science (WoS), formerly known as the Web of Knowledge, is a Clarivate Analytics platform that encompasses an extensive collection of bibliographic databases and references to scientific publications that can be used to analyse the scientific performance and quality of research (Paul & Singh, 2017). The Web of Science Core Collection (WoS CC) is a database included in WOS that contains comprehensive bibliographic references, citation indexes and h-indexes of authors from different disciplines, including the one covered in this study. Among other tasks, this database can be used to extract detailed information on the total number of published articles, number of citations and h-index and the citation thresholds and citations per article (Paul & Criado, 2020).

The aim of this study is to analyse WoS CC publications using a modern bibliometric approach (Goyal & Kumar, 2021) based on different indicators and using the Visualizing Scientific Landscape (VOS viewer) software. We mainly analyse the most cited papers and the main authors, institutions, countries and keywords in this field. The results are presented graphically with images obtained from the VOS viewer. Based on the mapping analysis by Merigó *et al.* (2018) we assess co-citation (Small, 1973), bibliographic linkage (Kessler, 1963), keyword co-occurrence and co-authorship (Merigó *et al.*, 2016).

The remainder of the paper is organised as follows. Section 2 presents the theoretical framework. Section 3 summarises the bibliometric methods. The results obtained from WoS CC are presented in section 4. Section 5 presents a graphic map of the bibliographic material produced with VOS viewer. Section 6 contains the discussion. Section 7 summarises the main conclusions, limitations and future lines of research.

2.2. Theoretical framework

Studies on data collection for social media marketing empirically investigate what data to collect, and for what purpose, in order to sell products and services, raise awareness of a brand, generate traffic to online platforms and/or create interactivity with companies and users on social media (Bianchi & Andrews, 2015; Schultz & Peltier, 2013).

Companies thus know their consumers more and better and can carry out more effective marketing actions. Schweidel and Moe (2014) point out that by capturing personal data, companies can gain greater control over the performance of their social media campaigns and improve the segmentation of their target audience. They also seek to focus their actions on social networks by trying to only reach users who are likely to be interested in the information. This is where the company's employees, who enter the information and share it, play a key role (Rokka *et al.*, 2014).

Indeed, the specific objectives and challenges of social media marketing in which personal data is captured will depend on factors such as whether the business model is B2B or B2C, the sector in which the company operates and the size of the company, but they all make it explicit that the capture of personal data and its inclusion in a database is essential. One way to create a database is through lead generation (Desai, 2019), which is based on generating requests for information from users and getting them to provide their personal data in order to send them commercial information of interest to them. Nowadays, lead generation is mostly done through social networks (Rothman, 2014).

The data must be as reliable as possible as they are subsequently used by companies to offer their products and services. The information will also be used to build consumer loyalty to the brand through engagement and sharing (Menon *et al.*, 2019). The effectiveness of social media marketing and the data that consumers are willing to share truthfully depends on the trust that consumers attribute to companies and brands in the social media sphere. Consumers may perceive companies and brands as untrustworthy (Fournier & Avery, 2011) and intrusive (Schultz & Peltier, 2013).

In contrast, other studies (Ashley & Tuten, 2015; Canhoto & Clark, 2013) show that users also want companies to be present on social media and quote or tag brands in their posts, thus tacitly offering their data. This discrepancy generates a duality among consumers. Some want brands to be active on social media, and others reject such practices.

This article reviews the literature on the collection of personal consumer data from social media for digital marketing purposes by applying methods derived from bibliometrics and content analysis to assess the current state of the literature in an objective and quantifiable manner. The analysis is based on the volume of publications, journals, impact factors, most cited articles and authors, and the most prolific countries. The aim of this review is to help identify the main current trends and future lines of research on the topic.

2.3. Methodology

Bibliometric analysis is a quantitative method used to classify and report bibliographic data. In the abundant literature on the bibliometric method (e.g. Broadus,1987; Goyal & Kumar, 2021), it is defined as a field of information science and library science that studies bibliographic material using quantitative methods (Gaviria-Marin *et al.* 2018). Although it is not a new methodology (Subramanyam, 1983), for it has existed in academia for more than a quarter of a century (Ding *et al.*, 2014), its use is booming among the scientific community because improvements in the required technology (Goyal & Kumar, 2021; Ruggeri *et al.*, 2019; Dao *et al.*,2017).
Many indicators can be calculated from a bibliometric analysis (Ruggeri *et al.*, 2019; Cancino *et al.*, 2017; Randhawa *et al.*, 2016; Merigó *et al.*, 2015) and several classifications have been accepted by academia, including the one by total

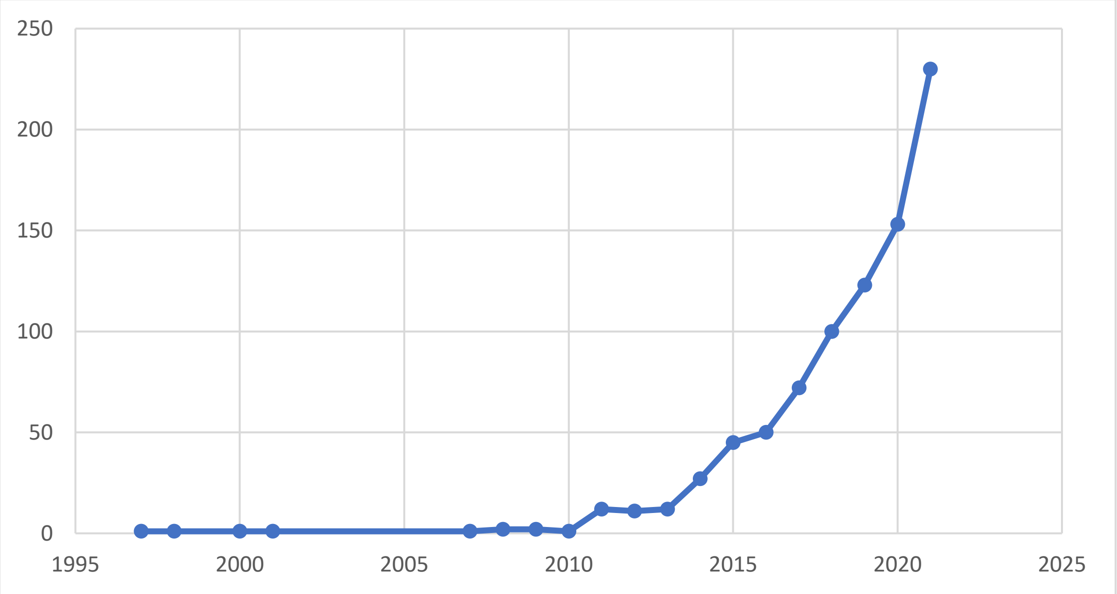
publications (Yi & Yang, 2014; Yu *et al.*, 2018). It is also common to sort results by citations received (Radicchi *et al.*, 2008; Valenzuela *et al.*, 2017) and by h-index (Costas & Bordons, 2007; Hirsch, 2005).

Most bibliometric studies apply all the aforementioned indicators to obtain the most complete, holistic picture of the results (Laengle *et al.*, 2018; Tur-Porcar *et al.*, 2018; Laengle *et al.*, 2017) as is the case with the present study, whose source database is the Web of Science Core Collection. Our research process began by searching for:

"data*" AND "social media*" AND "digital marketing*".

Only publications with consumer-related associations are considered in this study. The complete search and analysis procedure was conducted between February and March 2022. Figure 2 is a graphic representation on a general level of the annual evolution in the number of articles, with a total of 866 documents published (between January 1997 and April 2022). These documents received 17,243 citations during the period of analysis. Table 10 shows the annual citation structure of the publications.
Van Eck and Waltman (2010) highly recommend VOS software for better visualisation of the results of this type of study. Following Merigó *et al.* (2018), ours will also use VOS to display the bibliographic coupling of countries and to plot the co-citation results for authors and journals.

Figure 2. Evolution of number of articles from 1997 to 2022



Source: Own elaboration.

Table 10. Annual citation structure of publications

Year	>500	>200	>100	>50	>20	>10	>5	>1	Papers	Citations
2022	0	0	0	0	0	0	0	1	21	15
2021	0	0	0	0	0	1	1	1	230	819
2020	0	0	0	0	0	0	0	0	153	1,128
2019	0	0	0	0	0	0	0	0	123	1,142
2018	0	1	1	1	1	1	1	1	100	1,931
2017	0	0	1	2	2	2	2	2	72	1,988
2016	0	0	1	1	1	1	1	1	50	1,493
2015	0	0	0	0	0	0	1	1	45	2,308
2014	0	2	2	4	7	9	11	12	27	1,090
2013	2	2	2	3	6	6	7	7	12	396
2012	0	0	2	2	2	4	7	9	11	3,337
2011	0	1	3	8	14	15	19	25	12	1,020
2010	0	2	5	14	26	31	36	41	1	6
2009	0	1	1	7	25	32	39	45	2	116
2008	0	1	4	13	27	41	45	60	2	199
2007	0	1	3	6	27	53	70	88	1	237
2001	0	0	0	3	12	35	62	100	1	0
2000	0	0	0	2	12	34	59	108	1	0
1998	0	0	2	3	4	12	32	88	1	15
1997	0	0	0	0	0	0	0	5	1	3
Total	2	11	27	69	166	277	393	595	866	17,243
Percentage	0.23%	1.27%	3.12%	7.97%	19.17%	31.99%	45.38%	68.71%	100%	-

Abbreviations: >500, >200, >100, >50, >20, >10, >5, >1 = Number of papers with more than 500, 200, 100, 50, 20, 10, 5 and 1 citations.

Source: Own elaboration.

2.4. Results

This section analyses the results of the bibliometric analysis. It begins by presenting the publication and citation structure of the articles on the collection of consumers’ personal data from social networks. The most influential articles and top journals are then shown, followed by the main authors, institutions and countries.

2.4.1. Publication and citation structure on the collection of consumers' personal data from social networks

In bibliometric analyses, citations are analysed between two variables to determine the degree of citation between two publications (Wang *et al.*, 2018). When two different publications both cite a third publication, this is called bibliographic coupling (Kessler, 1963). On the other hand, when two different publications are cited by the same publication, this is called co-citation (Small, 1973). The most common keywords, which usually appear below the abstract, are measured through the co-occurrence of keywords. Network graphs are used to visualise the keywords that appear most frequently over time in the same types

of studies (Laengle *et al.*, 2018).
This section of the paper provides an overview of the publications and citations on the capture of consumers' personal data from social networks over the last 25 years to understand the general research trends in this field. A synthetic analysis is provided in order to better understand the situation of each global supra-region and region in terms of scientific contributions in this area of knowledge over the study period.

An initial overview of the journals with the highest number of publications and citations is also shown. Table 11 shows the twenty journals that have published the highest number of articles on the collection of consumers' personal data from social networks. The journal Sustainability has published the absolute number of articles in this field, and also the highest absolute number of articles (14,030 documents in 2021), although none of its articles are among the 25 with the highest number of citations.

In contrast, the Journal of Business Research, which has the second highest number of articles published in this category, and 985 articles published in absolute terms, does have three of the 25 most cited articles in this category. The highest proportion of citations per article per year in this area came from Information Communication & Society, with 234 citations/year. Table 12 shows the thirty most cited articles in this field of study, three of which were published by the Journal of Medical Internet Research.

The most cited is on "critical issues for big data", while the second most cited article is on "smart cities of the future". Table 10 shows that the first publication in this area appeared in 1997 and, from then on, we can observe a constant and exponential increase in publications. 2012 had the highest number of citations, namely 3,337, although the highest number of publications was in 2021, with 230 original contributions. The table also presents data on studies that received more than 500, 200, 100, 50, 20, 10, 5 and 1 citations. Of all the articles published in these 25 years, only 1.27% received 200 or more citations. 3.12% of them received 100 or more, while 68.71% of the articles received at least one citation.

Table 11.Top 20 journals

R	Journal	>2016	2017	2018	2019	2020	2021	TP
1	Sustainability	0	1	5	7	4	10	27
2	Journal of Business Research	3	0	0	1	1	10	15
3	Journal of Medical Internet Research	7	0	1	0	1	3	12
4	Industrial Marketing Management	3	0	1	0	1	6	11
5	Journal of Research in Interactive Marketing	2	0	1	6	1	1	11
6	European Journal of Marketing	2	1	1	4	1	1	10
7	Technological Forecasting and Social Change	2	0	1	2	2	3	10
8	International Journal of Information Management	1	1	1	2	1	3	9
9	Journal of Business Industrial Marketing	2	0	0	3	3	1	9
10	El Profesional de la Informacion	2	0	1	1	2	2	8
11	Journal of Theoretical and Applied Electronic Commerce Research	0	0	0	0	0	8	8
12	Ieee Access	0	0	0	2	3	2	7
13	Information Communication Society	2	2	0	1	1	1	7
14	International Journal of Environmental Research and Public Health	0	0	1	0	1	4	6
15	Journal of Retailing and Consumer Services	0	0	0	0	4	2	6
16	Online Information Review	1	2	2	0	1	0	6
17	Journal of Consumer Marketing	3	1	0	1	0	0	5
18	Management Decision	0	2	0	1	2	0	5
19	Marketing and Management of Innovations	0	0	0	0	3	2	5
20	Qualitative Market Research	0	0	0	2	3	0	5

Abbreviations: TP = total papers.

Source: Own elaboration.

2.4.2 *Journals with most cited articles on the capture of consumers' personal data from social media*

This part of the study presents the 30 most cited articles on the capture of consumers' personal data from social networks. The aim is to identify the most influential articles in this field. Table 12 shows the total number of citations received, the year of publication and the citations received per year for each article in this list including the name(s) of the author(s) and the journals in which they were published. As can be seen in this table, the article by Boyd & Crawford (2012), an in-depth analysis of critical issues for big data, is the leading article in terms of citations received, 2,341, and 234.10 citations per year. It is important to note the heterogeneity among the authors of the 30 most cited contributions.

Table 12. The 30 most cited documents

R	TC	Title	Author/s	Journal	Year	C/Y
1	2341	CRITICAL QUESTIONS FOR BIG DATA Provocations for a cultural, technological, and scholarly phenomenon	Boyd, Danah; Crawford, Kate	INFORMATION COMMUNICATION & SOCIETY	2012	234,10
2	817	Smart cities of the future	Batty, M.; Axhausen, K. W.; Giannotti, F.; Pozdnoukhov, A.; Bazzani, A.; Wachowicz, M.; Ouzounis, G.; Portugali, Y.	EUROPEAN PHYSICAL JOURNAL-SPECIAL TOPICS	2012	81,70
3	427	Closing the Marketing Capabilities Gap	Day, George S.	JOURNAL OF MARKETING	2011	38,82
4	396	Computer-based personality judgments are more accurate than those made by humans	Wu Youyou; Kosinski, Michal; Stillwell, David	PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA	2015	56,57
5	335	From Social to Sale: The Effects of Firm-Generated Content in Social Media on Customer Behavior	Kumar, Ashish; Bezawada, Ram; Rishika, Rishika; Janakiraman, Ramkumar; Kannan, P. K.	JOURNAL OF MARKETING	2016	55,83
6	315	Putting Education in Educational Apps: Lessons From the Science of Learning	Hirsh-Pasek, Kathy; Zosh, Jennifer M.; Golinkoff, Roberta Michnick; Gray, James H.; Robb, Michael B.; Kaufman, Jordy	PSYCHOLOGICAL SCIENCE IN THE PUBLIC INTEREST	2015	45,00
7	279	Features of Mobile Diabetes Applications: Review of the Literature and Analysis of Current Applications Compared Against Evidence-Based Guidelines	Chomutare, Taridzo; Fernandez-Luque, Luis; Arsand, Eirik; Hartvigsen, Gunnar	JOURNAL OF MEDICAL INTERNET RESEARCH	2011	25,36
8	237	Campaign ads, online messaging, and participation: Extending the communication mediation model	Shah, Dhavan V.; Cho, Jaeho; Nah, Seungahn; Gotlieb, Melissa R.; Hwang, Hyunseo; Lee, Nam-Jin; Scholl, Rosanne M.; McLeod, Douglas M.	JOURNAL OF COMMUNICATION	2007	15,80
9	228	Elements of strategic social media marketing: A holistic framework	Felix, Reto; Rauschnabel, Philipp A.; Hinsch, Chris	JOURNAL OF BUSINESS RESEARCH	2017	45,60

10	222	Social media analytics - Challenges in topic discovery, data collection, and data preparation	Stieglitz, Stefan; Mirbabaie, Milad; Ross, Bjorn; Neuberger, Christoph	INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT	2018	55,50
11	204	Challenges and solutions for marketing in a digital era	Leeflang, Peter S. H.; Verhoef, Peter C.; Dahlstroem, Peter; Freundt, Tjark	EUROPEAN MANAGEMENT JOURNAL	2014	25,50
12	198	Annual Research Review: Harms experienced by child users of online and mobile technologies: the nature, prevalence and management of sexual and aggressive risks in the digital age	Livingstone, Sonia; Smith, Peter K.	JOURNAL OF CHILD PSYCHOLOGY AND PSYCHIATRY	2014	24,75
13	197	Psychological targeting as an effective approach to digital mass persuasion	Matz, S. C.; Kosinski, M.; Nave, G.; Stillwell, D. J.	PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA	2017	39,40
14	183	On the Fintech Revolution: Interpreting the Forces of Innovation, Disruption, and Transformation in Financial Services	Gomber, Peter; Kauffman, Robert J.; Parker, Chris; Weber, Bruce W.	JOURNAL OF MANAGEMENT INFORMATION SYSTEMS	2018	45,75
15	175	Whose and what chatter matters? The effect of tweets on movie sales	Rui, Huaxia; Liu, Yizao; Whinston, Andrew	DECISION SUPPORT SYSTEMS	2013	19,44
16	174	Tourism analytics with massive user-generated content: A case study of Barcelona	Marine-Roig, Estela; Anton Clave, Salvador	JOURNAL OF DESTINATION MARKETING & MANAGEMENT	2015	24,86
17	166	Screen Media Exposure and Obesity in Children and Adolescents	Robinson, Thomas N.; Banda, Jorge A.; Hale, Lauren; Lu, Amy Shirong; Fleming-Milici, Frances; Calvert, Sandra L.; Wartella, Ellen	PEDIATRICS	2017	33,20
18	158	Digital transformation: A multidisciplinary reflection and research agenda	Verhoef, Peter C.; Broekhuizen, Thijs; Bart, Yakov; Bhattacharya, Abhi; Dong, John Qi; Fabian, Nicolai; Haenlein, Michael	JOURNAL OF BUSINESS RESEARCH	2021	158,00
19	154	CONTENT OR COMMUNITY? A DIGITAL BUSINESS STRATEGY FOR CONTENT PROVIDERS IN THE SOCIAL AGE	Oestreicher-Singer, Gal; Zalmanson, Lior	MIS QUARTERLY	2013	17,11

20	147	The impact of online brand community characteristics on customer engagement: An application of Stimulus-Organism Response paradigm	Ul Islam, Jamid; Rahman, Zillur	TELEMATICS AND INFORMATICS	2017	29,40
21	131	Digital marketing and social media: Why bother?	Melo Borges Tiago, Maria Teresa Pinheiro; Cristovao Verissimo, Jose Manuel	BUSINESS HORIZONS	2014	16,38
22	126	Social commerce: The transfer of power from sellers to buyers	Hajli, Nick; Sims, Julian	TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE	2015	18,00
23	125	Analyzing destination branding and image from online sources: A web content mining approach	Koeltringer, Clemens; Dickinger, Astrid	JOURNAL OF BUSINESS RESEARCH	2015	17,86
24	115	A Survey of Health-Related Activities on Second Life	Beard, Leslie; Wilson, Kumanan; Morra, Dante; Keelan, Jennifer	JOURNAL OF MEDICAL INTERNET RESEARCH	2009	8,85
25	111	Setting the future of digital and social media marketing research: Perspectives and research propositions	Dwivedi, Yogesh K.; Ismagilova, Elvira; Hughes, D. Laurie; Carlson, Jamie; Filieri, Raffaele; Jacobson, Jenna; Jain, Varsha; Karjaluoto, Heikki; Kefi, Hajer; Krishen, Anjala S.; Kumar, Vikram; Rahman, Mohammad M.; Raman, Ramakrishnan; Rauschnabel, Philipp A.; Rowley, Jennifer; Salo, Jari; Tran, Gina A.; Wang, Yichuan	INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT	2021	111,00
26	110	A Multimedia Mobile Phone-Based Youth Smoking Cessation Intervention: Findings From Content Development and Piloting Studies	Whittaker, Robyn; Maddison, Ralph; McRobbie, Hayden; Bullen, Chris; Denny, Simon; Dorey, Enid; Ellis-Pegler, Mary; van Rooyen, Jaco; Rodgers, Anthony	JOURNAL OF MEDICAL INTERNET RESEARCH	2008	7,86
27	108	The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 US Election	Bossetta, Michael	JOURNALISM & MASS COMMUNICATION QUARTERLY	2018	27,00

28	99	Development of a Risk Framework for Industry 4.0 in the Context of Sustainability for Established Manufacturers	Birkel, Hendrik S.; Veile, Johannes W.; Mueller, Julian M.; Hartmann, Evi; Voigt, Kai-Ingo	SUSTAINABILITY	2019	33,00
29	96	Harnessing marketing automation for B2B content marketing	Jarvinen, Joel; Taiminen, Heini	INDUSTRIAL MARKETING MANAGEMENT	2016	16,00
30	96	Desired Features of Smartphone Applications Promoting Physical Activity	Rabin, Carolyn; Bock, Beth	TELEMEDICINE AND E-HEALTH	2011	8,73

Abbreviations: C/Y = Citations per year. TC = total citations.

Source: Own elaboration.

2.4.3 *Main authors, institutions and countries of publications on the capture of consumers' personal data from social networks.*

This part of the research presents an analysis of the main authors, institutions and countries with regard to publications on the capture of consumers' personal data from social networks. These are presented in tables 13, 14, 15 and 16. Table 13 shows the top 30 authors who have published in this area of knowledge, based on the total number of publications. Where two or more authors had the same number of publications, the author with the highest number of citations is ranked higher. Additional information such as h-index, author affiliation and country of residence is also presented. According to these results, Kelly, Karjaluoto and Freeman are the top three authors in terms of number of publications, and Crawford is the leader in terms of total citations, with 2,429.

Table 13. Top 30 leading authors

R	Name	Institution	Country	TP	TC	H	TC/TP
1	Kelly B	Univ Wollongong, Fac Social Sci, Sch Hlth & Soc, Early Start, Northfields Ave, Wollongong, NSW 2522, Australia	Australia	8	294	40	36,75
2	Karjaluoto H	Univ Jyvaskyla, Sch Business & Econ, Mkt, Jyvaskyla, Finland	Finland	7	384	24	54,86
3	Freeman B	Univ Sydney, Sch Publ Hlth, Prevent Res Collaborat, Charles Perkins Ctr, Level 6,D17, Sydney, NSW 2006, Australia	Australia	6	207	6	34,50
4	Kar AK	Indian Inst Technol Delhi, Dept Management Studies, New Delhi 110016, India	India	6	47	3	7,83
5	Saura J R	Rey Juan Carlos Univ, Dept Business Econ, Fac Social Sci & Law, Paseo Artilleros S-N, Madrid 28032, Spain	Spain	5	118	4	23,60
6	Liu Y	Univ Connecticut, Dept Agr & Resource Econ, Storrs, CT 06260 USA	USA	5	215	4	43,00
7	Ramos CMQ	Univ Algarve, Escola Super Gestao Hotelaria & Turismo, P-8005139 Faro, Portugal	Portugal	4	15	2	3,75
8	Buchanan L	Univ Wollongong, Fac Social Sci, Sch Hlth & Soc, Early Start, Northfields Ave, Wollongong, NSW 2522, Australia	Australia	4	87	4	21,75
9	Casado-molina AM	Univ Evora, CEFAGE Ctr Adv Studies Management & Econ, Faro, Portugal	Portugal	4	7	2	1,75
10	Gupta S	Newcastle Univ, 102 Middlesex St, London E1 7EZ, England	UK	4	58	2	14,50
11	Kumar S	Indian Inst Technol Roorkee, Dept Comp Sci & Engn, Roorkee 247667, Uttarakhand, India	India	4	27	3	6,75
12	Laurell C	Stockholm Sch Econ, Inst Res, Box 6501, SE-11383 Stockholm, Sweden	Sweden	4	75	3	18,75
13	Levy S	Ariel Univ, Dept Econ & Business Adm, Mkt, Ariel, Israel	Israel	4	103	3	25,75
14	Mackey TK	Univ Calif San Diego, Sch Med, Dept Anesthesiol, San Diego, CA 92103 USA	USA	4	87	5	21,75
15	Weber I	HBKU, Qatar Comp Res Inst, Doha, Qatar	Qatar	4	9	2	2,25
16	Yeatman H	Univ Wollongong, Sch Hlth & Soc, Fac Social Sci, Northfields Ave, Wollongong, NSW 2522, Australia	Australia	3	87	4	29,00

17	Aydin G	Istanbul Medipol Univ, Hlth Management Dept, Kavacik Mah Ekinciler Cad 19 Kavacik Kavsagi, TR-34810 Istanbul, Turkey	Turkey	3	28	3	9,33
18	Crawford K	Microsoft Res, Cambridge, MA 02142 USA	USA	3	2.429	3	809,67
19	Dai HJ	Natl Taitung Univ, Dept Comp Sci & Informat Engn, Taitung 95092, Taiwan	Taiwan	3	32	3	10,67
20	Demant J	Univ Copenhagen, Dept Sociol, Oster Farimagsgade 5, DK-1353 Copenhagen, Denmark	Denmark	3	15	2	5,00
21	Dwivedi YK	Swansea Univ, Sch Management, Emerging Markets Res Ctr EMaRC, Bay Campus, Swansea, W Glam, Wales	UK	3	127	2	42,33
22	Gomez M	Adsmurai, Barcelona, Spain	Spain	3	13	2	4,33
23	Gvili Y	Ono Acad Coll, Sch Business Adm, Mkt, Kiryat Ono, Israel	Israel	3	100	2	33,33
24	Hong-jie DAI	Natl Taitung Univ, Dept Comp Sci & Informat Engn, Taitung 95092, Taiwan	Taiwan	3	32	3	10,67
25	Jacobson J	Ryerson Univ, Ted Rogers Sch Management, Toronto, ON, Canada	Canada	3	162	3	54,00
26	Kariippanon K	Univ Wollongong, Sch Hlth & Soc, Fac Social Sci, Northfields Ave, Wollongong, NSW 2522, Australia	Australia	3	58	3	19,33
27	Kauffman RJ	Singapore Management Univ, Sch Informat Syst, Informat Syst, Singapore, Singapore	Singapore	3	196	2	65,33
28	Kosinski M	Stanford Univ, Dept Comp Sci, Stanford, CA 94305 USA	USA	3	610	3	203,33
29	Krishen AS	Univ Nevada, Mkt & Int Business, Las Vegas, NV 89154 USA	USA	3	127	2	42,33
30	Kumar A	Aalto Univ, Sch Business, Mkt, Aalto, Finland	Finland	3	326	2	108,67

Abbreviations: TP = total papers; TC = total citations; H = *h-index*; TC/TP = ratio of citations divided by publications.

Source: Own elaboration.

Finally, this study analysed the leading countries in publications in the field of consumer data collection. Table 14 provides information on the most productive and influential countries in this area of study. According to this ranking, Asia, Europe and North America are the main regions for these indicators. The final item in this table is the proportion of each region among the 30 most cited articles, which is analysed in aggregate form in Table 14. The indicators used in this table are the same as in the others except that here the population of each country is added in order to calculate the publications per capita. As shown in table 14, the USA, the UK, China, Australia and Spain are the five countries that have published the most in this discipline. The USA tops the list, both in number of publications and number of citations. However, the indicator of total publications per capita offers another perspective, and a more homogeneous one as the total number of publications is fairly proportional to population size. As already mentioned, based on this analysis, Finland and Australia have the highest number of total publications per capita, while countries such as India and Pakistan, which rank 6th and 19th respectively, have one of the lowest total publication rates per capita, at 0.02 and 0.04. A similar case is Indonesia, which ranks 23rd in terms of total publications but only publishes 0.03 articles per member of its population. Table 14 details the top 30 publishing countries in the scope of this study.

Table 14. The most productive and influential countries

R	Country	Supraregion	TP	TC	H	TC/TP	Population	TP/POP	TOP 30
1	USA	North America	213	8,584	36	40	328,329,953	0.65	13
2	UK	Northern Europe	93	1,419	13	15	66,836,327	1.39	3
3	China	East Asia	56	738	15	13	1,397,715,000	0.04	-
4	Australia	Oceania	54	1,734	23	32	25,365,745	2.13	3
5	Spain	Southern Europe	41	708	13	17	47,133,521	0.87	1
6	India	Southern Asia	24	886	13	37	1,366,417,756	0.02	3
7	Italy	Southern Europe	23	1,291	13	56	59,729,081	0.39	3
8	Canada	North America	21	1,346	12	64	37,593,384	0.56	2
9	Germany	Western Europe	19	611	11	32	83,092,962	0.23	3
10	France	Western Europe	18	661	10	37	67,248,926	0.27	6
11	Finland	Northern Europe	15	1,089	15	73	5,521,606	2.72	2
12	Netherlands	Western Europe	14	909	16	65	17,344,874	0.81	2
13	South Korea	East Asia	12	372	9	31	51,709,098	0.23	1
14	Portugal	Southern Europe	11	252	4	23	10,286,263	1.07	-
15	UAE	Western Asia	10	131	6	13	9,890,400	1.01	1
16	Russia	Eastern Europe	10	85	5	9	144,406,261	0.07	-
17	Brazil	Latin America	10	34	3	3	211,049,519	0.05	-
18	Sweden	Northern Europe	9	249	8	28	10,278,887	0.88	1
19	Pakistan	Southern Asia	9	84	6	9	216,565,317	0.04	-
20	Denmark	Northern Europe	8	245	6	31	5,831,400	1.37	1
21	Taiwan	Eastern Asia	8	123	7	15	23,773,876	0.34	1
22	Romania	Eastern Europe	8	118	8	15	19,286,120	0.41	-
23	Indonesia	South-east Asia	8	32	3	4	273,523,620	0.03	-
24	South Africa	Africa	7	196	6	28	58,558,267	0.12	-
25	Turkey	Western Asia	7	88	4	13	83,429,607	0.08	-
26	Malaysia	South-east Asia	7	55	5	8	31,949,789	0.22	-
27	Saudi Arabia	Western Asia	7	49	5	7	34,813,870	0.20	-
28	Norway	Northern Europe	6	369	7	62	5,379,480	1.12	-
29	Austria	Western Europe	6	304	5	51	8,917,200	0.67	1
30	Iran	Southern Asia	6	87	4	15	83,992,950	0.07	-

Abbreviations: TP = total papers; TC = total citations; H = *h-index*; TC/TP = ratio of citations divided by publications; Population = thousands of inhabitants; TP/POP = Total papers per million inhabitants; TOP 30 = the 30 most cited papers.

Source: Own elaboration.

Table 15 shows the most productive and influential institutions. In this table, all bibliographic indicators have been calculated for each institution. The table also shows the position of these institutions based on two global rankings, the Academic Ranking of World Universities and Quacquarelli Symonds University Ranking (ARWU and QS) in order to facilitate a comparison between these two rankings and to be able to show the ranking of each university. The final column of this table shows the number of articles that each institution has among the 30 most cited. The University of Sydney tops this list, followed by the University of Jyvaskyla and the University of North Carolina. Note that 5 of the top 10 universities are in the USA, which is in line with the results shown in Table 15.

However, the most cited author in this field, Crawford, is not from any of the top 10 institutions. Detailed results for each sub-region are shown in Table 16. It analyses the trend in publications by geographical region of the countries for total publications, total citations, h-index and the ratio of total citations to total publications.

Table 15. The most productive and influential institutions

R	Institution	Country	TP	TC	H	C/P	>50	>25	>5	ARWU TOP 100	RANK QS TOP 100	TOP 30
1	University of Sydney	Australia	16	304	8	19.00	2	2	8	69	80.4	-
2	University of Jyvaskyla	Finland	12	575	11	47.92	4	7	0	-	-	12
3	University of North Carolina	USA	11	72	5	6.55	0	0	6	-	-	-
4	University of Wollongong	Australia	10	304	7	30.40	2	2	4	-	-	-
5	University of Pennsylvania	USA	8	741	7	92.63	2	2	3	15	90.7	8
6	University of texas Austin	USA	8	213	3	26.63	1	0	2	-	-	-
7	University of California San Diego	USA	8	173	7	21.63	1	2	4	18	76.1	-
8	University of Liverpool	UK	8	130	6	16.25	1	1	4	-	-	-
9	University of Oxford	UK	8	56	5	7.00	0	0	5	7	99.5	-
10	Stanford University	USA	7	789	4	112.71	3	0	1	2	98.7	7
11	University of Cambridge	UK	7	697	5	99.57	3	1	2	3	98.7	1
12	University of Groningen	Netherlands	7	509	6	72.71	3	0	2	64	-	-
13	University of Helsinki	Finland	7	205	6	29.29	1	1	4	82	-	-
14	Yonsei University	South Korea	7	149	5	21.29	1	1	3	-	65.5	7
15	Monash University	Australia	7	126	5	18.00	0	3	3	80	72.2	1
16	Johns Hopkins University	USA	7	91	4	13.00	0	2	1	16	85.9	-
17	University of Nottingham	UK	7	61	5	8.71	0	0	5	-	-	-
18	University of Manchester	UK	7	52	5	7.43	0	0	5	35	30	-
19	University of Melbourne	Australia	7	41	4	5.86	0	0	4	-	60	-
20	Universidad de Malaga	Spain	7	22	2	3.14	0	0	1	-	-	-
21	New York University	USA	6	289	4	48.17	2	1	1	27	78.9	1
22	University of Copenhagen	Denmark	6	184	3	30.67	2	0	1	30	65.5	1
23	Ryerson University	Canada	6	167	3	27.83	1	1	1	-	-	1
24	University of New South Wales Sydney	Australia	6	65	4	10.83	0	1	3	-	-	-
25	University of Queensland	Australia	6	57	5	9.50	0	0	5	-	76.6	-
26	Aalto University	Finland	5	386	4	77.20	1	0	3	-	77.7	1
27	Ariel University	Israel	5	117	3	23.40	1	1	1	-	-	-
28	McGill University	Canada	5	83	3	16.60	0	2	1	67	84	-
29	Deakin University	Australia	5	65	3	13.00	0	1	2	-	-	-
30	Macquarie University	Australia	5	42	3	8.40	0	0	3	-		-

Abbreviations are available in previous tables except for: ARWU and QS = Academic Ranking of World Universities and QS University Ranking.
Source: Own elaboration.

Table 16. Publications by supranational regions

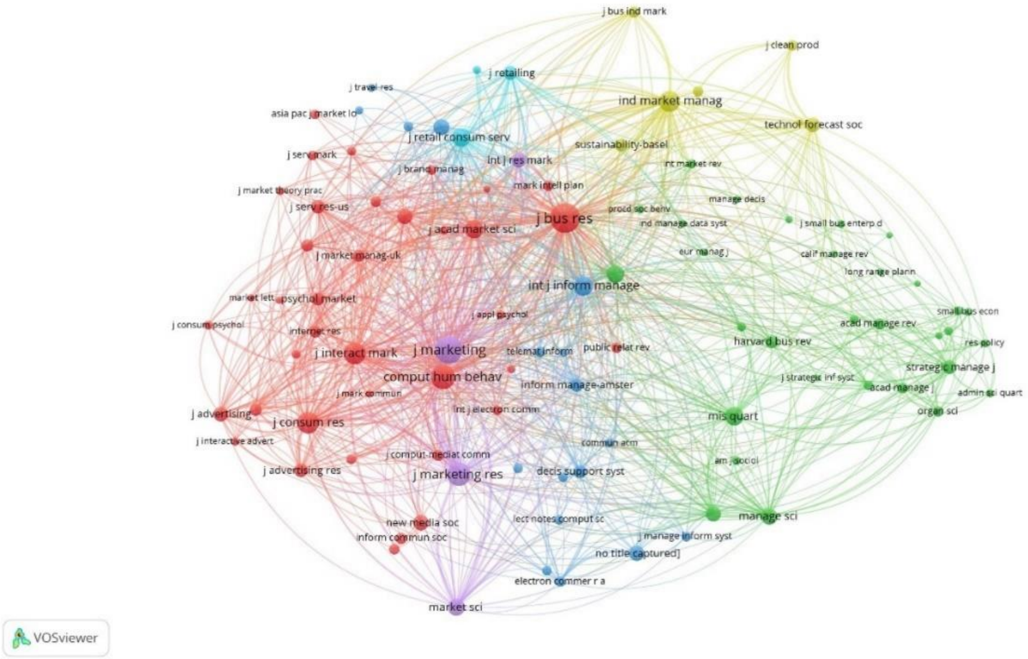
R	Supraregions	TP	TC	H	TC/TP	Top 30
1	North America	235	9962	16	42.39	9
2	Europe	354	10572	5	29.86	18
	Northern Europe	158	4273	9	27.04	4
	Western Europe	68	3564	6	52.41	7
	Southern Europe	86	2342	6	27.23	5
	Eastern Europe	42	393	3	9.36	2
3	Asia	177	4140	4	23.39	30
	East Asia	78	1274	8	16.33	5
	West Asia	37	1447	2	39.11	1
	South-east Asia	22	345	3	15.68	14
	Southern Asia	40	1074	6	26.85	10
4	Oceania	58	1972	15	34.00	2
5	Africa	21	491	1	23.38	-
6	Latin America	21	64	1	3.05	1

Abbreviations: H = *h-index*; TC/TP = ratio of citations divided by publications.
Source: Own elaboration.

2.5. Mapping results with Vos Visualisation Software

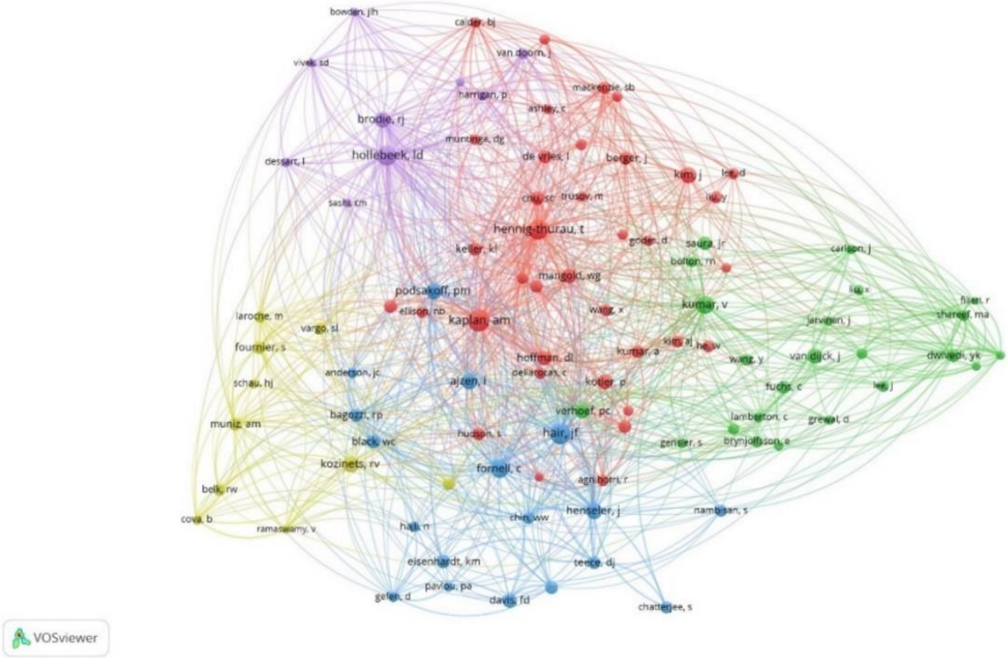
In this section, the VOS viewer software (Van Eck & Waltman, 2010) is used to view and graphically display the bibliographic coupling of countries and institutions, citations per university, co-citation of authors and journals, and co-occurrence of keywords defined by authors, as well as those extracted from article titles and abstracts. Figure 3 shows the co-citation of journals in the field of this study with a threshold of 10 and the 100 most representative co-citation connections. Figure 4 shows the co-citation of authors in the field of capture of consumers' personal data from social networks with a threshold of 10 and the 100 most representative co-citation connections. These results are in line with the previous results obtained from the analysis of authors in Table 13. The different clusters are shown in different colours, and the links between them are also indicated. Another item that was analysed using this software is the bibliographic coupling of institutions with a threshold of at least 3 publications and showing the 100 most representative connections. Figure 5 shows the results obtained, according to which and as can be seen in table 15, the University of Sydney and the University of Jyvaskyla are the most prominent institutions in this regard. The other universities in this ranking are mainly American. The following analysis performed with the VOS viewer software refers to the co-occurrence of author keywords to provide a complete understanding of the main keywords used in articles on consumers' personal data capture from social networks in the period of this study. As in the previous cases, this graph is in line with the results presented above. The following analysis performed using the VOS viewer software refers to the co-occurrence of keywords in titles and abstracts to facilitate a full understanding of the main keywords used in articles on the capture of consumers' personal data from social networks in the period of this study. Figure 6 shows the 100 strongest connections, with a threshold of five documents. As shown in the graph, the most prominent words are 'Social Media', 'Big Data' and 'Digital Marketing'. Figure 7 shows the 100 strongest connections, with a threshold of five documents. As can be seen, the expressions 'Social Media', 'Impact' 'Big Data' and 'Word of Mouth' are the most prominent. Another element that was analysed using this software is the bibliographic coupling of countries, with a threshold of at least 3 countries and showing the 100 most representative connections. Figure 8 shows the results where, as already seen in table 15, the USA, the UK, China, Australia and Spain are the most prominent countries.

Figure 3. Co-citation of journals



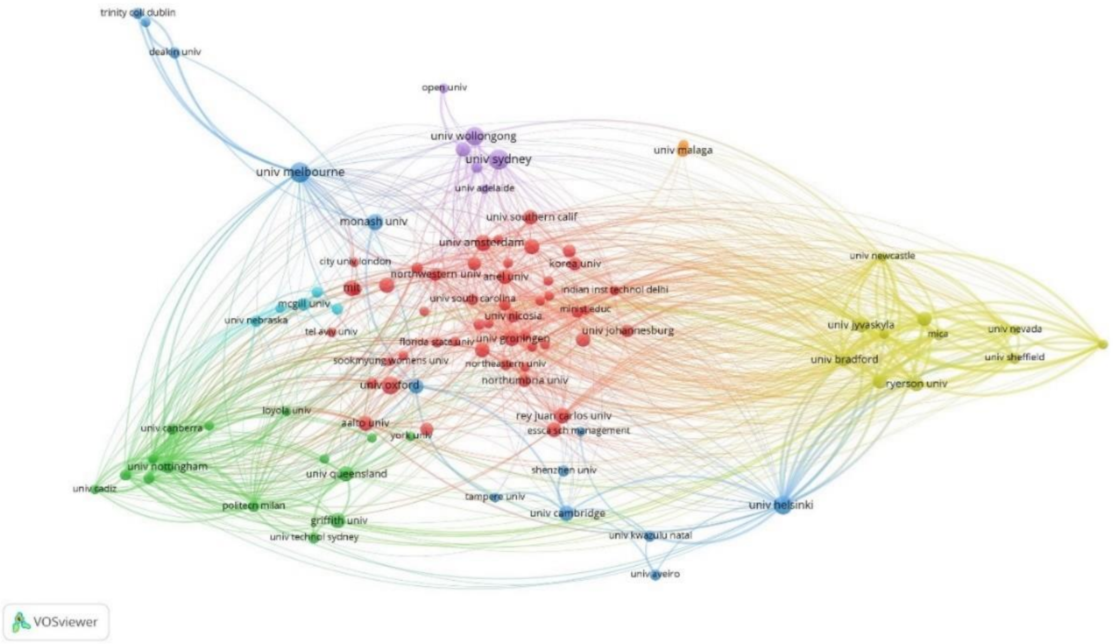
Source: Own elaboration with VOSviewer Software.

Figure 4. Co-citation of authors



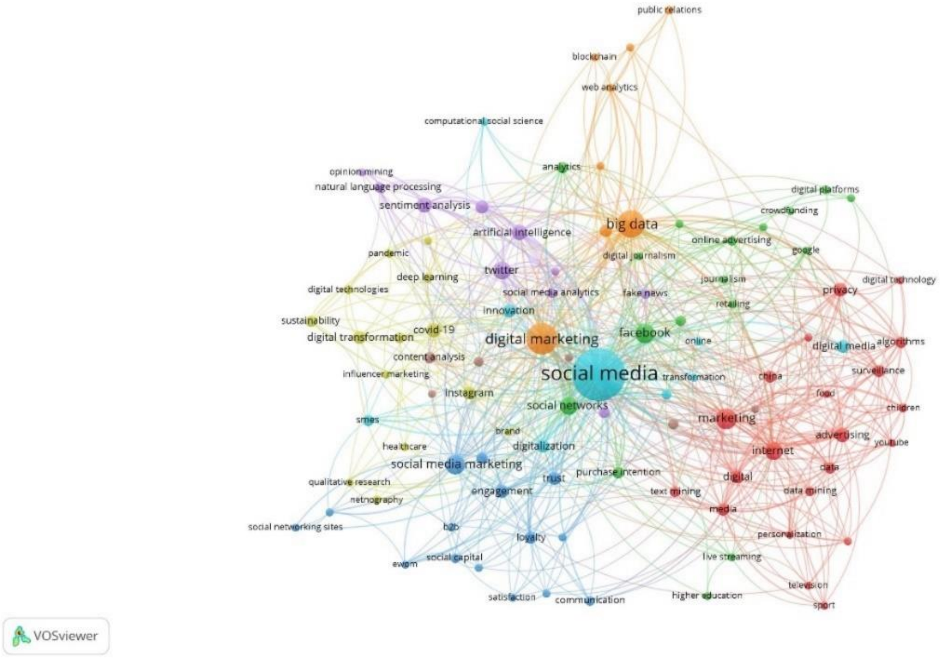
Source: Own elaboration with VOSviewer Software.

Figure 5. Bibliographic coupling of institutions



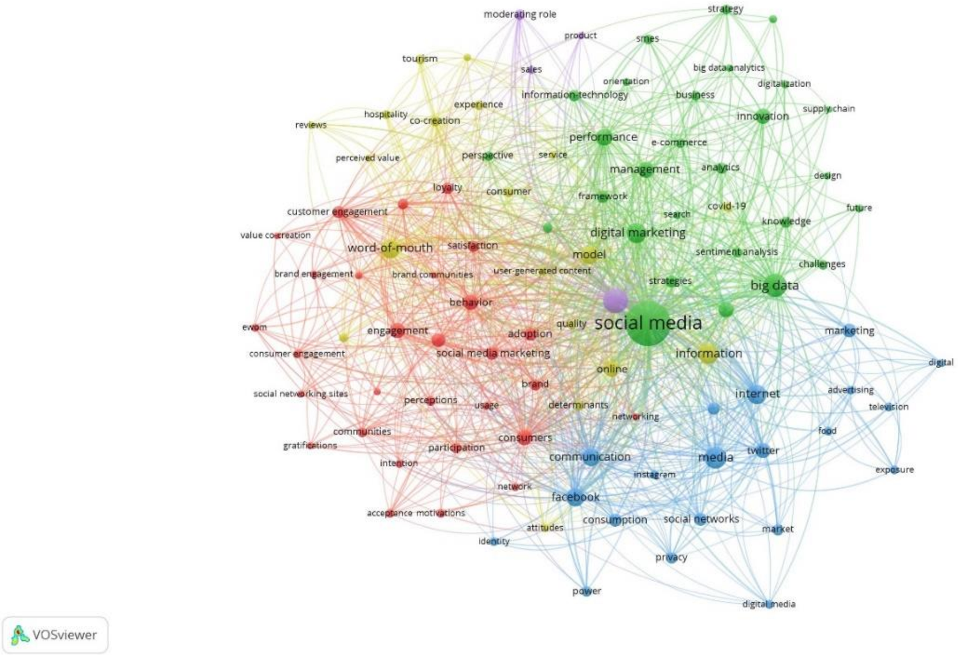
Source: Own elaboration with VOSviewer Software.

Figure 6. Co-occurrence of author keywords.



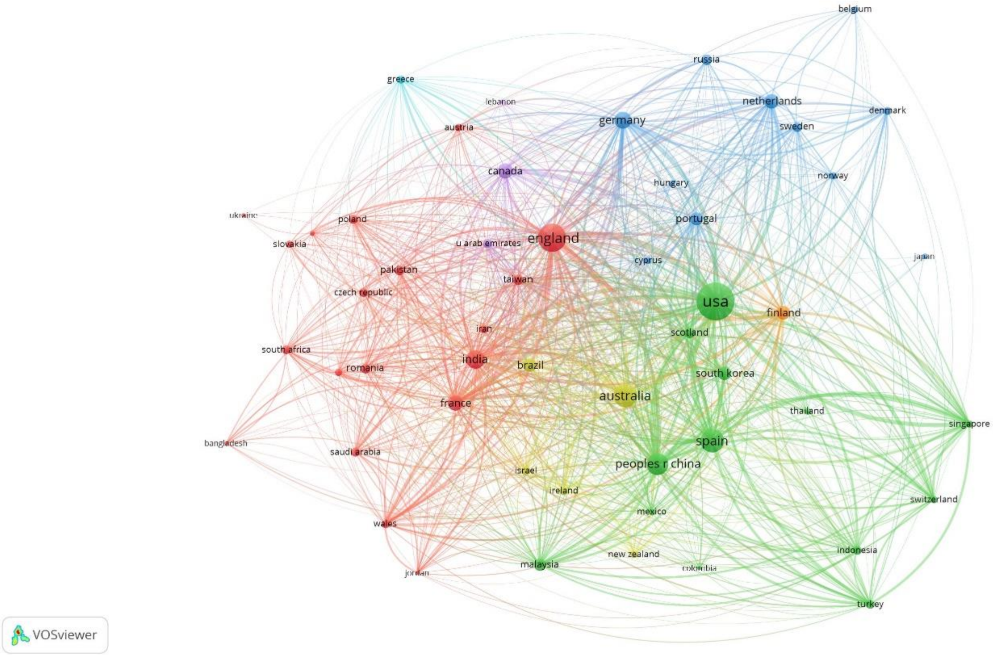
Source: Own elaboration with VOSviewer Software.

Figure 7. Co-occurrence of all keywords



Source: Own elaboration with VOSviewer Software.

Figure 8. Bibliographic coupling of countries



Source: Own elaboration with VOSviewer Software.

2.6 Conclusions, limitations and future lines of research

This study has analysed publications related to the collection of consumers' personal data from social networks over a period of twenty-five years, between 1997 and 2022. The bibliometric method was used to study the trends in publications in this field. The Web of Science Core Collection was used to extract the data on which the analysis was conducted. The study reveals an exponential increase in the number of publications and citations in this area during the period analysed.

This research contributes to the existing scientific literature on the collection of consumers' personal data from social networks. The main authors and journals have been identified, with two of the three authors with the most publications and who are most cited being attached to Australian institutions, while the other is attached to an institution in Finland.

The five most prolific countries in terms of publications are, in this order, the USA, the UK, China, Australia and Spain. Another relevant indicator is the number of publications per capita in each country, with Finland top of this ranking, followed by Australia.

The VOS viewer software was used to support the evidence obtained from the Web of Science Core Collection in order to map and graphically describe the results for the co-citation of journals and authors, the bibliographic coupling of countries and universities, and the co-occurrence of authors' keywords, thus making it visually easier to understand the relationship between the variables.

The tables in this study are based on various bibliometric measures to help readers to understand the trends in publications on the capture of consumers' personal data from social networks. This study is also useful to understand what lines of research have been pursued and what research remains to be done. The literature in this area is growing exponentially, although given the growing interest in this area of study, there are still many lines of research to be explored.

An attractive option for future research would be to perform the same analysis for the same period using different methods (g-index, p-index, article influence score, etc.) to make comparisons possible between the results of two papers.

There is also little literature on consumers' motivations for entering personal data on social networks in the context of digital marketing, which could be a promising line of research. It would also be interesting to analyse the generational and gender profiles of users who provide their personal data on social media.

The existing literature also seems to make the general assumption that users always enter truthful data. There is very little literature analysing fraudulent consumer data and its direct effects on digital marketing. A future line of research would be to analyse this phenomenon, detecting the profiles of these consumers, as well as their motivations for not providing truthful information.

Another interesting line of research would be to perform a cluster analysis of the users who enter their personal data in social media. Although authors on this topic offer keywords that can be clustered, no evidence has been found of clusters of consumers who disclose their personal information on social media within the framework of digital marketing. Analysis of these clusters would provide relevant information for a more detailed understanding of the consumer. Companies would be able to offer products and services that are much more aligned with their interests, increasing the level of satisfaction of both parties.

Regarding keyword analysis, an interesting future line of research would be to analyse the means used to capture personal data from social networks, and it would also be useful to know what kinds of rewards and incentives most motivate consumers to give up their personal data.

This article is not without its limitations. Undoubtedly, its clearest limitation stems from having limited the bibliometric analysis to a single database, the Web of Science Core Collection. However, this was intentional as such an in-depth analysis would not have been possible with a combination of various databases. It should be added that the limitations in term of scope that apply to the Web of Science Core Collection also apply to this research.

However, an interesting future line of research on the capture of personal data from social media would be to include a larger number of academic databases, such as the Scopus and Google Scholar databases, to establish even more complete classifications of journals, academics, academic institutions and countries (Koberg & Longoni, 2018) and hence extrapolate the results to the whole range of publications on this topic in relation to digital marketing.

CHAPTER 3. ONLINE CHEATERS: PROFILES AND MOTIVATIONS OF INTERNET USERS WHO FALSIFY THEIR DATA ONLINE[2]

[2] This chapter has been adapted from Sáez-Ortuño *et al*. (2023b).

Abstract

The digital environment, which includes the Internet and social networks, is propitious for digital marketing. However, the collection, filtering and analysis of the enormous, constant flow of information on social networks is a major challenge for both academics and practitioners. The aim of this research is to assist the process of filtering the personal information provided by users when registering online, and to determine which user profiles lie the most, and why. This entailed conducting three different studies. Study 1 estimates the percentage of Spanish users by stated sex and generation who lie the most when registering their personal data by analysing a database of 5,534,702 participants in online sweepstakes and quizzes using a combination of error detection algorithms, and a test of differences in proportions to measure the profiles of the most fraudulent users. Estimates show that some user profiles are more inclined to make mistakes and others to forge data intentionally, the latter being the majority. The groups that are most likely to supply incorrect data are older men and younger women. Study 2 explores the main motivations for intentionally providing false information, and finds that the most common reasons are related to amusement, such as playing pranks, and lack of faith in the company's data privacy and security measures. These results will enable academics and companies to improve mechanisms to filter out cheaters and avoid including them in their databases.

Keywords: Social networks, Online cheaters, False data, Sweepstakes, Data collection.

3.1. Introduction

As social media spreads, more and more people are using it to seek, consume and exchange information (Shu *et al.*, 2017), resulting in the generation of a massive amount of data (Kapoor *et al.*, 2018). The reason behind this trend lies in the very nature of social media, as it allows for more timely, easier and less costly consumption and dissemination of information than traditional news media (Shu *et al.*, 2017). This environment is propitious for digital marketing to understand new forms of online consumer behaviour and to promote and sell its products (Kumar *et al.*, 2016). Consumers can be analysed and segmented by referring to information about their demographic characteristics, consumption habits, etc., which can be captured from social networks (e.g. Instagram, Facebook, Twitter, TikTok, Pinterest, Snapchat, etc.) in order to generate leads. This can be done in a variety of ways, ranging from the publication of advertisements, participation in social networks, joining conversations, and creating online contests and sweepstakes (Desai, 2019). However, the information that users provide is not always correct. Many people take advantage of the anonymity offered by social networks to falsify their information, and to act in a dishonest manner (Allcott & Gentzkow, 2017; Bonald *et al.*, 2009; Vosoughi *et al.*, 2018; Wu *et al.*, 2022).

A cheater is someone who participates in a game but breaks the rules in order to gain an advantage. In other words, s/he wants to join in, but is not willing to play fair (Cosmides & Tooby, 2016). Since ancient times, cheating has been a perplexing problem for society and has been an especially huge obstacle for businesses (Cosmides & Tooby, 2016; Trivers, 1971). One form of cheating is to provide false information about oneself by misrepresenting or impersonating another person (Lwin *et al.*, 2016). Although the problems that this causes are recognised, little is known about the profiles of users who are most inclined to do so (Axelrod & Hamilton, 1981; Di Domenico *et al.*, 2020). In the digital marketing environment, where strategies depend on the provision of truthful, accurate information about consumers, it is essential to detect users who enter false information in order to remove them from databases (Blackburn *et al.*, 2014; Cosmides & Tooby, 2016; Pascual-Ezama at al., 2020). Knowledge about the user profiles that are more likely to misrepresent their data could help to refine detection methods and eliminate/reduce fraudulent practices (Ahmed, 2009). The incorporation of this information into artificial intelligence and machine learning algorithms that sift information could help improve their performance (Saura, 2021; Zhang *et al.*, 2020).

The global market for collection, storage and distribution of digital marketing-related data was worth nearly $17.7 billion in 2021, with the US being the largest market, accounting for 47% of the global value, around $24.7 billion (EMarketer-Statista, 2022). In Spain, digital advertising amounted to €3.03 billion in 2020 and the country is among the top ten in Europe with the highest spending in this area, with a figure that surpasses that for traditional advertising (EMarketer-Statista, 2019).

Digital marketers need accurate sources of data on potential consumers to target and optimise their marketing investments (Lee *et al.*, 2012). Therefore, it is essential for organizations that build databases from social media to eliminate as much pollution by fraudulent users as possible. The literature on social media marketing has focused more on data collection related to web traffic, on user engagement with each other and with the company or brand, than on the motivations or profiling of digital fraudsters (Chambers *et al.*, 2010). Certain studies have analysed the social and psychological motivations that lead consumers to provide their information online (Balint *et al.*, 2011; Fritsch *et al.*,2006), but fewer have focused on detecting the profiles of those who lie. One exception is Nazir *et al.* (2010), which analysed behaviour with false accounts used to play Facebook games. It found that users' main motivation for providing inaccurate personal data was to gain what they believed to be an advantage in the game, but the study did not profile these users.

One study that analysed a larger number of cases of cheating was carried out by Blackburn *et al.* (2014). These authors examined the cheaters flagged in an online game, finding that their number is not correlated with population density or the size of the game community. However, they did not provide information on cheaters' profiles, such as their age or sex, or the fields in which they were most likely to lie, although they do suggest that the costs of cheating are extremely significant, especially those to the industry as it seeks to detect and reduce the practice.

This paper aims to contribute by identifying the motivations and profiles of users who provide false data on the Internet. To this end, we requested the collaboration of one of the top lead generation companies in Europe, which has been operating in Spain since 2009. This company allowed us to study certain pieces of information from its database, which we used to estimate the amount of fraudulent data and to characterise cheaters by stated sex and age. We also sought to learn about their motivations for supplying false information. The data generation industry is particularly sensitive and exposed to cheaters, so early detection is critical to prevent them from contaminating the databases that will subsequently be used by companies to offer their products and services. In addition, good quality databases help to target commercial activity better, and generate greater acceptance, engagement and brand loyalty (Menon *et al.*, 2019). This study analyses the characteristics of users who have provided false information on the lead generation company's web pages, in order to fill the following gaps in the literature:

- detection of false information using AI algorithms and, in turn, the users who entered false information
- determination of whether users who entered false information did so intentionally or negligently
- examination of the fields where cheaters have entered the most false information
- characterisation by stated sex and generation of users who falsify their data the most, in order to incorporate these profiles into prediction algorithms, and
- understanding of the main motivations for intentionally providing false information.

To our knowledge, no previous work in the literature has analysed users' profiles and their motivations to enter false information when participating in online sweepstakes and quizzes with the aim of facilitating automatic detection of cheaters on social media. This research adopts a mixed-method approach that combines descriptive and exploratory research. For the descriptive research, we benefited from collaboration with the lead generation company CoRegistros, S.L.U., which provided us with several fields of a database of more than 5 million users.

The rest of the paper is organised as follows. First, a conceptual framework is presented focusing on the profiles of users who enter inaccurate personal data online and their motivations for doing so. Second, the methods and results of the two studies on which this research is based are presented. Following a discussion of the results, the implications for academia and management are addressed. The study concludes by proposing the key themes that emerged from the results, discussing its limitations, and suggesting certain avenues for future research.

3.2. Theoretical framework

3.2.1 Definition and types of fake information created by cheaters.

Although the concept of fake news originated in the 15th century (Shu *et al.*, 2017), that of online misinformation was coined six centuries later, in the early 21th century, to refer to a series of untruthful news stories and announcements generated and disseminated by websites (Mintz, 2002; Wendling, 2018) that affect most social domains (Allcott & Gentzkow, 2017; da Fonseca & Borges-Tiago, 2021). When information hits the web, cheaters, under protection of the anonymity afforded by the online environment, manipulate that information and re-distribute it, generating false content (Allcott & Gentzkow, 2017). Here, it is important to distinguish between misinformation and disinformation, as some studies have used them indiscriminately (Zubiaga *et al.* 2018). The terms have different meanings, for while misinformation refers to communications whose veracity is not yet confirmed and may or may not contain false information, disinformation involves deliberate manipulation to give the impression that the content is true (Tandoc *et al.*, 2018). That is, while the former concerns the authenticity of the information, the latter implies intentionality (Shu *et al.*, 2017).

This generation of misinformation, its ontology, detection methods and the motivations behind it have aroused much interest in the scientific community, which has carried out several studies to improve our understanding of the phenomenon. There have been studies such as the one by Habib *et al.* (2019), which endeavoured to classify misinformation into rumours, fake news, disinformation and hoaxes, and also described their characteristics to facilitate their detection and prevent cheating. Meanwhile, Tandoc *et al.* (2018) sought to categorise the purposes of false information that is disseminated online into satire, parody, political propaganda, advertising and manipulation.

Automated processes of online information dissemination are changing and increasingly attractive headlines and very limited and short-lived content are becoming more and more common, making manual monitoring impossible and thus favouring proliferation of fake news and the detection of cheaters (Conroy *et al.*, 2015). Although online misinformation is a recent phenomenon, some authors propose the adaptation of methods described in earlier literature to the detection of cheaters in different fields of application. Examples include Conroy *et al.* (2015), Parikh and Atrey (2018) and Shu *et al.* (2017), who focus on the automatic detection of false information once it has been generated, while others such as Zubiaga *et al.* (2018) address the problem more holistically, rather than merely detecting it once it has been produced. Along similar lines, Bondielli and Marcelloni (2019) approach false information from its origin, i.e. in terms of data sources and the way in which information is captured. Regardless of how cheaters are detected, all these methodologies recognise that they entail certain limitations and that they need to recurrently train their algorithms by means of behavioural and socio-demographic data.

3.2.2 Profiling types of errors: accidental or intentional

Since misinformation can be generated by accident, it is important to detect whether or not there is any malicious intent behind its creation (Pennycook *et al.*, 2021). To reduce the likelihood of error, it is proposed that robust protocols should be used to control the way that users complete registration forms (Karlova & Fisher, 2013). For example, they might be asked to enter the same data more than once, without being able to see what they typed previously, and the submission is only accepted if both entries match (Fallis, 2014). But these procedures are unable to prevent users from intentionally entering false data (Karlova & Fisher, 2013). For example, if a user gives his/her name as "Fool", and the system asks him/her to repeat it, s/he will do the same thing again. But if the algorithm detects that the word "Fool" is incorrect and lets him/her know, s/he is likely to use a fake, but apparently real, name on the second try, which is much harder to detect. It is therefore important to distinguish whether false data is provided due to error, misinformation, or where this is done intentionally, disinformation, and also to know which kinds of users are more likely to do so. In the former case, to improve the robustness of online forms, and, in the latter, in order to control and isolate such practices (Karlova & Fisher, 2013). Thus, the following research question is proposed:

RQ1. Are errors mainly produced accidentally, generating misinformation, or intentionally, generating disinformation, when filling in personal data online?

3.2.3 Fake information created by cheaters: detection of cheating through leads and user attitudes on registration.

Digital marketing often uses databases that gather information from potential consumers (leads) in order to target commercial offers better, and one way to create these databases is through lead generation (Desai, 2019). In the past, leads were acquired by making phone calls, usually without the respondent's authorisation or consent, but nowadays such processes are largely carried out through digital channels (Rothman, 2014) where the user gives their consent under a regulated framework (Spanish Data Protection Agency, 2022).

There are several ways to generate online leads, such as offering interesting content on blogs or websites (Bondarenko *et al.*, 2019), electronic requests made by social activists (Huang *et al.*, 2015), offering financial incentives such as prize draws or direct product discounts, or revealing the answers to a quiz in exchange for the user’s data. Another technique is snowballing, which consists of users winning rewards in exchange for recruiting friends and acquaintances, whose information thus becomes available to the company too (Baltar & Brunet, 2012). These all follow the principle of the social contract (Cosmides & Tooby, 2016), whereby participants are given the chance to win a prize (e.g. an iPhone, a gift voucher, a trip, etc.) or some kind of emotional reward, for getting the answers right to a quiz, test or challenge, in exchange for providing personal information. However, the information provided by participants often contains errors, and checks need to be performed to safeguard the quality of the database. This essentially involves input control and screening of the provided information. Robust data entry procedures are often used for the former, while the latter uses algorithms to detect patterns in transcription errors (incorrect names, missing phone numbers, etc.) regardless of whether they are accidental or intentionally malicious (Thakur *et al.*, 2017).

However, despite the importance of detecting the profiles and patterns of false registrations and, consequently, cheaters, no previous studies have been found that have considered the declared sex and/or age of participants that create false profiles (Pérez-Escoda *et al.*, 2021). Some, such as Sharif and Zhang (2014), did identify the main ways in which consumers could mislead and deceive on social media and how such deception can be detected. Others such as Viviani and Pasi (2017) identified and quantified a user's credibility when entering information on social media, while Conroy *et al.* (2015) demonstrated that some techniques are more effective than others in detecting online deception and identifying fraudsters. Although previous studies have addressed different aspects of the problem (Conroy *et al.*, 2015; Habib *et al.*, 2019; Parikh & Atrey, 2018; Shu *et al.*, 2017; Viviani & Pasi, 2017; Zubiaga *et al.*, 2018), they all highlight the need to create control mechanisms to ensure the quality of databases, and to use the knowledge extracted from them to compare approaches and profile cheaters better.

3.2.4 Profiling cheaters based on their generation and declared sex

Generational cohort marketing, first defined in the US at the turn of the last century, is still being used in marketing around the world (Meredith & Schewe, 1994). Cohorts are groups of individuals who are born around the same time and experience external events in a similar manner in their late teenage/early adult years. These "defining moments" influence their values, references, attitudes and purchasing behaviour in ways that persist throughout their lives (Meredith & Schewe, 1994). The experiences shared during the highly impressionable "coming of age" period [approximately 17-23 years of age] embody these values or "cohort effects" and remain relatively unchanged throughout life. Each generation is defined by its birth years and typically lasts 20 to 25 years, or about as long as it takes to grow up and have children. But a cohort can be as long or short as the external events that define it. Thus, the cohort defined by World War II might only be 6 years long (Meredith & Schewe, 1994). Schuman and Scott (1989) demonstrated that individuals of similar age have similar memories, related mainly to adolescence and young adulthood, and common experiences of major events, which they refer to throughout their lives. These characteristics mean that each cohort is a separate market segment and it can be particularly useful for marketing campaigns to target them in specific ways. In the US, seven distinct cohorts have been delineated as internally homogeneous in values yet heterogeneous across cohort groups (Meredith & Schewe, 2002). The most widespread classification of generational cohorts is usually: Silent Generation (also known as Mature, born between 1925 and 1942), Baby Boomers (born between 1943 and 1960), Generation X (born between 1961 and 1981), Millennial Generation (often referred to as Generation Y or Millennials, born between 1982 and 2000) (Brosdahl & Carpenter, 2011) and Generation Z (born between 2001 and 2009) (Yadav & Rai, 2017).

While there is growing interest in understanding the use of social media by different generations (Bolton *et al.*, 2013), little is known about which generations cheat the most. Thus, the following research question is proposed:

RQ2. Are there generational differences when entering incorrect personal data?

Another of the most common variables in marketing segmentation is the declared sex of users (Nickel *et al.*, 2020). Consumers have often been classified according to stated sex in order to optimise product design, as well as to create targeted communication and advertising campaigns (Meyers-Levy *et al.*, 2015). The selectivity hypothesis is based on using declared sex as a basic criterion to segment the market between male and female products (Moss, 2009). This theory suggests that most people who claim to be of certain sexes report different preferences and tastes and react differently to commercial stimuli (Nickel *et al.*, 2020). Although there is abundant literature on the different attitudes of men and women towards new technologies and internet use (Alalwan *et al.*, 2017) and even on their attitudes towards sweepstakes in the face of different types of stimuli and incentives (Schulten & Rauch, 2015), there is no evidence of studies that analyse sex differences among cheaters when entering personal data online. Therefore, the following research question is proposed:

RQ3. Are there any differences with regard to declared sex when entering incorrect personal data?

3.2.5 Motivations of users with fake identities

Although the instruments for collecting leads follow the logic of the social contract, in which financial and emotional incentives are offered in exchange for information, not everyone is willing to comply (Cosmides & Tooby, 2016). Previous studies have attempted to find the motivations for the generation of disinformation, and lack of understanding about the need to provide personal data, as well as privacy and security concerns, are cited as the main reasons (Sannon *et al.*, 2018). One of the ways to conceal information is the use of pseudonyms, whereby the participant can not only hide his/her identity, but might also impersonate someone else either for fun or as a joke, or for criminal reasons (harassment). Małgorzata *et al.* (2018) find that amusement is the main motivation for supplying false information. Also, when companies request a large amount of information this can generate distrust, and Keusch *et al.* (2019) showed that users feel more confident if data collection is limited to the minimum necessary and, moreover, if data protection rules are clearly explained. Although financial incentives also help, they are not completely decisive (Keusch *et al.*, 2019). Other authors, such as Sullivan *et al.* (2019), find that clearly describing the purpose of requesting information helps to prevent users from worrying that their privacy might be in jeopardy. Based on the evidence gathered, this study poses the following research question:

RQ4. What are the main motivations for intentionally entering incorrect online data, i.e. to generate disinformation?

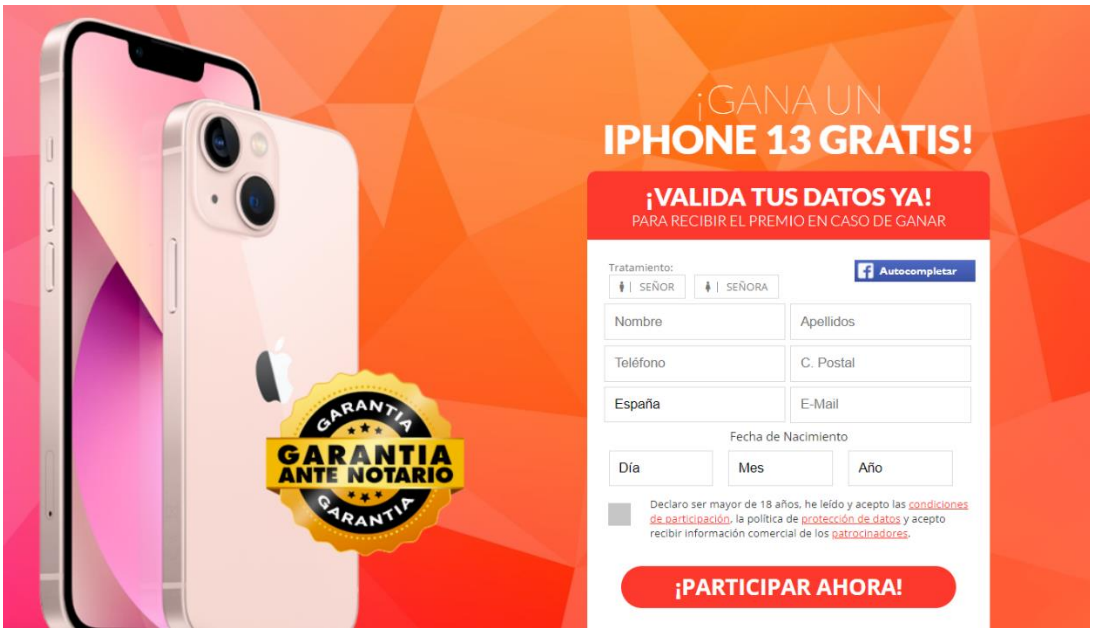
3.3. Overview of the studies

To address the research questions, this study has used triangulation, which is the combination of different methods to study the same phenomenon (Denzin, 1978). We used three methodologies in our two studies: Study 1 is quantitative and descriptive, and is used to estimate the volume of cheaters by stated sex and cohort of the database made up of the information provided by volunteer participants in online sweepstakes and tests. Meanwhile, Study 2 is mixed, combining qualitative exploratory research (2a) and quantitative descriptive research (2b), to determine the weight of the main motivators declared in Study 2a. Specifically, Study 1 used AI algorithms to estimate the amount of erroneous and falsified data in a sample provided by the lead generation company of 5,534,702 participants in online sweepstakes and quizzes between 2010 and 2021. Study 2 aimed to explore and estimate the main motivators for intentionally falsifying data provided to sweepstake sponsors. To this end, in the first stage, the exploratory Study 2a consisted of 33 in-depth interviews with participants to enquire about the main motivators for falsifying data and, in the second stage, the descriptive Study 2b used a choice-based conjoint analysis methodology with a sample of 269 participants to estimate the weight of the factors revealed in the first stage.

3.3.1 Study 1

In order to build a profile of cheaters, a descriptive analysis was proposed of certain fields of the database provided by the lead generation company after signing a confidentiality agreement. This database contains information provided by participants in sweepstakes (96%) and self-assessment quizzes (intelligence, geography, cooking, etc.) (4%) over a period of eleven years, from 2010 to 2021, and was collected through the use of landing pages (Figure 9 shows an example) that offer the possibility of winning an iPhone in exchange for the participant providing personal information. Other examples of landing pages can be found on the company's own websites:
(https://www.sorteopremios.com, https://www.mitest.de).

Figure 9. Example of a sample data collection form from
www.sorteopremios.com



¡VALIDA TUS DATOS YA!
PARA RECIBIR EL PREMIO EN CASO DE GANAR

Tratamiento:
SEÑOR | SEÑORA

Nombre	Apellidos
Teléfono	C. Postal
España	E-Mail

Fecha de Nacimiento

Día	Mes	Año

□ Declaro ser mayor de 18 años, he leído y acepto las condiciones de participación, la política de protección de datos y acepto recibir información comercial de los patrocinadores.

Source: www.sorteopremios.com. Retrieved: April 2022

On average, each user takes 3 minutes and 47 seconds to enter his/her data. To comply with the European Data Protection Regulation and the respective Spanish legislation, LOPD-RGPD, "Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y Garantía de Los Derechos Digitales" (2018), all users accessing the landing page were previously informed. They also had to opt-in by checking the consent box to agree to the different purposes for which their data was collected, declare that they were over 18 years old, confirm that they had read and accepted the entry conditions and data protection policy, and agree to receive commercial information from the sponsors.

Participation in the company's sweepstakes or quizzes does not entail any entry barrier, as the only requirement is to be over 18 years old and have an Internet connection, this latter requirement being met by more than 90% of the Spanish population (INE, 2020). The company has 5,534,702 registered users and it provided us with information on their names and surnames, emails, telephone numbers, and declared sex and ages. Data from users who did not declare their sex was not included in this study as there were only 55 such cases, amounting to just 0.001%. All users said they were of legal age and Spanish.

Based on their stated sex and age, the sample was divided into women and men, and into five generational cohorts: Silent Generation, Baby Boomers, Generation X, Millennials and Generation Z. The sample is somewhat asymmetrical, as there are more women, 65%, than men, 35%. The distribution of the cohorts is also heterogeneous, although all subsamples are representative (Rao *et al.*, 2021). Generation X (52%) is the most highly represented, followed by Millennials (40%), Baby Boomers (7%), Silent Generation (0.8%) and, finally, Generation Z (0.2%). Furthermore, in all cohorts, except for the Silent Generation, the relative frequency of women was almost double that of men.

3.3.1.1. *Measures*

Different procedures were followed to estimate the number of errors made unintentionally, misinformation, and intentionally, disinformation, in each of the available fields of the database. These ranged from developing a debugging algorithm for name and surname, to comparisons with official databases, verifications by means of chatbots, and automatic forwarding to registered email addresses. With all these tracking and control mechanisms, estimates of the amount of fraudulently entered data could be calculated.

Starting with the name and surname records, a debugging algorithm was developed from Node.js®, with which JavaScript is used to write command line tools (Escobar-Jeria *et al.*, 2007). This algorithm, which we cannot publish for copyright reasons, compares all the names and surnames registered by users who played sweepstakes and self-assessment quizzes with those in the databases of the repository of the National Statistics Institute and the IDA-Padrón, and detects all those words that do not match Spanish names and/or surnames or which appear fewer than 20 times in the country (or 5 per province). Once all records of suspicious names and surnames had been detected, an algorithm was applied to them to detect typographical errors, which are considered unintentional, while all other unusual names and surnames were considered fraudulent. In addition, to estimate the goodness of fit of the distribution of fraudulent names (Chi-square test), this was compared to the distribution of names in the company's standardised tables and blacklists. Subsequently, this frequency distribution of unusual names was incorporated into the algorithm as a contribution to machine learning. This means that the name registration software will not allow users to register names that have previously been identified as fraudulent.

Regarding the telephone number and e-mail address fields, the registration system is double-entry, which prevents typographical errors. Hence all errors made when entering this information were considered fraudulent. In order to detect bogus telephone numbers, even when they present a valid format according to the Spanish National Commission for Markets and Competition (2021), interactive voice response (IVR) systems were used (Dillman *et al.*, 2009). Call control samples were carried out automatically, by means of chatbots, and manually (control calls) to confirm that the supplied data exists and is valid.

Finally, for email address registration, verification simply consisted of sending automated messages and checking whether they were opened, click rates and other metrics. The bounce rate was measured to estimate fraudulent email addresses, aggregating soft bounces and hard bounces (Poulos *et al.*, 2020). While hard bounces occur when the e-mail indicator is incorrect and/or the user's name before the @ is false, soft bounces occur when, for example, a user cannot receive emails because their inbox is full, the sender's address has been blocked as spam, or the mail server is temporarily down (Maaß *et al.*, 2021).

3.3.1.2. *Analysis and results*

Having estimated the fraudulent data entered by users in the different fields of the database following the different procedures outlined above, an analysis was performed in different stages to determine the most fraudulent profiles. Following Saunders *et al.* (2009), we compared the differences between generations and declared sex for name and surname, telephone number and e-mail address information. We used the χ2 test (Chi-square) and then paired t-tests on the distributions and observed significant differences in terms of the results.

We estimated that 325,096 users included one or more errors in their registrations, representing 5.87% of the over 5 million unique users. Regarding the registration of names and surnames, the results show that 268,980 users (4.86% of the total) intentionally supplied fraudulent information, disinformation, a much higher figure than that for those who made errors due to inattention, misinformation, 55,903 (1.01% of the total). Hence, in response to RQ1, we conclude that intentionality and, therefore, the generation of disinformation is the main reason for errors.

In response to RQ2 and RQ3, the results also suggest different inclinations to provide incorrect information among the five generational and sex cohorts. Generation Z and Silent users are found to proportionally make the most unintentional errors and Millennials make the fewest. These are also the cohorts that make the most (Generation Z and Silent) and fewest (Millennials) intentional errors (see Table 17).

Table 17. Distribution of unintentional and intentional errors for name and surname compared to sample total

Type of error	Silent Generation	Baby Boomers	Generation X	Millennials	Generation Z	Total
Unintentional errors	1,038 (*421*) 59.4%	6,996 (*3,871*) 44.6%	30,456 (*28,887*) 5.1%	16,955 (*22,630*) -33.4%	458 (*91*) 79.9%	55,903
Intentional errors	5,743 (*2,027*) 64.7%	34,880 (*18,627*) 46.5%	131,426 (*138,994*) -5.7%	88,404 (*108,889*) -23.1%	8,527 (*441*) 94.8%	268,980
Total sample	41,710	383,282	2,860,029	2,240,587	9,094	5,534,702

Notes: In each cell: top figure, absolute frequencies; in brackets, expected values; and percentage deviation from the expected value. Unintentional errors χ^2 (4) = 6393.05, p = .000; Intentional errors χ^2 (4) = 173165, p = .000.

Source: Own elaboration

Regarding the comparative analysis by stated sex and generation, the male Silent Generation (72%), followed at a considerable distance by the male Baby Boomers (57%), are far more likely to make mistakes due to inattention. Similar figures were observed for disinformation among males: Silent Generation (69%) and Baby Boomers (58%). However, among females, it is the younger cohorts, Millennials (51%) and Generation X (50%), who have a slightly higher tendency to make errors due to inattention and these generations also make slightly more intentional errors, although Generation Z (58%) does so the most. To summarise, as shown in Table 18, men in the older cohorts and women in the younger cohorts are most likely to provide incorrect data both by accident and intentionally.

Table 18. Frequency distribution, expected frequency and relative frequency of inattentive and intentional errors for names and surnames by sex and generations

Errors due to lack of attention

Sex	Silent Generation	Baby Boomers	Generation X	Millennials	Generation Z	Total
Male	**745**	**4,014**	15,231	8,346	226	28,562
	(530.3)	*(3,574.4)*	*(15,560.6)*	*(8,662.7)*	*(234.0)*	*(28,562)*
	72%	57%	50%	49%	49%	51%
Female	293	2,982	**15,225**	**8,609**	**232**	27,341
	(507.7)	*(3,421.6)*	*(14,895.4)*	*(8,292.3)*	*(224.0)*	*(27,341)*
	28%	43%	50%	51%	51%	49%
Total	1,038	6,996	30,456	16,955	458	55,903
	100%	100%	100%	100%	100%	100%

Notes: In each cell: top figure, absolute frequencies; in brackets, expected values; and, below, relative frequency percentage. Unintentional errors χ^2 (4) = 326.70, p = .000. In bold, significant differences compared to the total.

Intentional errors

Sex	Silent Generation	Baby Boomers	Generation X	Millennials	Generation Z	Total
Male	**3,969.0**	**20,332.0**	64,556.0	44,156.0	3,593.0	136,606.0
	(2,916.7)	*(17,714.4)*	*(66,746.9)*	*(44,897.5)*	*(4,330.6)*	*(136,606)*
	69%	58%	49%	50%	42%	51%
Female	1,774.0	14,548.0	**66,870.0**	**44,248.0**	**4,934.0**	132,374.0
	(2,826.3)	*(17,165.6)*	*(64,679.1)*	*(43,506.5)*	*(4,196.4)*	*(132,374)*
	31%	42%	51%	50%	58%	49%
Total	5,743.0	34,880.0	131,426.0	88,404.0	8,527.0	268,980.0
	100%	100%	100%	100%	100%	100%

Notes: In each cell: top figure, absolute frequencies; in brackets, expected values; and, below, relative frequency percentage. Intentional errors χ^2 (4) = 2000.0, p = .000. In bold, significant differences compared to the total.

Source: Own elaboration

With regard to email address bounce rates, the results of the frequency analysis of hard and soft bounces (see Table 19) reveal that both the sex and generational cohorts affect the inclination to defraud. In the case of hard bounces, generated by entering invalid addresses, a higher frequency was observed among men of the Silent generation (50%) and Millennials (34%). As for soft bounces, which can be caused by the recipient's mailbox being full, the highest frequencies are also found among the Silent generation (70%) and also among Baby Boomers (24%). This reveals that older generations of men have a higher propensity to cheat when entering their email address. Among women, Generation X (71% for hard bounces and 82% for soft bounces) presents similar figures for both hard and soft bounces. The other generation of females that is most prone to soft bounces is Millennials (80%), while for hard bounces, it is Baby Boomers (70%). In the case of Generation Z, in both sexes the values were too low to be considered.

Table 19. Frequency distribution, expected frequency and relative frequency of the number of hard bounces and soft bounces of email by reported sex and generation

Hard bounces

Sex	Silent Generation	Baby Boomers	Generation X	Millennials	Generation Z	Total
Male	**52**	497	1.195	**737**	0	2.481
	(32)	*(507)*	*(1.270)*	*(672)*	*(0.30)*	*(2.481)*
	50%	30%	29%	34%	0%	31%
Female	53	**1.149**	**2.929**	1.444	1,00	5.576
	(73)	*(1.139)*	*(2.854)*	*(1.509)*	*(0.70)*	*(5.576)*
	50%	70%	71%	66%	100%	69%
Total	105	1.646	4.124	2.181	1	8.057
	100%	100%	100%	100%	100%	100%

Notes: In each cell: top figure, absolute frequencies; in brackets, expected values; and, below, relative frequency percentage. Hard bounces of email χ^2 (4) = 33.5953, p = .000. In bold, significant differences compared to the total.

Soft bounces

Sex	Silent Generation	Baby Boomers	Generation X	Millennials	Generation Z	Total
Male	**14**	**30**	80	45	3	172
	(4)	*(26)*	*(94)*	*(47)*	*(1)*	*(172)*
	70%	24%	18%	20%	60%	21%
Female	6	96	**373**	**182**	2	659
	(16)	*(100)*	*(359)*	*(180)*	*(4)*	*(659)*
	30%	76%	82%	80%	40%	79%
Total	20	126	453	227	5	831
	100%	100%	100%	100%	100%	100%

Notes: In each cell: top figure, absolute frequencies; in brackets, expected values; and, below, relative frequency percentage. Soft bounces of email χ^2 (4) = 37.7187, p = .000. In bold, significant differences compared to the total.
Source: Own elaboration

With regard to the number of fraudulent telephone numbers entered, there are also differences between sex and generation of users (see Table 20). Again, males from older cohorts (Silent Generation, 66%, and Baby Boomers, 51%) and, in this case, younger cohorts (Generation Z, 48%) are more likely to make mistakes when filling in their phone number. Among women, it is the middle-aged cohorts (Generation X, 55%, and Millennials, 54%) who have a higher propensity to disinform in this field. The greatest divergence between men and women is between the Silent Generation (with significant male participation) and Generation X (with high female participation).

Table 20. Frequency distribution, expected frequency and relative frequency of the number of telephone number errors by reported sex and generation

Sex	Silent Generation	Baby Boomers	Generation X	Millennials	Generation Z	Total
Male	**1,406**	**5,974**	35,898	27,502	**200**	70,980
	(990)	*(5,501)*	*(36,770)*	*(27,523)*	*(196)*	*(70,980)*
	66%	51%	45%	46%	48%	47%
Female	723	5,852	**43,147**	**31,666**		81,609
	(1,139)	*(6,325)*	*(42,276)*	*(31,645)*	*(225)*	*(81,609)*
	34%	49%	55%	54%	52%	53%
Total	2,129	11,826	79,045	59,168	421	152,589
	100%	100%	100%	100%	100%	100%

Notes: In each cell: top figure, absolute frequencies; in brackets, expected values; and, below, relative frequency percentage. Errors in phone χ^2 (4) = 440.9970, p = .000. In bold, significant differences by sex compared to the total.
Source: Own elaboration

In general, in the fields studied, there continues to be a tendency for older male generations (Silent Generation and Baby Boomers) and middle-aged female generations (Generation X and Millennials) to cheat when registering their personal data.

3.3.2. *Study 2a*

In order to discover users' main motivations for entering false information, disinformation, a sample of regular users of these online sweepstakes residing in Barcelona was invited to attend an in-depth interview at a central location, with a €50 cheque being offered as an incentive. This qualitative in-depth interview technique is used when a researcher wants to get a clearer idea about a phenomenon or when prior information is insufficient (Yin, 1994). To recruit participants, 650 telephone calls were made, of which 293 were answered, 163 expressed an intention to participate, and 33 were finally selected (16 stated male (M) and 17 female (F)). The saturation criterion was employed to select the sample size, i.e. the sample recruitment process ended when no new information was received from new sampled units (Lincoln & Guba, 1985).

3.3.2.1. *Data collection process*

Once participants arrived at the venue, they were welcomed, given a brief introduction to the study and told about the economic and social consequences of introducing false information in online communication. In order not to condition responses, we followed Sannon et al’s (2018) procedure of downplaying the importance of socially reprehensible behaviour. The respondents were told that we did not view the use of lies or falsification of data as good or bad, but that we were simply interested in analysing an important part of human communication. In order to contextualise the participants in the topic of the study, the interviewer presented a series of examples of false information supplied by participants in online sweepstakes and which had been collected by the lead generation company. The interviewees were then asked to try to explain the motives that might have led the entrants to provide incorrect data. After completing the consent form, the interviews were recorded. The interviews lasted an average of about 32 minutes.

3.3.2.2. *Data analysis and results*

The recorded data was transcribed and analysed sequentially following the principles of thematic analysis (Braun & Clarke, 2019). Coding was performed in three stages: First, the transcripts were read, the interesting sections were highlighted, and annotations were added in the margins. Second, the interesting parts were openly coded, and 24 codes were identified. Third, the codes were grouped hierarchically into a three-order structure. This axial condensation process (Tuomi *et al.*, 2021) ultimately resulted in three main themes: privacy concerns as a consequence of asking for too much information; trust in the company or website providing such quizzes and sweepstakes, and amusement.

To check the analytical consistency of the coding process, the codebook, the descriptions of each main theme, and selected paragraphs from the interviews were emailed to an independent reviewer for recoding. Following the instructions of Tuomi et al (2021), this reviewer was not connected in any way to the research and also came from a different university background (Coder: 37 years old, Computer Engineer). The reliability between the two proposed codifications was determined by Cohen's Kappa, indicating very good inter-coder agreement (>0.80) (Landis & Koch, 1977).

Regarding the results, the participants stated that they use online sweepstakes and quizzes as a source of entertainment, and that transcription errors (‘typos’) are indeed common, as this is the least entertaining or interesting part of the activity. They also comment that intentional mistakes (giving a different name to their own) are made in order to preserve their anonymity. In addition, issues such as the topics of greatest interest (history, geography, celebrities, music, etc.) were raised, including whether the prize was more or less attractive. Regarding the motives for providing false information, the different topics were grouped into three categories that were labelled Privacy, Trust and Amusement.

(1) Privacy. Respondents express concern about the loss of anonymity and that websites ask for too much information, which conveys a sense of risk.

(2) Trust. Participants expressed some doubts as to who is sponsoring the online sweepstakes and tests. It was commented that advertising should offer guarantees that it is safe and should also engender trust. There was consensus that the site from which data is requested is important. If it belongs to a public body, so much the better.

(3) Amusement. Some users impersonate the names of acquaintances for fun. There is also talk of minors, whose participation is not allowed by the system, so they do so by entering false information.

The results of the categorisation from the open coding are shown in Table 21.

Table 21. Interview results and categorisation from the open coding

Themes	Description	Sample Quotations	Listing Key Phrases
Privacy	On the Digital Marketing side, the privacy component that is considered by users is feeling that their privacy might be jeopardised (Sannon *et al.*, 2018)	"At first you get really excited when you see the prize, but then you think, why do they need all this data to give me the prize? I don't mind giving my email address but my phone number!" A, 32 "Why do you ask me for so much information, and what use will you make of it?" R, 43 "Could it be a scam?" O, 51	Too much information Excessive amount of information requested Risk
Trust	Trust refers to the data that users must provide to sponsors and raises questions about who is sponsoring online sweepstakes (Lwin *et al.*, 2016)	"I can understand why Social Security asks for your data, such as your national ID number, but why does a private entity need it? Either you are very clear that it is a necessary requirement to obtain the prize, and they guarantee me security, or I will only offer my valid email address, the rest of the data will be invented" J, 22 "Who is behind the sweepstake, can I trust them, will it be a scam?" D, 27 "Is it worth giving all this information for the prize I might possibly get?" R, 37	Phishing Hackers Security Possible Benefit vs. risk
Joke	Impersonating other people is form of amusement (Małgorzata et alt.,2018)	"You put the name of an acquaintance for a laugh, you hope they call and that it will be a surprise" R,19 "Minors, who cannot enter because the system does not allow it, can impersonate adults" X, 18 "Surprise a friend" A, 24 "I often get bored and don't know what to do with my time, so I enjoy playing jokes" N, 18	Playing pranks Kill boredom Waiting times

Source: Own elaboration

3.3.3. *Study 2b*

Based on the results obtained from the qualitative study, a quantitative study was used to measure the importance of the three factors revealed by the thematic content analysis (Braun & Clarke, 2019). Since the aim was to measure the weight of factors of socially reprehensible behaviour (Sannon *et al.*, 2018), instead of asking direct questions, a decomposition methodology (conjoint analysis) was used to estimate the users' preference structure.

This consists of forming scenarios by combining the motivations that arose from the exploratory research (privacy, trust and amusement) and asking participants to choose the scenario that best identifies them.

3.3.3.1. *Process of data collection, measurement and analysis*

Discrete choice-based decomposition methods require five steps:

(1) Determine the number of factors and levels. In this study, three factors at two levels were considered: F1 (Information: +1 excessive information, -1 not too much information), F2 (Distrust: +1 high distrust, -1 trust), F3 (Amusement: +1 for fun, -1 not for fun).

(2) Create the experimental design. Considering three factors at two levels, the number of possible scenarios is 2 3 = 8. However, instead of asking participants to compare 8 scenarios and choose the one that best identified with them, as is usual in classical conjoint analysis, we used an adaptive conjoint analysis (ACA) design consisting of twelve blocks of two profiles (Huertas-García *et al.*, 2016). Each respondent was randomly assigned a scenario consisting of three blocks of two profiles each (e.g. block 1 consisting of profiles 1 and 7, block 2 consisting of 2 and 5, and block 3 consisting of 5 and 8). In total, four 3-block scenarios with two profiles each were assigned (24 profiles). From each block, the respondent had to choose one of the two, so three pieces of information were collected from each respondent to allow estimation not only of the weight of the main factors but also between two-factor interactions. This ACA experimental design was proposed by Huertas-García *et al.* (2016) and a practical application can be found in Perdiger *et al.* (2019).

(3) Develop the appropriate question to elicit the choice in each choice set. The proposal was: "Imagine that you are participating in an online quiz and you have to fill in the data shown below (Figure 9) in order to win the prize. Which of the following sentences best describes your opinion regarding the supply of false information?" An example choice set is: "Please choose only one of the following options":

> Option 1 (+ 1 excessive information; -1 confidence; -1 not for fun). "Because an excessive amount of information is requested, although I trust the site, and I do not create false names for fun".

Option 2 (- 1 not too much information, + 1 great mistrust, - 1 not for fun). "Because, although the amount of information requested is not excessive, I am very suspicious of the site, and I am not in favour of creating false names for fun".

Option 3. None of the options identifies me.

(4) Implement the choice sets following the experimental design with a sample of consumers. A purposive sampling strategy was used by sending emails and using Google Forms to create and share online forms and analyse responses in real time. 13,500 emails were sent to regular users of the online sweepstakes inviting them to complete the questionnaire and encouraging them to enter the IPhone 13 sweepstake, of which 2,929 were opened, 336 questionnaires were completed, and 269 were valid.

(5) Analyse the data with an appropriate analytical model. The results were estimated using the Multinomial Logit Model (Rao, 2014). Data was collected in May 2022.

3.3.3.2. *Results*

Table 22 summarises the results and shows the weight of main factors and two-factor interactions that motivate the supply of false information.

Table 22. Relevance of the factors that motivate the introduction of false information obtained through statistical regression inference.

Interception	*Coefficients*	*Standard error*
F1	**2,60359717857956*****	0,827215655
F2	**1,94519835354427****	0,827215655
F3	**2,90268521016663*****	0,827215655
F12	-1,960548827	1,169859598
F13	-1,439019901	1,169859598
F23	-1,44745917	1,169859598
Coefficient of determination R^2	0,808261437	
Standard error	1,547578782	

**p<.05 *** p< .01. F1 means "Not trusting enough", F2 = "Safeguarding one's privacy" and F3= "Amusement"; F12 means the interaction between F1 and F2, and so on subsequently.
Source: Own elaboration

The results show that the main motivation for users to enter false information was amusement, F3 (Amusement: +1 for fun), followed by not having enough trust in the company's website, F2 (Distrust: +1 high distrust), and, finally, the desire to maintain their privacy and considering that too much information was being requested, F1 (Information: +1 excessive information). Furthermore, the results indicate that the three factors act independently, as no interaction between two factors reached significant values. Therefore, the response to RQ4 on the main motivations for entering false information are: amusement, lack of trust in the company's website, and the desire to maintain one’s privacy and not reveal an excessive amount of information.

Recently, the problems generated by the proliferation of misinformation and disinformation on social networks, and the need to detect it, have attracted a great deal of attention (Di Domenico *et al.*, 2020; Pascual-Ezama at al., 2020). Existing approaches to cheater detection are mainly based on the use of certain user characteristics, such as unusual names, offensive words, and non-existent phone numbers or email addresses, and the configuration of blacklists of users, which artificial intelligence algorithms detect quickly and accurately (Saura, 2021; Zhang *et al.*, 2020). However, knowing which user profiles are more inclined to misinform can boost the performance of these bots. In this study, it has been detected that men of older generations and women of younger generations are more likely to falsify their data. In addition, the main reasons for this socially reprehensible behaviour are fun, lack of trust in the website requesting the data and safeguarding privacy. Identification of cheaters and their motivations can help academics and practitioners to try to improve methods for capturing information, and also ways of detecting cheaters on social networks.

3.4. General discussion and conclusions

The emergence of social networks and the information flows generated between them have created an enabling environment for digital marketing. However, it is not easy to synthesise the enormous volume of information that circulates on networks in a way that can help academics and digital marketers to make decisions. One of the ways to analyse relevant information is to use databases of potential consumers collected by lead-generating companies (Desai, 2019). However, in order for these databases to fulfil their function, they must be as reliable as possible, i.e., they must contain real data that is as clean as possible of misinformation.

This study describes the profiles of users who enter false information when registering for online sweepstakes and quizzes, based on estimates of negligent, misinformation, and intentional inaccuracy, disinformation. The results suggest that most errors are made intentionally, at a ratio of almost 5:1 with regard to unintentional actions. Furthermore, men of older generations and women of younger generations are more likely to falsify their data. However, and in line with the findings of Dabija *et al.* (2018), small differences are also observed regarding the disclosure of names and surnames, emails and telephone numbers. We found that the most repeated motivations for producing disinformation were, in the following order, amusement, lack of trust in the site requesting the data, and safeguarding privacy. These results are in line with previous research, which has shown that trust is a key aspect and can be considered a predictor of whether or not the users of social networks will provide false information (Gefen *et al.*, 2003). This study furthers our knowledge about the process of capturing data from internet users, in this case by means of online sweepstakes and quizzes, and the problems arising from the volume of fraud committed by users. Indeed, we have not found any previous study that estimates and analyses such practices when users register their information on websites with such a large sample (more than five million) and over such a long period of time (eleven years). Although each of the fields requires a different method for estimating errors, there are common trends among some generational and sex profiles.

In the analysis of names and surnames, the cohorts with a higher propensity to enter incorrect data are Generation Z, Silent and Baby Boomers. However, when crossing the data with declared sex, we find that it is men from the older generations and younger women who are most inclined to misrepresent. However, actions when filling in the email address and phone number fields do not follow the exact same pattern as the previous ones, results that are in line with those obtained by Dabija *et al.* (2018). The estimation using hard bounces highlights male Silents and Millennials and female Baby Boomers and Generation X as the most fraudulent. Finally, using call-backs, male Silents and Baby Boomers and female Generation X and Millennials were found to be the most fraudulent generations.

Theoretical implications

While this research provides evidence of and support for the tendency of users to enter fraudulent information on social networks (Islam *et al.*, 2020; Pennycook & Rand, 2019), we also find that this occurs in less than 6% of cases (and less than 5% for intentional errors). However, while there is room for improvement in mechanisms to reduce unintentional errors, mechanisms to control for intentional errors should be directed towards cheater profiling (Cosmides & Tooby, 2016).

This study presents evidence that some user profiles are more inclined than others to enter false information when registering on the Internet, so their identification can help predict such behaviour and target measures to control these practices better (Song *et al.*, 2021). For example, it would be logical to assume that older people (Silent and Baby Boomers) are more inclined to make unintentional errors than younger people, as they are more affected by deterioration in physical condition and cognitive abilities (problems with sight, memory loss, difficulty typing letters correctly on a keyboard, etc.). However, this is only true when they are compared with middle-aged generations, but not with younger people (Generation Z) who are the most inclined towards such practices. Maybe, although young people are more accustomed to the Internet, they write in a hurried manner without checking that the information is correct (Valentine & Powers, 2013). Regarding intentional errors, one might assume, on the one hand, that more mature people, with more life experience and who have adopted these technologies much later, use them for a clear purpose and to obtain a specific outcome (Dabija & Grant, 2016). However, the results only partially confirm this assumption, as older, self-trained male cohorts tend to be more likely to enter false information. On the other hand, nor does the assumption hold that younger generations, who were born in the age of the Internet and social networks, behave differently to other generations, for the results of this study do not point in that direction (Lenhart *et al.*, 2010). In fact, in the analysis of the name and surname fields, younger users behave similarly to older generations.

Implications for management

Although the introduction of false personal data does not occur in alarming proportions, it does affect both individual users and businesses (Shu *et al.*, 2020). Given that the proliferation of cheaters is inversely correlated with good practices in tacit or explicit negotiations (Axelrod & Hamilton, 1981), it is important to detect them in order to prevent and eliminate fraud. Moreover, online environments with a large number of users facilitate such practices (Allcott & Gentzkow, 2017). Although psychological mechanisms have been developed in offline environments to dissuade cheaters (Mealey *et al.*, 1996), these mechanisms are not directly transferable to online environments, so AI and technology play a key role in developing devices to mitigate the consequences of fraudulent information (Zhang & Ghorbani, 2020). Tackling these problems creates opportunities in the innovation and development of tools for detecting, preventing and monitoring potential fraud, with significant economic benefits through value creation and capture. Therefore, having a clear profile of cheaters as well as knowledge of their motivations for cheating can be very valuable (Nambisan *et al.*, 2019), as it helps to filter the information that is fed into the databases used by companies and decision-makers, and directly affects the outcome of their decisions (Zhang et al., 2016; Ogilvie *et al.*, 2017; Bondarenko *et al.*, 2019; Lin et al., 2021). The determination of the most common cheater profiles (in terms of generation and stated sex) can help to filter databases so that companies can offer better personalised services to their customers (Zhang et al., 2016), preventing them from receiving information that is of no interest to them (Agrawal *et al.*, 2011), and instead increasing the likelihood of making attractive offers and maximising returns (Zhang et al., 2016). For companies, more reliable databases will improve productivity (Lin et al., 2021), ensure they do not miss out on business opportunities (Bondarenko *et al.*, 2019) and, ultimately, raise their profits (Tripathy *et al.*, 2013). In turn, this will increase employee satisfaction, as they will achieve better sales, loyalty and more personalised customer services (Ogilvie *et al.*, 2017). In short, our research helps companies to develop more targeted and effective communication strategies, which will have a positive impact on customer value and loyalty, as well as on the company's profits.

Limitations and future studies

The results of this study were based exclusively on the data contained in the database provided by the lead generation company. However, they would need to be validated against data supplied by other such companies (Jung *et al.*, 2020), as well as other – even unstructured – data on user behaviour (Choudrie *et al.*, 2021). It would also be interesting to contrast the results with other web data collection formats (Cruz-Benito *et al.*, 2018). Also, the data was analysed globally without taking into account recruitment sources or methodologies, or different origins and social networks (Parekh *et al.*, 2018). This additional information could enrich studies in this field. As indicated by authors such as Borges-Tiago *et al.*, (2020) attitudes differ depending on the country that users come from. Our data was collected in Spain, and it remains to be seen whether its conclusions can be extrapolated to other countries and cultures and whether future generations will continue to behave in the same way (Altman and Bland, 1998).

There is also no evidence of exploratory research into the sectoral clustering of profiles that enter their data online and whether there are differences in behaviour by generation or declared gender. It would be especially interesting for future research to examine how different profiles behave in terms of decision-purchase-post-purchase behaviour. It would also be useful to study the clustering of consumer profiles by sector and thus analyse how the resulting algorithm is affected by the false information entered, which would to help to devise mechanisms to correct or eliminate such practices.

The particular casuistry of cohorts that are more prone to unintentional errors, such as older people who are more affected by health conditions and accessibility issues, leads to an ethical debate that could be explored further, namely that on mechanisms to avoid penalising the participation of these older users just because they might find it harder to read, write or remember information. In other words, it would be very interesting to look in depth at the ethical implications of systematically excluding or limiting the participation of certain users in prize draws and tests, simply because they may make mistakes due to health conditions, and to investigate why female members of the same cohorts do not seem to be affected by such difficulties to the same extent.

CHAPTER 4. MARKET SEGMENTATION METHODS: A COMPARATIVE ANALYSIS BETWEEN SUPERVISED AND UNSUPERVISED LEARNING

Abstract

One of the basic instruments for marketing planning is segmentation into groups that are as homogeneous as possible, and one of the usual techniques for their division is clustering. However, technological changes that have driven digital marketing have allowed unprecedented amounts of data to be collected, which traditional techniques have difficulty analyzing. The purpose of this research is to address this challenge by proposing the use of two AI algorithms, a supervised algorithm based on a hierarchical decision tree structure and an unsupervised clustering algorithm, to segment large databases of lead-gathering companies and compare their effectiveness. For this purpose, the XGBoost algorithm has been considered as supervised algorithmic methods and K-means as unsupervised. This experiment was carried out with a sample of 5 million Spanish users captured between 2010 and 2022. The results show that supervised learning with this type of data is more useful for segmenting consumer markets than unsupervised learning, as it provides more reliable, optimal, and cost-effective solutions.

Keywords: Social media, Clusters, Unsupervised algorithms, Supervised algorithms, XGBoost, K-means, Lead generation.

4.1. Introduction

Market segmentation is one of the fundamental phases in today's strategic marketing (Liu *et al*., 2010), as it allows companies to divide the market into homogeneous subsets and focus their efforts on specific customer groups, thus increasing the effectiveness of their marketing policies (Stead *et al.*, 2007). The segmentation process and targeting are among the most studied activities in academic literature, as well as applied by marketing practitioners (DeSarbo & Grisaffe 1998; Wedel & Kamakura, 2000).

The need to segment the market responds to its divergent and heterogeneous nature. In fact, Smith (1956, p.6) used these terms in his definition of market segmentation: "Market segmentation, on the other hand, consists of viewing a heterogeneous market (one characterized by divergent demand) as a number of smaller homogeneous markets in response to differing product preferences among important market segments." Very soon, however, the descriptive technique became prescriptive. Not only was the market divided into more homogeneous groups, but segmentation became a target in itself, allowing marketing policies to be applied more effectively, such as designing a personalized communication campaign directed at a particular segment to contribute to the positioning of a product (Myers, 1996; Liu *et al.,* 2010).

Grouping algorithms are usually used for constructing market segments, and undoubtedly the most popular among marketing researchers and practitioners are clustering techniques (Wedel & Kamakura, 2000). The application of clustering algorithms requires the collection of data on some attributes of consumers such as their demographics, purchasing habits and product preferences (Kaufman & Rousseeuw, 2009), and aims to group consumers into as homogeneous groups as possible around a centroid, and, in turn, the centroids maintain a sufficient distance between them so that they can be considered distinct groups (Wedel & Kamakura, 2000). However, in other to obtain valuable consumer groups, it is essential that the data are as reliable as possible, that they have been cleaned of false or erroneous information (Shu *et al*., 2017), and that the resulting groups are as homogeneous as possible (Boone & Roehm, 2002).

In recent years, with the development of new technologies, the internet and social networks have created an environment of social interaction where everything is recorded, offering the possibility of collecting a multitude of information of different nature, from multiple sources (websites, blogs or posts), but at the same time very noisy (Ali *et al.*, 2022). This environment poses a challenge to market researchers, who find that traditional clustering algorithms are not adapted to working with such complex and noisy databases (Wedel & Kamakura, 2000).

Although there are numerous clustering techniques and methodologies, for example, Milligan & Cooper (1985) considers that the five dominant ones are Forgy's method, Jancey's method, MacQueen's method (K-means), the convergence method, and the Exchange algorithm, there is a need for new clustering techniques capable of generating effective segmentation solutions with information from multi-source, multi-natured, noisy, and data-rich markets (Boone & Roehm, 2002). It has been suggested that the use of AI algorithms may be best suited to analyse and classify such large databases (Ezugwu *et al*., 2022), and, in this study, two of them will be considered: supervised and unsupervised clustering algorithms. Both types have been considered, since there is no a priori consensus on the most appropriate algorithms for clustering large databases. On the one hand, some researchers have pointed out that the computational burden involved makes the use of unsupervised techniques more appropriate (Punj & Stewart, 1983; Wedel & Kamakura, 2000), but, on the other hand, other authors consider that supervised machine learning algorithms would be the most appropriate (Tukey, 1962; Kaufman & Rousseeuw, 2009). Finally, other researchers note that the most appropriate algorithm for clustering subjects depends on the characteristics of the database to be analysed (Arabie *et al*., 1996; Wedel & Kamakura, 2000).

The main objective of this article is to shed some light on the problems outlined above. Respectively, we propose to analyse and compare the performance of two algorithms, one supervised, XGBoost, and one unsupervised, K-Means, in clustering a lead-generated database. Although XGBoost, proposed by Chen and Guestrin (2016), is an AI algorithm that has been applied to a wide variety of engineering problems (e.g., Chen *et al.*, 2021), its application in marketing has been much less frequent and, although it is not a specific clustering algorithm, it offers this possibility when working with defined labels (Liang *et al*., 2019).

This study aims to fill this gap in the literature, using the XGBoost algorithm to cluster the data generated through leads, from a sample of over 5 million Spaniards, and compare its performance with that of the unsupervised K-Means algorithm, proposed by MacQueen (1967), and common in comparative studies as a control element (Boone & Roehm, 2002). The data analysed come from users who have registered to participate in sweepstakes and online tests and have been provided by the lead generation company CoRegistros, S.L.U., one of the leading companies in Europe, which has been operating in Spain since 2009.

It is expected that the segments created by XGBoost can provide valuable information to guide marketers in either tailoring their offer for each specific segment (Punj & Stewart, 1983), or focusing their offer on the target most likely to buy it (Yankelovich & Meer, 2006). In this way, it can use the information gathered to improve audience, identify market opportunities, optimise profitability (Cavusgil *et al.*, 2004) and evaluate performance (Ailawadi *et al.*, 2001).

In summary, using clusters allow market researchers classifying customers into homogeneous groups that led them to better understand consumers' preferences and desires. However, the origin, type, and degree of data cleansing, as well as the technique (supervised or unsupervised) and type of algorithm used, will provide different groupings with varying levels of confidence to ensure campaign effectiveness and optimize return on investment (Mobasher *et al.*, 2000).

The rest of the document is organized as follows. First, a conceptual framework focused on cluster algorithmics in marketing is presented. Second, the methods and results of the study on which this research is based are presented. After an analysis of the results, the implications for academia and management are addressed. The study concludes by proposing the key themes that emerged from the results, discussing their limitations, and suggesting certain avenues for future research.

4.2. Literature review

4.2.1. Methods for market segmentation and cluster analysis in marketing.

According to Wedel & Kamakura (2000), there are more than 50 methods for grouping data that could be used for market segmentation. One of the pioneers was Ball & Hall (1967), who defined ISODATA. This was a practical computational method aimed at grouping multivariate data to find patterns with complex interactions, whose resulting solutions were a set of cluster centroids that tended to minimize the sum of the squared distances of each piece of information with respect to the nearest centroid.

Cluster methods can be classified as nonoverlapping, overlapping, and fuzzy. Nonoverlapping analysis assigns everyone to only one cluster, overlapping allows the same individual to be in several clusters at once, while fuzzy assigns proportions of individuals to different segments (Wedel & Kamakura, 2000). However, the most common is nonoverlapping, which starts from individual data and groups them based on their similarities and differences until a single group is formed (Ezugwu *et al.*, 2021). As for the method of grouping individuals, algorithms can be hierarchical or non-hierarchical. However, previous literature has already pointed out that for processing large databases, hierarchical-based clusters are not recommended due to problems derived from the computational load they require, and the biases associated with the selection of centroids, so the use of non-hierarchical methods is recommended (Wedel & Kamakura, 2000). So far, the literature has not been able to identify a technique that generally prevails over the rest (Arabie *et al.*, 1996; Boone & Roehm, 2002; Wedel & Kamakura, 2000). For example, in an investigation carried out by Vriens *et al.* (1996) in which they compared nine segmentation methods, coming from conjoint metric, using a Monte Carlo study, they found that differences in predictive accuracy were small. Each of the methods has its own strengths and limitations (Dayan *et al.*, 2021), which indicates that for each database (depending on the information it contains), it is possible to find a technique that provides better performance than another.

As noted above, the new social interaction framework provided by the internet and social networks, where every exchange of information is recorded, offers a wide range of possibilities for market researchers to gather information about the business-customer relationship in volumes never seen before (Hoffman & Novak, 2009). Collecting and analysing unstructured information of a diverse nature, multi-sourced and noisy represents a paradigm shift (Ali *et al.*, 2023), and requires the search for new efficient heuristic methods (Liu *et al.*, 2010).

A possible solution can be sought in the adaptation of new AI algorithms for market research (Zhu *et al.*, 2016). Machine learning algorithms used for classification are usually divided into supervised and unsupervised. Supervised algorithms assume the availability of a supervisor, which is the result of training the algorithm on a collection of representative data known as a corpus, and then the trained algorithm can be applied to the dataset. For the construction of the supervisor in the training phase, the algorithm uses data vectors and label vectors and associates them by building a model that will manage the rest of the data. While the unsupervised ones do not require prior training, and therefore do not use labels, the clustering is generated by the data's own internal features (Chaovalit & Zhou, 2005).

Another challenge related to the use of AI algorithms is the need for high-quality databases. Data collected from the Internet is usually very noisy, and requires a prior screening process, which is often costly in both time and effort (Sáez-Ortuño *et al.*, 2023b). In the specific case of clustering algorithms, it is also challenging to work with very diverse audiences or made up of very varied characteristics, which makes it more complex to find common patterns among consumers (Lund *&* Ma., 2021).

4.2.2. Unsupervised and non-hierarchical machine learning techniques: the K-means algorithm.

The K-means algorithm, proposed by MacQueen (1967), is an unsupervised and non-hierarchical machine learning clustering technique used to divide a dataset into k clusters or groups of similarity. It is one of the most widely used techniques (Wedel & Kamakura, 2000) and is often used as a control algorithm in comparative studies (Boone & Roehm, 2002; Hruschka & Natter, 1999) due to its simplicity and efficiency (Kuo *et al.,* 2002).

The K-means algorithm requires the number of target clusters to be specified a priori, which may lead to suboptimal results if the data have complex shapes or if there are outliers (Voges *et al.,* 2002). It is an iterative algorithm represented by the function J, which aims to minimize the variance within each cluster at each iteration, or the quadratic error function for all points and for each cluster (see equation 1).

$$J = \sum_{m}^{i=1} \sum_{K}^{k=1} w_{ik} |x_i - \mu_k|^2$$

where w_{ik} equals 1 if the point x_i belongs to the cluster k, and 0 in any other case and μ_k it is the centroid for cluster k. Interpret: The K-means algorithm works by randomly assigning k data points, as initial centroids, and then assigning each data point to the closest cluster based on its Euclidean distance. In a second iteration, the centroids are recalculated as the mean of the data points assigned to the cluster, and the data points are reassigned. This process is repeated until the centroids no longer change or a predetermined number of iterations is reached (Lloyd, 1982).

The K-means algorithm is fast and easy to implement as it does not require a model training phase, and it assumes that clusters are circular, which can be a drawback as it may not work well for clusters of other shapes. However, as previously noted, there are no dominant and conclusive techniques in this field and the algorithm's performance varies depending on the database to be clustered (Arabie *et al.,* 1996; Boone & Roehm, 2002; Wedel & Kamakura, 2000).

Thus, the following research question is proposed:

RQ1. Are unsupervised algorithms efficient for clustering consumers captured through leads from sweepstakes and online tests?

4.2.3. Supervised and hierarchical machine learning techniques: the XGBoost algorithm.

Although they were not designed for this purpose, supervised machine learning algorithms are often used to cluster data (Mitchell & Frank, 2017; Gultom *et al.*, 2018). However, the origin of these algorithms is not clear as they have evolved over time with numerous contributions from many researchers and theorists (Amoozad *et al.*, 2022).

The earliest known studies in the field of machine learning date back to the 1950s, with the development of information theory and reinforcement learning (Samuel, 1959). Since then, this field has evolved rapidly thanks to the incorporation of new algorithms, different techniques, and numerous practical applications in various areas. For example, Platt (1998) proposed an efficient algorithm for training support vector machines, or LeCun (1989) demonstrated the use of backpropagation (a machine learning algorithm) for recognizing zip codes. Additionally, some authors, not only proposed some basic concepts, but also warned of some dangers such as bias or preference deviations in the machine learning process (Mitchell, 1997).

One of these algorithms is XGBoost (Extreme Gradient Boosting), which is used for both classification and estimation through regression (Chen & Guestrin, 2016) and is common in market research (Liang *et al.*, 2019). The algorithm works by creating a set of decision trees, called weak trees, which are then combined to create a stronger model. Due to its ability to handle large volumes of data with numerous features (e.g., sparse data), the clusters achieve performance comparable to that of more complex machine learning algorithms (Liang *et al.*, 2019).

As a supervised machine learning algorithm, XGBoost requires defined target classes or labels to train the model. The model begins by constructing a decision tree where each node is split into subnodes based on a specific feature and assigned a score, as well as a pruning threshold (Liang *et al.*, 2019). Trees are built sequentially, and XGBoost uses a technique called gradient boosting that adjusts the weights of each tree based on the errors of the previous tree (Hastie *et al.*, 2009). For clustering tasks, the "one-hot encoding" technique is used (Tang, 2020), meaning a different column is created for each target class label and then XGBoost is applied to train the function that separates the data into different clusters. This can be useful when data clusters have complex shapes or when the number of clusters is unknown a priori (Liang *et al.*, 2019).

The objective function (Equation 2) that the XGBoost algorithm minimizes in each iteration is as follows:

$$J^{(t)} = \sum_{n}^{i=1} j\left(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)\right) + \Omega(f_t)$$

where y_i is the target label of point x_i known from the dataset and $\widehat{y_i}$ is the predicted label. We can observe that the objective function J of the XGBoost algorithm is a function of functions j, which in turn is a differentiable convex loss function that measures the difference between the prediction $\widehat{y_i}$ and the target y_i. The regularized term $\Omega(f_t)$ penalizes the complexity of the model (set of trees) and helps to smooth the learned final weights to avoid overfitting. Intuitively, it tends to select a model using simple and predictive functions. To optimize this function of functions, we must do it iteratively. Therefore, we must calculate the function $J^{(t)}$ at iteration t from the prediction of labels $y_i^{(t-1)}$ in the previous iteration (t-1) and greedily add the tree $f_t(\boldsymbol{x_i})$ to the model in such a way that it improves it.

This algorithm is based on boosting (Hastie *et al.*, 2009), which consists of generating multiple sequential models of weak predictions, so that each one takes the results from the previous model, generating a stronger model with greater predictive power and greater stability in its results (Schapire, 2013). The optimization process follows gradient descent (Zou & Hastie, 2005), as the XGBoost algorithm learns from groups that maximize the difference between them using the target (conversions) as a supervisor (Chen & Guestrin, 2016). During XGBoost training, the parameters of each weak model are iteratively adjusted. Thus, as each model is compared to the previous one, if a new model has better results, it is taken as the basis for making modifications. If, on the contrary, it has worse results, it returns to the best previous model (Chen & Guestrin, 2016).

Although the XGBoost algorithm can process a large volume of data with multiple features, the fact that it requires labeling the data beforehand can be a drawback. In particular, the way in which labels are defined can condition the result and, therefore, can generate bias if not done correctly. Certainly, as previously noted, there is no technique that prevails and is conclusive in this field (Arabie *et al.*, 1996; Boone & Roehm, 2002; Wedel & Kamakura, 2000), so the following research question is proposed:

RQ2. Are supervised algorithms efficient for clustering in marketing using labeled data from giveaways and online tests?

4.2.4. Unsupervised vs Supervised training algorithms: K-means vs XGBoost.

Both K-Means and XGBoost are common clustering or classification algorithms in market research (Henriques *et al.,* 2020). However, each one has its own characteristics, and their choice will depend on the particularities of the data and the specific needs of each problem. While K-Means can be used to divide customers into groups or clusters, based on their demographic characteristics, purchase behaviors, or other factors according to the specialist's perceptions; XGBoost can be used to predict the probability that a customer will make a purchase or the value of a transaction, based on a labeled dataset to train the model (Henriques *et al.,* 2020).

Since both techniques may be appropriate for clustering conglomerates in market research, despite coming from different purposes, they do not always offer the same performance (Zhu *et al.,* 2019). Thus, the following research question is proposed:

RQ3. Which of the two algorithmic methods, supervised or unsupervised, is more efficient for clustering in marketing with data from online surveys and tests?

4.3. Overview of the study

To address the research questions, this study tests the two proposed algorithmic methods, K-means and XGBoost, to determine their effectiveness in generating homogeneous market segments and to compare their performance in reproducing the market structure. However, when analysing data collected from the market, the market structure is not known a priori, and it is therefore difficult to estimate its external validity. In these cases, the degree of effectiveness in generating homogeneous solutions and the explained variability of the groups formed is usually used as an estimator (Boone & Roehm, 2002). In this case, the real-world database consists of a sample of 5,185,857 participants in online draws and contests collected between 2010 and 2022 in Spain by a lead generation company.

4.3.1. Data set

The database was obtained after reaching an agreement and signing a confidentiality commitment with the company CoRegistros, S.L.U. The data matrix contains 37 fields (columns) from data provided by online sweepstakes participants (96%) and self-assessment questionnaires on topics such as intelligence, geography, cooking, among others (4%) collected over twelve years (2010-2022), according to information provided by the company. To reduce noise, the data was screened and cross-checked for accuracy (for more information see Sáez-Ortuño et al., 2023b). The database is made up of 37 fields (see Table 25) grouped into five blocks: 1. Users, 2. Marketing, 3. Conversions, 4. Ads, and 5. Sweepstakes (see Table 23). The users block contains descriptive data about the consumers, the marketing block describes how the user has provided their information, the conversion block contains the history of users who have become buyers by purchasing a product, the campaigns block contains the marketing actions in which the user has participated, and finally, the sweepstakes block shows information linked to the sweepstakes column. The last block establishes several links, with user table through the id_prom variable, with the marketing table through id_prom, since each marketing campaign is assigned a sweepstakes (the same sweepstakes may be assigned to different marketing campaigns). Finally, the variable that identifies the user is id_user. Table 23 describes the items that correspond to each block.

Table 23. Description of the tables provided by the company for the study

Table name in the database	Description of the content of the database table
1. users	Master table of users. Contains all fields with descriptive information about the user.
2. marketing	Master table of marketing campaigns through which users are registered. It relates to the users table through the id_m field.
3. conversions	Master table of conversions. Contains the historical data of users who have converted to a product in the past. It relates to the users table through the id_user field.
4. ads	Master table of client campaigns. These campaigns are sent to users who are registered in the database with the aim of converting them to the offered product. It relates to the conversions table through the id_ad field.
5. sweepstakes	Master table of sweepstakes. It relates to the sorteo column in the users table through the id_prom column. It also relates to the marketing table through the id_prom field since each marketing campaign is assigned a sweepstakes (the same sweepstakes can be assigned to different marketing campaigns).

Source: Own elaboration

4.3.2. Measures

The user block was used as the database to be grouped and the rest of blocks as dimensions. After a first analysis of database, it was considered that it would be relevant to know, on the one hand, the description of the different marketing campaigns that had been carried out, as well as the product that had been raffled and, in addition, more information about conversions into purchases. To do so, new information was requested from the company, which was delivered in three more blocks: 1. ads_tipo.csv, 2. clasificacion_sorteos.csv, and 3.clasificacion_conversions.csv. Table 24 shows the items contained in the blocks.

Table 24. Description of the auxiliary tables provided by the company for the study.

Name of file	**Description of block content**
1. ads_type.csv	Analyzing the ads table, the need to know the description of different campaign types was identified. To solve this issue, the ads_type file was created as a master of campaign descriptions (with a tab as a separator). This file is related to the ads table through ad_type.
2. clasification_sweepstakes.csv	Analyzing the sorteos table, the need to classify the raffles according to the raffled product was identified. To solve this issue, the clasification_sweepstakes file was created (with a tab as a separator). This file is related to the sweepstake table through id_prom. The created categories are: beauty, content, electronics, home, iPhone, leisure, test, and travel.
3. clasification_conversions.csv	Analyzing the conversions table, the need to classify the client's campaigns (id_ad) that appear in that table (i.e., campaigns that have resulted in at least one conversion) according to the final product to which each user converted was identified. To solve this issue, the clasification_conversions file was created. This file is related to the conversions and ads tables through id_ad. The created categories are: hearing aids, energy, finance, games, NGO, insurance, and telcos.

Source: Own elaboration

Finally, to complete the data, information was sought about some external variables from public sources based on the postal code, such as geographic longitude and latitude, which was incorporated into the database. With all this information, the company was asked to perform an Extract-Transform-Load (ETL) to transform the table into its final format before applying the algorithms. Finally, the variables to be considered in the models were determined and those that were transformed into Boolean logic. Table 25 indicates the list of variables, the type of variable (string, Boolean, and interval), and those that participated in the comparative study are marked with an X.

Table 25. List of final columns of the user's table.

Index	Column	Tipo	K-Means	XGB
1	**producto_conv**	String	-	-
2	**id_producto_conv**	Int (*)	-	X
3	**id_user**	Int (*)	X	X
4	**email**	String	-	-
5	**dominio_email**	String	-	-
6	**id_dominio_email**	Int (*)	-	-
7	**sexo**	String	-	-
8	**id_sexo**	Bool	X	X
9	**nombre**	String	-	-
10	**edad**	Int	X	X
11	**codigopostal**	String	-	-
12	**latitude**	Float	X	X
13	**longitude**	Float	X	X
14	**telefono**	Int (*)	-	-
15	**comp_telf**	String	-	-
16	**grupo_comp_telf**	String	-	-
17	**valido**	Bool	X	X
18	**finaliza**	Bool	X	X
19	**espactividad**	Bool	X	X
20	**estado_telf**	Bool	X	X
21	**cla_sorteo**	String	-	-
22	**id_cla_sorteo**	Int (*)	-	-
23	**dominio_email_gmail**	Bool	X	X
24	**dominio_email_hotmail**	Bool	X	X
25	**dominio_email_outlook**	Bool	X	X
26	**dominio_email_yahoo**	Bool	X	X
27	**dominio_email_live**	Bool	X	X
28	**dominio_email_msn**	Bool	X	X
29	**dominio_email_otros**	Bool	X	X
30	**cla_sorteo_belleza**	Bool	X	X
31	**cla_sorteo_contenido**	Bool	X	X
32	**cla_sorteo_electronica**	Bool	X	X
33	**cla_sorteo_hogar**	Bool	X	X
34	**cla_sorteo_iphone**	Bool	X	X
35	**cla_sorteo_ocio**	Bool	X	X
36	**cla_sorteo_test**	Bool	X	X
37	**cla_sorteo_viajes**	Bool	X	X

Source: Own elaboration

4.3.3. Methodological study of the unsupervised algorithm: K-means.

To test the unsupervised K-means algorithm with the sample data, the following steps were followed: (1) selection of the dataset, (2) data standardization (mean = 0 and variance = 1), (3) centroid selection, (4) application of the algorithm, and (5) validation and estimation of the effectiveness and efficiency of the algorithm.

First, 24 variables (2 int, 2 float, and 20 Boolean) were selected from the set of 37, and although most of the variables are Boolean, they were standardized using the Python StandarScaler library (Zamri *et al.*, 2022). Although it is not necessary to standardize binary Boolean variables, according to Stead *et al.* (2007), it is recommended to do so in clustering processes. Additionally, in this case, not all variables were binary, as there were four non-Boolean variables (Chakraborty *et al.*, 2009). That is, combining binary variables with scales or ratios is not recommended as one of them may contain higher variances than the others, and it could dominate over the remaining ones incorrectly (Stead *et al.*, 2007).

To apply the K-means algorithm, it is necessary to define the number of target groups or clusters, represented by the variable *k*. Based on the study by Kodinariya and Makwana (2013), it was considered that the number of groups should be related to the users collected by the lead, who had transformed into buyers of some product (id_producto_conv ≠ 0). The elbow rule was applied to this criterion, and five groups were considered (Likas *et al.*, 2003).

The elbow method is a rule of thumb used to select the number of clusters in a dataset using clustering analysis. According to Han *et al.* (2011), it consists of observing the distribution curve of the explained variance as a function of the number of clusters and choosing the point where a significant decrease occurs. As mentioned by Jain and Dubes (1988), to apply this rule the researcher should plot the sum of intra-cluster distances as a function of the number of clusters and observing where an elbow occurs in the plot. In other words, where the figure of a elbow is reproduced on the graph (Milligan & Cooper, 1985).

Once the number of centroids was selected, the K-means algorithm from the Python library was applied as follows: (1) The *k* centroids are initialized at random coordinates; (2) The distance between each user and each centroid is calculated, and each user is grouped around the nearest centroid based on the minimum distance between the points and the centroid; (3) The centroids are updated by recalculating their new position, and steps (2) and (3) are repeated; (4) The process stops when the stopping criterion is reached, which in this study occurred when the centroids stopped changing (Likas *et al.*, 2003).

Next, the results were analyzed to verify their validity and effectiveness. Since labels were available, we could calculate the accuracy to determine whether K-means had "matched" the labels of the target class. In general, these groups will be considered good if *k* has been chosen correctly.

4.3.3.1. Results

To determine the number of segment groups, a principal component analysis was carried out. The results are shown in Figures 10 and 11, which show up to 12 principal components (PC) and served as a guide for selecting the optimal number of PCs.

Figure 10. Variance explained by the principal components (PC)

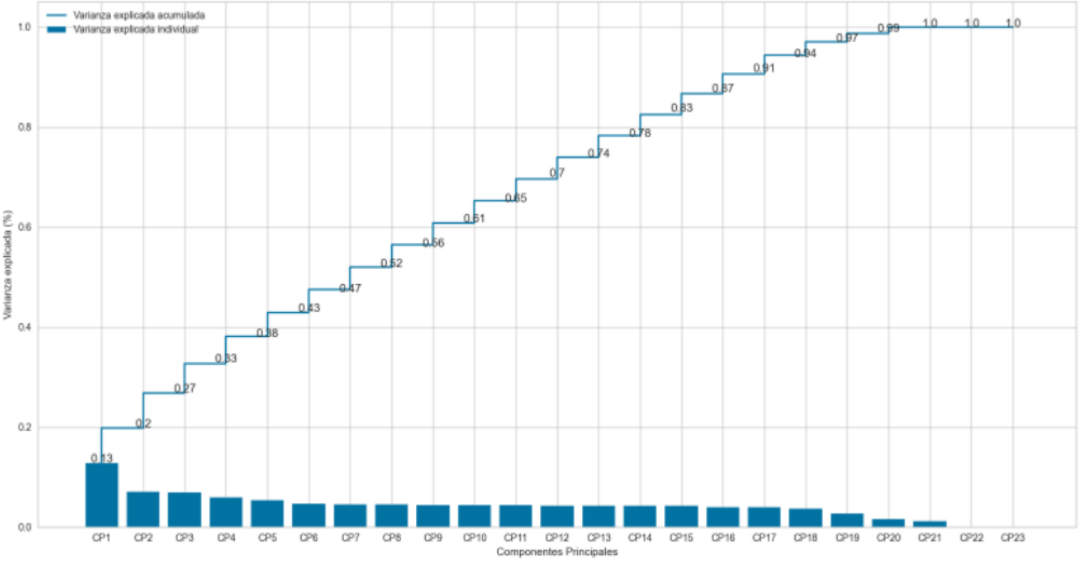
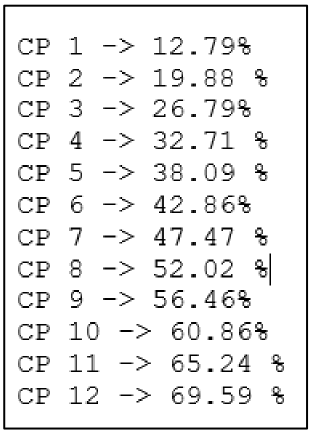


Figure 11. Output of Python script: Cumulative explained variance

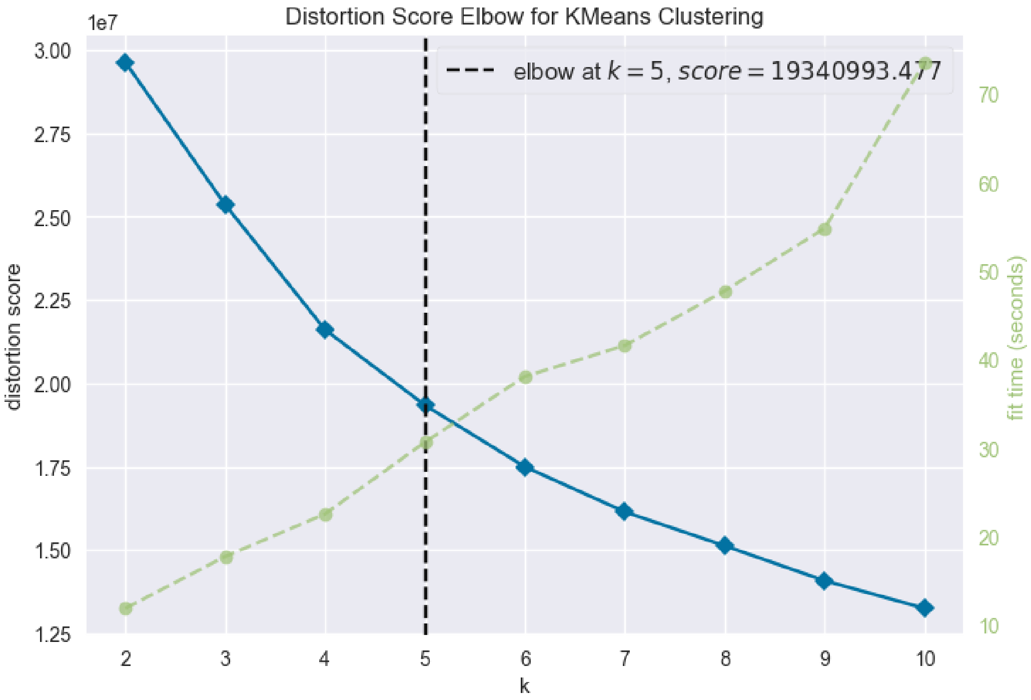


```
CP 1 -> 12.79%
CP 2 -> 19.88 %
CP 3 -> 26.79%
CP 4 -> 32.71 %
CP 5 -> 38.09 %
CP 6 -> 42.86%
CP 7 -> 47.47 %
CP 8 -> 52.02 %
CP 9 -> 56.46%
CP 10 -> 60.86%
CP 11 -> 65.24 %
CP 12 -> 69.59 %
```

From the result of the principal component analysis, a smaller sample of data was taken, and the elbow method was applied to determine the optimal number of clusters resulting in a value of k=5 (Sreedhar *et al.*, 2017).

The elbow method, as shown in Figure 12, contrasts two functions: the degree of homogeneity achieved by subjects assigned to the clusters, measured by the sum of intra-cluster distances (upward dashed line), and the number of clusters (downward solid line). Therefore, the intersection point provides an optimal solution by combining the number of clusters and the intra-cluster distance in the dataset (Tibshirani *et al.*, 2001).

Figure 12. Method of elbow applied to a reduced sample of the dataset (1,000,000 users).



The results obtained, illustrated in Figure 13, provide a disappointing result for the objectives of the study. Figure 13 shows the distribution of conversions in each of the clusters, which corresponds to the values taken by the Conversion (%) column in Table 26 (Jain *et al.*, 1999). The K-means algorithm has generated five rather heterogeneous groups, with overlapping attributes (e.g., the Insurance variable participates in all five groups and, in four of them, is the dominant one) and, moreover, very unbalanced. In other words, the algorithm has not found clear, homogeneous, and distinct groups among them as it was intended (Kamthania *et al.*, 2018).

Since the characteristics of the database determine the most appropriate algorithm for grouping subjects (Arabie *et al.*, 1996; Wedel & Kamakura, 2000), in this study, for data collected through lead capture, the use of the K-means algorithm does not provide the desirable results for cluster identification (Syakur *et al.*, 2018). Finally, to corroborate this result, the K-means algorithm was run again with the same data, and the results changed significantly, generating inconsistent and unstable results, corroborating the unsuitability of the algorithm for that dataset (Murray *et al.*, 2017).

Figure 13. Distribution of conversions by cluster.

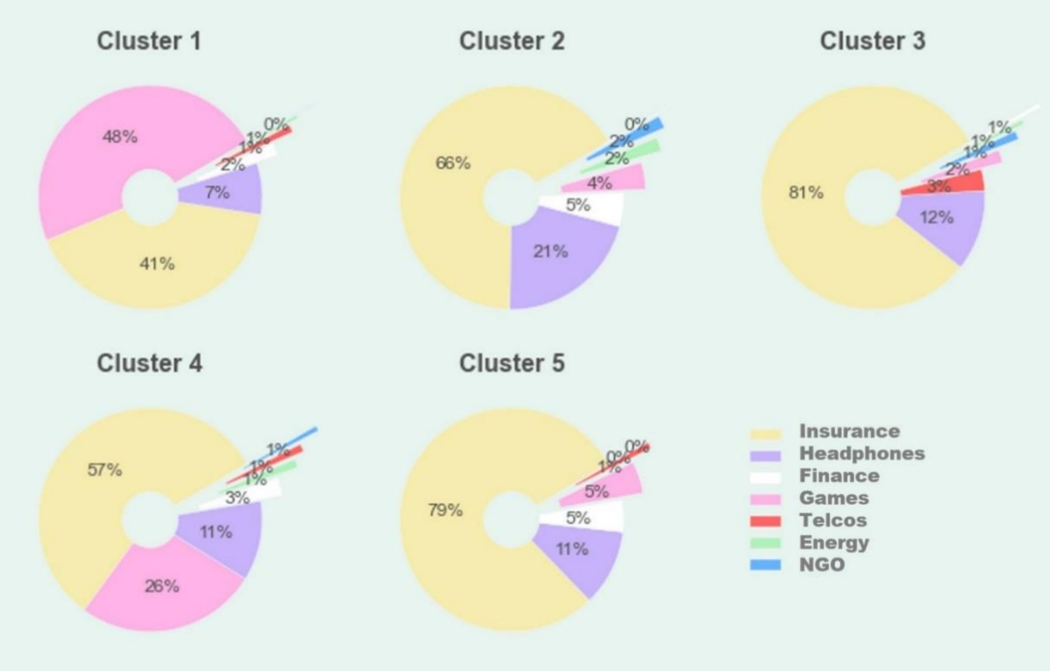


Table 26 shows the information that each generated cluster holds. This table shows a confusion matrix, which is a tool for evaluating the performance of a classification algorithm. In this case, it is a text classification model that categorizes texts into different product categories. The product categories include "Headphones", "Energy", "Finance", "Games", "NGO", "Insurance", and "Telcos". The diagonal cells (from top left to bottom right) represent correct predictions, where the actual category matches the predicted category. For example, the model correctly classified 318 texts in the "Headphones" category and 1 text in the "Energy" category. The off-diagonal cells represent incorrect predictions. For example, the model incorrectly classified 184 texts from the "Headphones" category as "Insurance". The totals in the last column and last row indicate the total number of texts in each actual and predicted category, respectively. For example, the model classified a total of 515 texts as "Headphones", while there were actually 599 texts in that category. As shown in the data, it can be concluded that there is no notable feature that distinguishes one cluster from another (Unnikrishnan & Hebert, 2005).

Table 26. Confusion matrix

Product	**Audiphones**	**Energy**	**Finance**	**Games**	**NGO**	**Insurance**	**Telcos**	**Totals**
Audiphones	**318**	0	0	12	1	184	0	**515**
Energy	4	**1**	0	1	0	26	0	**32**
Finance	0	0	**3**	10	0	1	0	**14**
Games	16	0	1	**449**	0	96	0	**562**
NGO	28	0	0	3	**0**	43	0	**74**
Insurance	232	1	0	113	5	**996**	1	**1.348**
Telcos	1	0	0	0	0	16	**0**	**17**
Totals	599	2	4	588	6	1.362	1	**2.562**

Source: Own elaboration

Figures 14 and 15 show the graphical representation of clusters in two dimensions, without highlighting conversions and highlighting conversions, respectively. The representation of clusters in 2D is a visualization technique used to show the distribution of data in a two-dimensional space. This technique is often used in clustering analysis to show how data is grouped into different clusters (Strehl & Ghosh, 2003). In this 2D representation of clusters, each point represents an observation and is colored according to the cluster to which it belongs. Points belonging to the same cluster are grouped together and separated from points belonging to other clusters. In the 2D representation of clusters highlighting users with conversions, an additional layer of information has been added in which users who have made conversions are visually highlighted using a different colour. The clusters observed in these representations were not conclusive again. This was mainly due to the fact that the identification of clusters depends on the data being clustered and did not offer conclusive results. It was noted that the 2D representation may not fully capture the structure of the data in a high-dimensional space. While cross-validation techniques were considered to determine the quality of the clustering solution, it was ultimately dismissed due to consistent results across multiple tests, and these representations confirmed that the algorithmic solution was not suitable for segmenting the market with the type of data being used.

Figure 14. Distribution of conversions by cluster.

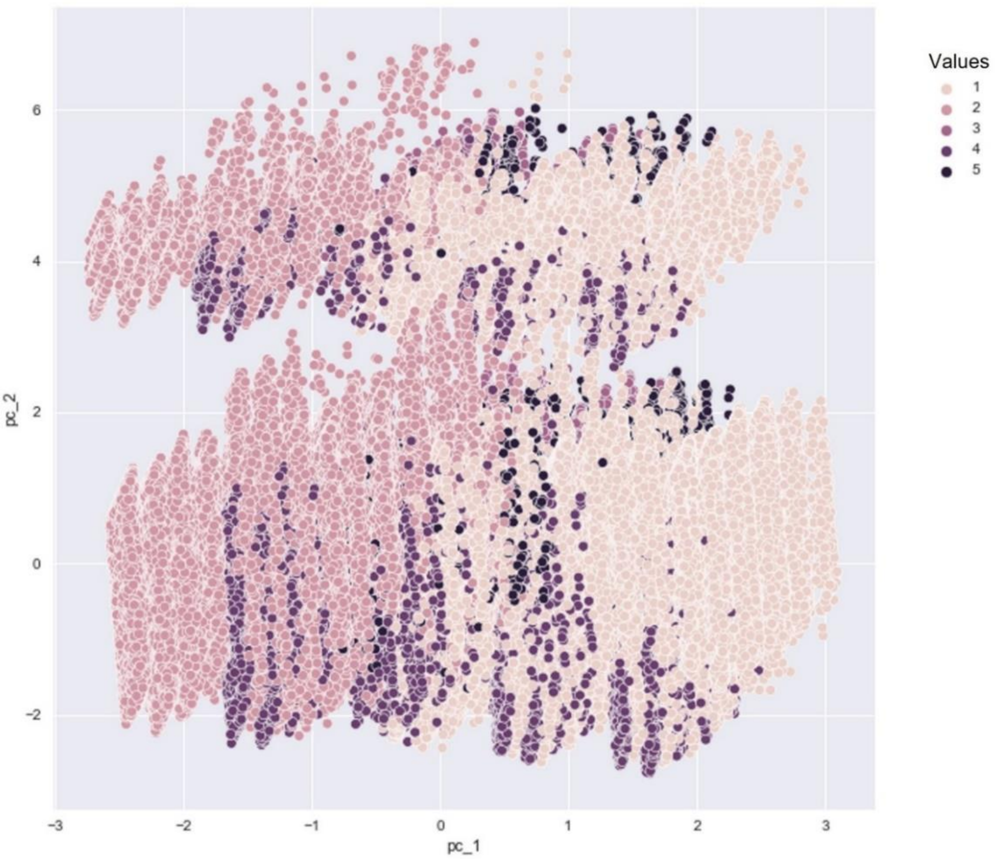
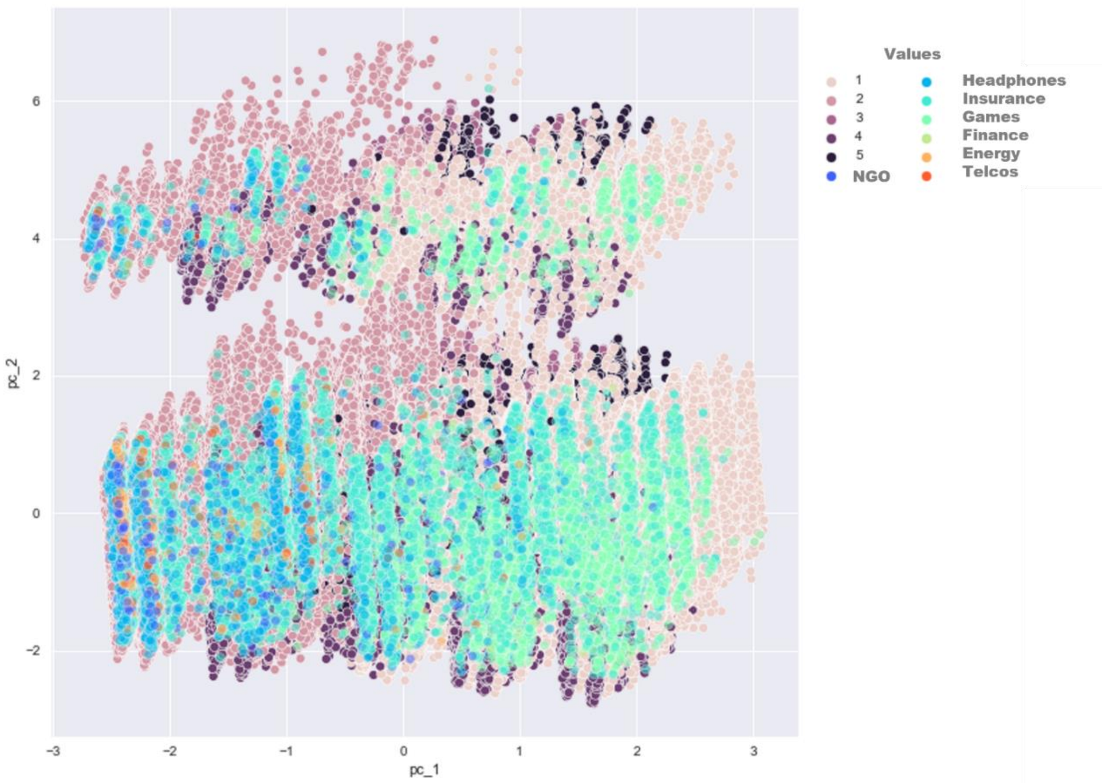


Figure 15. Representation of the clusters in 2D highlighting the users with conversions.



Therefore, in response to RQ1, "Are unsupervised algorithms efficient for clustering in marketing with data from online tests and surveys?", we conclude that they are not sufficiently efficient, in line with the findings of Dasgupta (2016).

4.3.4. Methodological study of the supervised algorithm: XGBoost.

To address this issue, a supervised learning approach was applied using the Extreme Gradient Boosting (XGBoost) algorithm. For the implementation of this algorithm, the XGBClassifier library of the Python xgboost package was used (Chen & Guestrin, 2016). It should be noted that XGBoost has high levels of confidence for large datasets with a mixture of categorical and numerical variables that are in different units of measurement (Chen & Guestrin, 2016), which is why data standardization is not necessary. In this approach, the conversion target variable was used to train the model and clusters were identified that maximized the difference in conversion rates between groups. To apply the XGBoost algorithm, a supervised learning method based on decision trees (Chen & Guestrin, 2016), the following steps were followed: (1) selection of the dataset; (2) application of the algorithm to train and fit the model; (3) selection of hyperparameters; (4) application of the algorithm and evaluation of the obtained performance; (5) visualization of the results (through graphics such as the learning curve and predictor variable importance); and (6) cross-validation.

(1) Firstly, 25 variables (3 int, 2 float, and 20 boolean) were selected from the set of 37 variables. Before applying the algorithm, it was necessary to divide the dataset into several subsets as shown in Table 27.

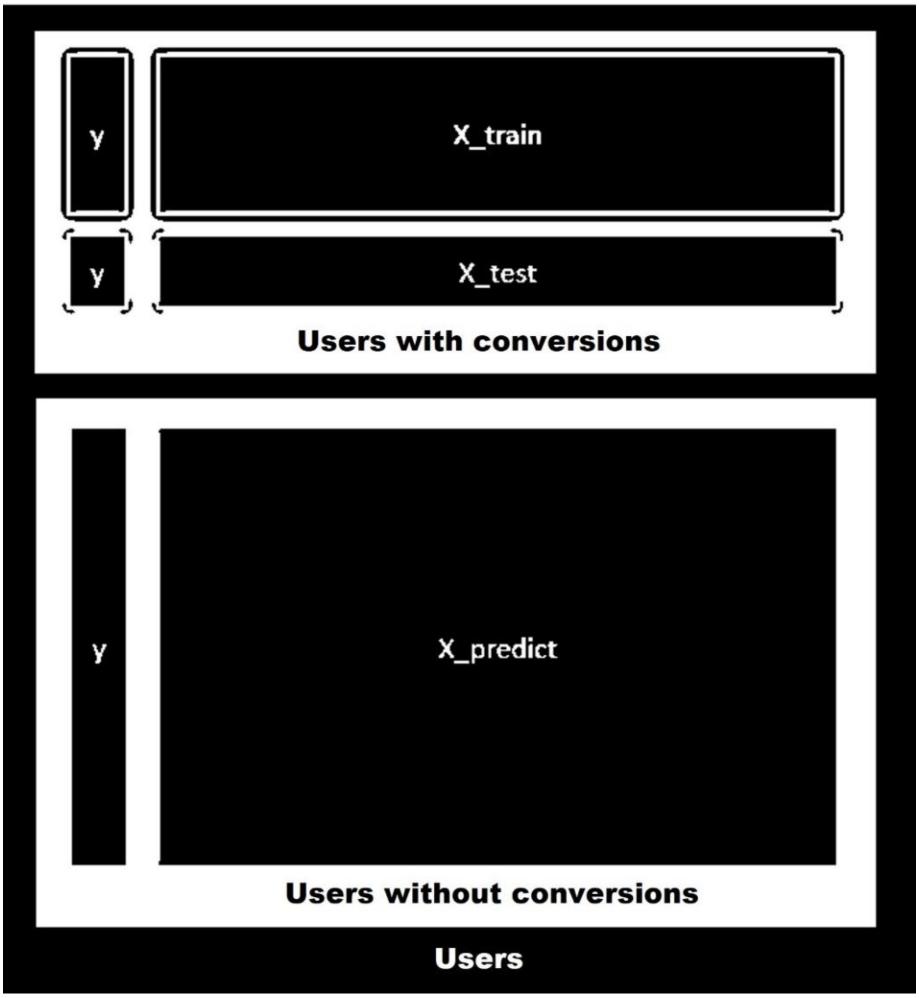
Table 27. Subsets of data.

Subsets of data.	**Description**	**Size**
X	Users with conversions. The variables in this matrix are those indicated in section 3.6 Final Structure, with the exception of id_producto_conv and id_user.	[25.612 x 23]
Y	**id_producto_**conv corresponds to the users of **X**.	[25.612 x 1]
X_train	90% of users with conversions.	[23.050 x 23]
y_train	**id_producto_**conv corresponds to the users of **X_train.**	[23.050 x 1]
X_test	10% of users with conversions.	[2.562 x 23]
y_test	**id_producto_conv** corresponds to the users of **X_test.**	[2.562 x 1]
X_predict	Users without conversions.	[5.160.245 x 23]
y_predict	A value that is unknown at the beginning of the study (**id_producto_conv = 0)** and will be predicted after applying this algorithm	[5.160.245 x 1]

Source: Own elaboration

(2) Next, the algorithm was applied to train and adjust the model. This second stage needed to be developed in two parts: the training of the algorithm through the function X_train and the evaluation of its performance through X_test, which estimated the level of confidence (Goodfellow *et al.*, 2016). Next, X_predict was applied. For the training process, only users who had previously converted to a product, that is, users with conversions (Bishop & Nasrabadi, 2006), were taken into account. In this sense, the training data had to contain some information about the correct response or what would be the study's target variable (Hastie *et al*., 2009). Thus, the learning algorithm (see Figure 16) found patterns in X_train, assigned the input data attributes to the target (Y_train), and generated a Machine learning (ML) model that captured those patterns (Jordan & Mitchell, 2015).

Figure 16. Schema of the different subsets of data.



Source: Own elaboration

(3) Next, the hyperparameters were defined, which were the number of iterations (n_estimators=100) and the maximum depth of each tree (max_depth=8). To do this, several permutations were analyzed

(n_estimators = 50, 100, 200, 500, 750, 1,000 and max_depth = 4, 6, 8, 10, 15, 20). Additionally, since the frequencies of the products from the conversions were clearly imbalanced, as can be seen in table 28, weights had to be adjusted, a parameter that assigned a weight or weighting to each group or product.

Table 28. Characteristics of each cluster.

Cluster	Number of users	Average age	Product	Nº Conversions	Conversions rate (%)
1	2.760.626	44,45	Audiphones	732	7,42
			Energy	50	0,51
			Finance	103	1,04
			Games	4729	**47,9**
			NGO	173	1,75
			Insurance	4068	41,2
			Telcos	8	0,08
2	2.158.265	45,23	Audiphones	2740	21
			Energy	258	1,98
			Finance	2	0,01
			Games	503	3,86
			NGO	630	4,84
			Insurance	8643	**66,4**
			Telcos	237	1,82
3	174.126	42,38	Audiphones	19	11,6
			Energy	1	0,61
			Finance	5	3,07
			Games	3	1,84
			NGO	1	0,61
			Insurance	132	**81**
			Telcos	2	1,22
4	229.770	52,13	Audiphones	93	11,5
			Energy	10	1,24
			Finance	9	1,11
			Games	212	26,2
			NGO	21	2,59
			Insurance	458	**56,6**
			Telcos	6	0,74
5	41.356	46,18	Audiphones	12	11
			Energy	0	0
			Finance	1	0,91
			Games	5	4,59
			NGO	5	4,59
			Insurance	86	**78,9**
			Telcos	0	0

Source: Own elaboration

To calculate these weights, the Python library class_weight was used, which internally performed the following calculation:

$$\text{product weight } \boldsymbol{i} \ = \frac{\text{Size of } \mathbf{y\_trai}\boldsymbol{n}}{number\ of\ products \cdot \ \text{Frequency of product } \mathbf{i} \text{ in } \mathbf{y\_trai}\boldsymbol{n}}$$

(4) Next, the algorithm was applied and its performance was evaluated. After adjusting the parameters and training the model with X_train, the resulting algorithm was applied to X_test and the values of y_test were predicted. To determine if the resulting predictions could be considered optimal, some metrics were calculated, which are presented below. In order to determine the percentage of correct predictions, that is, the accuracy with which the values of y_test were predicted, the Python accuracy_score metric was used. On the other hand, to determine the percentage of correct predictions, but by product, the recall_score metric was used. Once the parameters were adjusted and the confidence of the defined algorithm was known, the algorithm was trained again, but this time with 100% of the records with conversions. Subsequently, it was applied to X_predict and y_predict was obtained. y_predict saved the index of the highly recommended product (highest probability percentage) for each of the users in X_predict. This data was saved in a column (called id_pro_recomendacion_1) of a new table called "recomendador". The next step was to save another new column: "id_pro_recomendacion_2", which contains the indices of the second recommended product. Since in many cases the second product had a small percentage (less than 10%), two new columns were added: "pb_recomendacion_1" and "pb_recomendacion_2", which saved the percentages with which each product is recommended. And a recommendation table was generated (see table 29), which shows the probability that a user will convert in two of the categories.

Table 29. Conversion frequencies and accuracy percentages by product.

Product	Frequencies totals	Frequencies y_train	Frecuencies y_test	Correct predictions rate (%)
Audiphones	5.363	4.848	515	61.74
Energy	318	286	32	3.12
Finance	116	102	14	21.42
Games	5.447	4.885	562	79.89
NGO	831	757	74	0.00
Insurance	13.286	11.938	1.34 8	73.88
Telcos	251	234	17	0.00

Source: Own elaboration

As an example, for the product "telcos" (id_producto_conv=8), the table "recomendador_telcos" was obtained, which was formed by those users who verified the following condition: CONFIDENCIAL 44 id_producto_conv = 8 OR id_pro_recomendacion_1 = 8 OR id_pro_recomendacion_2 = 8. It should be noted that the three mentioned conditions could not occur simultaneously. That is, each user in the "usuarios" table verified one or none of them. Then, the column "pb_recomendacion" was created, whose value is defined as follows: (1) 1 IF id_producto_conv = 8, (2) pb_recomendacion_1 IF id_pro_recomendacion_1 = 8, (3) pb_recomendacion_2 IF id_pro_recomendacion_2 = 8. The "pb_confianza" column was also added, which groups the percentages of "pb_recomendacion" into different intervals. The values it took were: [0, 0.1), [0.1, 0.2), [0.2, 0.3), ..., [0.8, 0.9), [0.9, 1), 1 where 1 corresponds to users who had converted to a "telcos" product in the past.

That is, the clustering technique sought to identify groups of users who present significant differences in terms of the products they acquire, with the aim of maximizing the probability of conversion (Grbovic *et al.*, 2015).

(5) Subsequently, the results were visualized using graphs such as the learning curve and the importance of predictor variables. Visualization of results through graphs is a fundamental tool in data analysis as it allows for a visual and understandable representation of patterns and trends in the data. Specifically, the learning curve is a graph that shows how the accuracy of a machine learning model improves as the size of the training dataset increases. On the other hand, the importance of predictor variables refers to how much they influence the final outcome of the model, which can be visualized graphically (Tufte, 2001).

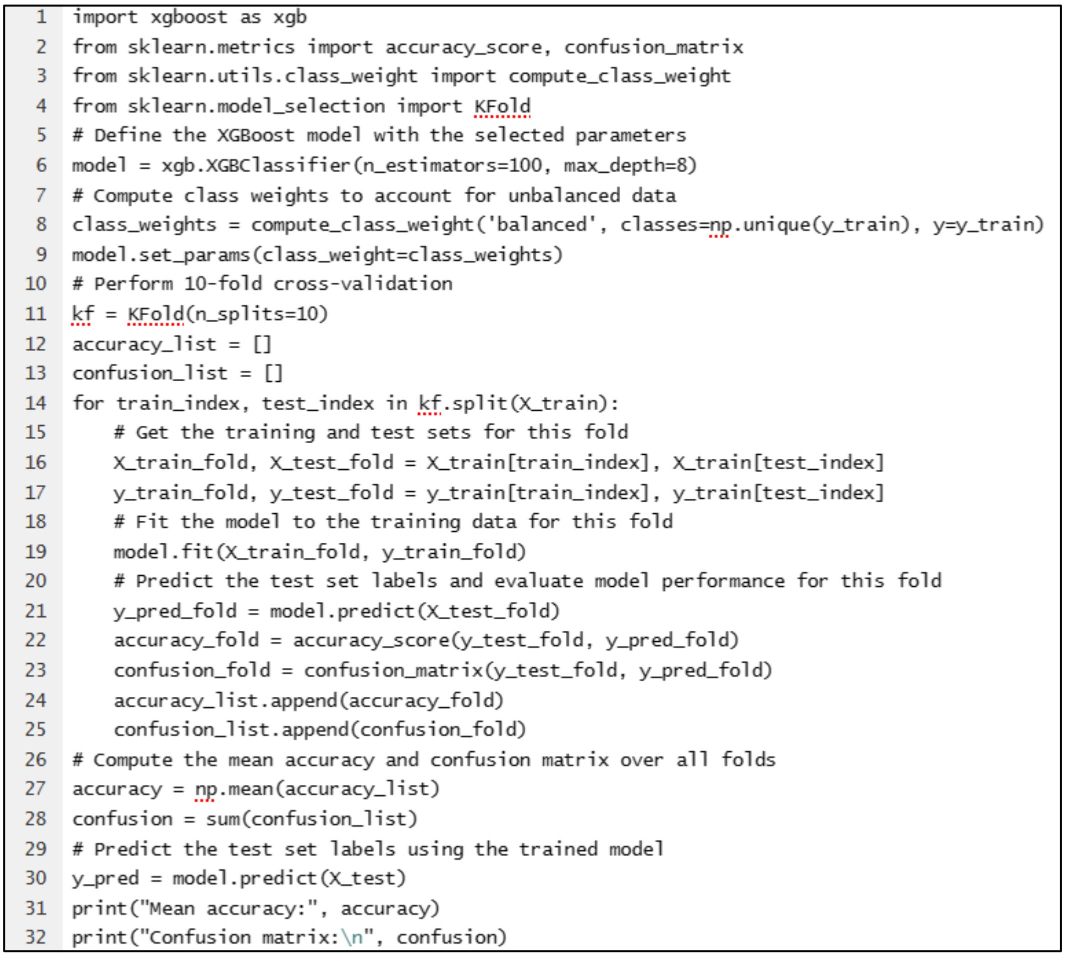
(6) As a validation criterion for the experiment, cross-validation was used. Specifically, ten-fold cross-validation was used, which means that the model worked with 90% of the records of users who became buyers, and from the model fitting, the behavior of the remaining 10% of users was predicted (Kohavi, 1995).

4.3.4.1. Results

The results with XGBoost showed a significant improvement in the ability to create clusters to predict product conversion. In addition, greater homogeneity was observed within the clusters, suggesting that the identified groups are more coherent and useful for customer segmentation. Overall, these findings suggest that the supervised learning approach was more effective than the unsupervised clustering approach for segmenting customers based on their propensity for conversion using the available data.

Specifically, once the supervised algorithm was applied and the model was trained with XGBoost, it provided the following results through the y_predict output (the prediction of id_producto_conv) for those users who had not converted in the past to any product or service (Chen & Guestrin, 2016). That is, it provided the products that best fit each user. As can be seen in Figure 17, some metrics were found through Python scripts that allowed the level of reliability of the resulting predictions to be determined (Friedman, 2002), which are presented below.

Figure 17. Python script for Performance evaluation of XGBoost algorithm on imbalanced data with class weighting and cross-validation.

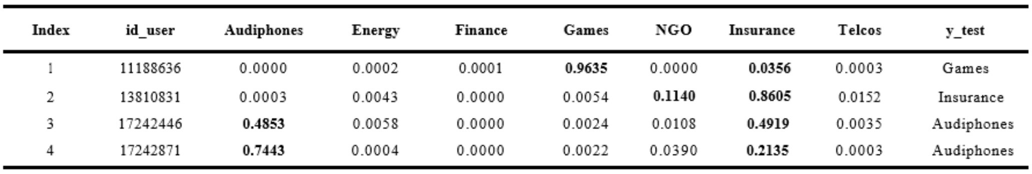


```
import xgboost as xgb
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.utils.class_weight import compute_class_weight
from sklearn.model_selection import KFold
# Define the XGBoost model with the selected parameters
model = xgb.XGBClassifier(n_estimators=100, max_depth=8)
# Compute class weights to account for unbalanced data
class_weights = compute_class_weight('balanced', classes=np.unique(y_train), y=y_train)
model.set_params(class_weight=class_weights)
# Perform 10-fold cross-validation
kf = KFold(n_splits=10)
accuracy_list = []
confusion_list = []
for train_index, test_index in kf.split(X_train):
    # Get the training and test sets for this fold
    X_train_fold, X_test_fold = X_train[train_index], X_train[test_index]
    y_train_fold, y_test_fold = y_train[train_index], y_train[test_index]
    # Fit the model to the training data for this fold
    model.fit(X_train_fold, y_train_fold)
    # Predict the test set labels and evaluate model performance for this fold
    y_pred_fold = model.predict(X_test_fold)
    accuracy_fold = accuracy_score(y_test_fold, y_pred_fold)
    confusion_fold = confusion_matrix(y_test_fold, y_pred_fold)
    accuracy_list.append(accuracy_fold)
    confusion_list.append(confusion_fold)
# Compute the mean accuracy and confusion matrix over all folds
accuracy = np.mean(accuracy_list)
confusion = sum(confusion_list)
# Predict the test set labels using the trained model
y_pred = model.predict(X_test)
print("Mean accuracy:", accuracy)
print("Confusion matrix:\n", confusion)
```

*Note: this code, X_train and y_train are the training data and labels, respectively, and X_test and y_test are the test data and labels, respectively.

The results are shown in the following Table 30. This table displays the probabilities (from 0 to 1) of a user belonging to each of the categories (Headphones, Energy, Finance, Games, NGO, Insurance, Telcos), along with the true value of the category to which that user belongs (y_test). For example, for the first user (id_user 11183636): The probability of belonging to Headphones is 0.0000, the probability of belonging to Energy is 0.0002, the probability of belonging to Finance is 0.0001, the probability of belonging to Games is 0.9635 (very high), the probability of belonging to NGO is 0.0000, the probability of belonging to Insurance is 0.0356, the probability of belonging to Telcos is 0.0003. And the true value of the category to which that user belongs is Games. For the second user, the probabilities indicate that it is highly likely to belong to Insurance, and the true value is indeed Insurance. And so on for the remaining users. That is, the table shows the model's predictions in the form of probabilities, along with the true value, for some examples.

Table 30. Extraction of the probability table along with the true value of y_test.



Index	id_user	Audiphones	Energy	Finance	Games	NGO	Insurance	Telcos	y_test
1	11188636	0.0000	0.0002	0.0001	**0.9635**	0.0000	**0.0356**	0.0003	Games
2	13810831	0.0003	0.0043	0.0000	0.0054	**0.1140**	**0.8605**	0.0152	Insurance
3	17242446	**0.4853**	0.0058	0.0000	0.0024	0.0108	**0.4919**	0.0035	Audiphones
4	17242871	**0.7443**	0.0004	0.0000	0.0022	0.0390	**0.2135**	0.0003	Audiphones

Source: Own elaboration

Another metric that was analyzed was the confusion matrix (Provost *et al.*, 1998). Each column of the matrix represents the number of predictions for each product, while each row represents the instances in the real class (Chlebus *et al.*, 2011). As an example, as shown in table 26, for the category of Headphones, these were the results: 318 records were correct, and 197 records were incorrect. These failures were distributed as follows (see table 30): 12 in Games, 1 in NGOs, and 184 in Insurance. Additionally, 4 records whose real value was Energy were classified as Headphones, 16 records whose real value was Games were classified as Headphones, 28 records whose real value was NGOs were classified as Headphones, 232 records whose real value was Insurance were classified as Headphones, and 1 record whose real value was Telcos was classified as Headphones.

The probability matrix (Ravikumar *et al.*, 2010) was also calculated, which reports the probability percentage that a user will convert to a specific product. When the algorithm predicts the value of y_test, what it does internally is to take the product that has obtained the highest probability (Provost *et al.*, 1998). As shown in the probability matrix (table 31), it was observed that sometimes the algorithm was very confident in its prediction (Provost *et al.*, 1998), while other times it was not so sure and offered similar percentages for two products (Ravikumar *et al.*, 2010).

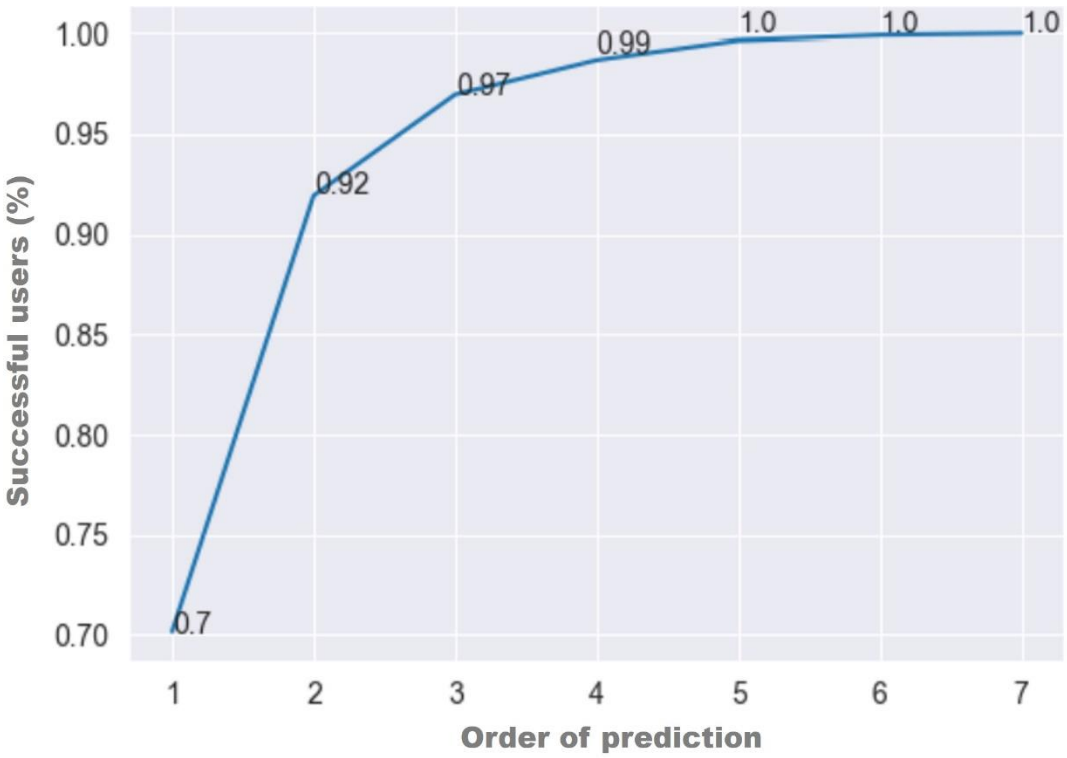
Table 31. Sample of the content of the 'recommendation' table.

id_user	id_pro_ recomendation_1	id_pro_ recomendation_2	pb_ recomendation_1	pb_ recomendation_2
154063	7	5	0.6797	0.3158
287605	6	2	0.6833	0.2454
329118	3	5	0.9552	0.0329
473911	7	4	0.9027	0.0493

Source: Own elaboration

Analyzing the probability table alongside the actual value of y_test, it was observed that most of the time when the algorithm failed, the product to which the second highest probability was assigned was the correct one. Thus, if only the product with the highest probability was taken, 70% accuracy was achieved. But if the accuracy of the second product with the highest probability was added to this percentage, 92% accuracy was achieved. This can be seen graphically in Figure 18.

Figure 18. Cumulative accuracy percentage



After analyzing the results obtained in the algorithm training and adding the second product with the highest probability to the initial result, confidence levels of 92% were achieved, and it was considered that the optimal solution to the study is to assign each user the two products with the highest probabilities (Ravikumar *et al.*, 2010), indicating the degree of probability of each one (Provost *et al.*, 1998). In response to RQ2 about whether unsupervised algorithms are efficient for clustering in marketing with data from drawings and online tests, the results of this study show that they offer high levels of confidence that can be considered valid for grouping users in digital marketing.

Regarding RQ3, on which of the two types of algorithms, unsupervised or supervised, the results concluded that with the data we started with, from drawings and online tests, and after applying K-Means, unsupervised, and XGBoost, supervised, only the supervised one offered valid results.

4.4. Conclusions

Digital marketing is currently booming (Sáez-Ortuño *et al.,* 2023a), and in particular, market research for segmentation through clustering (Boone & Boehm, 2002). This is one of the common tools used by both academia and industry to divide consumers into groups or "clusters" based on similarities in their buying behaviors, attitudes, or demographic characteristics (Wedel & Kamakura, 2000). These groups are used to develop specific marketing strategies with the aim of maximizing the effectiveness of communication and marketing efforts (James *et al.,* 2013). In this research, the performance of two commonly used algorithms in cluster analysis is compared: K-Means (unsupervised learning) and XGBoost (supervised learning). K-Means is an iterative clustering algorithm that divides a data set into k groups based on the similarity between data points. It is fast and easy to implement but may be prone to suboptimal results due to the dependence on the choice of k and the initialization of centroids. Additionally, K-Means can only handle numerical datasets and cannot process categorical data (MacQueen, 1967). XGBoost, on the other hand, is a machine learning algorithm based on decision trees that is often used in classification and regression problems. In summary, XGBoost offers better predictive performance as it can leverage the information of the target variables to generate better predictive models, although there may be cases where K-Means may be more suitable than XGBoost for data clustering in the marketing context depending on the segmentation objective and problem characteristics being addressed. Studies such as (Chen and Guestrin, 2016) show that XGBoost often outperforms other algorithms in prediction accuracy. XGBoost detects more complex patterns. Decision trees in XGBoost allow for detecting interactions between variables and non-linear patterns. K-Means uses a Euclidean distance measure that can only represent linear patterns. According to (Ke *et al.,* 2017), this gives XGBoost a greater ability to represent complex relationships in the data. Greater robustness to noise and outliers should also be taken into consideration. XGBoost is more robust to outliers and variability, thanks to boosting and optimization of the objective function. K-Means is very sensitive to these effects, as indicated by Sculley (2010) and (Raschka & Mirjalili, 2019). Outliers can bias cluster centroids in K-Means. XGBoost provides significant advantages by allowing greater predictive capacity, detecting more complex patterns in the data, and greater robustness to noise, compared to K-Means. This translates into more accurate and useful clustering models for decision making. In this study, we have analyzed these two algorithms applied to the clustering of data recorded on websites with a large sample size over a long period of time. The results of this study are consistent with previous research on the phenomenon (MacQueen, 1967; James *et al.*, 2013).

4.4.1. Theoretical implications

The results obtained in this study reinforce the idea that supervised learning algorithms, such as XGBoost (Chen & Guestrin, 2016), may be more effective than unsupervised algorithms, such as K-Means, in customer segmentation based on propensity to conversion, especially when working with data from online surveys and tests. This finding supports the growing trend in the literature to use supervised algorithms in the field of marketing and customer segmentation (Friedman, 2002; (Provost *et al.*, 1998).

Additionally, this study expands the understanding of how to use supervised learning algorithms in customer segmentation by analyzing metrics such as confusion matrix and probability matrix (Provost *et al.*, 1998; Ravikumar *et al.*, 2010) to evaluate the reliability of resulting predictions.

Specifically, in the context of data from online surveys and tests, the choice between supervised (XGBoost) and unsupervised (K-means) algorithms may have several theoretical implications depending on the nature of the problem. If the goal is to predict a target variable, such as the probability that a user will convert in purchasing a good or service, XGBoost has been shown to be the most appropriate option (Chen & Guestrin, 2016). On the other hand, if the goal is to group users based on their characteristics and preferences and there is an adequate database available, K-means could also be a valid option as it has provided informative clusters without being predictive (MacQueen, 1967). Performance and accuracy should also be taken into account. XGBoost has also shown to be efficient in classification and regression tasks, outperforming K-Means in this problem (Chen & Guestrin, 2016). However, this advantage only applies if the goal of the analysis is to predict a target variable. In clustering tasks, K-means is a widely used and efficient algorithm although it requires the initial choice of centroids that may not always be available (Jain, 2010). Finally, interpreting results appropriately is important. Decision tree-based models, such as XGBoost, may be easier to interpret than clustering models, such as K-means. Decision trees provide a graphical representation of the decisions made by the model, facilitating the understanding of the relationship between features and the target variable (Breiman, 2017). Conversely, K-means results may be harder to interpret as the formed groups may not always have a clear meaning and may be sensitive to the initial choice of centroids (Arthur & Vassilvitskii, 2007). In summary, XGBoost is the appropriate option if data comes from online surveys and tests and the goal of the study is to predict a target variable and the nature of the data does not present an initial centroid.

4.4.2. Management implications

The results of this study suggest that companies and digital marketing professionals can benefit from implementing supervised learning algorithms, such as XGBoost, to segment their customers based on their propensity to convert. By doing so, they can identify more coherent and homogeneous groups of customers, which could facilitate the creation of more effective and personalized marketing strategies.

Furthermore, this study highlights the importance of considering not only the product with the highest probability of conversion but also the second product with the highest probability. By doing so, companies can increase their confidence levels in the resulting predictions and ultimately improve the accuracy of their marketing campaigns (Ravikumar *et al.*, 2010).

In summary, this study provides a strong foundation for digital marketing professionals to use supervised learning algorithms in customer segmentation and identifying products with a higher propensity to convert. By adopting these approaches, companies can improve their marketing strategies and ultimately increase their conversion rates and revenue (Eskerod, 2020).

4.5. General Discussion

The results of this study suggest that companies and digital marketing professionals can benefit from implementing supervised learning algorithms, such as XGBoost, to segment their customers based on their propensity to convert. By doing so, they can identify more coherent and homogeneous groups of customers, which could facilitate the creation of more effective and personalized marketing strategies.

Additionally, this study highlights the importance of considering not only the product with the highest probability of conversion but also the second product with the highest probability. By doing so, companies can increase their confidence levels in the resulting predictions and ultimately improve the accuracy of their marketing campaigns (Ravikumar *et al.*, 2010).

In summary, this study provides a solid basis for digital marketing professionals to use supervised learning algorithms in customer segmentation and in identifying products with higher propensity to convert. By adopting these approaches, companies can improve their marketing strategies and ultimately increase their conversion rates and revenue (Eskerod, 2020).

4.6. Limitations and future lines of research

Although clustering algorithms can be very useful in the field of marketing (Forgy, 1989), this study presented some limitations. First, the collection, filtering, and analysis of the constant stream of information from social media is a significant challenge that requires continuous monitoring (Jansen *et al.*, 2009). This study collected data from over a decade, and while social media platforms are constantly evolving, the study's findings may be specific to the moment and not generalizable to those platforms over time (Kaplan & Haenlein, 2010). A promising future line of research could focus on developing sophisticated methods of collecting and analyzing social media data to gain deep insights over time (Chen & Guestrin, 2016). For example, using machine learning techniques and sentiment analysis to monitor social networks (Pang & Lee, 2008), conducting longitudinal studies to track how trends on social media evolve over time, and making systematic comparisons between different social media platforms to identify singularities and general trends that may vary over time (Kietzmann *et al.*, 2011).

Another limitation of the study is that it was based on a database of participants in online contests and giveaways, which may not represent the entire population, limiting the generalization of the results (Rubin & Babbie, 2016). A line of research could be to use samples from more diverse origins and representative of the population to improve generalization. For example, samples where data come from other types of incentives or motives for registering online (Kraut *et al.*, 2004).

It is also important to consider that the sample used in the studies was from Spain, which may limit the generalizability of the findings to other geographic regions (Ravikuma *et al.*, 2010). Replicating the studies in different geographic and cultural locations to determine the extent to which the findings are specific to Spain would be an interesting line of inquiry.

Finally, the study was based solely on social media data, which may provide a limited view of participants' behavior and communication patterns when leaving their data on social networks. Mixed research methods that include interviews, surveys, participant observation, and social network analysis could be used (Creswell & Clark, 2017). It is important to note that each individual is unique, and a deeper understanding of their needs and preferences is necessary before making marketing decisions (Lloyd, 1982). It would be interesting to collaborate with interdisciplinary teams that can leverage a variety of perspectives and methodological experiences and involve stakeholders and study participants to identify significant and relevant research questions (Han & Tong, 2022).

CHAPTER 5. CONCLUSIONS

The structure of this chapter is as such: Firstly, the overall findings of the thesis are summarised, as the previous chapters have already presented individual conclusions. Secondly, the significant contributions of the research will be highlighted, along with their academic and practical implications. Finally, the chapter will elaborate on the limitations of the study and propose potential areas for future research.

5.1. General conclusions

The preliminary conclusion drawn from this thesis is that there has been an exponential increase in the number of publications and citations related to the capture of consumers' personal data from social networks for digital marketing purposes from 1997 up to nowadays. The study identifies the most relevant trends through the analysis of the most significant articles, keywords, authors, institutions and countries. The United States and Australia are the countries that publish the most in this field, while Finland and Australia have the highest number of publications per capita. The thesis not only presents an analysis of the current state of research on the capture of consumers' personal data in social networks for digital marketing purposes, but also addresses two crucial questions that emerge from the first study and that have not been addressed in depth in the existing literature to date. First, the thesis explores consumers' motivations for volunteering personal data on social networks, as well as the direct impact of fraudulent consumer data on digital marketing, providing valuable insights into the complex interplay between consumer behaviour, privacy concerns and digital marketing. This information facilitates the development of more effective strategies for the collection and use of consumer data in the future. It concludes that a significant number of users intentionally provide false information when signing up for online sweepstakes and contests, with the most common motivations being fun, lack of trust in the site requesting the data, and privacy concerns. In addition, this research highlights the importance of trust in online marketing and the need for advertising to provide assurances that it is safe and trustworthy. The site from which data is requested is also important, with public bodies being preferable.

Secondly, cluster marketing has been investigated. Given the current boom in digital marketing, this research concludes that cluster marketing has become an accepted tool in academia and industry for dividing consumers into groups based on similarities in purchasing behaviour, attitudes or demographic characteristics. Two algorithms commonly used in cluster analysis, K-Means and XGBoost, were studied. K-Means, an unsupervised algorithm, is fast and easy to implement, but can be prone to suboptimal results due to the choice of k and the initialisation of centroids. On the other hand, XGBoost, a supervised algorithm, can handle both numerical and categorical data and has better predictive performance, but may require more computational resources. XGBoost offers better predictive performance and can detect more complex patterns than K-Means. It also has a better ability to represent complex relationships in the data and is more robust to outliers and variability.

5.1.1. Profile of Cheaters in Online

This research concludes that various factors such as age, gender, familiarity with technology, and personal motivations influence the tendencies of some profiles of cheaters to provide false information when filling out personal data forms. Among, the following that influence their behavior should be highlighted:

- Older men generation

Older men are more likely to make mistakes when filling in their phone number and to enter false information in general. This thesis concludes that this may be due to a variety of factors, such as a lack of familiarity with technology, a greater sense of anonymity provided by the internet, and a willingness to take risks in providing false information. Additionally, the research found that older, self-trained male cohorts tend to be more likely to enter false information intentionally, possibly because they have more life experience and are using these technologies for a clear purpose and to obtain a specific outcome. However, the study also notes that the assumption that younger generations, who were born in the age of the internet and social networks, behave differently to other generations does not hold, as the results of the study do not point in that direction. Overall, the reasons why older men may be more likely to lie when registering their personal data may be complex and multifaceted, and may depend on individual factors such as age, gender, and personal motivations (Jacobsen *et al*., 2018).

-Women of younger generations

The thesis found that younger women have a higher propensity to provide false information in the field of phone numbers. The reasons for this are not explicitly stated in the study, but it is possible that younger women may feel more comfortable with the anonymity provided by the internet and may be more willing to take risks in providing false information. Additionally, younger generations may be more accustomed to using technology and may be more likely to make mistakes due to inattention or haste when filling out online forms. It is also possible that younger women may be more concerned about their privacy and may be more likely to provide false information in order to protect their personal information. According to Jacobsen *at al.,* (2018), the issue of younger women providing false information during personal data registration could be influenced by a variety of factors, such age or gender, and personal motivations. Therefore, the reasons behind this behavior are likely to be intricate and diverse.

- Privacy

Privacy is the most important factor in motivating users to provide false information in online sweepstakes and quizzes. Participants expressed concern about the loss of anonymity and the risk associated with providing too much personal information. They also questioned why so much information was being requested and what it would be used for. Some participants were willing to provide their email address but not their phone number, indicating a desire to protect their personal information. The study concludes that maybe measures to address privacy concerns and build trust with users may be effective in reducing the incidence of false information. This could include providing clear and transparent information about how personal data will be used, ensuring that data is stored securely, and offering users the option to provide only the minimum amount of information necessary to participate in the sweepstakes or quiz. Gefen *et al.,* (2003) suggest that taking privacy concerns into account during the design phase of online sweepstakes and quizzes could be a useful strategy in reducing the occurrence of false information. The study emphasizes the significance of this approach in promoting a more secure and trustworthy online environment.

- Trust

Trust is a significant factor in motivating users to provide false information in online sweepstakes and quizzes. Participants expressed doubts about who was sponsoring the sweepstakes and whether they could trust the website requesting their data. The research also concludes that measures that could build trust with users may be effective in reducing the incidence of false information. This could include providing clear and transparent information about how personal data will be used, ensuring that data is stored securely, and offering users the option to provide only the minimum amount of information necessary to participate in the sweepstakes or quiz. In general, the research emphasizes the significance of taking trust issues into account while creating online contests and questionnaires and proposes that dealing with these concerns can be a productive approach to minimize the prevalence of inaccurate data (Gefen *et al.*,2003).

-Amusement

A considerable motivation for users to provide false information in online sweepstakes and quizzes is amusement. Participants reported that they sometimes impersonated the names of acquaintances for fun, and that they enjoyed playing jokes and pranks. Some participants also mentioned that they entered false information to kill boredom or to pass the time. The study concludes that addressing the motivation of amusement may be a viable strategy to decrease the prevalence of false information in online sweepstakes and quizzes. This could include designing sweepstakes and quizzes that are engaging and entertaining, and that offer users a sense of fun and enjoyment. Additionally, adds that addressing these motivations may be an effective way to reduce the incidence of false information. The issue of users providing false information for amusement purposes may involve a range of complex and diverse factors that depend on individual traits, such as personality and personal motivations, and may interact with other underlying factors. Zannettou *at al.,* (2019) highlight the complexity of this phenomenon and suggest that a deeper understanding of the underlying factors is necessary to develop effective interventions that can reduce the prevalence of false information in online contexts.

5.1.2. Clustering online users

This thesis concludes that various factors, such as the ability to handle numerical and categorical data, performance, and robustness, influence the choice of clustering algorithm for data analysis. Among these factors, the following should be highlighted as having a significant impact on the choice of algorithm:

-Handling data

One important factor that influences the choice of clustering algorithm is its ability to handle both numerical and categorical data (Rodriguez *et al.*, 2019). While some algorithms, like K-Means, can only handle numerical data, others like XGBoost can handle both types of data (Chen & Guestrin, 2016). Categorical data represents variables that are not numerical in nature, such as gender, occupation, or marital status. Such variables are typically transformed into binary variables (0 or 1) to be used in K-Means. However, this transformation may not always be appropriate and can lead to the loss of important information in the data (Rodriguez *et al.*, 2019). XGBoost, on the other hand, can handle categorical data directly, which enables it to capture more complex relationships in the data and improve its predictive performance (Chen and Guestrin, 2016). Thus, the ability to handle both numerical and categorical data is an important consideration when choosing a clustering algorithm (Rodriguez *et al.*, 2019).

- Performance

Performance is another key factor that influences the choice of clustering algorithm. In this regard, it is important to consider the complexity of the data, the size of the dataset, and the computational resources available. K-Means is a fast and efficient algorithm that can handle large datasets, but its performance may be suboptimal in certain scenarios due to its dependence on the choice of k and the initialization of centroids (Arthur & Vassilvitskii, 2006). On the other hand, XGBoost is a more sophisticated algorithm that can handle both numerical and categorical data, and has been shown to offer higher predictive performance (Chen & Guestrin, 2016). However, it may be slower to train and requires more computational resources than K-Means. Therefore, when selecting a clustering algorithm, it is important to consider the balance between speed and accuracy, as well as the availability of computational resources (Arthur & Vassilvitskii, 2006; Chen & Guestrin, 2016).

- Robustness

The last conclusion of this thesis refers to robustness that is defined as the ability of an algorithm to handle noise and outliers in the data. In the case of clustering algorithms, this means that the algorithm can still generate accurate clusters even if there are some data points that deviate significantly from the rest. XGBoost is considered more robust than K-Means because it utilizes boosting and objective function optimization, which helps to improve the algorithm's performance and reduce the impact of outliers (Chen & Guestrin, 2016). On the other hand, K-Means is more sensitive to noise and outliers, which can skew the centroids of the clusters and result in suboptimal cluster assignments (Arthur & Vassilvitskii, 2007). Thus, XGBoost is a more reliable choice when dealing with data that may contain noise or outliers.

5.1.3 Future of the data collection in social media

The future of data collection in social media for digital marketing is poised to undergo significant changes in the coming years. With the increasing adoption of social media platforms, there is a growing need for companies to collect and analyze large amounts of data to gain insights into consumer behavior and preferences (Van Esch & Stewart, 2021)

One of the key trends that will shape the future of data collection is the rise of artificial intelligence (AI) and machine learning (ML) technologies (Grover *et al*., 2022). AI and ML can help companies to analyze vast amounts of data quickly and efficiently, allowing them to identify patterns and trends that might be difficult or impossible to detect manually (Rudin, 2019). In addition, these technologies can help companies to automate data collection processes, reducing the need for human intervention and improving data accuracy (Arora *et al.*, 2019).

Another trend that will shape the future of data collection is the growing importance of privacy and data security (Dwivedi *et al.*, 2021). As consumers become more aware of the risks associated with sharing personal information online, companies will need to be more transparent about how they collect and use data. They will also need to implement more robust security measures to protect sensitive data from unauthorized access or theft (Dwivedi *et al.*, 2021).

To remain competitive in the digital marketing landscape, companies will also need to be more innovative in their data collection methods. For example, they may need to explore new sources of data, such as wearable devices or smart home systems, to gain a more comprehensive understanding of consumer behavior (Pew Research Center, 2019). Additionally, companies may need to use more sophisticated data analysis techniques, such as predictive analytics or natural language processing, to gain deeper insights into consumer preferences and behavior (Halevy *et al.*, 2009).

It is important to note that the collection and use of consumer data also raises ethical concerns, such as the potential for data breaches, invasion of privacy, and bias in decision-making (Floridi, 2013). Companies must be transparent in their data collection practices and ensure that they are compliant with regulations and ethical guidelines.

In conclusion, the future of data collection in social media for digital marketing is set to be shaped by AI and ML technologies, a growing focus on privacy and data security, and a need for more innovative data collection methods. Companies that are able to adapt to these changes and develop effective data collection strategies while also being mindful of ethical concerns will be better positioned to gain a competitive edge in the digital marketplace.

5.2. Contribution

This thesis has provided a unique perspective on social media data collection in digital marketing by exploring various aspects that have not yet been thoroughly investigated in the existing literature. Firstly, address the gap in the literature in the collection of consumers' personal data from social media for digital marketing purposes. The thesis conducts a bibliometric analysis to identify the current trends and future lines of research on the topic and provides a theoretical framework and discusses the potential of social networks to facilitate relationships between subjects from different backgrounds. It also highlights the challenges faced by digital marketers in handling the enormous volume of information generated by social media (Kumar *et al.*,2016). The study proposes a narrower focus to increase the value of information to researchers in the field and concludes by summarizing the main findings, limitations, and future lines of research. Overall, contributes to the understanding of the collection of consumers' personal data from social media for digital marketing purposes and provides insights for future research in this area.

Based on the results obtained from the bibliometric research, this thesis provides insights into the motivations and characteristics of users who intentionally provide false information when registering for online sweepstakes and quizzes and offers suggestions for improving mechanisms to filter out cheaters and avoid including them in databases. Additionally, it highlights the challenges of filtering and analyzing the enormous flow of information on social networks for digital marketing purposes. Overall, provides valuable information for academics and practitioners interested in understanding user behavior in the digital environment and improving data privacy and security measures (Addae *et al.*, 2019). Hence, the analysis shows that intentional disinformation is the main reason for errors in online sweepstakes and quizzes, and there are differences in tendencies to provide incorrect information among different generations and sexes. The research found that older male generations and middle-aged female generations being more likely to cheat when registering their personal data. Regarding, the motivations behind fraudulent data entry were related to privacy concerns, trust in the company or website, and amusement.

This thesis also has contributed to highlighting that some kind of algorithms have better performance to run cluster market analysis to predict their future behaviour when data from social media is collected. Supervised learning algorithms, concretely XGBoost, had appear to be the more appropriate algorithm for the analysis of online test and sweepstakes data. XGBoost outperforms, non-supervised algorithms, specifically, K-means for this type of data. XGBoost has shown superior performance in classification and regression tasks across a variety of domains, particularly when dealing with large amounts of features and imbalanced data. Additionally, XGBoost only requires labelled data to train the model, making it a suitable option for predicting a target variable, such as the probability of a user converting to a purchase (Zhao *et al*., 2022). Thus, this research has demonstrated that XGBoost has been a more appropriate algorithm for the analysis of online test and sweepstakes data and demonstrated high performance in classification and regression tasks and is particularly suitable for dealing with large amounts of features and imbalanced data. Moreover, XGBoost provides a graphical representation of the decision-making process, making it easier to interpret the relationship between features and the target variable.

In general terms, this thesis has contributed to highlighting some of the most relevant aspects of the collection and analysis of personal data of consumers from social networks for digital marketing purposes. The comprehensive review of the literature presented in this study has identified the most relevant trends in data collection methods, which will undoubtedly be useful for future research in this area. Additionally, this thesis sheds light on the motivations behind the provision of false information by certain user profiles, emphasizing the need for stronger data privacy and security measures. Finally, the study presents a clear comparison between supervised and non-supervised algorithms in clustering consumers based on their social media data, concluding that XGBoost outperforms K-means in this regard. Overall, this thesis makes a valuable contribution to the field of digital marketing by providing new insights into the complex world of personal data collection and analysis.

5.3 Implications

This thesis provides implications both for academia and practitioners in the field. On the academic side, this study contributes to the existing literature by providing a comprehensive analysis on Social Media Data Collection. By conducting thorough research and using advanced methods, this thesis provides valuable insights into the subject matter, offering a theoretical framework that can be used for further research. Additionally, this study fills the gap in the literature on this topic, providing a foundation for future research in the field.

On the other hand, the practical implications of this study are equally significant. By offering insights that can be applied to real-world situations, practitioners in the field can use this research to improve their practices. This thesis provides suggestions and recommendations for practitioners on how to improve their strategies in their respective fields. By highlighting the challenges and opportunities that exist, this study equips practitioners with the necessary tools and knowledge to make informed decisions.

The theoretical framework and practical recommendations offered in this Thesis can be used to advance knowledge in the field and improve practices in the industry.

5.3.1. Academic implications

Concretely, in terms of research implications, this thesis addresses a gap in the existing literature on social media data collection for digital marketing purposes. By conducting a bibliometric analysis, the study identifies current trends and future research lines in this area, providing a theoretical framework to discuss the potential of social networks. The study also highlights the challenges faced by digital marketers in dealing with large volumes of data and offers insights into the motivations and characteristics of users who intentionally provide false information. Additionally, the study demonstrates the superiority of XGBoost over K-means for analyzing online test and sweepstakes data, contributing to the understanding of user behavior, data privacy, and security measures.

5.3.2. Practical implications

Specifically, the practical implications of this thesis are equally significant, as it offers digital marketers, suggestions for improving their data collection strategies on social media. By identifying differences in tendencies to provide incorrect information among different generations and sexes, the study helps marketers filter out cheaters and avoid their inclusion in databases. Furthermore, the study informs marketers on the motivations behind fraudulent data entry and recommends the use of XGBoost for better market analysis and predicting user behavior. The graphical representation of XGBoost's features and target variables enhances understanding of their relationship. Lastly, the study encourages stronger data privacy and security measures in digital marketing practices to protect user data.

Overall, this thesis provides valuable insights for both academia and industry, advancing knowledge in the field of social media data collection and digital marketing and are summarized in table 32.

Table 32. Summarize Academic and Practical Implications

Implications	*Description*
Academic	1. Addresses the gap in literature. 2. Conducts a bibliometric analysis. 3. Provides a theoretical framework. 4. Highlights challenges in handling data. 5. Offers insights into user motivations. 6. Demonstrates XGBoost's superiority. 7. Contributes to understanding user behavior.
Practical	1. Improves data collection strategies. 2. Provides suggestions for filtering cheaters. 3. Identifies differences in tendencies. 4. Informs on motivations behind fraudulent data. 5. Recommends XGBoost for market analysis. 6. Enhances understanding of feature relationships. 7. Encourages stronger data privacy and security measures.

Source: Own elaboration

5.4. Limitations

When reflecting on the scope of this thesis, it's important to note that some limitations may arise from the research questions and methods employed, which could potentially limit a full understanding of the research problem. Additionally, constraints like resources available for data collection and analysis may impact the breadth and depth of the research investigation, while external factors like social, cultural, or political contexts could influence the interpretation and application of the study's findings.

More specifically, in chapter 2, certain limitations should be acknowledged. The keyword analysis was restricted to publications that correspond to the intersection of three key words, which could have led to different results if other combinations had been considered. Additionally, the bibliometric analysis was limited to the Web of Science Core Collection, which might not provide a complete analysis of the field. Furthermore, the choice of data was restricted to articles published in journals, neglecting other sources like books and conference papers. Finally, the study recognizes that the results might have varied if additional sources had been considered.

As for the studies presented in chapters 3 and 4, it's important to recognize that collecting, filtering, and analyzing the constant stream of information from social networks is a significant challenge that requires continuous monitoring. Additionally, the studies relied on a database of participants in online sweepstakes and quizzes, which may not represent the entire population, limiting the generalizability of the results. Lastly, the sample used in the studies was from Spain, which could limit the generalizability of the findings to other geographical regions.

5.5. Future research

Science advances are usually conducted incrementally. Acknowledging limitations can provide direction for further investigation and indicate potential avenues for future research. This section presents the limitations of this research and suggests ideas for future lines of research.

Firstly, to broaden the scope of the study, it may be beneficial to incorporate additional academic databases, such as Scopus and Google Scholar, in order to create more comprehensive categorizations of journals, scholars, academic organizations, and countries (Koberg & Longoni, 2018). This would allow for the extension of findings to a wider range of publications about digital marketing, and the drawing of conclusions that are applicable to the entire spectrum of literature on this topic. Another interesting future line of research would be to analyse the means used to capture personal data from social networks, and it would also be useful to know what kinds of rewards and incentives most motivate consumers to give up their personal data.

Future research should also include validating the results against data from other lead generation companies and unstructured data on user behavior (Jung *et al.*, 2020; Choudrie *et al.*, 2021). Also, contrasting the results with other web data collection formats (Cruz-Benito *et al.*, 2018), enriching studies by taking into account recruitment sources or methodologies, different origins and social networks, and attitudes that differ depending on the country that users come from (Parekh *et al.*, 2018; Borges-Tiago *et al.*, 2020), examining how different profiles behave in terms of decision-purchase-post-purchase behavior and studying the clustering of consumer profiles by sector to analyze how the resulting algorithm is affected by false information entered (Altman & Bland, 1998). At the same time, the ethical implications of systematically excluding or limiting the participation of certain users in prize draws and tests due to unintentional errors, such as older people who are more affected by health conditions and accessibility issues, should be investigated, and why female members of the same cohorts do not seem to be affected by such difficulties to the same extent (Altman & Bland, 1998), comparing intentional errors made by more mature people with more life experience and who have adopted these technologies much later, with younger generations who were born in the age of the Internet and social networks (Valentine & Powers, 2013; Dabija & Grant, 2016; Lenhart *et al.*, 2010). Implications for findings managemen~~t~~, such as improving mechanisms to filter out cheaters and avoid including them in databases, should also be studied (Bolton *et al*., 2013; Bondarenko *et al.*, 2019; Di Domenico & Visentin, 2020).

Moreover, future research could investigate the use of other machine learning algorithms for data analysis. For instance, researchers could examine how other algorithms perform in different contexts and compare their results with XGBoost and K-means algorithms (Chen & Guestrin, 2016; MacQueen, 1967). Another interesting research line could be exploring the use of hybrid approaches. In some cases, combining different algorithms to analyze data could lead to more accurate and nuanced results (Chen & Guestrin, 2016; MacQueen, 1967). Researchers could investigate the impact of different initialization strategies to develop new methods that minimize the effect of initialization on clustering results (MacQueen, 1967).

Additionally, researchers could investigate the relative interpretability of different algorithms in different contexts and develop methods for improving the understandable of less comprehensible algorithms (Chen & Guestrin, 2016; Breiman, 2017). Finally, exploring the impact of different data pre-processing techniques is also essential because machine learning algorithms can be heavily influenced by the quality of the data being analyzed (Chen & Guestrin, 2016).

REFERENCE LIST

Addae, J. H., Sun, X., Towey, D., & Radenkovic, M. (2019). Exploring user behavioral data for adaptive cybersecurity. *User Modeling and User-Adapted Interaction*, 29, 701-750.

Agrawal, D., Bamieh, B., Budak, C., El Abbadi, A., Flanagin, A. J., & Patterson, S. (2011, August). *Data-Driven Modeling and Analysis of Online Social Networks*. In WAIM (pp. 3-17).

Ahmed, D. T., & Shirmohammadi, S. (2009). An Algorithm for Measurement and Detection of Path Cheating in Virtual Environments. In 2009 *IEEE International Conference on Virtual Environments, Human-Computer Interfaces and Measurements Systems*, 138–142, Hong Kong, China.

Ailawadi, K. L., Neslin, S. A., & Gedenk, K. (2001). Pursuing the value-conscious consumer: store brands versus national brand promotions. *Journal of marketing*, 65(1), 71-89.

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

Ali, I., Balta, M., & Papadopoulos, T. (2022). Social media platforms and social enterprise: Bibliometric analysis and systematic review. *International Journal of Information Management*, 102510.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 211-236.

Altman, D. G., & Bland, J. M. (1998). Generalisation and extrapolation. British Medical Journal, 317(7155), 409-410.

Ambler, T. (2003). Marketing and the bottom line: the marketing metrics to pump up cash flowPearson Education.

Amoozad Mahdiraji, H., Hafeez, K., Kord, H., & Abbasi Kamardi, A. (2022). Analysing the voice of customers by a hybrid fuzzy decision-making approach in a developing country's automotive market. *Management Decision*, 60(2), 399–425.

Appel, G., Grewal, L., Hadi, R., & Stephen, A. T. (2020). The future of social media in marketing. *Journal of the Academy of Marketing science*, 48(1), 79-95.

Arabie, P., Hubert, L., & De Soete, G. (Eds.). (1996). Clustering and classification. NJ. World Scientific Publishing.

Arora, A., Bansal, S., Kandpal, C., Aswani, R., & Dwivedi, Y. (2019). Measuring social media influencer index-insights from Facebook, Twitter and Instagram. *Journal of retailing and consumer services*, 49, 86-101.

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 1027-1035).

Ashley, C., & Tuten, T. (2015). Creative strategies in social media marketing: An exploratory study of branded social content and consumer engagement. *Psychology & Marketing, 32*(1), 15–27.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211(4489), 1390-1396.

Bala, M., & Verma, D. (2018). A critical review of digital marketing. M. Bala, D. Verma (2018). A Critical Review of Digital Marketing. *International Journal of Management, IT & Engineering*, 8(10), 321-339.

Balint, M., Posea, V., Dimitriu, A., & Iosup, A. (2011). An analysis of social gaming networks in online and face-to-face bridge communities. *In Proceedings of the 3rd International Workshop on Large-Scale System and Application Performance*. ACM, 35–42. California, USA.

Ball, G. H., & Hall, D. J. (1967). A clustering technique for summarizing multivariate data. *Behavioral science*, 12(2), 153-155.

Baltar, F., & Brunet, I. (2012). Social research 2.0: virtual snowball sampling method using Facebook. *Internet research*, 22(1), 57-74.

Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. *Ai & Society*, 30, 89-116.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., ... & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37), 9216-9221.

Bianchi, C., & Andrews, L. (2015). Investigating marketing managers' perspectives on social media in Chile. *Journal of Business Research*, *68*(12), 2552–2559.

Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.

Blackburn, J., Kourtellis, N., Skvoretz, J., Ripeanu, M., & Iamnitchi, A. (2014). Cheating in online games: A social network perspective. *ACM Transactions on Internet Technology*, 13(3): 9:1-9:25.

Boer, M., Stevens, G. W., Finkenauer, C., de Looze, M. E., & van den Eijnden, R. J. (2021). Social media use intensity, social media use problems, and mental health among adolescents: Investigating directionality and mediating processes. *Computers in Human Behavior*, 116, 106645.

Bolton, R. N., Parasuraman, A., Hoefnagels, A., Migchels, N., Kabadayi, S., Gruber, T., Komarova Loureiro, Y., & Solnet, D. (2013). Understanding Generation Y and their use of social media: a review and research agenda. *Journal of Service Management*, 24(3), 245-267.

Bonald, T., Feuillet, M., & Proutiere, A. (2009). Is the "Law of the Jungle" Sustainable for the Internet?”. In IEEE INFOCOM 2009, pp. 28-36, Rio de Janeiro, Brazil.

Bondarenko, S., Laburtseva, O., Sadchenko, O., Lebedieva, V., Haidukova, O., & Kharchenko, T. (2019). Modern lead generation in internet marketing for the development of enterprise potential. *International Journal of Innovative Technology and Exploring Engineering*, 8(12), 3066-3071.

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences*, 497*,* 38–55.

Boone, D. S., & Roehm, M. (2002). Retail segmentation using artificial neural networks. *International Journal of Research in Marketing*, 19(3), 287-301.

Borges-Tiago, T., Tiago, F., Silva, O., Guaita Martinez, J. M., & Botella-Carrubi, D. (2020). Online users' attitudes toward fake news: Implications for brand management. *Psychology & Marketing*, 37(9), 1171-1184.

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication*, 13(1), 210-230.

Boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662-679.

Breiman, L. (2017). Classification and regression trees. New York: Routledge.

BPS (The British Psychological Society). (2009). Code of ethics and conduct: Guidance published by the Ethics Committee of the British Psychological Society. Leicester: The British Psychological Society.

Broadus, R. N. (1987). Toward a definition of "Bibliometrics". *Scientometrics,12* (5-6), 373-379.

Brosdahl, D. J. C., & Carpenter, J. M. (2011). Shopping orientations of US males: A generational cohort comparison. *Journal of Retailing and Consumer Services,* 18(6), 548-554.

Bruns, A., & Burgess, J. E. (2011). The use of Twitter hashtags in the formation of ad hoc publics. *Paper presented at the European Consortium for Political Research General Conference,* Reykjavik, Iceland.

Brynjolfsson, E., & McAfee, A. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies. New York, US. WW Norton & Company.

Buchanan, E. A., & Hvizdak, E. E. (2009). Online survey tools: Ethical and methodological concerns of human research ethics committees. *Journal of empirical research on human research ethics*, 4(2), 37-48.

Bughin, J., Seong, J., Manyika, J., Chui, M., & Joshi, R. (2018). Notes from the AI frontier: Modeling the impact of AI on the world economy. McKinsey Global Institute, 4.

Bullas, J. (2021). The top 10 social media metrics you should track. Retrieved on 4th April 2023 from https://www.jeffbullas.com/top-10-social-media-metrics-you-should-track/

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine,* 9(2), 48-57.

Cancino, C. A., Merigó, J. M., & Coronado, F. (2017). Big names in innovation research: A bibliometric view. *Current Science,* 13(8), 1507-1518.

Canhoto, A. I., & Clark, M. (2013). Customer service 140 characters at a time: The users' perspective. *Journal of marketing Management,* 29(5-6), 522-544.

Cavusgil, S. T., Kiyak, T., & Yeniyurt, S. (2004). Complementary approaches to preliminary foreign market opportunity assessment: Country clustering and country ranking. *Industrial Marketing Management*, 33(7), 607-617.

Chakraborty, H., Moore, J., Carlo, W. A., Hartwell, T. D., & Wright, L. L. (2009). A simulation based technique to estimate intracluster correlation for a binary variable. *Contemporary clinical trials*, 30(1), 71-80.

Chambers, C., Feng, W.-C., Sahu, S., Saha, D., & Brandt, D. 2010. Characterizing online games. *IEEE/ACM Transactions on Networking*, 18(3), 899–910.

Chaovalit, P., & Zhou, L. (2005, January). Movie review mining: A comparison between supervised and unsupervised classification approaches. *In Proceedings of the 38th annual Hawaii international conference on system sciences* (pp. 112c-112c). IEEE.

Chen, J., & Dibb, S. (2010). Consumer trust in the online retail context: Exploring the antecedents and consequences. *Psychology & Marketing*, 27(4), 323-346.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Chen, W., Qiu, X., Cai, T., Dai, H. N., Zheng, Z., & Zhang, Y. (2021). Deep reinforcement learning for Internet of Things: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 23(3), 1659-1692.

Chen, Y., Fay, S., & Wang, Q. (2011). The role of marketing in social media: How online consumer reviews evolve. *Journal of interactive marketing*, 25(2), 85-94.

Chlebus, E., Chrobot, J., Gąbka, J., & Susz, S. (2011). Clusters as a modern pattern of running business supporting innovation. *Management and Production Engineering Review*, 2(2), 71-79.

Choudrie, J., Banerjee, S., Kotecha, K., Walambe, R., Karende, H., & Ameta, J. (2021). Machine learning techniques and older adults processing of online information and misinformation: A covid 19 study. *Computers In Human Behavior*, *119*, 106716.

Conroy, N.J., Rubin, V.L., & Chen, Y. (2015). Automatic deception detection: methods for fnding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1), 1–4.

Cosmides, L., & Tooby, J. (2016). Adaptations for reasoning about social exchange. In D.M. Buss (Eds.), The Handbook of Evolutionary Psychology (pp. 625-668). New Jersey: John Wiley & Sons, Inc.

Costas, R., and Bordons, M. (2007). The h-index: Advantages, limitations and their relationship with other micro-level bibliometric indicators. *Journal of Informetrics, 1*(3), 193-203.

Creswell, J. W., & Clark, V. L. P. (2017). Designing and conducting mixed methods research. Sage publications.

Cronin, B., Perra, N., Rocha, L. E. C., Zhu, Z., Pallotti, F., Gorgoni, S., ... & De Vita, R. (2021). *Ethical implications of network data in business and management settings. Social Networks*, 67, 29-40.

Cruz-Benito, J., Sánchez-Prieto, J. C., Vázquez-Ingelmo, A., Therón, R., García-Peñalvo, F. J., & Martín-González, M. (2018). How different versions of layout and complexity of web forms affect users after they start it? A pilot experience. In Trends and Advances in Information Systems and Technologies: Volume 2 6 (pp. 971-979). Springer International Publishing.

da Fonseca, J. M. R., & Borges-Tiago, M. T. (2021). Cyberbullying From a Research Viewpoint: A Bibliometric Approach. In Handbook of Research on Cyber Crime and Information Privacy (pp. 182-200). IGI Global.

Dabija, D.-C., & Grant, D. B. (2016). Investigating Shopping Experience and Fulfillment in Omnichannel Retailing: A Proposed Comparative Study in Romanian and UK of Generation Y Consumers. *In The 21st LRN Annual Logistic Network Conference*, *Hull*, UK.

Dabija, D.-C., Bejan, B. M., & Tipi, N. (2018). Generation X versus Millennials communication behaviour on social media when purchasing food versus tourist services. *E+M Ekonomie a Management*, 21(1), 191-205.

Dao, S. D., Abhary, K., & Marian, R. (2017). A bibliometric analysis of Genetic Algorithms throughout history. *Computers & Industrial Engineering*, *110*, 395-403.

Dasgupta, S. (2016, June). A cost function for similarity-based hierarchical clustering. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing* (pp. 118-127).

Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. *Harvard business review*, 96(1), 108-116.

Dayan, N., Twitto, M., Rochman, Y., Beitler, U., Zion, I. B., Bortnikov, E., ... & Rabinovich, N. (2021). The end of Moore's law and the rise of the data processor. *Proceedings of the VLDB Endowment*, 14(12), 2932-2944.

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Social media as a measurement tool of depression in populations. *In Proceedings of the 5th Annual ACM Web Science Conference* (pp. 47-56).

De Luca, L. M., Herhausen, D., Troilo, G., & Rossi, A. (2021). How and when do big data investments pay off? The role of marketing affordances and service innovation. *Journal of the Academy of Marketing Science, 49*(4).

De Vries, N. J., & Carlson, J. (2014). Examining the drivers and brand performance implications of customer engagement with brands in the social media environment. *Journal of Brand Management*, 21, 495-515.

Denzin, N.K. (1978). The Research Act: A Theoretical Introduction to Sociological Methods (2nd ed.). McGraw-Hill, New York, NY.

Desai, V. (2019). Digital marketing: A review. *International Journal of Trend in Scientific Research and Development*, *5*(5), 196-200.

DeSarbo, W. S., & Grisaffe, D. (1998). Combinatorial optimization approaches to constrained market segmentation: An application to industrial market segmentation. *Marketing Letters*, 9, 115-134.

Di Domenico, G., & Visentin, M. (2020). Fake news or true lies? Reflections about problematic contents in marketing. *International Journal of Market Research, 62*(4), 409-417.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, B. L. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research*, 38(1), 1-18.

Ding, Y., Rousseau, R., & Wolfram, D. (2014). *Measuring academic impact: Methods and practice*. Cham, Switzerland: Springer.

Diouf, R., Sarr, E. N., Sall, O., Birregah, B., Bousso, M., & Mbaye, S. N. (2019, December). Web scraping: state-of-the-art and areas of application. *In 2019 IEEE International Conference on Big Data (Big Data)* (pp. 6040-6042). IEEE.

Dwivedi, Y. K., Ismagilova, E., Hughes, D. L., Carlson, J., Filieri, R., Jacobson, J., Jain, V., Karjaluoto, H., Kefi, H., Krishen, A. S., Kumar, V., Rahman, M. M., Raman, R., Rauschnabel, P. A., Rowley, J., Salo, J., Tran, G. A., & Wang, Y. (2021). Setting the future of digital and social media marketing research: Perspectives and research propositions. *International Journal of Information Management*, 59, 102168.

Epstein, M. J., & Buhovac, A. R. (2014). Making sustainability work 2nd edition: Best Practices in managing and measuring corporate social, environmental, and economic impacts. San Francisco (CA, US): Berrett-Koehler Publishers.

Escobar-Jeria, V. H., Martín-Bautista, M. J., Sánchez, D., & Vila, M.-A. (2007). Analysis of Log Files Applying Mining Techniques and Fuzzy Logic. H.G. Okuno, & M. Ali (Eds.). BT - *New Trends in Applied Artificial Intelligence* (pp. 483-492), Berlin Heidelberg. Springer

Eskerod, P. (2020). A stakeholder perspective: Origins and core concepts. In Oxford Research Encyclopedia of Business and Management.

Ess, C., & Jones, S. (2004). Ethical decision-making and Internet research: Recommendations from the AoIR ethics working committee. *In Readings in virtual research ethics: Issues and controver*sies (pp. 27-44). IGI Global.

Everitt, B. S. (1979). Unresolved problems in cluster analysis. Biometrics, 169-181.

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110, 104743.

Ezugwu, A. E., Shukla, A. K., Agbaje, M. B., Oyelade, O. N., José-García, A., & Agushaka, J. O. (2021). Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature. *Neural Computing and Applications*, 33, 6247-6306.

Fallis, D. (2014). The Varieties of Disinformation. In L. Floridi, & P. Illari (Eds.). The Philosophy of Information Quality (pp. 135-161). Cham: Springer.

Fatima, R., Yasin, A., Liu, L., Wang, J., Afzal, W., & Yasin, A. (2019). Sharing information online rationally: An observation of user privacy concerns and awareness using serious game. *Journal of Information Security and Applications*, 48, 102351.

Felix, R., Rauschnabel, P. A., & Hinsch, C. (2017). Elements of strategic social media marketing: A holistic framework. *Journal of business research*, 70, 118-126.

Floridi, L. (2013). The ethics of information. Oxford University Press.

Forgy, C. L. (1989). Rete: A fast algorithm for the many pattern/many object pattern match problem. *In Readings in Artificial Intelligence and Databases* (pp. 547-559).

Fournier, S., & Avery, J. (2011). The uninvited brand. *Business Horizons*, *54*(3), 193–207.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4), 367-378.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2), 337-407.

Friggeri, A., Adamic, L., Eckles, D., & Cheng, J. (2014, May). Rumor cascades. *In proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 101-110).

Fritsch, T., Voigt, B., & Schiller, J. (2006). Distribution of online hardcore player behavior: (how hardcore are you?). *In Proceedings of 5th ACM SIGCOMM Workshop on Network and System Support for Games*. ACM, USA.

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

Gaviria-Marin, M., Merigo, J. M., & Popa, S. (2018). Twenty years of the Journal of Knowledge Management: A bibliometric analysis. *Journal of Knowledge Management*, *22*, 1655-1687.

Gefen, D., Karahanna, E., & Straub, D.W. (2003). Trust and TAM in online shopping: an integrated model. *MIS Quarterly*, 27(1), 51-90.

Ghorbani, Z., Kargaran, S., Saberi, A., Haghighinasab, M., Jamali, S. M., & Ale Ebrahim, N. (2021). Trends and patterns in digital marketing research: bibliometric analysis. *Journal of Marketing Analytics*, *10*(5), 158-172.

Godes, D., & Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing science*, 23(4), 545-560.

Golder, S., Ahmed, S., Norman, G., & Booth, A. (2017). Attitudes toward the ethics of research using social media: a systematic review. *Journal of medical internet research*, 19(6), e195.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Goyal, K., & Kumar, S. (2021). Financial literacy: A systematic review and bibliometric analysis. *International Journal of Consumer Studies*, 45(1), 80-105.

Grbovic, M., Radosavljevic, V., Djuric, N., Bhamidipati, N., Savla, J., Bhagwan, V., & Sharp, D. (2015, August). E-commerce in your inbox: Product recommendations at scale. *In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1809-1818).

Grewal, D., Herhausen, D., Ludwig, S., & Villarroel Ordenes, F. (2021). The future of digital communication research: Considering dynamics and multimodality. *Journal of Retailing*, *98*(2), 224-240.

Grishikashvili, K., Dibb, S., & Meadows, M. (2014). Investigation into Big Data Impact on Digital Marketing. *Online Journal of Communication and Media Technologies*, *4*(Special Issue), 26-37.

Grover, P., Kar, A. K., & Dwivedi, Y. K. (2022). Understanding artificial intelligence adoption in operations management: insights from the review of academic literature and social media discussions. *Annals of Operations Research,* 308(1-2), 177-213.

Gruzd, A., Wellman, B., & Takhteyev, Y. (2011). Imagining Twitter as an imagined community. *American Behavioral Scientist,* 55(10), 1294-1318.

Gultom, S., Sriadhi, S., Martiano, M., & Simarmata, J. (2018, September). Comparison analysis of K-means and K-medoid with Ecluidience distance algorithm, Chanberra distance, and Chebyshev distance for big data clustering. *In IOP Conference Series: Materials Science and Engineering* (Vol. 420, No. 1, p. 012092). IOP Publishing.

Habib, A., Asghar, M.Z., & Khan, A. (2019). False information detection in online content and its role in decision making: a systematic literature review. *Social Network Analysis and Mining,* 9(1), 1-20.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE intelligent systems*, 24(2), 8-12.

Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan kaufmann.

Hanelt, A., Bohnsack, R., Marz, D., & Antunes Marante, C. (2021). A systematic review of the literature on digital transformation: Insights and implications for strategy and organizational change. *Journal of Management Studies*, 58(5), 1159-1197.

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.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

Henriques, J., Caldeira, F., Cruz, T., & Simões, P. (2020). Combining k-means and xgboost models for anomaly detection using log datasets. *Electronics*, 9(7), 1164.

Hirsch, J. E. (2005). An index for quantifying an individual's scientific research output. *Pnas,*102 (46) 16569-16572.

Hoffman, D. L., & Novak, T. P. (2009). Flow online: lessons learned and future prospects. *Journal of Interactive Marketing*, 23(1), 23-34

Homburg, C., Jozić, D., & Kuehnl, C. (2017). Customer experience management: toward implementing an evolving marketing concept. *Journal of the Academy of Marketing*

Science, 45, 377-401.

Hootsuite (2021). Social media trends 2021. Retrieved on 4th April 2023 from https://blog.hootsuite.com/social-media-trends-2021/

Hruschka, H., & Natter, M. (1999). Comparing performance of feedforward neural nets and K-means for cluster-based market segmentation. *European Journal of Operational Research*, 114(2), 346-353.

Huang, R., & Sarigöllü, E. (2012). How brand awareness relates to market outcome, brand equity, and the marketing mix. *Journal of Business Research*, 65(1), 92-99.

Huang, S.-W., Suh, M., Hill, B. M., & Hsieh, G. (2015). How Activists Are Both Born and Made: An Analysis of Users on Change.org. *In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 211-220), Seoul, Republic of Korea.

Huertas-Garcia, R., Gázquez-Abad, J.C., & Forgas-Coll, S. (2016). A design strategy for improving adaptive conjoint analysis. *Journal of Business & Industrial Marketing*, 31(3), 328-338.

INE. (2020). Survey on Equipment and Use of Information and Communication Technologies in Households 2019, INEbase, National Institute of Statistics. Retrieved 6/7/2022 from https://www.ine.es/dynt3/inebase/en/index.htm?padre=6898&capsel=6933

Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020). Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(1), 1-20.

Jacobsen, C., Fosgaard, T. R., & Pascual-Ezama, D. (2018). Why do we lie? A practical guide to the dishonesty literature. *Journal of Economic Surveys*, 32(2), 357-387.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651-666.

Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Prentice-Hall, Inc.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3), 264-323.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11), 2169-2188.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255-260.

Jung, W.-J., Yang, S., & Kim, H.-W. (2020). Design of sweepstakes-based social media marketing for online customer engagement. *Electronic Commerce Research*, 20(1), 119-146.

Kamthania, D., Pawa, A., & Madhavan, S. S. (2018). Market segmentation analysis and visualization using K-mode clustering algorithm for E-commerce business. *Journal of computing and information technology*, 26(1), 57-68.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, *53*(1), 59-68.

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

Karlova, N. A., & Fisher, K. E. (2013). A social diffusion model of misinformation and disinformation for understanding human information behaviour", *Information Research*, 18(1), 573.

Karpińska-Krakowiak, M., & Modliński, A. (2018). The Effects of Pranks in Social Media on Brands. *Journal of Computer Information Systems*, 58(3), 282-290.

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis. Hoboken, NJ, USA. John Wiley & Sons.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. A*merican Documentation*, *14*(1).

Keusch, F., Struminskaya, B., Antoun, C., Couper, M., & Kreuter, F. (2019). Willingness to participate in passive mobile data collection. *Public Opinion Quarterly*, 83, 210-235.

Kharchenko, T. (2019). Modern lead generation in internet marketing for the development of enterprise potential. *International Journal of Innovative Technology and Exploring Engineering*, *8*(12).

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business horizons*, 54(3), 241-251.

Ko, S., Cho, I., Afzal, S., Yau, C., Chae, J., Malik, A., ... & Ebert, D. S. (2016, June). A survey on visual analysis approaches for financial data. *In Computer Graphics Forum* (Vol. 35, No. 3, pp. 599-617).

Koberg, E., & Longoni, A. (2018). A systematic review of sustainable supply chain management in global supply chains. *Journal of Cleaner Production*, *207*, 1084-1098.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal*, 1(6), 90-95.

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. *In Ijcai* (Vol. 14, No. 2, pp. 1137-1145).

Kramer, A. D., & Guillory, J. E. (2016). Automatic detection of emotional contagion from massive social networks. *PloS one,* 11(1), e0148390.

Kraut, R. E., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: Report of Board of Scientific Affairs' Advisory Group on the Conduct of Research on the Internet. *American Psychologist*, 59(2), 105-117.

Krishen, A. S., Dwivedi, Y. K., Bindu, N., & Kumar, K. S. (2021). A broad overview of interactive digital marketing: A bibliometric network analysis. *Journal of Business Research*, *131*(C), 183-195.

Kross, E., Verduyn, P., Demiralp, E., Park, J., Lee, D. S., Lin, N., ... & Ybarra, O. (2013). Facebook use predicts declines in subjective well-being in young adults. *PloS one*, 8(8), e69841.

Kumar, A., Bezawada, R., Rishika, R., Janakiraman, R., & Kannan, P. K. (2016). From social to sales: The Effects of Firm-Generated Content in Social Media on Customer Behavior. *Journal of Marketing*, *80*(1), 7-25.

Kumar, V., Petersen, J. A., & Leone, R. P. (2010). Driving profitability by encouraging customer referrals: Who, when, and how. *Journal of Marketing*, 74(5), 1-17.

Kuo, R. J., Ho, L. M., & Hu, C. M. (2002). Integration of self-organizing feature map and K-means algorithm for market segmentation. *Computers & Operations Research*, 29(11), 1475-1493.

Laengle, S., Merigó, J. M., Miranda, J., Slowinski, R., Bomze, I., Borgonovo, E., Teunter, R. (2017). Forty years of the European Journal of Operational Research:A bibliographical overview. *European Journal of Operational Research*, *262* (3), 803-816.

Laengle, S., Modak, N. M., Merigó, J. M., & Zurita, G. (2018). Twenty-five years of group decision and negotiation: A bibliometric overview. *Group decision and negotiation*, *27*(4), 505-542.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4), 541-551.

Lee, K.-C., Orten, B., Dasdan, A., & Li, W. (2012). Estimating conversion rate in display advertising from past performance data. *In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, Beijing, China.

Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. *Journal of marketing*, 80(6), 69-96.

Lenhart, A., Purcell, K., Smith, A., & Zickuhr, K. (2010). Social Media & Mobile Internet Use among Teens and Young Adults. Millennials. Retrieved 20/6/2022 from https://eric.ed.gov/?id=ED525056

Leung, X. Y., Sun, J., & Bai, B. (2017). Bibliometrics of social media research: A co-citation and co-word analysis. *International Journal of Hospitality Management*, 66, 35-45.

Liang, Y., Wu, J., Wang, W., Cao, Y., Zhong, B., Chen, Z., & Li, Z. (2019, August). Product marketing prediction based on XGboost and LightGBM algorithm. *In Proceedings of the 2nd international conference on artificial intelligence and pattern recognition* (pp. 150-153).

Liao, S. H., Widowati, R., & Hsieh, Y. C. (2021). Investigating online social media users' behaviors for social commerce recommendations. *Technology in Society*, 66, 101655.

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 and Artificial Intelligence*, *5*(5), 134.

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition,* 36(2), 451-461.

Lin, S.-Y., Chen, Y.-W., Kang, H.-C., Wu, Y.-J., Chen, P.-Z., Wu, C.-W., Lin, C.-S., Wu, F.-L. L., Shen, L.-J., Huang, Y.-M., & Huang, C.-F. (2021). Effects of a pharmacist-managed anticoagulation outpatient clinic in Taiwan: evaluation of patient knowledge, satisfaction, and clinical outcomes. *Postgraduate Medicine,* 133(8), 964-973.

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. Sage Publications, London, UK.

Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., & Lin, Z. (2021). When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys* (CSUR), 54(2), 1-36.

Liu, Y., Ram, S., Lusch, R. F., & Brusco, M. (2010). Multicriterion market segmentation: a new model, implementation, and evaluation. *Marketing Science,* 29(5), 880-894.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory,* 28(2), 129-137.

ln, L. I. S. (2013). Data collection techniques a guide for researchers in humanities and education. *International Research Journal of Computer Science and Information Systems* (IRJCSIS), *2*(3), 40-44.

Lomborg, S., & Bechmann, A. (2014). Using APIs for data collection on social media. *The Information Society*, 30(4), 256-265.

Lund, B., & Ma, J. (2021). A review of cluster analysis techniques and their uses in library and information science research: k-means and k-medoids clustering. *Performance*

Measurement and Metrics, 22(3), 161-173.

Lwin, M.O., Wirtz, J., & Stanaland, A.J.S. (2016). The privacy dyad: Antecedents of promotion- and prevention-focused online privacy behaviors and the mediating role of trust and privacy concern. *Internet Research*, 26(4), 919-941.

Maaß, M., Clement, M.-P., & Hollick, M. (2021). Snail Mail Beats Email Any Day: On Effective Operator Security Notifications in the Internet. *In The 16th International Conference on Availability, Reliability and Security*, Vienna, Austria.

MacQueen, J. (1967, June). Classification and analysis of multivariate observations. In 5th Berkeley Symp. Math. Statist. Probability (pp. 281-297). Los Angeles, USA: University of California.

Mealey, L., Daood, C., & Krage, M. (1996). Enhanced memory for faces of cheaters. *Ethology and Sociobiology*, 17(2), 119-128.

Menon, R. V., Sigurdsson, V., Larsen, N. M., Fagerstrøm, A., Sørensen, H., Marteinsdottir, H. G., & Foxall, G. R. (2019). How to grow brand post engagement on Facebook and Twitter for airlines? An empirical investigation of design and content factors. *Journal of Air Transport Management*, *79*, 101678.

Meredith, G., & Schewe, C. (1994). The power of cohorts. *American Demographics*, 16, 22-31.

Merigó, J. M., Cancino, C. A., Coronado, F., & Urbano, D. (2016). Academic research in innovation: a country analysis. *Scientometrics*, *108*(2), 559–593.

Merigó, J. M., Mas-Tur, A., Roig-Tierno, N., & Ribeiro-Soriano, D. (2015). A bibliometric overview of the Journal of Business Research between 1973 and 2014. *Journal of Business Research*, *68*(12), 2645-2653.

Merigó, J. M., Pedrycz, W., Weber, R., & de la Sotta, C. (2018). Fifty years of Information Sciences: A bibliometric overview. *Information Sciences*, *432*, 245-268.

Meyers-Levy, J., & Loken, B. (2015). Revisiting gender differences: What we know and what lies ahead. *Journal of Consumer Psychology*, 25(1), 129-149.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50, 159-179.

Milne, G. R., & Culnan, M. J. (2004). Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices. *Journal of interactive marketing*, 18(3), 15-29.

Mintz A.P. (2002). Web of deception: Misinformation on the Internet. New Jersey, USA .Information Today, Inc.

Mitchell, R., & Frank, E. (2017). Accelerating the XGBoost algorithm using GPU computing. *PeerJ Computer Science*, 3, e127.

Mitchell, T. M. (1997). Artificial neural networks. *Machine learning*, 45(81), 127.

Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8), 142-151.

Morwitz, V. G., & Schmittlein, D. (1992). Using segmentation to improve sales forecasts based on purchase intent: Which “intenders” actually buy?. *Journal of marketing research*, 29(4), 391-405

Moss, G. (2009). Gender, design, and marketing. How Gender Drives our Perception of Design and Marketing. London, Routledge.

Murray, P. W., Agard, B., & Barajas, M. A. (2017). Market segmentation through data mining: A method to extract behaviors from a noisy data set. *Computers & Industrial Engineering*, 109, 233-252.

Nadaraja, R., & Yazdanifard, R. (2013). Social media marketing: advantages and disadvantages. Center of Southern New Hempshire University, 1-10.

Nambisan, S., Wright, M., & Feldman, M. (2019). The digital transformation of innovation and entrepreneurship: Progress, challenges and key themes. *Research Policy*, 48(8), 103773.

Nazir, A., Raza, S., Chuah, C.-N., & Schipper, B. (2010). Ghostbusting facebook: Detecting and characterizing phantom profiles in online social gaming applications. *In Proceedings of the 3rd Workshop on Online Social Networks (WOSN'10)*, Boston MA, USA.

Nickel, K., Orth, U. R., & Kumar, M. (2020). Designing for the genders: The role of visual harmony. *International Journal of Research in Marketing* 37(4), 697-713.

Ogilvie, J., Rapp, A., Bachrach, D. G., Mullins, R., & Harvey, J. (2017). Do sales and service compete? The impact of multiple psychological climates on frontline employee performance. *Journal Of Personal Selling & Sales Management*, *37*(1), 11-26.

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2, 13.

Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature human behaviour*, 3(2), 173-182.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2), 1-135.

Parekh, D., Amarasingam, A., Dawson, L., & Ruths, D. (2018). Studying jihadists on social media: A critique of data collection methodologies. *Perspectives on Terrorism*, 12(3), 5-23.

Parikh S.B, & Atrey P.K. (2018). Media-rich fake news detection: a survey. *In 2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, Miami, FL, USA.

Pascual-Ezama. (2020). Cheaters, Liars, or Both? A New Classification of Dishonesty

Profiles. *Psychological Science*, 31(9), 1097–1106.

Paul, J., & Criado, A. R. (2020). The art of writing literature review: What do we know and what do we need to know?. *International Business Review*, *29*(4), 101717.

Paul, J., & Singh, G. (2017). The 45 years of foreign direct investment research: Approaches, advances and analytical areas. *The World Economy*, *40*(11), 2512-2527.

Paul, N., & DeHart, J. L. (2020). Social Media Use, Political Participation, and Civic Engagement in Election 2016. *The Journal of Social Media in Society*, 9(2), 275-305.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. Cognition, 188, 39-50,

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, 592, 590-595.

Peppers, D., & Rogers, M. (1993). The one to one future: Building relationships one customer at a time. New York: Currency Doubleday.

Pérez-Escoda, A., Barón-Dulce, G., & Rubio-Romero, J. (2021). Mapping media consumption among youngest: Social networks, fake news and trustworthy in pandemic times. *Index Comunicacion*, 187-208.

Pew Research Center. (2019). Tech Adoption Climbs Among Older Adults. Retrieved on 10th April 2023 from https://www.pewresearch.org/internet/2017/05/17/tech-adoption-climbs-among-older-adults/

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. *Advances in Kernel Methods-Support Vector Learning*. 208.

Poulos, M., Korfiatis, N., & Papavlassopoulos, S. (2020). Assessing stationarity in web analytics: A study of bounce rates. *Expert systems*, 37(3), e12502.

Primack, B. A., Shensa, A., Sidani, J. E., Whaite, E. O., yi Lin, L., Rosen, D., ... & Miller, E. (2017). Social media use and perceived social isolation among young adults in the US. *American journal of preventive medicine*, 53(1), 1-8.

Provost, F., & Fawcett, T. (1998, July). Robust classification systems for imprecise environments. *In AAAI/IAAI* (pp. 706-713).

Provost, F. J., Fawcett, T., & Kohavi, R. (1998, July). The case against accuracy estimation for comparing induction algorithms. *In ICML* (Vol. 98, pp. 445-453).

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research*, 20(2), 134-148.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Towards an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, *105*(45), 17268-17272.

Rana, N. P., Chatterjee, S., Dwivedi, Y. K., & Akter, S. (2022). Understanding dark side of artificial intelligence (AI) integrated business analytics: assessing firm’s operational inefficiency and competitiveness. *European Journal of Information Systems*, 31(3), 364-387.

Randhawa, K., Wilden, R., & Hohberger, J. (2016). A bibliometric review of open innovation: Setting a research agenda. *Journal of Product Innovation Management*, *33*(6), 750-772.

Rao, G., Wang, Y., Chen, W., Li, D., & Wu, W. (2021). Matching influence maximization in social networks. *Theoretical Computer Science*, 857, 71-86.

Rao, V. R. (2014). Applied conjoint analysis. New York. Springer.

Raschka, S., & Mirjalili, V. (2019). Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2. Packt Publishing Ltd.

Ravikumar, P., Wainwright, M. J., & Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ 1-regularized logistic regression. *The Annals of Statistics*, 38(3), 1287–1319.

Reinartz, W., & Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of Marketing*, 64(4), 17-35.

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. D. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PloS one*, 14(1), e0210236.

Rokka, J., Karlsson, K., & Tienari, J. (2014). Balancing acts: Managing employees and reputation in social media. *Journal of Marketing Management*, *30*(7/8), 802–827.

Roma, P., & Aloini, D. (2019). How does brand-related user-generated content differ across social media? Evidence reloaded. *Journal of Business Research*, 96, 322-339.

Rothman, D. (2014), *Lead Generation for Dummies*. Hoboken, New Jersey: Wiley.

Rubin, A., & Babbie, E. R. (2016). Empowerment series: Research methods for social work. Cengage Learning.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence,* 1(5), 206-215.

Ruggeri, G., Orsi, L., & Corsi, S. (2019). A bibliometric analysis of the scientific literature on Fairtrade labelling. *International Journal of Consumer Studies*, *43*(2), 134-152.

Sáez-Ortuño, L., Forgas-Coll, S., Huertas-Garcia, R., & Sánchez-García, J. (2023a). What's on the horizon? A bibliometric analysis of personal data collection methods on social networks. *Journal of Business Research*, 158, 113702.

Sáez-Ortuño, L., Forgas-Coll, S., Huertas-Garcia, R., & Sánchez-García, J. (2023b). Online cheaters: Profiles and motivations of internet users who falsify their data online. *Journal of Innovation & Knowledge*, 8(2), 100349.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM *Journal of research and development*, 3(3), 210-229.

Sannon, S., Bazarova, N. N., & Cosley, D. (2018). Privacy lies: Understanding how, when, and why people lie to protect their privacy in multiple online contexts. *In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-13), Montreal, Canada.

Sarstedt, M., Wilczynski, P., & Melewar, T. C. (2013). Measuring reputation in global markets—A comparison of reputation measures' convergent and criterion validities. *Journal of World Business*, 48(3), 329-339.

Saunders, M., Lewis, P., & Thornhill, A. (2009). Research Methods for Business Students. Essex, England. Pearson Education,

Saura, J. R. (2021). Using Data Sciences in Digital Marketing: Framework, methods, and performance metrics. *Journal of Innovation and Knowledge*, 6(2), 92–102.

Schapire, R. E. (2013). Explaining adaboost. In *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, (pp. 37-52). Springer Berlin Heidelberg.

Schulten, M. B., & Rauch, M. (2015). Ready to Win? Generating High-Quality Leads Through Online Sweepstakes and Conquizzes. *Journal of Marketing Theory and Practice*, 23(1), 21-37.

Schultz, D. E., & Peltier, J. (2013). Social media's slippery slope: Challenges, opportunities and future research directions. *Journal of Research in Interactive Marketing*, *7*(2), 86–99.

Schuman, H., & Scott, J. (1989). Generations and collective memories. *American Sociological Review*, 54(3), 359–381.

Schweidel, D. A., & Moe, W. W. (2014). Listening in on social media: A joint model of sentiment and venue format choice. *Journal of Marketing Research*, *51*(4), 387–402.

Sculley, D. (2010, April). Web-scale k-means clustering. *In Proceedings of the 19th international conference on World wide web* (pp. 1177-1178).

Sharif, S.M. & Zhang, X. (2014). A survey on deceptions in online social networks. *In Computer and information sciences (ICCOINS), 2014 International Conference on IEEE*, (pp. 1–6),

Shu, K., Mahudeswaran, D., Wang, S. H., Lee, D., & Liu, H. (2020). FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data*, 8(3), 171-188.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: a data mining perspective. *ACM SIGKDD Explorations Newsletter* 19(1), 22–36.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *In IOP conference series: materials science and engineering (Vol. 336, p. 012017)*. IOP Publishing.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4). 265-269.

Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. *Journal of marketing*, 21(1), 3-8.

Song, C., Ning, N., Zhang, Y., & Wu, B. (2021). A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing and Management*, 58(1), 102437.

Sponder, M. (2018). Social media analytics: Effective tools for building, interpreting, and using metrics. McGraw Hill Professional.

Sreedhar, C., Kasiviswanath, N., & Chenna Reddy, P. (2017). Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop. *Journal of Big Data*, 4(1), 27.

Statista (2021). Digital lead generation advertising spending in the United States from 2019 to 2023 (in billion U.S. dollars). https://www.statista.com/statistics/190328/us-online-lead-generation-spending-forecast-2010-to-2015/

Statista. (2021). Spain: digital advertising spending 2020. Retrieved 25/6/2022 from https://www.statista.com/statistics/281254/spain-digital-advertising-spending/

Statista. (2022). Data usage in marketing and advertising - statistics & facts. Retrieved 12/9/2022 from: https://www.statista.com/topics/4654/data-usage-in-marketing-and-advertising/

Stead, M., Gordon, R., Angus, K., & McDermott, L. (2007). A systematic review of social marketing effectiveness. *Health education*, 107(2), 126-191.

Stewart, M. C., & Arnold, C. L. (2018). Defining social listening: Recognizing an emerging dimension of listening. *International Journal of Listening*, 32(2), 85-100.

Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social media analytics: An interdisciplinary approach and its implications for information systems. *Business & Information Systems Engineering*, 6, 89-96.

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.

Strehl, A., & Ghosh, J. (2003). Relationship-based clustering and visualization for high-dimensional data mining. INFORMS *Journal on Computing*, 15(2), 208-230.

Subramanyam, K. (1983). Bibliometric studies on research collaboration: A review.

Journal of Information Science, 6(1), 33-38.

Sue, V. M., & Ritter, L. A. (2012). Conducting online surveys. Thousand Oaks California. Sage.

Sullivan, E., Bountouridis, D., Harambam, J., Najafian, S., Loecherbach, F., Makhortykh, M., Kelen, D., Wilkinson, D., Graus, D., & Tintarev, N. (2019). Reading news with a purpose: explaining user profiles for self-actualization. In Publication of the 27th Conference on User Modeling, *Adaptation and Personalization, Association for Computing Machinery* (pp. 241-245), New York NY, USA.

Tandoc, E.C. Jr., Lim, Z.W., & Ling, R. (2018). Defning fake news: a typology of scholarly definitions. *Digital Journalism*, 6(2), 137–153.

Tang, P. (2020, December). Telecom customer churn prediction model combining k-means and xgboost algorithm. *In 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)* (pp. 1128-1131). IEEE.

Thakur, S., Meenakshi, Er., & Priya, A. (2017). Detection of malicious URLs in big data using ripper algorithm, *In 2017 2nd IEEE international conference on Recent Trends in Electronics, Information and Communication Technology* (RTEICT)(pp. 1296-1301).

Thelwall, M. (2009). Introduction to webometrics: Quantitative web research for the social sciences. *Synthesis lectures on information concepts, retrieval, and services,* 1(1), 1-116.

Tibshirani, R. J., & Efron, B. (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1).

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 411-423.

Tripathy, R. M., Bagchi, A., & Mehta, S. (2013). Towards combating rumors in social networks: Models and metrics. *Intelligent Data Analysis*, 17(1), 149-175.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity, and other methodological pitfalls. *In Proceedings of the 8th International Conference on Weblogs and Social Media*, ICWSM 2014 (pp. 505-514). AAAI Press.

Tufte, E. R. (2001). The visual display of quantitative information. Cheshire, Connecticut: Graphics Press.

Tukey, J. W. (1962). The future of data analysis. *The annals of mathematical statistics*, 33(1), 1-67.

Tuomi, A., Tussyadiah, I. P., & Hanna, P. (2021). Spicing up hospitality service encounters: the case of Pepper™. *International Journal of Contemporary Hospitality Management*, 33(11), 3906-3925.

Tur-Porcar, A., Mas-Tur, A., Merigó, J. M., Roig-Tierno, N., & Watt, J. (2018). A Bibliometric History of the Journal of Psychology between 1936 and 2015. *Journal of Psychology, 152*(4), 199-225.

Tyagi, A. K., & Chahal, P. (2022). Artificial intelligence and machine learning algorithms. *In Research Anthology on Machine Learning Techniques, Methods, and Applications* (pp. 421-446). IGI Global.

Unnikrishnan, R., & Hebert, M. (2005, January). Measures of similarity. *In 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)-* (Vol. 1, pp. 394-394). IEEE.

Valentine, D. B., & Powers, T. L. (2013). Generation Y values and lifestyle segments. *Journal of Consumer Marketing,* 30(7), pp. 597-606.

Valenzuela, L. M., Merigo, J. M., Johnston, W. J., Nicolas, C., & Jaramillo, J. F. (2017). Thirty years of the Journal of Business & Industrial Marketing: A bibliometric analysis. *Journal of Business & Industrial Marketing*, *32*(1), 1-17.

Van Dijck, J. (2013). The culture of connectivity: A critical history of social media. New York. Oxford University Press.

Van Eck, N., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523-538.

Van Esch, P., & Stewart Black, J. (2021). Artificial intelligence (AI): revolutionizing digital marketing. *Australasian Marketing Journal*, 29(3), 199-203.

Van Raan, A.F. (2004). Measuring Science. In: Moed, H.F., Glänzel, W., Schmoch, U. (eds) Handbook of Quantitative Science and Technology Research. Springer, Dordrecht.

Verhoef, P. C., Kannan, P. K., & Inman, J. J. (2015). From multi-channel retailing to omni-channel retailing: introduction to the special issue on multi-channel retailing. *Journal of retailing*, 91(2), 174-181.

Viviani, M., & Pasi, G. (2017). Credibility in social media: opinions, news, and health information-a survey. *Wires Data Mining and Knowledge Discovery*, 7(5), e1209.

Voges, K. E., Pope, N. K., & Brown, M. R. (2002). Cluster analysis of marketing data examining on-line shopping orientation: A comparison of k-means and rough clustering approaches. *In Heuristic and Optimization for Knowledge Discovery* (pp. 208-225). IGI Global.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146-1151.

Vriens, M., Wedel, M., & Wilms, T. (1996). Metric conjoint segmentation methods: A Monte Carlo comparison. *Journal of Marketing Research*, 33(1), 73-85.

Wang, D., & Zhang, S. (2020). Unsupervised person re-identification via multi-label classification. *In Proceedings of the IEEE/CVF conference on computer vision and* *pattern recognition* (pp. 10981-10990).

Wang, T., Duong, T. D., & Chen, C. C. (2016). Intention to disclose personal information via mobile applications: A privacy calculus perspective. *International journal of information management*, 36(4), 531-542.

Wang, W. Y. C., & Wang, Y. (2020). Analytics in the era of big data: The digital transformations and value creation in industrial marketing. *Industrial Marketing Management, 86*(3), 12-15.

Wang, W., Laengle, S., Merigó, J. M., Yu, D., Herrera-Viedma, E., Cobo, M. J., & Bouchon-Meunier, B. (2018). A bibliometric analysis of the first twenty-five years of the International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 26*, 169-193.

Wang, X., & Song, Y. (2020). Viral misinformation and echo chambers: the diffusion of rumors about genetically modified organisms on social media. *Internet Research*, 30(5), 1547-1564.

Wedel, M., & Kamakura, W. A. (2000). Market segmentation: Conceptual and methodological foundations. Norwell. Kluwer Academic Publishers Group.

Wendling, M. (2018). The (almost) complete history of "fake news". Retrieved 30/11/2022 from https://www.bbc.com/news/blogs-trending-42724320

Wilson, R. E., Gosling, S. D., & Graham, L. T. (2012). A review of Facebook research in the social sciences. *Perspectives on psychological science*, 7(3), 203-220.

Wu, Y., Ngai, E.W.T., Wu, P. and Wu, C. (2022). Fake news on the internet: a literature review, synthesis and directions for future research. *Internet Research*, 32(5), 1662-1699.

XGBoost (2022) XGBoost Documentation. Retrieved 05/03/2023.

https://xgboost.readthedocs.io/en/stable/

Yadav, G. P. and Rai, J. (2017). The Generation Z and their SocialMedia Usage: A Review and a Research Outline. *Global Journal of Enterprise Information System*, 9(2), 110-116.

Yankelovich, D., & Meer, D. (2006). Rediscovering market segmentation. *Harvard Business Review*, 84(2), 98-108.

Yi, H., and Yang, J. (2014). Research trends in post disaster reconstruction: The past and the future. *Habitat International*, *42*, 21-29.

You, Z., Si, Y. W., Zhang, D., Zeng, X., Leung, S. C., & Li, T. (2015). A decision-making framework for precision marketing. *Expert Systems with Applications*, 42(7), 3357-3367.

Yu, D., Xu, Z., Kao, Y., & Lin, C. T. (2018). The structure and citation landscape of IEEE Transactions on fuzzy systems (1994-2015). *IEEE Transactions on Fuzzy Systems*, 26(2), 430-442. https://doi.org/10.1109/TFUZZ.2017.2672732

Zamri, N., Pairan, M. A., Azman, W. N. A. W., Abas, S. S., Abdullah, L., Naim, S., ... & Gao, M. (2022). A comparison of unsupervised and supervised machine learning algorithms to predict water pollutions. *Procedia Computer Science,* 204, 172-179.

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.

Zhang, H., Alim, M. A., Li, X., Thai, M. T., & Nguyen, H. T. (2016). Misinformation in Online Social Networks: Detect Them All with a Limited Budget. *Acm Transactions On Information Systems*, 34(3), 1-24.

Zhang, W., Du, W., Bian, Y., Peng, C. H., & Jiang, Q. (2020). Seeing is not always believing: an exploratory study of clickbait in WeChat. *Internet Research*, 30(3), 1043-1058.

Zhang, X. C., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2), 102125.

Zhang, Z., & Gupta, B. B. (2018). Social media security and trustworthiness: overview and new direction. *Future Generation Computer Systems*, 86, 914-925.

Zhao, H., Lyu, F., & Luo, Y. (2022). Research on the effect of online marketing based on multimodel fusion and artificial intelligence in the context of big data. *Security and Communication Networks*, 2022, 1-9.

Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, 17, 100179.

Zhu, J. J., Chang, Y. C., Ku, C. H., Li, S. Y., & Chen, C. J. (2021). Online critical review classification in response strategy and service provider rating: Algorithms from heuristic processing, sentiment analysis to deep learning. *Journal of Business Research*, 129, 860-877.

Zhu, L., Chen, J., Liao, X., & Wen, Y. (2014). An empirical study on microblog sentiment analysis using multi-level hierarchical classifier. *Journal of Information Science*, 40(6), 815-828.

Zhu, Y. Q., & Chen, H. G. (2015). Social media and human need satisfaction: Implications for social media marketing. *Business horizons*, 58(3), 335-345.

Zimmer, M. (2010). “But the data is already public”: On the ethics of research in Facebook. *Ethics and Information Technology,* 12(4), 313-325.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2), 301-

320.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: a survey. *ACM Computing Surveys*, 51(2), 1-36.