INSTITUTO
UNIVERSITÁRIO
DE LISBOA

A DATA SCIENCE BASED FRAMEWORK TO ENHANCE USER ENGAGEMENT, ACQUISITION AND RETENTION IN NEWS

Elizabeth Silva Fernandes

PhD in Information Science and Technology

Supervisor(s):

PhD Sérgio Moro, Assistant Professor with Habilitation, ISCTE-IUL - Instituto Universitário de Lisboa

PhD Paulo Cortez, Full Professor, University of Minho

April, 2023

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

A DATA SCIENCE BASED FRAMEWORK TO ENHANCE USER ENGAGEMENT, ACQUISITION AND RETENTION IN NEWS

Elizabeth Silva Fernandes

PhD in Information Science and Technology

Supervisor(s):

PhD Sérgio Moro, Associate Professor with Habilitation, ISCTE-IUL - Instituto Universitário de Lisboa

PhD Paulo Cortez, Full Professor, University of Minho

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

A DATA SCIENCE BASED FRAMEWORK TO ENHANCE USER ENGAGEMENT, ACQUISITION AND RETENTION IN NEWS

Elizabeth Silva Fernandes

PhD in Information Science and Technology

Jury:

April, 2023

To Beatriz and Bruno for all the inspiration

Acknowledgments

The Doctor of Philosophy is a personal goal that I defined for myself to be the best in class in the industry, and to be recognized by peers as a complete data professional. Of course, as I usually say it requires lots of Patience, Persistence, Perseverance, and Passion. But fortunately, I had the right combination between family, friends, and supervisors to achieve this challenging goal. Thus, I must express my gratitude to all those that helped me.

The story of how I decided to complete my PhD dates to 2008 at the Oporto Economics Faculty in my Master thesis defense. At that time, Professor Paulo Cortez was a jury member and, after analyzing my work he said that my master's thesis was close to be a PhD work. He felt me believe, and be confident, that I could do the PhD. At that time, I thought, "*if I'll do the PhD, this Professor will be my supervisor*". The years passed. In 2017, in the Master Executive in Busyness Analytics classes, I was in a lunch conversation with Professor Sérgio Moro that shared "*my PhD supervisor is Professor Paulo Cortez"*. I thought "*this is the sign; I will do the PhD with both*". That was the moment that I decided to start this journey. I knew that the only way to achieve the goal was to have a strong commitment to my supervisors and have the right supervisors.

Life brought me some challenges across the journey. I should admit that there were several moments when I felt terribly tired with a strong desire to give up. But the commitment to my supervisors made me feel ashamed to admit that I wanted to give up despite having medical order to stop. Here, I want to express my gratitude to my supervisors Professor Sérgio Moro and Professor Paulo Cortez. They were always giving me the right directions, recommendations, and guidance. And more than that, they motivated me at the right moment, they applied the right words and guided me to achieve each milestone.

This work was developed in a real context in the Portuguese newspaper Público Comunicação Social S.A. with all the advantages and disadvantages that this entails. I would like to thank Público for providing the dataset used in this research, the computational power, and data storage space. It was necessary to allocate colleagues' resources despite limited resources. I have excellent colleagues that helped me at Público to do this project and to drive value from data. I am deeply grateful to Cristina Soares, Maria José Palmeirim, Manuel Carvalho, Sónia

Matos, Joana Villas, Amilcar Correia, Fabiola Mousinho, Natacha Cabral, Vitor Ferreira, António Ferreira and, Pedro Vasquez.

Moreover, to face personal challenges and keep working on several projects, my husband and my in-laws played a core role by supporting me and giving me all the help to continue. In the middle of this journey, I became the mother of a beautiful girl, Beatriz. She doesn't know, but she taught me to say no and to prioritize what is important to me. That was crucial to increase my focus on the goal.

Finally, I also would like to offer my special thanks to Nuno António for his insightful comments, and suggestions. He was there even when he was full of work. Also, I would like to extend my sincere thanks to Professor Fernando Batista and Professor Ricardo Ribeiro for their guidance through the first paper, it was fundamental to improve my writing skills.

Lastly, I also had the financial support of Público and the FCT - Fundação para a Ciência e Tecnologia, under Projects UIDB/04466/2020, UIDP/04466/2020, UID/CEC/00319/2019, and UIDB/50021/2020.

For all, I express my deepest gratitude.

Abstract

In the last decade, the decline of print advertising revenue, the distribution of free digital content and the change of reader's behaviour induced a need of new sources of revenue. The Portuguese newspaper Público was no exception. The digital subscription business model become a solution to assure profits. In the digital landscape, each second of time results in millions of readers interactions across digital platforms, which provides huge volumes of data to be collected and stored. This new Big Data era demanded the development of new technologies and brought Data Science (DS) capabilities to the newsrooms.

Motivated by the increasing interest in DS methods, and the necessity to drive value from data at Público, the present research provides a framework to enhance user engagement, acquisition, and retention. This research presents exploratory and empirical research using DS methods applied to real-world data. In particular, by focusing on a multichannel strategy, editorial and/or marketing data-driven actions applied, namely: an Instagram strategy that resulted in an increase of 156% on website visits and +5 p.p. on subscribers from Instagram; a reader's segmentation model that avoided the provider's costs and helped the marketing team to improve the subscriptions strategy; an automated newsletter for readers with high propensity to churn that induced an engagement increase on 55% of those subscribers. Several of these DS actions produced an impact at Público, contributing to the global company goals to become the national leader of readers website visits (2020 and 2021) and keeping a sustainable increase on subscriptions.

Keywords: Data Science, Digital Journalism, Big Data, Reader Engagement, Audience Funnel, Machine Learning

Resumo

Na última década, a queda das receitas publicitárias da versão impressa, a distribuição de conteúdo digital gratuito e a mudança de comportamento dos leitores induziram a necessidade de novas fontes de receita. O jornal português Público não foi exceção. O modelo de negócio de assinatura digital (AD) tornou-se uma solução para garantir receita. Neste cenário digital, a cada segundo acontecem milhares de interações dos leitores nas plataformas digitais, o que origina um elevado volume de dados. Esta nova era de *Big Data* exigiu o desenvolvimento de novas tecnologias e introduziu as competências da Ciência de Dados (CD) nas redações.

Motivados pelo crescente interesse dos métodos da CD e pela necessidade de gerar valor a partir dos dados no Público, a presente dissertação visa fornecer uma estratégia para melhorar a fidelização, aquisição e retenção do leitor. Esta investigação apresenta análises exploratórias e empíricas utilizando métodos de CD aplicados em dados reais. Em particular são aplicadas ações editoriais e/ou de marketing baseadas em dados, nomeadamente: estratégia de Instagram que resultou num aumento de 156% em visitas ao site e +5 p.p. em AD do Instagram; modelo de segmentação de leitores que evitou custos de fornecedor e permitiu à equipa de marketing melhorar a estratégia de AD; *newsletter* editorial para leitores com propensão de *churn* que induziu um aumento de fidelização em 55% desses assinantes. Estas acções produziram impacto, contribuindo para os objetivos globais da empresa em ser líder nacional de visitas ao site (2020 e 2021), e mantendo um aumento sustentável de AD.

Keywords: Ciência de Dados, Jornalismo Digital, *Big Data*, Fidelização do leitor, Funil de audiência, *Machine Learning*

Contents

List of figures

List of tables

Acronyms

AD - Active days

AI - Artificial Intelligence

BQ - Big query

CDP - Changepoint detection problem

CTR - Click through rate

CRISP-DM - Cross-Industry Standard Process model for Data Mining

CRIPS-ML(Q) - CRoss-Industry Standard Process model for the development of Machine Learning applications with Quality assurance methodology

DJ - Digital Journalism

DL - Deep Learning

DS - Data Science

DSBM - Digital subscription business models

DSLM - Data Science lifecycle management

ED - External data

GA - Google analytics

HTC - Hierarchical text mining

IG - Instagram

INMA - International News Media Association

KPI - Key performance indicator

MRQ - Main research questions

ML - Machine learning

NL - Newsletter

OR - Open Rate

RFV - Recency, frequency, and volume

RQ - Research questions

RS - Recommendation systems

SLR – Systematic literature review

SM - Social media

SME - Social media editor

SNSM - Simple news selection mechanism

UGC - User-generated content

CHAPTER 1

Introduction

This PhD study is motivated by the interest of researchers and practitioners in the application of data science methods to increase customer loyalty by maximizing the value delivery from data. At the beginning of this research, the PhD project aimed to present a "Decision Support System in the Restaurant Industry" that was replaced into a "Data Science Based Framework to Enhance Reader Engagement, Acquisition and Retention in News" as result of researcher career change. Despite that both industries are different, the common problem faced by managers and companies, among others, it is to define strategies to increase customer loyalty, i.e., to increase the willingness to advocate the brand (Kotler et al., 2016). In other words, managers are concerned about improving product and service strategies to maximize customer retention and guarantee companies' profitability.

In the connectivity era, the customer path is described by Kotler et al. (2016) into five A's: *aware*, *appeal*, *ask*, *act*, and *advocate* (Kotler et al., 2016). That is a simple model to describe the straightforward funnel process that customers go through. Across the funnel, data driven strategies improve customer engagement, advocacy, and conversions (see Figure 1). Furthermore, the Big Data environment induces companies to use data science algorithms to be more competitive, as it happens in the media industry (Rußell et al., 2020).

In the last decade, publishers faced a digital transformation that introduced new challenges and opportunities (Arrese, 2016). The decline of print advertising revenue, the distribution of free digital content, and the change of reader's behaviour induced a need of new digital business models (Arrese, 2016; Rußell et al., 2020). Digital subscription business models (DSBMs), usualy in the form of paywall models (Pattabhiramaiah et al., 2019) are documented in the academic literature mainly into three types: metered (consumption limited by a few articles), freemium (a selection of content is premium, i.e., only available for subscribers), and hard paywalls (all the content only accessible for subscribers) (Myllylahti, 2019).

Despite the wide use in other industries, the use of Machine Learning (ML) and Artificial Intelligence (AI) algorithms are a relatively new area of study in the media landscape (Davoudi, 2018). Nowadays, the adoption of data science models for segmentation, personalization, or recommendation attracted the attention of several media publishers (INMA, 2022; Meguebli et al., 2017). Effective management and engagement analysis allows managers to understand reader behavior by providing insights to conduct data-driven strategies.

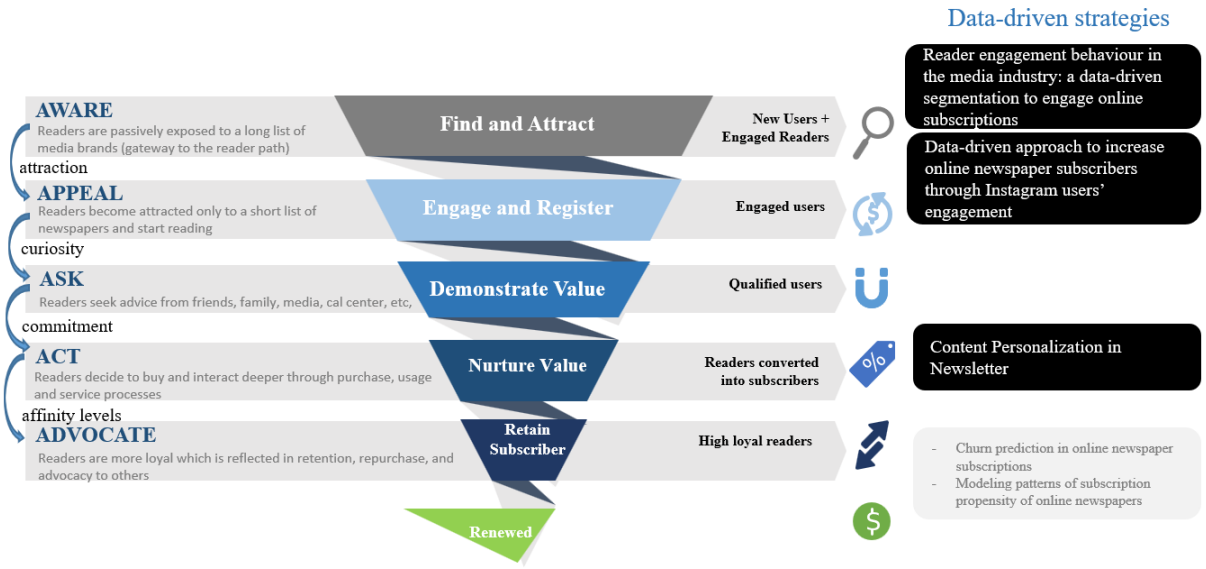


Figure 1. Main data-driven strategies implemented across the reader conversion funnel.

Furthermore, studies argue that reader retention is related to how often a user reads (regularity) (Davoudi, 2018; Kim et al., 2021). Moreover, reader engagement is a multidimensional phenomenon (Steensen et al., 2020) related to the level of attention and involvement (emotional, cognitive and behavioral) with media (Attfield et al., 2011; Ksiazek et al., 2016; Mersey et al., 2010). In fact, researchers acknowledged the relationship between customer engagement and purchase intention in DSBMs (Clement Addo et al., 2021).

The present research aims to present a state of the art of DS studies in the media industry, followed by examples of real data-driven strategies to increase retention and conversion across the reader funnel (see Figure 1). The funnel approach was inspired on Kotler five A's (Kotler et al., 2016; Piechota, 2020). Thus, this work focuses on the effectiveness of the data-driven strategy used by the Portuguese publisher Público to increase engagement, retention and acquisition.

Público website presents a reach higher that 4 million users and verifies more than 200 million user events by month. The main goal in the digitalization and platformization of *Público* journalism was to raise readers engagement along the reader funnel (Kotler et al., 2016) by distributing the right content at the right time at the right format. Our research progress was accelerated and driven by the need to find solutions in the Portuguese newspaper *Público*. As responsible for the analytics department, the PhD researcher developed several activities and experiments to provide the best insights and recommendations, to the marketing and editorial teams, that consequently increased reader engagement and conversion. The derived experiments and results are actionable, interpretable, and replicable.

1.1. Motivation

Digital innovation introduced a dramatic change in media companies. The decline of print advertising revenue, the distribution of free digital content and the change of reader's behavior induced a need of new sources of revenue (Arrese, 2016; Rußell et al., 2020). Subscription business models, usually in the form of paywall models (Pattabhiramaiah et al., 2019; Rußell et al., 2020), become a solution to assure companies' sustainability (Davoudi & Edall, 2018; Simon & Graves, 2019). Consequently, high level data-based models have emerged (Davoudi, 2018).

Currently, each second of time results in millions of readers interacting on digital platforms, which provides huge volumes of data to be collected and stored by media companies (Lewis, 2015). This new Big Data era in journalism demanded the development of new technologies and brought Data Science (DS) and AI capabilities to the newsroom (Borges et al., 2021).

Moreover, the adoption of ML methods is mentioned in the Reuters Digital Report as the new trend in media companies, especially for personalization and content recommendation (Newman, 2019; Yeung & Yang, 2010; Zihayat et al., 2019).

Comment analysis, event mining, and journalism automation have attracted a great attention and nowadays continue being an outstanding research area. Currently, ML and Deep Learning (DL) approaches have been successfully applied to diverse fields, such as Natural Language Processing, Social Network Analysis or business models development (Davoudi, 2018) . As argued by (Suárez, 2020) and (Rußell et al., 2020) reader retention, personalization and paywall models are some of the major points of concern in the industry.

Motivated by the increase of interest in DS (including AI and ML) in Digital Journalism, this research aims to present a DS framework to increase reader engagement, acquisiton and retention. The researcher aims to define the research design and process that will allow to propose a model to improve reader engagement and consequently subscription propensity.

As presented at Figure 2, the research was planned to be performed between 2019 and 2023 by following five steps as follows: research idea and direction definition, literature review, research design followed by the research process, and finally the contributions, dissertation discussion and conclusion.

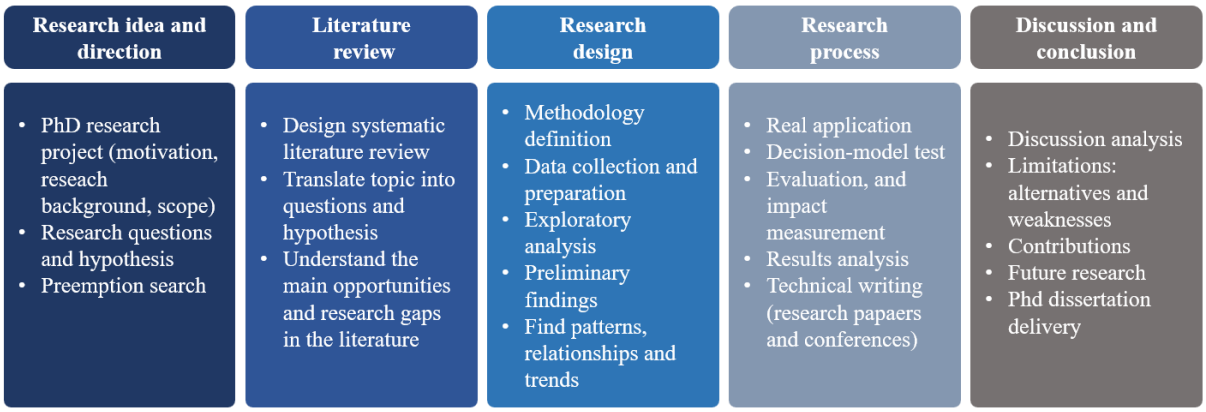


Figure 2. The conducted research plan.

1.2. Objectives

Across the five stages of the reader conversion funnel presented at Figure 1, the goal it is to convert the maximum number of anonymous users into high loyal subscribers. However, at each stage, conversion rates need to be measure. Low levels of conversion rates result in low customer attraction, low curiosity, low commitment, and low affinity (Kotler et al., 2016). Furthermore, along the customer path many events are generated that allow to understand customer behaviour and consequently improve service quality and increase customer loyalty. Nonetheless, the customer path can be a spiral as result to the multiple touchpoints and engagement features across the website. This results in a high volume of data that culminate in a complex group of DS challenges (Van Der Aalst, 2011). Thus, in the case of *Público*, the main goal was to drive value from data by applying actionable DS methods through the implementation of a data driven culture.

1.3. Research methodology

According to Kothari (2004), the purpose of research is to discover answers to questions through the application of scientific procedures. Each study has its own specific purpose (Kothari, 2004). Furthermore, research methods usually follow the principles of the scientific method that generates knowledge by creating an inquiry and resolving it through systematic experimentation (Checkland & Holwell, 1998; Kothari, 2004). An integral part of scientific method is the experimentation that it is done to test hypotheses and to discover relationships (Kothari, 2004). Moreover, one of the main principles of the scientific method is the reductionism, i.e., by decomposing a complex phenomenon to simpler elements we can explain the complex. However, (Checkland & Holwell, 1998) argue that the application of this process is problematic in more complex phenomena such social ones.

In order to solve the PhD goals, we employed an exploratory and empirical research by using DS and ML methods applied to real-world data regarding the analysed digital news domain. Due to the lack of a methodology for ML applications, (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, & Rudiger, 2000) presented the Cross-Industry Standard Process model for Data Mining (CRISP-DM) that consists of a cycle that comprises six stages (see Table 1). Since the publication of the CRISP-DM, new methodologies, and various Machine Learning (ML) solutions are proposed due to the problems complexity (Abonyi et al., 2022). Recently, (Studer et al., 2021) proposed a new methodology to face the need for guidance throughout the life cycle of ML application to meet business expectations. The CRoss-Industry Standard Process model for the development of Machine Learning applications with Quality assurance methodology CRIPS-ML(Q) is a process that cover six phases (see Table 1) from the scope to maintaining the deployed machine learning application (Studer et al., 2021). Quality assurance is introduced in each phase and task of the process model. A *monitoring and maintenance* phase is introduced to address risks of model degradation in a changing environment.

At each step of the research plan (Figure 2) a methodology was applied and detailed in the respective chapter. Firstly, to understand the business and the state of the art, we perform a literature review by analysing the literature published between 2010 and 2021. A 4-step systematic method was presented, as detailed at Section 2.2. Secondly, we implemented a data-driven strategy to acquire Instagram subscribers, following the steps: editorial strategy understanding, data gathering and preparation, results achieved through data analysis, results discussion, and limitations and conclusions. Finally, to segment readers (see Chapter 4), we follow the six phases of CRISP-DM methodology combined with business validation similarly to the Business Success Criteria of the CRISP-ML(Q) (Abonyi et al., 2022).

Furthermore, from an operational perspective, at each model implementation a Data visualization was provided to the users to monitor results, to evaluate marketing or editorial strategies, and at least find ways to improve the data-driven strategy or model proposed. All the framework was built in the Google ecosystem allowing the teams to monitor values in real-time, implements website tests, and measure daily results and actions' impact.

Thus in this research, to guarantee the data value delivery to the business, we were inspired on the CRISP-DM methodology (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, & Wirth, 2000) combined with the Data Science lifecycle management (DSLM) (Abonyi et al., 2022), i.e., a framework to support the development of ML algorithms to develop, maintain and manage the whole lifecycle (see Figure 3).

The DSLM starts with the business understanding, goals, and criteria of success definition. Then, at each experience a strong effort is invested in data understanding, acquisition, and exploration. Following by a modelling stage that comprises three sub-parts: development of the model, to compete and select different models; deployment that includes monitor and install models; and operation, which includes data visualization and the development of intelligent applications. The arrows in Figure 3 illustrate the cyclically of the development.

Table 1 Comparing process models for DM and ML projects.

Model	**Stages**						
CRISP-DM (Chapman et al., 2000)	Business understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment	---
CRISP_ML(Q) (Studer et al., 2021)	Business understanding and Data Understanding		Data Preparation	Modeling	Evaluation	Deployment	Monitoring and Maintenance

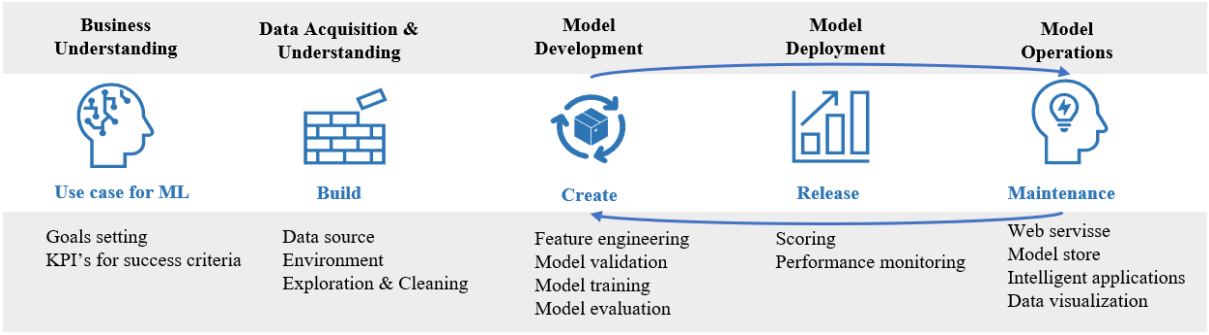


Figure 3 Data Science lifecycle management and CRISP-ML.

1.4. Contributions

The main outcome of this dissertation is to prove the value delivery to the business by combining a challenging and competitive digital business context, in a Big Data ecosystem, with the application of data-driven strategies (see Figure 1). As argued by Suarez (2020) some of the major points of concern in the media industry are reader retention, personalization, and

paywall models. Thus, the present approach provides, to a real digital publisher, solutions to improve reach, retention and, consequently the propensity to subscribe. On a managerial level, the presented outcome proposes: a strategy to increase reach by increasing readers from Instagram; a user's segmentation to increase reader engagement and revenue from two sources; and a content personalization suggestion for newsletters that it is a channel with high engagement level, i.e., readers present more propensity to subscribe and less propensity to churn (Mcmullan, 2018).

Throughout this research, experiment results and findings were published or submitted in peer-reviewed journals, to journals indexed in the well-established Web of Science (WoS) and Scopus databases, ranked in Quartile 1 (best ranked category), and presented in renowned industry conferences. The list of documents written, that resulted from the research developed, are presented at Table 2.

Table 2 Documents written.

Deliverables	**Title**	**Journal - Publisher – Conference (status)**	**IF (Impact Factor)**	**SJR score**
Directly related to this PhD research	Data Science, Machine Learning and Big Data in Digital Journalism: A survey of state-or-the-art, challenges, and opportunities	Expert Systems with Applications (Published 5th March 2023)	8.665	2.07
	A data-driven approach to increase online newspaper subscribers through Instagram users' engagement	Submitted to an International Journal	--	--
	Reader engagement behaviour in the media industry: a data-driven segmentation to engage online subscriptions	Submitted to an International Journal	--	--
	Towards a News Recommendation System to increase Reader Engagement through Newsletter Content Personalization	Submitted to an International Conference	--	--

In addition to the resulted research articles that directly contribute to this PhD work (presented in Table 2), during the execution of this PhD thesis there were other interesting publication outcomes. For instance, in the beginning of this PhD execution, a different application domain was initially targeted, related with the tourism area and that resulted in the article: "*A data-driven approach to measure restaurant performance by combining online reviews with historical sales data*" published in the top-tier International Journal of Hospitality Management, with Impact factor of 10.427 in the Web of Science database (ranked 5th out of 57 in Hospitality, Leisure, Sport & Tourism.

Moreover, when approaching the current PhD application domain (online news), there were several research opportunities and collaborations that were considered interesting to pursuit, although being out of scope of the PhD main theme, thus not fully described in this document.

As result of the positive impact delivered by the analytics strategy at Público, the work developed by the researcher was mentioned at the International News Media Association (INMA) report "*The Benefits and Risks of Media Data Democratisation*" as a case study (INMA, 2022). At the report, INMA gives high prominence to the data culture performed by information delivery thought actionable dashboards, also the strategic communication of results to all teams and departments, and the focus on engagement strategies through data analysis that retain valuable readers. The researcher was invited to present at the following worldwide conferences:

- "*La adopción de los datos en un medio de comunicación: El caso de Público*", 2021 INMA Virtual LATAM Conference.
- *"Digital transformation roadmap: Product and Data Analytics at Público"*, 2021 INMA Product and Data Summit.

Furthermore, the researcher was invited by INMA to be "*Accelerator mentor"* on the Audience Analytics Accelerator LATAM program, a Facebook Journalism Project. The two online newspaper "*Diário do Nordeste*" and "*Gazeta do Povo*" from Brazil were mentored to implement a data-driven strategy. The "*Diário do Nordeste*" results were also published as a case study by INMA (INMA, 2022).

Across the research timespan, the researcher attended conferences, workshops and, trainings to increase her knowledge and keep up with industry trends. Moreover, to better understand the role of DS and how to prove the value delivered in the company, the researcher got certified on "*Data Strategy and Data Governance*" by the Business School "*Instituto de Empresa*" in Madrid. Thus, at Chapter 6 an analysis is presented on the company Data Strategy diagnosis and the impact on the success of DS methods.

1.5. Thesis organization

After present a description of the research background and motivation, this section details the document structure and the framework designed. A visual comprehensive summary of the PhD study is shown in Figure 4. The main goal of this PhD research is to present a DS framework to enhance reader engagement, acquisition, and retention in online news. Across the *Público* audience funnel this research delivers value from data by applying a data-driven strategy

between the period of study. Thus, the main goal it is to maximize reader engagement by applying DS methods, coupled with a data culture implementation.

This study presents business applications instead to be a descriptive study. Hence, to define the PhD research framework, detailed at Figure 4, we started by present a deep description of the title by answering the: what, where, how, why, and when of this study. What does this study do? Drives value from data across the reader funnel (Kotler et al., 2016). Driving value from data, means the contribution of the data to the company strategy. According to (Campos & Rodríguez, 2020), across the data lifecycle, the internal value of the data (VID) is a function of data volume (V), data quality (Q), number of use cases (N), the use cases utility (U), and the strategic relevance (SR) as presented in the following equation:

$$VID = \sum [V * Q] * [N * U] * SR$$

How is this done? As we are in a Big Data context (V) with a good data quality (Q), we aim to present Data Science methods with high utility and high strategic relevance for the Portuguese publisher *Público* (the where) which is facing a competitive landscape and aims to increase reader engagement (the why). Thus, between 2019 and 2022 (the when) we applied advanced analytics experiments, that are presented in the visual representation at Figure 4, to provide valuable information to increase user engagement, acquisition, and retention.

Regarding the motivations of this research, the following Main Research Questions (MRQs) are addressed to organize the study:

- MRQ 1: What are the main motivations and the major topics when adopting DS in DJ?
- MRQ 2: How can a data-driven strategy engage readers and fights misinformation through Instagram?
- MRQ 3: How to segment readers to define a strategy based on engagement in a subscriptions business model?
- MRQ 4: How increase reader engagement by channel, for example, in newsletters (NL)?

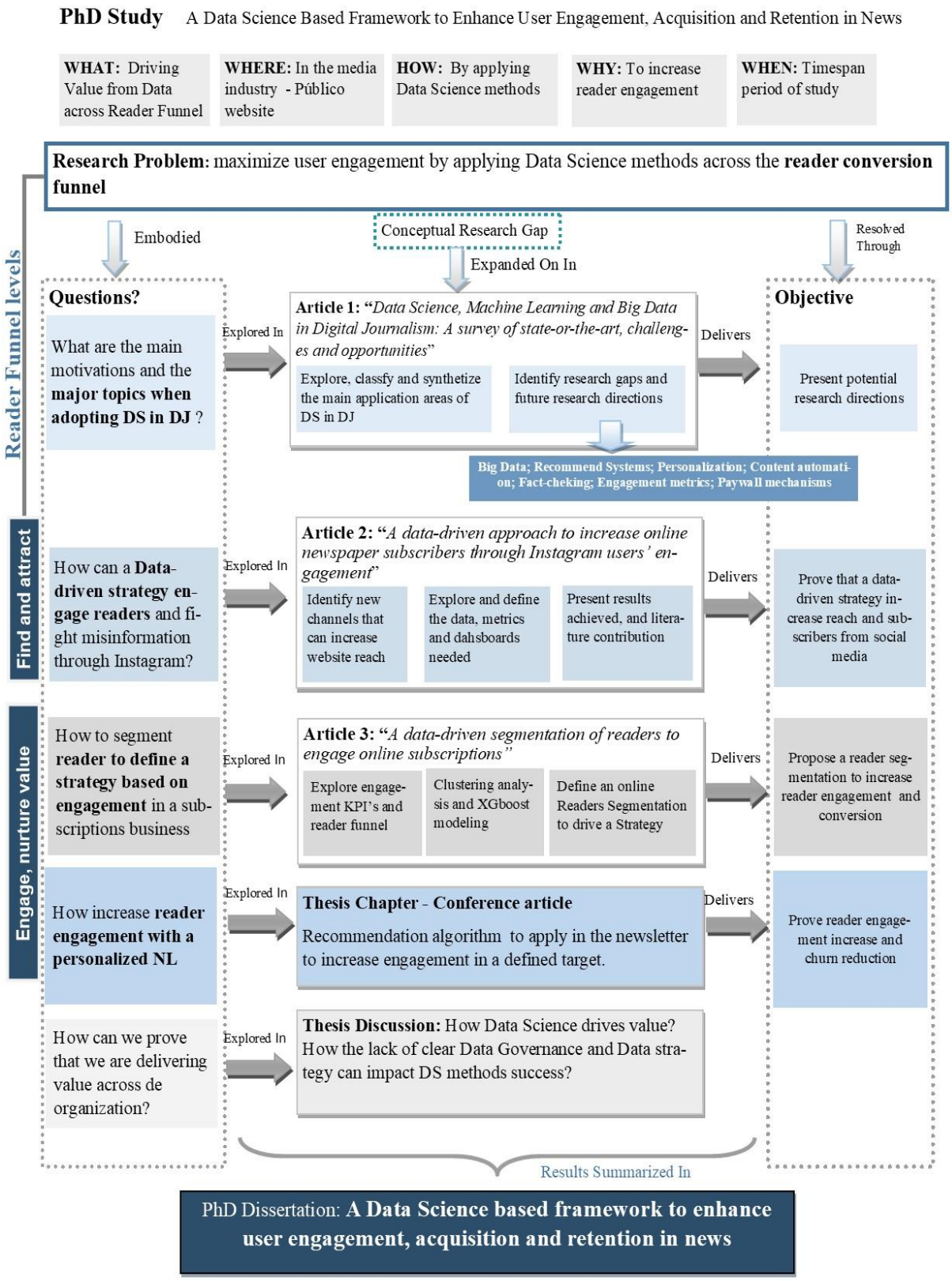


Figure 4 Research framework: a visual comprehensive summary of the PhD.

To answer the MRQs, the present document presents the following structure: Chapter 2 details a literature review that is presented in the first article mentioned at Table 2. Across the Chapter 2 it is presented a critical literature review with a synthesis of the main application areas of DS in DJ. Furthermore, an analysis of the ongoing research trends with gaps, challenges or opportunities for future studies are detailed. Then, Chapter 3 shows how a data-driven strategy allowed Público to engage new readers from Instagram in a pandemic context (MRQ2) that resulted into the second article presented at Table 2. Then, to answer MRQ3, at Chapter 4 we propose a reader's segmentation. Readers were segmented into eight clusters. A twofold strategy is proposed to impact on acquisition, retention, and conversion of online readers across the audience funnel. Finally, regarding the results obtained in a churn model, a personalized newsletter was developed to increase reader engagement and to reduce churn propensity (Chapter 5).

CHAPTER 2

Related work

2.1. Research context

This section presents a research literature analysis on the role of DS in DJ. Specifically, the aim is to present a critical literature review, synthetizing the main application areas of DS in DJ, highlighting research gaps, challenges, and opportunities for future studies. Through a Systematic literature review (SLR) integrating bibliometric search, text mining, and qualitative discussion, the relevant literature was identified and extensively analyzed. The derived results were compiled into an article as presented at Table 2 "*Data Science, Machine Learning and Big Data in Digital Journalism: A survey of state-or-the-art, challenges and opportunities*".

2.2. Article context and motivation

Nowadays, media companies driven by economic pressures are investing in data and technological solutions to achieve business results. According to the INMA report (INMA, 2022), data is critical to create reader-centric products. Furthermore, the report argues that bringing data to the centre of the decision-making process is a current and an ongoing process in media companies. Moreover, as discussed by Kotler et al. (2016), companies should map the customer path to purchase, understand customer touchpoints, and improve critical touchpoints. Consequently, across the reader's conversion funnel the goal it is to maximize reader's engagement (Lagun & Lalmas, 2016), retention and consequently increase revenue (Sapian & Vyshnevska, 2019).

Despite the lack of a clear definition of reader engagement, authors agree that engagement is a multidimensional phenomenon (Steensen et al., 2020) related to the level of attention and involvement (emotional, cognitive and behavioral) with media (Attfield et al., 2011; Ksiazek et al., 2016; Mersey et al., 2010). Furthermore, to measure reader engagement a range of engagement metrics are available on the literature (Davoudi et al., 2019; Ksiazek et al., 2016; Lehmann et al., 2012; Peterson & Carrabis, 2008). However, to the best of our knowledge, there is a lack of studies that analyze the large body of knowledge on how DS can improve reader engagement. Furthermore, publishers are using DS methods to understand media consumers and their consumption patterns (Villi & Picard, 2019) to increase engagement levels.

Some examples can be listed: audience monitoring (Myllylahti, 2017), recommendation algorithms (Gonzalez Camacho & Alves-Souza, 2018; Yeung & Yang, 2010; Zihayat et al., 2019), news performance or engagement prediction models (Fernandes et al., 2015; Jääskeläinen et al., 2020; Zihayat et al., 2019), fake news detection (Antoun et al., 2020; Shim et al., 2021; Souza Freire et al., 2021) or, algorithms for paywall design (Davoudi et al., 2018; Rußell et al., 2020). Zhou and Liou (2020) presented a bibliometric analysis of communication research on AI and Big Data, which proved an increase of publications in since 2013 (Zhou & Liao, 2020). However, to the best of our knowledge, none intensive survey on the role of DS in DJ has been published recently.

Hence, by examining the existing research literature of the last decade, this research surveys what has been done with DS methods in media. Moreover, one of the main contributions of this research it is to present research gaps in the current literature and opportunities for future research.

2.3. Systematic literature review (SLR)

Synthesizing past research findings is a complex task that requires a detailed methodological approach (Aria & Cuccurullo, 2017; Zupic & Cater, 2015). Thus, to examine the existing literature, this research assumes a Systematic Literature Review (SLR) (Abdelmageed & Zayed, 2020; Aria & Cuccurullo, 2017), which consists of a 4-step methodology. As presented at Table 3, the presented approach combines three widely known methodologies resulting in four steps that guided our research. Firstly, the study design, then data collection and selection, followed by data analysis and findings, and finally, discussion and results presentation. This well-defined process allows us to identify, evaluate and interpret the literature to answer relevant research questions (RQs) that are detailed at Section 2.5.

Table 3 Comparison of distinct Literature Review stages.

SLR stages	**Standard Science Mapping Workflow**	**Data Analytics Approach in SLR**	**Proposed approach**
Kitchenham and Ebse (2007)	Zupic and Cater (2015)	Haneem et al. (2017)	
Planning the review	Study Design	Purpose of the Literature Review	Study Design
		Protocol and Training	
	Data Collection	Search the literature	Data Collection and Selection
		Practical Screening	
		Quality Assessment	
Conducting the review	Data Analysis	Analysis and Findings	Data Analysis and Findings
	Data Visualization		
Reporting the review	Interpretation	Writing the review	Results and discussion

2.4. Contribution

In pursuit for the state of the art on a subject, several researchers present different techniques (Brous, Janssen, & Herder, 2020; Donthu et al., 2021). Table 4 presents four literature review frameworks that were chosen to represent different and recent literature analysis on research areas related to DJ. For each framework, the table mentions the keywords' selection criteria, the methodology followed, as well as, the tools used. The first two works (Engelke, 2019; O'Brien et al., 2020) present a manual analysis, while the remaining two (Zhou & Liao, 2020; Zhou & Zhou, 2020) conducted a three-step bibliometric analysis by using the VOSviewer tool (Donthu et al., 2021; Van-Eck & Waltman, 2010). Finally, the last row presents the proposed approach. Our approach is the only literature review study that includes Text Mining (TM) automated methods and a closed criteria for the keywords' selection, followed by a Hierarchical Clustering to define exclusion criteria's. Thus, this approach reduces the SLR manual effort, resulting in a more easily replicable semi-automated methodology.

This study differs from others, firstly, because it presents a literature review that investigates the relation between DS and DJ, a broader theme than the research presented by (Zhou & Liao, 2020). Secondly, at each step of the process the human intervention was minimized by reducing the subjectivity in the keywords' selection or document exclusion criteria. Finally, the process combines TM methods developed using the open source R statistical tool, thus benefiting from a community of supporters contributing with packages for a myriad of data analysis tasks (Cortez, 2021), as well as, science mapping analysis (SMA) by using VOSviewer (Donthu et al., 2021; Van-Eck & Waltman, 2010) and *bibliometrix*, the R-tool for comprehensive science mapping analysis (Aria & Cuccurullo, 2017). Moreover, the use of TM for synthesizing existing literature enables to efficiently extract insights from a large body of knowledge (Moro et al., 2015). Thus, the richness of the text of published articles combined with TM enables deeper analysis beyond keywords analysis. Resulting in an approach that, to the best of our knowledge, is innovative on a DJ survey (Zhou & Zhou, 2020).

Table 4 Examples of relevant frameworks for literature analysis and the proposed approach.

Author	Areas of Research	Literature sources, timespan and number of articles	Keywords selection (query strings)	Methodology	Approach and Tools
(Engelke, 2019)	Online participatory jouralism	**Data base:** Scopus **Timespan:** 1997 to 2017 **Nr. Articles:** 378	Previous literature analysis to achieve content validity.	SLR based on (Cooper, 1998). Steps: problem formulation, data collection, data evaluation, analysis and interpretation, public presentation.	Manual selection and inspection of the articles. Biblimetric analysis conducted manually.
(O'Brien et al., 2020)	Factors that contribute to consumer's pay intentiont in DJ	**Data base:** Google Scholar, EBSCOhost, Web of Science and ProQuest **Timespan:** 2000 to 2019 **Nr. Articles:** 37	Authors comprised combinations of phrases related to the field.	SLR based on (Webster & Watson, 2002). Steps: identify literature, structure the review, theoretical development, theory evaluation, discussion and conclusion.	Manual selection and inspection of relevant Journals and articles. Biblimetric analysis conducted manually.
(Zhou & Liao, 2020)	Artificial Intelligence and Big Data in communicatio n research	**Data base:** Web of science **Timespan:** Until February 2020 **Nr. Articles:** 685	Authors defined the keywords without previous research.	Bibliometric analysis Steps: data collection, analysis and intepretation, discussion and conclusion.	Data analysis conducted with Pyhon. Bibliometric analysis conducted by using VOSviewer.
(Zhou & Zhou, 2020)	Human-Computer interaction in journalism	**Data base:** Web of science **Timespan:** Until 2020 **Nr. Articles:** 2156	Authors defined the keywords without previous research.	Bibliometric analysis Steps: data collection, analysis and intepretation, discussion and conclusion.	Data analysis conducted with Pyhon. Bibliometric analysis conducted by using VOSviewer.
Proposed approach	Data Science in digital journalism	**Data base:** Scopus **Timespan:** 2010 to 2021 **Nr. Articles:** 514	Combination of the top keywords of two journals and top terms in the Document term matrix (TM method).	SLR Data Analysis approach that combines science mapping analysis workflow and text minng.	Document's agglomerative hiearchical clustering to define exclusion criterias (R statistical tool) and reduce the size of search-space. Bibliometric analysis conducted by using bibliometrix and VOSviewer.

2.5. SLR methodology

This section presents the 4-step systematic method for reviewing the literature. The SLR process begins, comprising study design, data collection and selection. Each stage encompasses several activities, as outlined in Figure 5. The following subsection describes each stage of the SLR.

2.5.1. Study design, data collection and selection

This stage involves the preparation of the research work to conduct the review that includes the objective and research questions definition. According to the motivation of this research, the following research questions (RQs) and motivations are addressed to organize the study:

- RQ1 - What are the main motivations and the major topics when adopting DS in DJ?
- Motivation: Identify the most significant publications in the field.
- RQ2 - What are the benefits or positive impacts of using DS in the DJ domain?
- Motivation: Identify the DS approaches and applications domains in DJ.
- RQ3 - What gaps exist in the current literature that provide new research paths?
- Motivation: Identify challenges and research opportunities.

In the first step, the RQs were broken into thematic areas according to the bibliometric technique: co-citation, co-author, co-word and bibliographic coupling (Cobo et al., 2011; Zupic & Cater, 2015). In the search process a database was chosen, in contrast to focusing on specific journals to not limit the review's comprehensiveness. Data pre-processing and cleaning was performed (Jin et al., 2019). The digital database considered was Scopus which is the largest abstract and citation database of peer-reviewed literature (Ballew, 2009) and it is used by multiple researches (Amado et al., 2018; Borges et al., 2021).

As the SLR is a semi-automated process, some human-led tasks (HLT) were performed. Thus, across the text we use the abbreviation HLT to signal a human-led task and ALT to signal an automated-led task. Therefore, the data collection and selection process followed the procedures described below:

- The inclusion and exclusion criteria were applied. The first inclusion criteria consisted of terms that appeared in the titles, abstracts, and keywords.
- The initial keywords selection was based on filter the top keywords of the Journals "Decision Support Systems" and "Digital Journalism". We selected the 20 most frequent keywords by year in the last 5 years for both Journals. Then, we saved the keywords that are in the top 20 more than one year (ALT). This resulted in two lists of 26 and 24 keywords. Despite that we aimed to minimize human intervention, in both lists some keywords were still considered out of the scope of our research. For example, the first list comprises some of the following keywords: **Information Systems**, Electronic Commerce, **Artificial Intelligence**, Commerce, Sales, **Decision Making**,

Investments, Finance, **Big Data** and Costs. Thus, the authors saved those related to the scope of the research that are highlighted in bold (HLT). The same rationale was applied to the second list, where for instance the keywords "facebook" and "twitter" were excluded. Moreover, in the second list, the keyword "news" was considered often commonly used by other scientific branches, thus the term was replaced by "digital news", "news media" and "news industry" (HLT). To reduce the subjectivity, the three authors analyzed all HLT decisions, reaching a consensus. It should be mentioned that one author is an analytics and audience insights manager in a national newspaper since 2015. Finally, the first query, with 25 keywords resulted in 1,689 documents, as presented at Figure 5.

- Then, after a preliminary analysis of the dataset, by using *bibliometrix*, some topics not related to our study appeared, for example, "health" or "security", thus an enhancement of keywords was required.
- The second keywords selection was improved by excluding the top terms that are out of the research scope. As presented in the next section, TM methods were used to find the top terms presented in the sample (ALT). Then, non-related documents were removed from the collection by adding an exclusion condition in the second search query as result of a manual selection of top terms out of research scope (HLT).
- Concerning the research literature type, only articles from journals, conferences, and review researchs were included.
- The search focused only on articles published in English to avoid any misperception and efforts in translation.
- Concerning the timeline, a period between 2010 and 2021 was chosen, as it contains the period of "Paywalls Popularization" (Arrese, 2016), an adequate period to see the recent evolution of DS in DJ.
- The bibliographic search resulted in 514 documents. For each publication, we retrieved the following data elements: title, authors, abstract, publication year, keywords, source title, document type and language.

As result of a semi-automated process, we further note that the final dataset can contain documents not directly related to the scope of research, nevertheless, we decided not to skim the article title and abstract to avoid a human bias.

Keywords enhancement with text mining

The selection of terms to exclude in the second research query encompasses three steps. Firstly, we extracted the information from the database; then punctuation, numbers or stopwords were removed, as well as, text was stemmed (António et al., 2018; Welbers et al., 2017). The matrix with the frequency of each term by document (DTM) was calculated. Furthermore, to avoid non-informative terms, the matrix DTM-tf-idf was also calculated. The term frequency-inverse document frequency (tf-idf) measures the relative importance of a word to a document (Silge & Robinson, 2019; Welbers et al., 2017). Finally, agglomerative hierarchical clustering (AHC) was performed to find the main clusters in the sample. The AHC is an unsupervised algorithm that starts by assigning each document to its own cluster and then the algorithm iteratively joins at each stage the most similar document until there is only one cluster (Gordon, 1999).

In order to obtain compact and well-separated groups we calculate four measures: average distances within and between clusters, Dunn index and average Silhouette (Rendón et al., 2011). Thus, the number of clusters that optimizes the four measures was nine (ALT). Then, we explored the clusters by inspecting the word clouds (HLT). As each cluster contains information related to the research scope, we cannot exclude any cluster. Thus, to refine the query, the 20 most frequent words by cluster were analysed to find non-related terms. As result, non-related documents were excluded from the Scopus search query by removing the words highlighted in bold (see Figure 6).

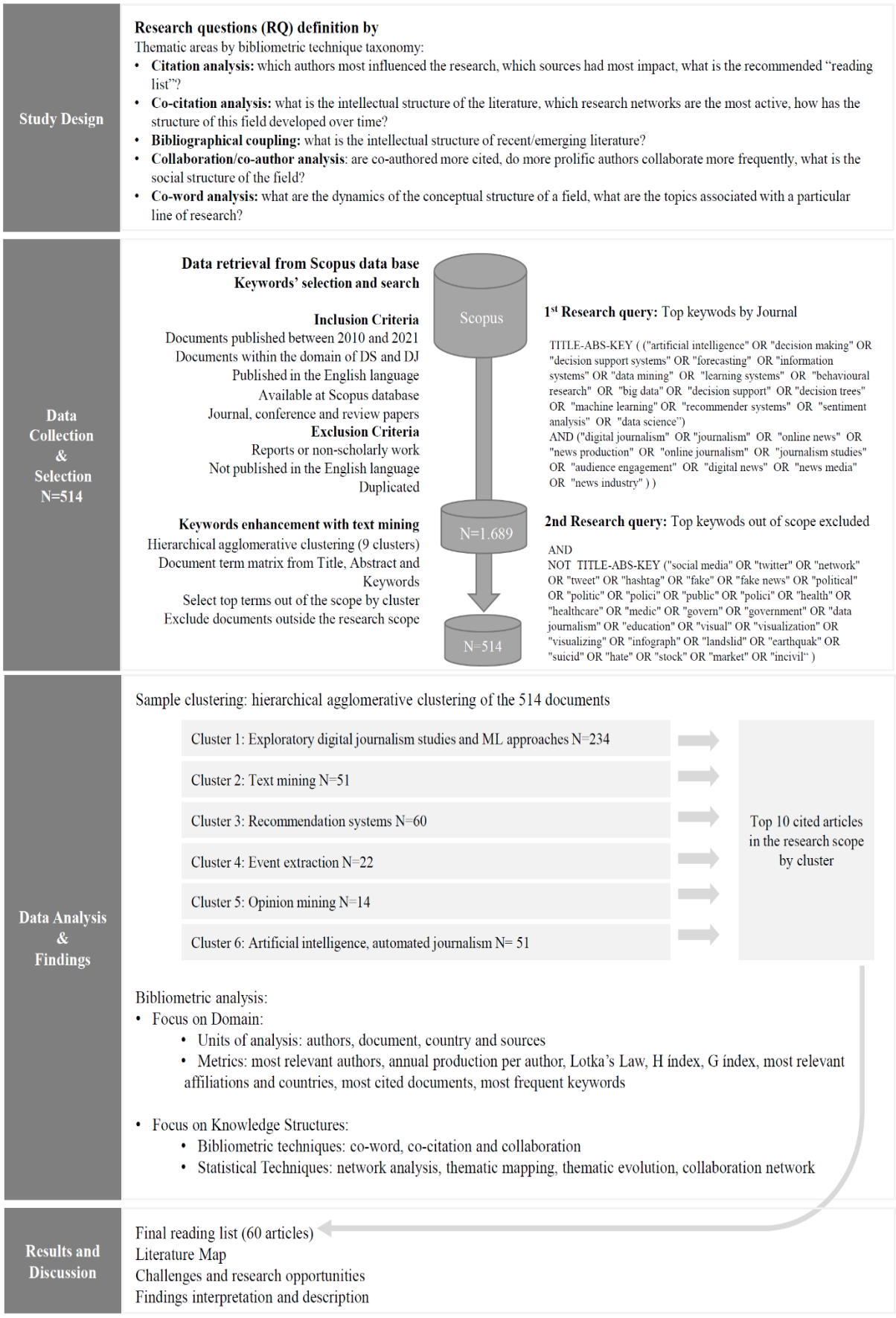


Figure 5 Framework of the systematic literature review process.

Information Extraction: Text from Title, Abstract and Keywords

Pre-processing

- stemming, normalization, stopwords and noise removal.
- document term matrix (DTM) calculation: only words with length greater than three were selected.
- term frequency-inverse document frequency calculation (tf-idf)
- DTM-tf-idf matrix calculation

R packages: tm, NLP, qdap Rweka

Agglomerative hierarchical clustering

- simililaritude measure: cosine distance
- Ward's agglomerative method ward.D2
- Overall validity, cluster's compactness and separation measures

R packages: stylo, factoextra

Cut level	Average distance between clusters	Average distance within clusters	Dunn Index	Average Silhouette
8	0.7970	0.8477	0.9675	0.02729
9	**0.7970**	**0.8418**	**0.9675**	**0.02835**
10	0.7616	0.8366	0.9519	0.02800
11	0.7616	0.8320	0.9519	0.02775

9 clusters

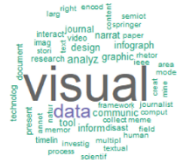
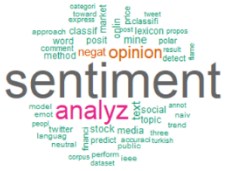
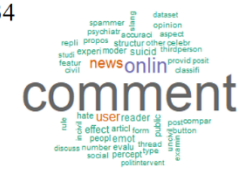
N° of documents by cluster	164	344	582
Wordcloud			
Top 20 most frequent words by cluster	**social, media, twitter, network, tweet,** analyz, data, content, inform, platform, **hashtag**, sourc, mine, share, autom, user, link, servic, credibl, differ	news, predict, model, learn, text, topic, extract, featur, detect, classif, onlin, machin, method, articl, propos, **fake**, popular, document, mine, languag	media, **polit, health, public**, newspap, report, **polici**, engag, risk, behavior, frame, studi, **govern**, coverag, particip, decis, chang, audienc, search, examin
N° of documents by cluster	232	48	95
Wordcloud			
Top 20 most frequent words by cluster	**journal, journalist, data**, technolog, digit, **educ,** scienc, practic, communic, ethic, profession, intellig, comput, autom, develop, artifici, open, author, product, research	**visual**, data, analyz, communic, graphic, design, tool, inform, **infograph**, narrat, journal, interact, research, video, document, disast, present, technolog, imag, timelin	recommend, user, news, system, person, onlin, algorithm, read, interest, model, item, articl, filter, base, propos, prefer, collabor, profil, divers, approach
N° of documents by cluster	91	99	34
Wordcloud			
Top 20 most frequent words by cluster	event, extract, detect, news, report, futur, inform, semant, time, **landslid**, framework, spatiotempor, locat, unrest, cluster, **twitter, earthquak**, sourc, present, relat	sentiment, analyz, opinion, negat, text, mine, **stock,** onlin, media, social, **market,** lexicon, classif, topic, word, posit, method, detect, polar, classifi	comment, onlin, news, user, effect, emot, reader, **suicid,** moder, **hate**, peopl, articl, public, evalu, percept, structur, experi, type, number, **incivil**

Figure 6 Framework of the text mining process to find the keywords to exclude in the second search query.

2.5.2. Data analysis and findings

The present section aims to explore the thematic areas presented in Figure 6. Hence, by performing citation, co-citation, collaboration and co-word analysis, coupled with an hierarchical clustering of the collection, the RQ1 presented at Section 3.1 is answered. Then, at Section 5 a co-word analysis coupled with keywords co-occurrences maps is also presented, which answers RQ2 and RQ3.

The statistical analysis was performed by using two open-source tools: *biblioshiny* that is a shiny app providing a web-interface for bibliometrix (Aria & Cuccurullo, 2017) and VOSviewer (Cobo et al., 2011; Van-Eck & Waltman, 2010).

The sample comprises three types of documents: 228 articles/journal researchs (44%), 278 conference researchs (54%) and 8 review researchs (2%) (see Table 5). Furthermore, 47% of total sample was published between 2018 and 2020 (see Figure 7). In fact, in the last decade, there is an increasing interest in DS along with the popularization of paywall models (Arrese, 2016; Rußell et al., 2020). Moreover, we have 1,161 authors (87%) with a single contribution, which indicates that a diverse group of researchers is interested in this research field. Besides, that it is also corroborated by the high number of sources (324) proving that most editors consider the subject relevant.

The worldwide spreading of authors, obtained from *biblioshiny* (see Figure 8 a)), indicates that northern hemisphere is more representative, i.e., researchers from North America, Asia and European Union (including UK) published 25%, 34% and 35% of the total number of documents, respectively. Furthermore, as presented at Figure 8 b), the most cited countries are USA, China and Singapore. However, the country with higher average article citations is Switzerland, followed by Singapore and Portugal that have an average year of publication 2017 and 2018 respectively; while Switzerland has older publications. Figure 8 c) illustrates a bibliometric VOSviewer network visualization map of co-authorship (international collaboration) using country by average year of publication and number of publications (Eck & Waltman, 2013; Romero & Portillo-Salido, 2019). The distance between countries approximately indicates the relatedness of the countries in terms of co-authorship.

Citation analysis intends to identify the authors and journals that most influenced the research (Donthu et al., 2021). It also provides the authors and journals that consequently contributed to the major topics of research on DS in DJ presented in the final reading list at Table 7 (thus answering to RQ1). In particular, Table 6 enables to identify the 20 most productive authors. The vast majority contributed or started the contribution after 2015. Three authors present more than four publications: firstly, Nicholas Belkin, affiliated with Rutgers University, presents contributions in 6 of the 10 years under analysis. Followed by Duen-Ren Liu affiliated with National Chiao Tung University and Nello Cristianini affiliated with Bristol University, each one contributed with 5 documents. Nicholas Belkin contributed with studies in the field of information retrieval, information search and user behavior, which appear in the first research domain shown at Table 7 (Cole et al., 2011, 2015; Liu et al., 2010). Liu D.-R. research focuses on recommendation systems (D. R. Liu et al., 2018) while Cristianini focuses on news content analysis and readers preferences (Flaounas et al., 2013).

Table 5 Main information about the collection (source: bibliometrix).

	Description	Results
Main information about data	Timespan	2010:2021
	Sources (Journals, Books, etc)	324.00
	Documents	514.00
	Average years from publication	4.44
	Average citations per documents	8.35
	Average citations per year per doc	1.24
Document types	Article	228.00
	Conference research	278.00
	Review	8.00
Document contents	Author's Keywords (DE)	1,476.00
Authors	Authors	1,330.00
	Author Appearances	1,777.00
	Authors of single-authored documents	87.00
	Authors of multi-authored documents	1,243.00
Authors collaboration	Single-authored documents	90.00
	Documents per Author	0.39
	Authors per Document	2.59
	Co-Authors per Documents	3.07
	Collaboration Index (the average number of co-authors noted solely in multi-authored publications (Gil et al., 2020))	2.93

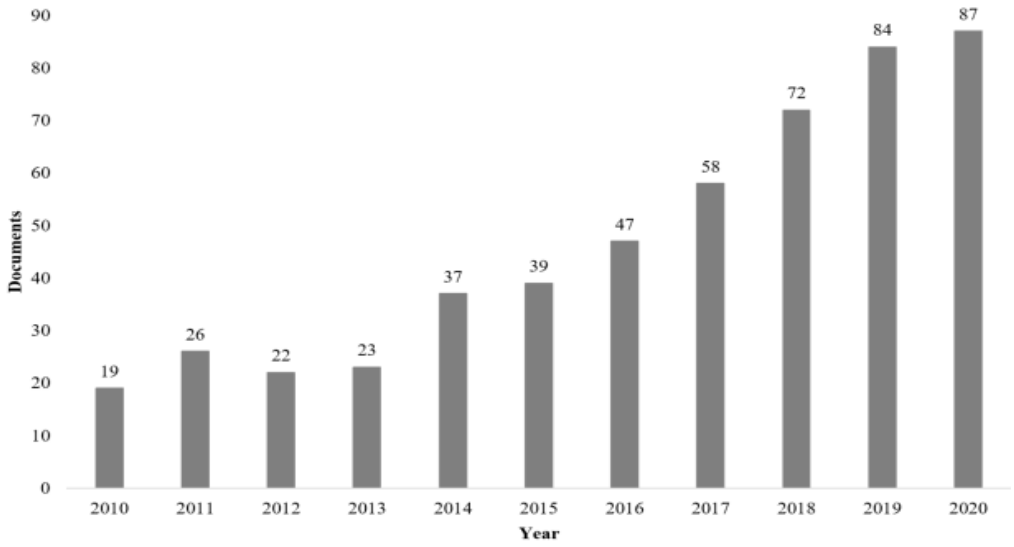


Figure 7 Year-wise distribution of publications.

In terms of researcher's impact measure, at Table 6 we present H-index and G-index that are based on the number of publications and the number of citations of the bibliographic collection (Egghe, 2006; Hirsch, 2005). In the overall sample, the author's with the highest H-index and G-index are Nicholas Belkin and Nello Cristianini. Nicholas Belkin ranked top in the list where 6 of his articles have been cited at least 6 times each.

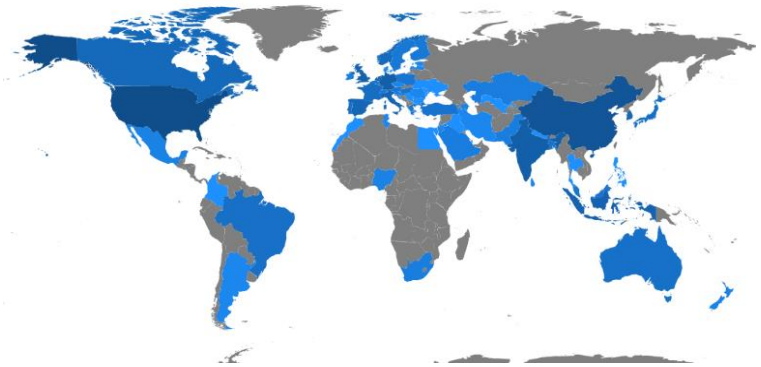
In order to perform collaboration analysis, it was identified 26 clusters of collaboration network. Figure 9 illustrates 11 of the 26 clusters and their main fields of research. Cluster 10, 2 and 16 present the highest network with 6, 5 and 4 researchers. Nicholas Belkin and Michael Cole published documents together (cluster 10) (see Table 6) and two of those are in the top most 20 cited articles of the sample (Cole et al., 2011; Liu et al., 2010). Furthermore, Nello Cristianini, Ilias Flaounas, Omar Ali and Tijl De Bie (cluster 2) have one article in the top 20 most cited (Flaounas et al., 2013) as presented at Table 7.

Seeking to investigate **RQ1** regards, the analysis of keywords allow us to understand the boundaries of the research domain, to find trends and to identify some relationships (Abdelmageed & Zayed, 2020). Thus, Figure 10 presents the wordcloud of the top 50 author's keywords and highlights the most common keywords of the articles of the database (Donthu et al., 2021). In the most frequent keywords we find "text mining" that occurs 31 times in our collection, followed by "machine learning", "big data" and "sentiment analysis", which appears 26, 23 and 22 times respectively. Next, the words "artificial intelligence", "data mining", "news recommendation" and "natural language processing" occur between 18 to 13 times.

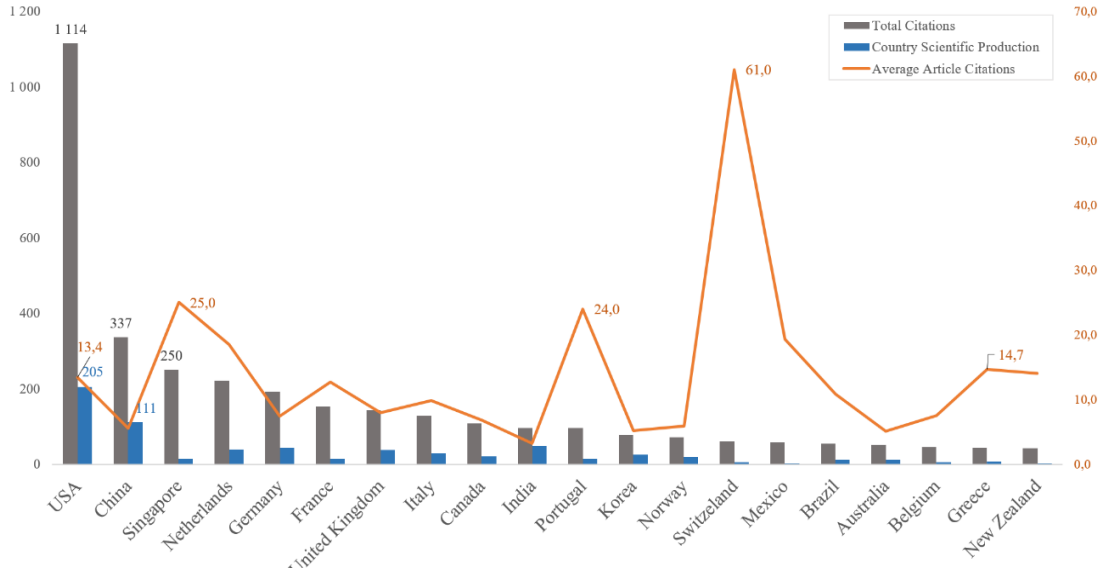
Moreover, Figure 11 illustrates the overlay visualization mode of author keywords, the full-counting method was applied to calculate keyword weights. We considered the minimum number of occurrences of a keyword of 5. The distance between keywords approximately indicates the relatedness of the keywords in terms of co-citation links, and each color represents a moment in the timespan. Accordingly, to the diagram colours, from blue to yellow, during the period of study (from 2010 to 2021), keywords such as “text mining”, or “information extraction” were more frequent at the beginning of the period. Followed by “news recommendation” or “sentiment analysis” (at green colour) and, recently, “artificial intelligence”, “big data” or “automated journalism”.

Furthermore, by exploring the thematic evolution map (see Figure 12) to complement the data presented at Figure 11, we can note that “text mining”, “svm” and “computational journalism” are important keywords between 2010 and 2017. Moreover, both stages have little connection, as the number of common keywords is low. The focus between the first and second stages evolved to other DS domains such as, audience engagement, machine learning or artificial intelligence, which is also corroborated by Figure 11. As an example, “text mining” evolved into “online news”, “machine learning” or “sentiment analysis”.

In addition, a clustering of our collection help us to explore the main domains of research. By repeating the previous AHC algorithm, the collection was partitioned into six groups (see Figure 13). Each cluster allows us to identify the major research domains to adopt DS in DJ (RQ1), that are: exploratory studies and detached ML approaches, text mining analysis, recommendation systems, event extraction, opinion mining, and automated journalism. In accordance to the previous network analysis, the period between 2018 and 2020 presented an increase on exploratory studies and detached ML studies as well as, as increase on research on text mining, recommendation systems and artificial intelligence.

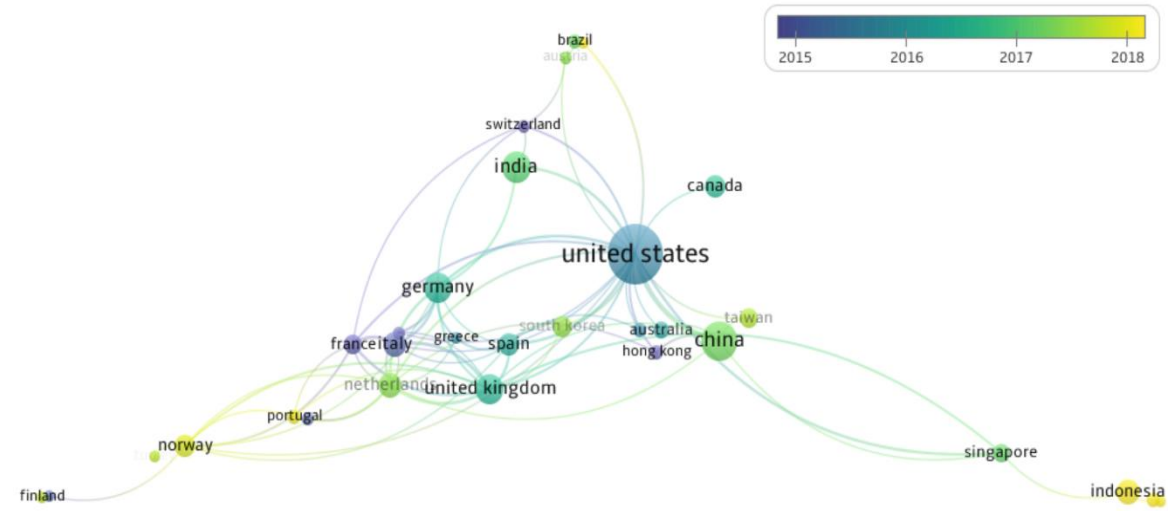


(a)



Note: total citations (left y-axis at grey colour), country scientific production (blue) and average article citations (right y-axis at orange).

(b)



Note: network visualization map of country co-authorship by average year of publication and number of publications (documents weights).

(c)

Figure 8 (a) Geographic distribution of published articles by country-based scientific production (b) the 20th most cited countries (c) VOSviewer network visualization map.

Table 6 Authors' production over time from the top 20 authors that contributed with 73 documents.

Authors' Production over time (Top 20)	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Total (#)	H-index	G-index	Cluster of collaboration
Nicholas Belkin	1	1				1	1	1	1			6	6	6	10
Duen-Ren Liu								1	1	3		5	2	2	9
Nello Cristianini	1	2		1		1						5	4	5	2
Abhijnan Chakraborty								1	1	1	1	4	2	3	14
Jaron Harambam									1	3		4	3	4	7
Andreas Lommatzsch						1	1	2				4	1	2	17
Simon Fong			1	1	1		1					4	3	4	3
Ralf Steinberger			1	2	1							4	3	4	4
Ilias Flaounas	1	2		1								4	3	4	2
Heidar Davoudi								1	1		1	3	2	2	11
Dimitrios Bountouridis										3		3	3	3	7
Nicholas Diakopoulos							1			2		3	2	2	12
Yun-Cheng Chou								1	1	1		3	2	2	9
Saptarshi Ghosh						1		1		1		3	2	3	14
Marcel Broersma				1					2			3	2	3	8
Miriam Boon						1		1	1			3	1	1	6
Bich-Liên Doan					2			1				3	1	2	16
Michael J. Cole	1	1				1						3	3	3	10
Omar Ali	1	1		1								3	3	3	2
Tijl De Bie	1	1		1								3	3	3	2
Total (#)	6	8	2	8	4	6	4	10	9	14	2	73			
								62%							

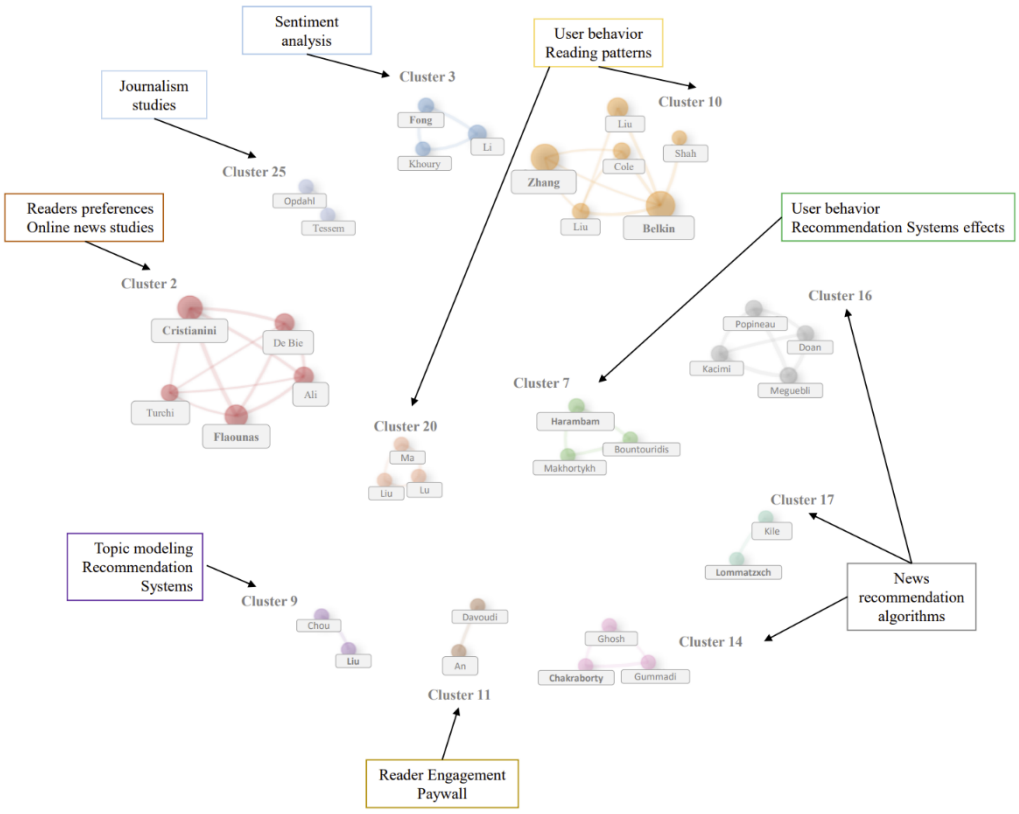
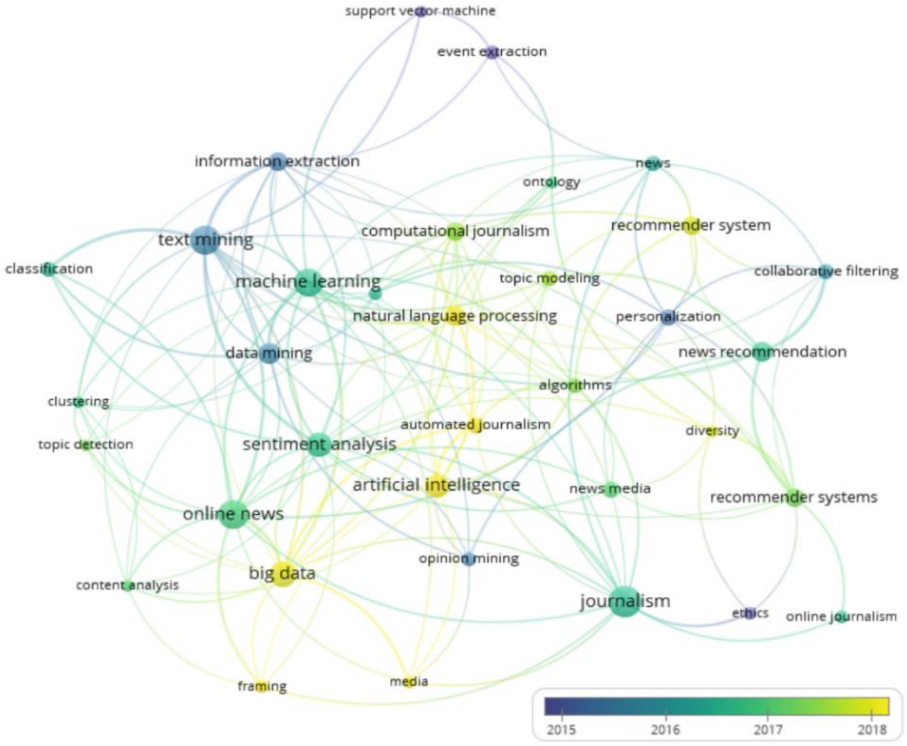


Figure 9 *Bibliometrix* collaboration network map between authors from 11 of the 26 clusters of authors.

Figure 10 Wordcloud of top 50 author's keywords, the word size depends on word occurrence.



Note: co-occurrences map of keywords based on the full-counting method with a minimum number of occurrences of a keyword 5. The size of the nodes represents the relevance of the terms in the researchs. The thickness of the lines means the bonding force between them. Finally, the colours indicate the average year of articles publication that mention those keywords.

Figure 11 VOSviewer co-occurrences map of keywords.

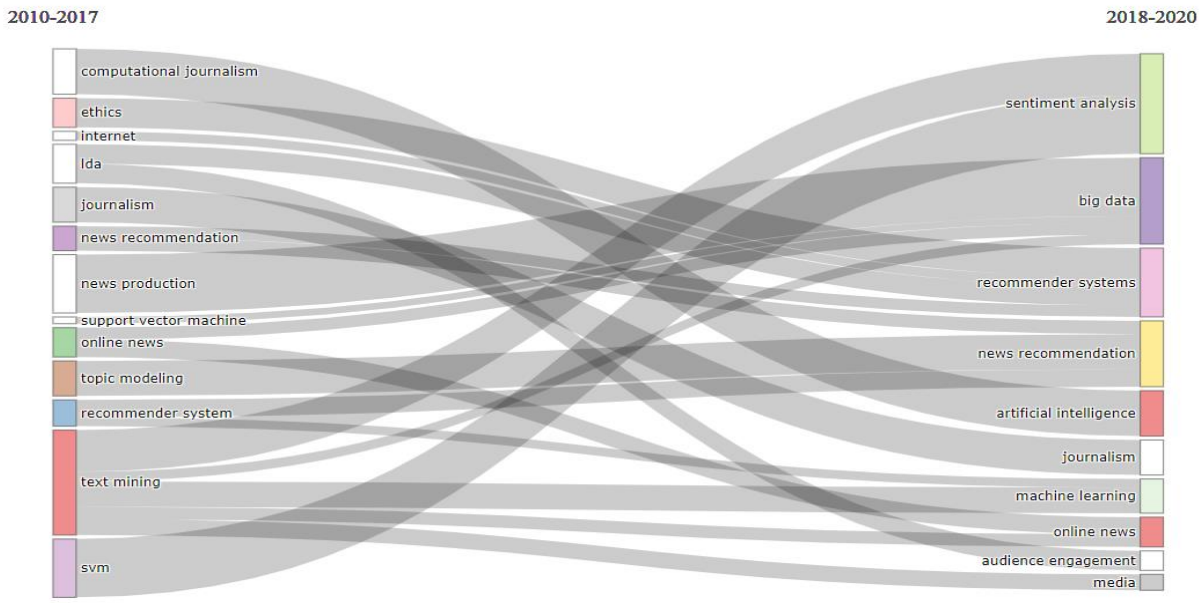


Figure 12 *Bibliometrix* thematic evolution map that demonstrates the evolution of keywords in two different stages (2010-2017, 2018-2020).

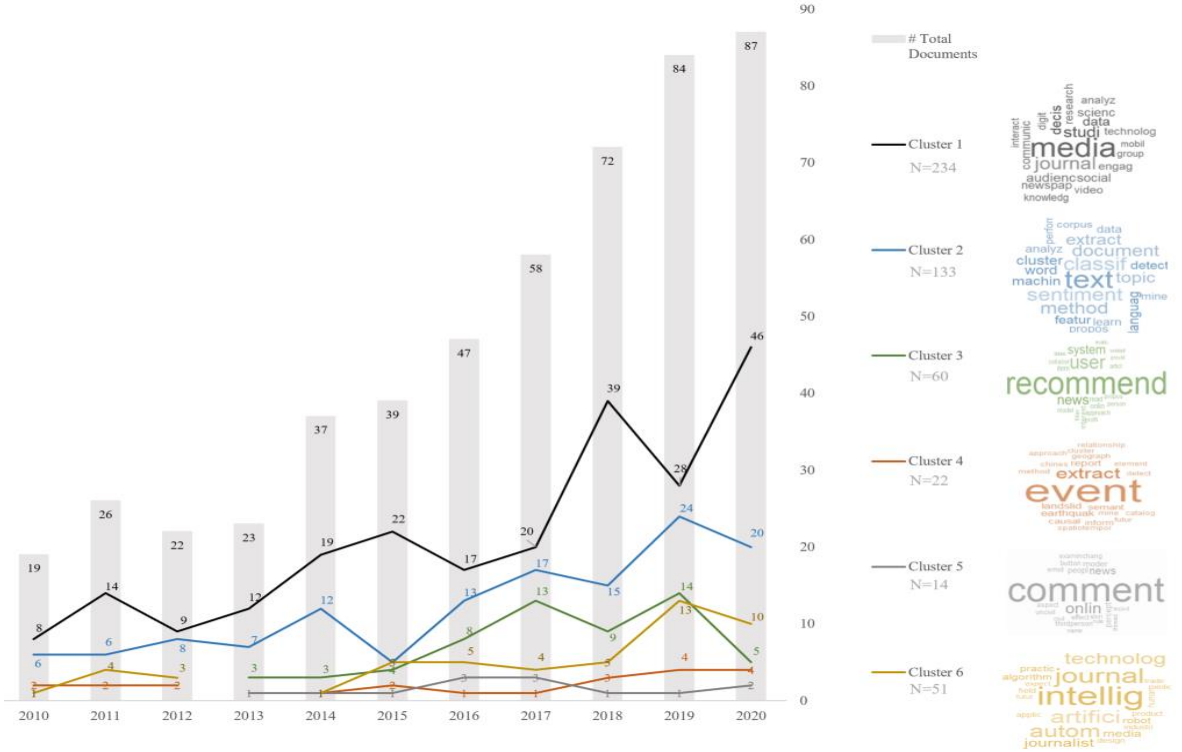


Figure 13 Scientific Production by year and by cluster.

The top cited research, which presents a news recommendation system, was published by (J. Liu et al., 2010) and it is also the top cited per year (32.9). Followed by, (Tandoc Jr, 2014) with 21 citations per year, that studies the impact of web analytics in the gatekeeping process. And, the third most cited per year, (Carlson, 2014) presents a case study analysis about automated journalism. These studies indicate the most significant research domains (RQ1) of the collection that contain some of the most important keywords between 2018 and 2020 (see Figure 11), such as, "recommender systems" or "artificial intelligence".

In the collection, only 23 articles (4.4%) have more than five TC per year as the interest on DS in DJ is recent. As result, the comparison of older articles with newer only based in citations could exclude influential documents. Furthermore, *bibliometrix* presents the normalized citation score of a document (NCS) calculated by dividing the actual count of citing items by the expected citation rate for documents with the same year of publication (Aria & Cuccurullo, 2017). In our collection, 45 documents (8.6%) present value higher than 3 and 360 documents (70%) less than one. Thus, the top three highest NCS (13.8, 12.1 and 11.1) were published by (Haim et al., 2018; S. Lewis et al., 2019; Schonlau & Zou, 2020), related to "personalization", "journalism automation" and "statistical learning". However, the last two are not at Table 7, as their TC is lower than the top ten articles of their cluster. Nevertheless, both are mentioned in the literature map (see Figure 16), as they present promising future research trends in journalism.

2.6. Discussion and challenges

In this section, the analysis conducted is based on the outcome from the procedure illustrated at Figure 5. To answer the RQs, a deeper analysis across each cluster (see Figure 13), allows us to summarize the major topics (RQ1), benefits (RQ2) and gaps (RQ3) of DS applications in DJ. Furthermore, we summarize the research by presenting a literature map (see Figure 16) that contains different levels of interactions, which are: the main domains found in the six clusters, DS topics of research in DJ and some of the most relevant studies found in the SLR. The characteristics of the detected clusters are summarized as follows (across the text we use the abbreviations RQ1, RQ2 and RQ3 to signal each RQ answer):

- Cluster 1 ("**Exploratory studies and ML approaches**") contains 234 articles (46%), with the top author keywords being "online news", "big data", "machine learning", "opinion mining", "personalization" and "audience engagement". This cluster presents

approaches for personalization on information retrieval that include user behaviour analysis (Cole et al., 2011, 2015) (RQ1). Those studies use engagement metrics, such as dwell time (Liu, Cole, et al., 2010) to analyze reader preferences and satisfaction (Lu et al., 2018), or to measure live events engagement (Sanz-Narrillos et al., 2020). Moreover, articles proposing new engagement metrics are presented in this cluster, such as *viewport* time (Lagun & Lalmas, 2016). On the other hand, approaches based on ML algorithms include linear log prediction model (Tatar et al., 2014) or random forests to predict news popularity (Fernandes et al., 2015; Obiedat, 2020) and to predict news' shares (Schonlau & Zou, 2020) (RQ2). However, other engagement metrics could be used in predictive models, such as the number of comments that could be an opportunity for future research (Davoudi et al., 2018) (RQ3).

Furthermore, this cluster contains the only sample article about an objective function for optimal paywall decision making that shows the relevance of user engagement to increase subscription possibility (RQ2). Such result indicates that low research has been done about paywall solution's and their optimal design (Olsen et al., 2020; Rußell et al., 2020). Thus, improve digital business models in DJ is an opportunity for future research (Rußell et al., 2020) (RQ3).

The thematic evolution map (see Figure 12) shows that, the most frequent keyword in the first cluster, "online news" evolved to "big data". In fact, big data technologies make the management of online news big data feasible (RQ2). However, the exponential increase of data and the changes of reader behavior (Rußell et al., 2020) make some of the presented approaches limited with regard to their input. When dealing with real data, the future can be completely different from the past. Indeed, one of the three types of uncertainties when dealing with real forecasting situations is data uncertainty (Makridakis et al., 2020). Thus, research on data sources and data quality in DJ can help to improve DS models and results (RQ3).

This group contains 10 of the 20 articles with highest Normalized TC. They are not at Table 7, as they are recent, TC is less than the minimum of the top 10 in the cluster. Personalization (Haim et al., 2018), automation (S. Lewis et al., 2019), predict news shares (Schonlau & Zou, 2020), topic analysis in news (Canito et al., 2018), content analysis (Burggraaff & Trilling, 2017) are the main topics in the articles to be considered in the literature map presented at Figure 16.

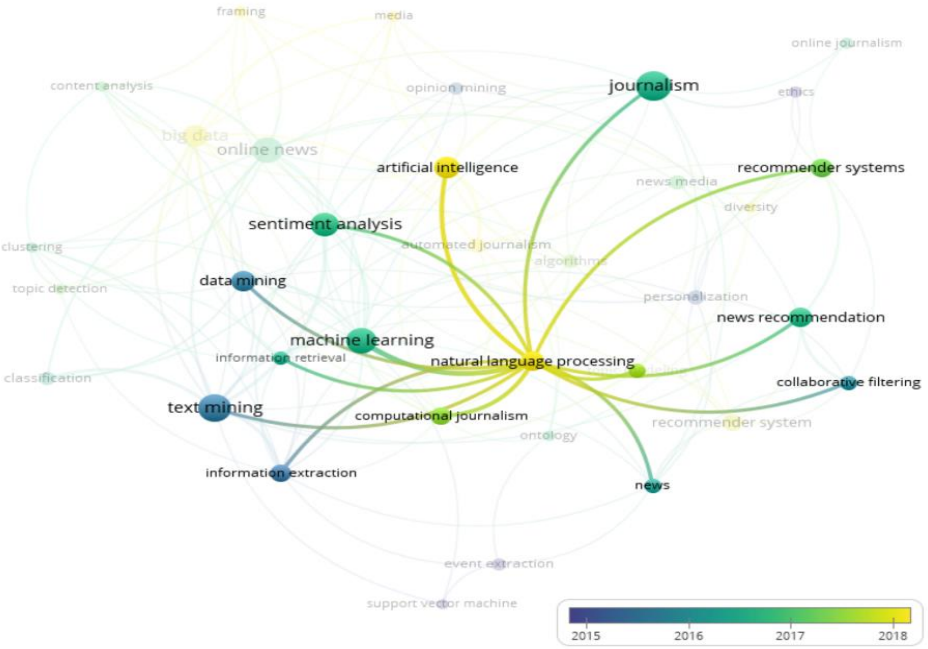
From the remaining five clusters, three are related to text analysis, the third main domain on the literature map (Figure 16).

- Cluster 2 ("**text mining - sentiment analysis**") contains 51 articles with the top author keywords being "text mining", "sentiment analysis" and "natural language processing". Those keywords are presented in blue, green and yellow at Figure 13 indicating a line of research in DJ across the timespan (RQ1). Furthermore, 82% of the articles were published since 2015 (see Figure 13) proving the increasing interest on TM approaches in the last five years. Approaches include topic modeling methods to build emotional dictionaries (Rao et al., 2014), classification algorithms (Li et al., 2016; Manjesh et al., 2017; Rivera et al., 2014) or predictive models (X. Bai, 2011) (RQ2). Besides, this cluster contains two articles that show the increasing interest on ML methods for automatic fact-checking (Azevedo, 2018; Indurthi et al., 2018) (RQ2). Both authors agree that in the big data era there is an imperative need and a research opportunity on fake news detection to build reader confidence (RQ3).
- Clusters 4 and 5 ("Orange" and "Grey" at Figure 16), defined as "**event extraction**" and "**opinion mining**" (see Figure 13), contain 22 and 14 documents, respectively. Each cluster present less than four publications by year. Two of the top 20 most cited articles belong to these clusters, one about event extraction (Hogenboom et al., 2011) and the other one about news comments modeling (Tsagkias et al., 2010) (as presented at Table 7). Approaches for **event mining** include the use spatiotemporal features to provide localized future suggestions to the reader (Ho et al., 2012), the development of semantic information extraction to track occurrences and evolution of event dynamics (W. Wang & Stewart, 2015), and research on methods for event semantic extraction to relieve information overrun (W. Wang, 2012; W. Wang et al., 2010) (RQ2). Furthermore, approaches for **opinion mining** include multiple classifiers (Häring et al., 2018; Lee & Ryu, 2019), meta-comments or ERIC's (engaging, respectful, and informative conversations) identification (Balali et al., 2013) (RQ2). Those studies prove the increasing importance to better understand reader comments to improve reader engagement (Häring et al., 2018) (RQ3). In fact, co-occurrences map (see Figure 15) present a clear line of research related to text mining fields, machine learning algorithms, natural language processing and big data.
- Clusters 3 and 6 ("Green" and "Yellow") mainly focused on **news recommendation** and **automated journalism**, respectively (RQ1). Both, present a slight increase of

publications since 2016 that demonstrates the increasing interest in simplify the content discovery (RQ2) and advanced analytics approaches (Gonzalez Camacho & Alves-Souza, 2018; Mizgajski & Morzy, 2019). Furthermore, there is an increasing interest in understanding how AI can help to improve DJ (Carlson, 2014; Lehmkuhl & Peters, 2016; S. Wu et al., 2019) (RQ3).

News recommendation systems development is a line of research that evolved from algorithms based on click behaviour (J. Liu, Dolan, et al., 2010) to more advanced methods (Babanejad et al., 2020; Hazrati & Elahi, 2021). Approaches that use temporal features (Muralidhar et al., 2015), movie and mobile solutions (Tewari et al., 2016; Viana & Soares, 2016), collaborative filtering applications (Saranya & Sadasivam, 2017; H. Wang et al., 2017) or neural networks to solve the cold-start problem (Misztal-Radecka et al., 2021) (RQ2). However, to explore other features such as, the article cost, the author level of engagement or the content propensity to induce subscription, can be relevant in future research (RQ3).

In what **automated journalism** concerns, the most part of the articles focus on exploratory studies (RQ1). Approaches focus on understanding ethical issues and the impact on the working practices of journalists in digital newsrooms (Carlson, 2014; García-Avilés, 2014a), on the potentialities and pitfalls for news organizations (S. C. Lewis et al., 2019), as well as analyze the user perception to automated news (Zheng et al., 2018). Finally, there are other studies related with specific topics, such as: AI techniques to improve the organization, management and distribution of content (Barriuso et al., 2016); or intelligent news robots (W. Yang, 2020) to reduce routine tasks to prove the positive impact of AI in DJ (RQ2). Moreover, there seems to exist a low emphasis on the use of AI to increase levels of reader engagement (RQ3). This is an interesting finding, revelling a gap on the research on how AI can affect readers' engagement (RQ3).



Note: co-occurrences map based on the full-counting method (cluster 2 - "text mining"). The weight being visualized is the occurrence, thus when a keyword has a greater weight the label and bubble are bigger (Eck & Waltman, 2013).

Figure 14 VOSviewer keywords co-occurrences map.

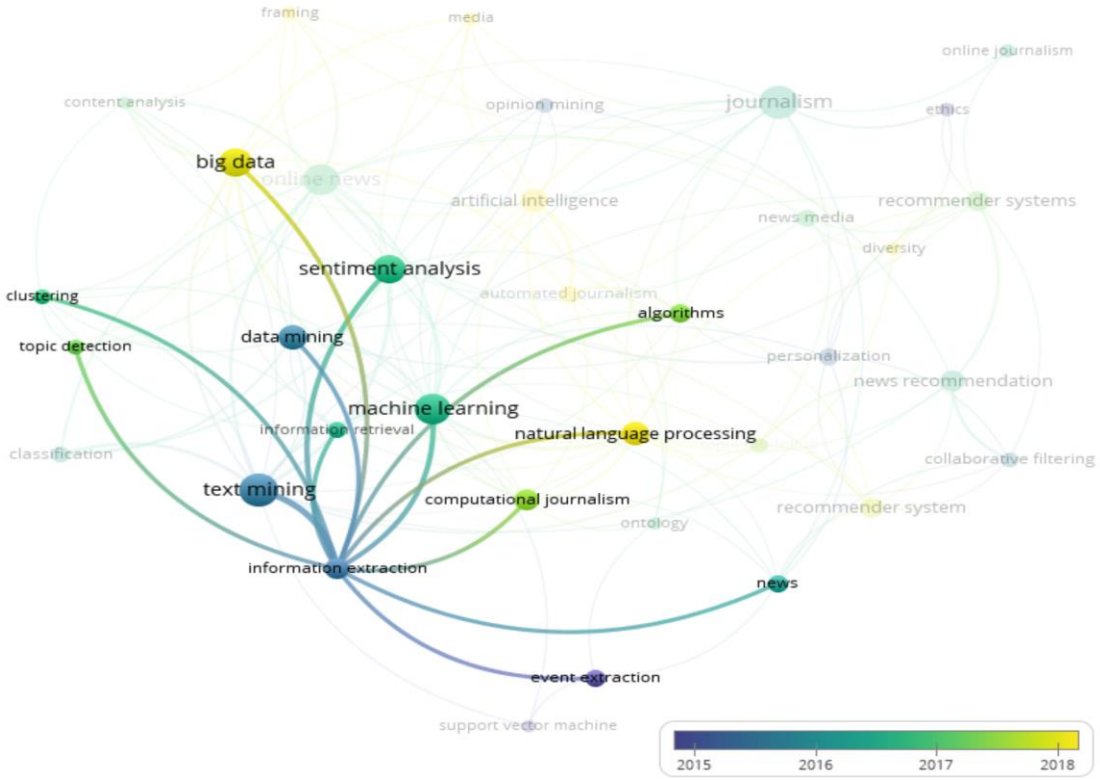


Figure 15 VOSviewer keywords co-occurrences map based on the full-counting method (cluster 4 - "event extraction").

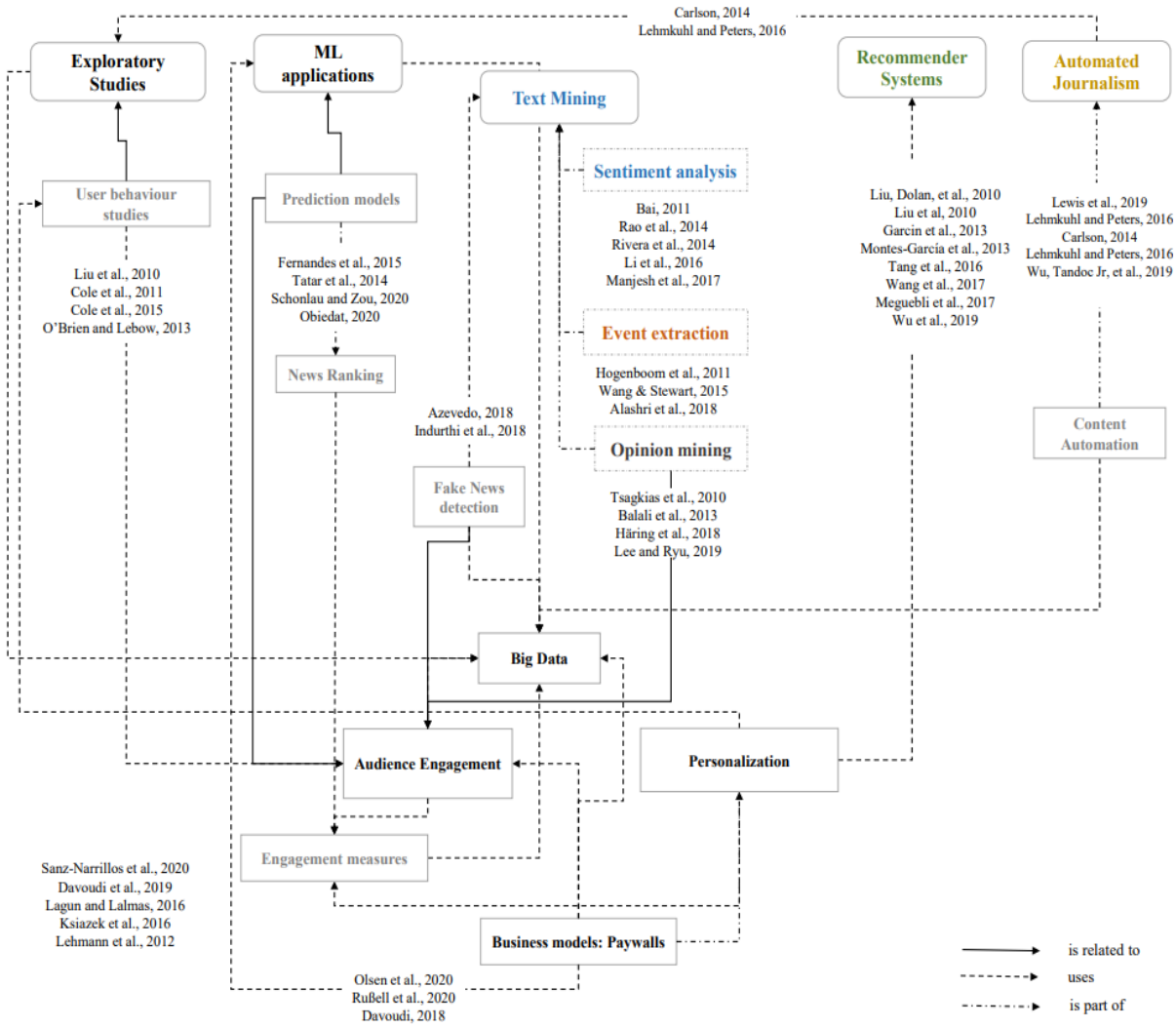


Figure 16 Literature map.

Across the SLR, we have demonstrated the main motivations and positive impacts of DS use in DJ to improve reader engagement (RQ1 and RQ2). For instance, exploratory web analytics studies and practical ML applications improve reader experience and simplify content discovery, consequently, increases engagement metrics, such as time, interactivity or *viewport* time. Furthermore, applications on news popularity (Y. Yang et al., 2020) forecast helps media companies to optimize homepage decisions and maximize content distribution to acquire and retain more readers. Moreover, TM applications by using sentiment analysis methods (Greco & Polli, 2020), event mining or opinion mining allow in understanding reader's interests, helps to provide better recommendation according to readers' opinions and consequently media platforms provide more content increasing recirculation and time per visit. We further note the increasing relevance of recommendation systems to improve personalization (Gonzalez Camacho & Alves-Souza, 2018). As well as the use of automated journalism to reduce routine tasks and improve truly journalism.

Table 7 The ten most cited articles related to the field under study by cluster.

Cluster	Authors, Year	Title	TC (rank number)	TC per Year	Normalized TC (rank number)	Source (highlighted top 10 sources)	IF	SJR 2019
1 - Exploratory research and detached ML approaches N=234	(Tandoc Jr, 2014)	Journalism is twerking? How web analytics is changing the process of gatekeeping	168 **(2nd)**	**21.0**	**10.7 (6th)**	New Media and Society	4.577	2.96
	(Liu et al., 2010)	Search behaviors in different task types	82 (6th)	**6.8**	2.3 (69th)	Proceedings of the ACM International Conference on Digital Libraries	---	---
	(Fernandes et al., 2015)	A proactive intelligent decision support system for predicting the popularity of online news	73 (7th)	**10.4**	4.7 (22nd)	**Lecture Notes in Computer Science**	---	0.43
	(Leetaru, 2011)	Culturomics 2.0: Forecasting Large-Scale human behavior using global news media tone in time and space	63 (11th)	**5.7**	3.1 (44th)	First Monday	---	0.7
	(Tatar et al., 2014)	From popularity prediction to ranking online news	61 (12th)	**7.6**	3.9 (30th)	Social Network Analysis and Mining	0.398	0.4
	(Haim et al., 2018)	Burst of the Filter Bubble?: Effects of personalization on the diversity of Google News	60 (14th)	**15**	**13.8 (1st)**	**Digital Journalism**	4.476	2.69
	(Cole et al., 2011)	Task and user effects on reading patterns in information search	50 (17th)	4.5	2.5 (62th)	Interacting with Computers	1.036	0.42
	(Reis et al., 2015)	Breaking the news: First impressions matter on online news	48 (18th)	**6.9**	3.1 (45th)	Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015	---	---
	(Flaounas et al., 2013)	Research methods in the age of digital journalism: Massive-scale automated analysis of newscontent—topics, style and gender	48 (19th)	**5.3**	2.8 (51st)	**Digital Journalism**	4.476	2.69
	(Lagun and Lalmas, 2016)	Understanding and measuring user engagement and attention in online news reading	45 (20th)	**7.5**	5.2 (16th)	WSDM 2016 - Proceedings of the 9th ACM International Conference on Web Search and Data Mining	---	0.78
2 - Text mining N=51	(X. Bai, 2011)	Predicting consumer sentiments from online text	141 **(3rd)**	**12.8**	**7.1 (9th)**	Decision Support Systems	4.721	1.92
	(Rao et al., 2014)	Building emotional dictionary for sentiment analysis of online news	96 **(5th)**	**12.0**	6.1 (15th)	World Wide Web	2.892	0.53
	(Christin, 2017)	Algorithms in practice: Comparing web journalism and criminal justice	64 (10th)	**12.8**	**10.8 (5th)**	Big Data and Society	4.577	3.25
	(Du et al., 2015)	Dirichlet-hawkes processes with applications to clustering continuous-time document streams	57 (15th)	8.1	3.7 (31st)	**Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**	---	---
	(Burrows et al., 2013)	Paraphrase acquisition via crowdsourcing and machine learning	41 (24th)	4.6	2.4 (64th)	ACM Transactions on Intelligent Systems and Technology	---	1.05
	(Li et al., 2016)	Hierarchical classification in text mining for sentiment analysis of online news	25 (41st)	4.1	2.9 (46th)	Soft Computing	3.050	0.71
	(Steinberger, 2012)	A survey of methods to ease the development of highly multilingual text mining applications	23 (46th)	2.3	2.6 (57th)	Language Resources and Evaluation	1.014	0.44
	(Flaounas et al., 2010)	The structure of the EU mediasphere	23 (47th)	1.9	0.6 (197th)	PLoS ONE	---	1.02
	(Rivera et al., 2014)	A text mining framework for advancing sustainability indicators	19 (57th)	2.4	1.2 (129th)	Environmental Modelling and Software	4.807	1.9
	(Zhu, Zhu, et al., 2014)	Tracking the Evolution of Social Emotions: A Time-Aware Topic Modeling Perspective	18 (60th)	2.3	1.1 (146th)	**Proceedings - IEEE International Conference on Data Mining, ICDM**	---	0.79

3 - Recommendation systems N=60	(J. Liu, Dolan, et al., 2010)	Personalized news recommendation based on click behavior	395 (1st)	**32.9**	**11.1 (4th)**	International Conference on Intelligent User Interfaces, Proceedings IUI	---	0.59
	(Garcin et al., 2013)	Personalized news recommendation with context trees	61 (13th)	6.8	3.6 (35th)	RecSys 2013 - Proceedings of the 7th ACM Conference on Recommender Systems	---	---
	(O'Brien & Lebow, 2013)	Mixed-methods approach to measuring user experience in online news interactions	42 (23rd)	4.7	2.4 (63th)	Journal of the American Society for Information Science and Technology	2.410	---
	(Montes-García et al., 2013)	Towards a journalist-based news recommendation system: The Wesomender approach	32 (30th)	3.6	1.9 (85th)	Expert Systems with Applications	5.452	1.49
	(J. A. Yang, 2016)	Effects of popularity-based news recommendations ("most-viewed") on users' exposure to online news	31(32nd)	5.2	3.6 (34th)	Media Psychology	2.397	1.863
	(Tang et al., 2016)	An empirical study on recommendation with multiple types of feedback	20 (55th)	3.3	2.3 (68th)	**Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**	---	---
	(H. Wang et al., 2017)	Hybrid recommendation model based on incremental collaborative filtering and content-based algorithms	15 (71st)	3.0	2.5 (59th)	**Proceedings of the 2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design, CSCWD 2017**	---	---
	(Mizgajski & Morzy, 2019)	Affective recommender systems in online news industry: how emotions influence reading choices	11 (93rd)	3.7	4.3 (25th)	User Modeling and User-Adapted Interaction	4.682	1.57
	(C. Wu et al., 2019)	Neural news recommendation with attentive multi-view learning	10 (100th)	3.3	3.9 (29th)	IJCAI International Joint Conference on Artificial Intelligence	---	1.21
	(Chakraborty et al., 2019)	Optimizing the recency-relevancy trade-off in online news recommendations	9 (108th)	1.8	1.5 (114th)	26th International World Wide Web Conference, WWW 2017	---	---
4 - Event extraction N=22	(Hogenboom et al., 2011)	An overview of event extraction from text	66 (9th)	6.0	3.3 (38th)	**CEUR Workshop Proceedings**	---	0.18
	(W. Wang & Stewart, 2015).	Spatiotemporal and semantic information extraction from Web news reports about natural hazards	30 (35th)	4.2	1.9 (81st)	Computers, Environment and Urban Systems	4.655	1.36
	(Ho et al., 2012)	Mining future spatiotemporal events and their sentiment from online news articles for location-aware recommendation system	28 (36th)	2.8	3.2 (40th)	Proc. of the 1st ACM SIGSPATIAL Int. Workshop on Mobile Geographic Inf. Systems, MobiGIS 2012 - In Conjunction with the 20th ACM SIGSPATIAL Int. Conf. on Advances in Geographic Inf. Systems, GIS 2012	---	---
	(W. Wang, 2012)	Chinese news event 5W1H semantic elements extraction for event ontology population	17 (63rd)	1.7	1.9 (82nd)	WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web Companion	---	---
	(W. Wang et al., 2010)	Extracting 5W1H event semantic elements from Chinese online news	14 (79th)	1.2	0.4 (237th)	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	---	0.43
	(Wei Wang et al., 2012)	Chinese news event 5W1H elements extraction using semantic role labeling	7 (131st)	0.6	0.2 (280th)	Proceedings - 3rd International Symposium on Information Processing, ISIP 2010	---	0.58
	(Tessem & Opdahl, 2019)	Supporting journalistic news angles with models and analogies	5 (150th)	1.7	1.9 (80th)	Proceedings - International Conference on Research Challenges in Information Science	---	---

	(Zhang et al., 2015)	RCFGED: Retrospective Coarse and Fine-Grained Event Detection from Online News	5 (159th)	0.6	0.3 (200th)	Proceedings - 2015 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2015	---	0.00
	(Fu et al., 2019)	Mining Newsworthy Events in the Traffic Accident Domain from Chinese Microblog	3 (210th)	1	1.2 (136th)	International Journal of Information Technology and Decision Making	1.894	0.41
	(Alashri et al., 2018)	Snowball: Extracting Causal Chains from Climate Change Text Corpora	2 (250th)	0.5	0.5 (217th)	Proceedings - 2018 1st International Conference on Data Intelligence and Security	---	0.21
5 – Opinion mining N=14	(Tsagkias et al., 2010)	News comments: Exploring, modeling, and online prediction	73 (8th)	6.1	2.0 (75th)	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	---	0.43
	(Chung et al., 2015)	Triggering participation: Exploring the effects of third-person and hostile media perceptions on online participation	24 (44th)	3.4	1.6 (113th)	**Computers in Human Behavior**	5.003	2.17
	(G. M. Chen & Ng, 2016)	Third-person perception of online comments: Civil ones persuade you more than me	23 (45th)	1.9	2.7 (197th)	**Computers in Human Behavior**	5.003	2.17
	(G. M. Chen & Ng, 2017)	Nasty online comments anger you more than me, but nice ones make me as happy as you	13 (80th)	2.6	2.2 (72nd)	**Computers in Human Behavior**	5.003	2.17
	(Napoles et al., 2017)	Automatically identifying good conversations online (yes, they do exist!)	9 (109th)	1.8	1.5 (115th)	Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017	---	0.55
	(Balali et al., 2013)	A supervised approach for reconstructing thread structure in comments on blogs and online news agencies	5 (164th)	0.6	0.3 (257th)	Computacion y Sistemas	0.620	0.19
	(Häring et al., 2018)	Who is addressed in this comment? Automatically classifying meta-comments in news comments	3 (205th)	0.8	0.7 (187th)	Proceedings of the ACM on Human-Computer Interaction	5.120	0.54
	(Meguebli et al., 2017)	Towards better news article recommendation: With the help of user comments	3 (212th)	0.6	0.5 (206th)	World Wide Web	2.892	0.46
	(Riedl et al., 2020)	The downsides of digital labor: Exploring the toll incivility takes on online comment moderators	2 (235th)	1	3.1 (42nd)	**Computers in Human Behavior**	5.003	2.17
	(Lee & Ryu, 2019)	Exploring characteristics of online news comments and commenters with machine learning approaches	2 (237th)	0.7	0.8 (174th)	Telematics and Informatics	4.139	1.44
6 - Automated journalism N=51	(Carlson, 2014)	The Robotic Reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority	137 (4th)	19.6	8.9 (8th)	**Digital Journalism**	4.476	3.69
	(García-Avilés, 2014b).	Online Newsrooms as Communities of Practice: Exploring Digital Journalists' Applied Ethics	22 (50th)	2.8	1.4 (121st)	**Journal of Mass Media Ethics: Exploring Questions of Media Morality**	0.867	0.549
	(Melki & Mallat, 2016)	Block Her Entry, Keep Her Down and Push Her Out: Gender discrimination and women journalists in the Arab world	15 (74th)	2.5	1.7 (93rd)	**Journalism Studies**	2.345	1.51
	(Lehmkuhl & Peters, 2016)	Constructing (un-)certainty: An exploration of journalistic decision-making in the reporting of neuroscience	13 (81th)	2.2	1.5 (116th)	Public Understanding of Science	2.338	1.14
	(Gravengaard, G. Rimestad, 2012)	Elimination of ideas and professional socialization:	12 (91st)	1.2	1.4 (125th)	**Journalism Practice**	1.542	1.26

		Lessons learned at newsroom meetings						
	(Y. Yang et al., 2017)	Perceived emotional intelligence in virtual agents	11 (94th)	2.2	1.9 (86th)	Conference on Human Factors in Computing Systems - Proceedings	5.23	0.67
	(S. C. Lewis et al., 2019)	Libel by Algorithm? Automated Journalism and the Threat of Legal Liability	10 (98th)	3.3	3.9 (27th)	Journalism and Mass Communication Quarterly	1.706	1.66
	(S. Wu et al., 2019)	When Journalism and Automation Intersect: Assessing the Influence of the Technological Field on Contemporary Newsrooms	9 (106th)	3.0	3.5 (36th)	**Journalism Practice**	1.542	1.26
	(Zheng et al., 2018)	When algorithms meet journalism: The user perception to automated news in a cross-cultural context	9 (107th)	2.3	2.1 (74th)	**Computers in Human Behavior**	5.003	2.17
	(Galily, 2018)	Artificial intelligence and sports journalism: Is it a sweeping change?	7 (123th)	1.8	1.6 (95th)	Technology in Society	2.414	0.566

2.7. Potential research opportunities

While there is an increasing need for data-driven approaches in journalism, the translation into ML approaches is still a complex task (Davoudi, 2018). Our findings rise in the form of a list of key topics with enhancements areas and future research opportunities (RQ3) listed as follows:

- **Big data**: the establishment of new datasets sources in DJ is required as most of the research is being done with limited datasets (Von Bloh et al., 2020). External data, like weather data or financial information, can help to better understand readers' patterns and behaviours, as well as, to improve DS models that consequently improve readers' engagement (Renó & Renó, 2015; Z. Yang, 2020).
- **Recommend Systems**: most of the existing approaches focuses on user's clicks as the indicator to understand users' interests either in, for example, engagement business indicators. Therefore, further research is required to explore innovative solutions, for example, to handle cold start problems, for multimedia content recommendations or to improve real-time recommendations (Ficel et al., 2021; Hazrati & Elahi, 2021; Zihayat et al., 2019).
- **Personalization**: as users present a wide range of reader behaviours (J. Liu et al., 2010), the development of innovative DS algorithms to better personalize user experience by website page, by device, by channel or by content type, will improve engagement levels (Haim et al., 2018; Omar et al., 2020). Furthermore, AI can be a solution to increase content engagement optimization (Kulkarni et al., 2019; Lim & Zhang, 2022).

- **Content automation**: one of the issues in the journalism ethics analysis is to explore the advantages of content automation (Danzon-Chambaud, 2021; S. C. Lewis et al., 2019; S. Wu et al., 2019). However, to invest in content automation can reduce routine tasks to improve journalism to the full potential (Carlson, 2014; Zheng et al., 2018). Furthermore, further research is required to automate content in other languages such as Spanish or Portuguese (Campos et al., 2020).
- **Fact-checking:** as the information increases, the information credibility and readers' trustworthiness become a matter of concern. Thus, explore new models on fake news detection is an opportunity of research (Azevedo, 2018; Meel & Vishwakarma, 2020; Shim et al., 2021).
- **Engagement metrics**: further research is required to bridge the gap between reader engagement metrics and business goals (Davoudi et al., 2019). Thus, to explore others metrics, such as, sentiment perceived in the comments to develop better predictive models (chum prediction or propensity to subscribe) can have a positive impact in the business model (Davoudi, 2018; Lehmann et al., 2012; Seale, 2021).
- **Paywall mechanism/business model**: as to the best of our scrutiny, only (Davoudi, 2018) investigate an adaptive paywall mechanism by using advanced analytics. ML and AI can help to design and improve more efficient paywall mechanisms. Furthermore, our study has shown that there is still a research gap concerning to the use of more advanced DS methods (e.g., Deep Learning (Goldani et al., 2021)). In fact, these findings are consisted with the work of (Davoudi, 2018), which argued that there is a gap between journalism and ML communities.

2.8. Conclusions of the SLR

In this section, we present a SLR analysis focused on the interaction between journalism, technology and data through the use of DS methods (including AI and ML) to improve reader engagement, attempt to identify trends, knowledge gaps and to indicate propositions to future researches. A total of 541 articles gathered from the Scopus database and published from 2010 to 2021 were scrutinized. The large number of articles makes the usage of TM convenient for a better selection and analysis of the literature. Bibliometric research and Hierarchical text mining (HTC) were combined to answer the RQs.

Generally, the findings show the hype of DS in DJ research, especially in the last three years, due to its potential to extract valuable information from big data. The SLR suggests that the literature about DS in DJ puts more emphasis on studying TM methods followed by recommendation systems. Furthermore, exploratory studies, web analytics and the impact of analytics in newsrooms are popular in the research. Finally, we note there is still a research gap concerning to the use of more advanced DS methods (RQ3), e.g., Deep Learning (Goldani et al., 2021). In fact, these findings are consisted with the work of (Davoudi, 2018), which argued that that there is a gap between journalism and ML communities.

Currently, big data challenges (Z. Yang, 2020), reader retention (Suárez, 2020), personalization and paywall models (Rußell et al., 2020) are some of the major points of concern in the industry. Furthermore, more research is required to improve data sources, to explore engagement metrics, to develop models for fake news detection (Goldani et al., 2021), and to investigate innovative paywall models.

In terms of theoretical contributions, this section presents an intensive literature review on the state of the art of DS in DJ, something that, to the best of our knowledge, none intensive SLR in this field of research has been published before. Nevertheless, this SLR has some limitations that also provides future research opportunities. Firstly, the literature search was carried out only on documents published at Scopus. Furthermore, non-English research and book chapters were neglected. Thus, future research can consider other scientific databases. Moreover, this study proposes three research questions, other researchers may add other questions. Then the final reading list can exclude important recent research as DS is a recent research field in DJ. Finally, non-scientific literature published by respectful entities in the area, such as INMA could be included in future research to explore recent successful DS use cases.

Hopefully, the results of this SLR can guide researchers in their collaboration with media companies in order to help publishers to improve readers' engagement through DS. In this PhD study, the research opportunities were also combined with Público analytical needs. Focused on engagement, Público aimed to increase reach from new targets, analyse solutions to segment readers to improve editorial and marketing actions, and also work on personalization to increase engagement and consequently reduce churn.

CHAPTER 3

Data-driven approach to increase online newspaper subscribers through Instagram users' engagement

3.1. Research context

This section presents the role of a data-driven strategy used to engage Instagram followers, and consequently, convert them into newspaper subscribers. Furthermore, this refers to a project implemented at Público to act in the first level of the reader funnel, as presented in Figure 1.

Smartphone popularization improved the trend toward visual social media platforms, which introduced new challenges to firms' social media strategies. Despite Instagram users' increase, research on its strategic use is rather limited, especially in the media industry. To fill this gap, an empirical analysis was performed through statistical analysis, text mining, and a quantitative discussion of internal and external data. Data was gathered between January 2020 and September 2021. Findings highlighted the positive effect of interactive content and vivid content. Futhermore, publishers should also take full advantage of narrative content that involves readers. Moreover, an effective information delivery to the newsroom through commented reports and useful dashboards was pivotal to efficiently delivering credible content, that drove the company to achieve business goals. The message strategy analyzed in the study provided guidelines for social media editors to create Instagram messages without affecting journalistic output. Thus, a multidisciplinary team composed of design, social media, and analytics' elements proven the advantage of applying a data-driven decision making process to optimize content mobile-first distribution in Instagram by ensuring journalistic values.

3.2. Article introduction

In the last decade, publishers faced a digital transformation that introduced challenges and opportunities to the media industry (Arrese, 2016). The decline of print revenue induced a move to digital business models, usualy in the form of paywall models (Pattabhiramaiah et al., 2019). As people engage with social media (SM) platforms (Newman, 2022), news organizations had to adapt their strategies to engage followers and boost readers's engagement into their own platforms (Rußell et al., 2020). Authors argue that, audiences are sensitive to the tone and style of communication (Denisova, 2022). Furthermore, narrative engagement plays an important role in involving audiences (Dahmen et al., 2021). Moreover, news organizations have been embracing the use of analytics as part of editorial decision-making in a big data environment

(Neilson & Gibson, 2022). Technological tools, organizational structures and cultural components were necessary parts to build a prodata culture in the newsroom (Lamot & Paulussen, 2020; Moyo et al., 2019). Therefore, Social Media Editors (SME) are intermediaries between audiences, newsrooms, and marketing departments (Neilson & Gibson, 2022) that need to improve strategies as they have a strategic position within the news organization (Moyo et al., 2019).

According to the Reuters Institute Digital News Report (Newman, 2022), Portugal remains in the countries with higher levels of trust in news. However, the rise in COVID-19 related misinformation in SM platforms induced new actions from media organizations. Furthermore, the literature showed a lack of research focusing on studying SM strategies in media (Hermida & Mellado, 2020; Larsson, 2018). Having recognized this gap, and motivated by the increase of interest on SM strategy (Neilson & Gibson, 2022; Yu et al., 2021), speacially in the media industry (Overgaard, 2021). Thus, the present research focuses on the effectiveness of the data-driven strategy used by the publisher Público to engage Instagram followers.

In this exploratory research we aim to address the following research questions (RQs): Are new approaches and formats of Instagram posts more engaging? Which type of message is the most effective? Is there a constant tone on caption posts that ensures authenticity and trustfulness? How can a data-driven strategy leverage website readers from Instagram? Are Instagram users converting into subscribers? By answering these RQs, this study aims to contribute to the literature on SM communication and to guide SMEs on how to effectively promote content that users engage and consequently convert into subscribers. Drawing on data from external and internal databases, collected between January 2020 and September 2021, we analyse the strategy followed.

3.3. Literature review

3.3.1. Reader engagement and social media journalism

Authors agree that reader engagement is a multidimensional phenomenon (Steensen et al., 2020) related to the level of attention and involvement emotional, cognitive and behavioral, with media (Attfield et al., 2011; Ksiazek et al., 2016; Manosevitch & Tenenboim, 2017; Mersey et al., 2010). Furthermore, previous studies have drawn attention to the relationship between SM strategy and reader engagement (Aydin, 2020; Shahbaznezhad et al., 2021). Moreover, studies reveal that there is a shift in the news production process, with a reliance increase on analytics and SM for multi-platform news dissemination (Moyo et al., 2019). However, in the case of Instagram, literature is rather scarce, especially in the media industry.

By consulting the digital database Scopus, which is the largest abstract and citation database of peer-reviewed literature (Ballew, 2009), the following query to articles' titles, abstracts and keywords only returns 27 documents (as of March 2023).

(TITLE-ABS-KEY (instagram AND (newspaper OR engagement OR online OR publisher) AND news AND strategy))

During the COVID-19 pandemic, SM has become the mainstream communication tool for information generation, dissemination, and consumption (Denisova, 2022; Newman, 2022; Volkmer, 2021) due to non-pharmaceutical interventions by governments that restricted physical mobility and led to confinement (Piccinelli et al., 2021). However, SM platforms also spread misinformation and biased information (Cage et al., 2020; Newman, 2022).

In fact, authors proved the association between misinformation and mental distress, self-harm and suicide increase (Rosenberg et al., 2020). Moreover, constructive journalism has a positive impact in times of crisis (Overgaard, 2021).

A systematic review of literature on health-related misinformation (Y. Wang et al., 2019) presents the main psychological aspects that make snowballing of misinformation occur. The authors concluded that potential interventions should act on cultivating critical thinking (Wang et al., 2019). Thus, journalism plays a crucial role in society and democracy by providing verified knowledge about current events (Carlson, 2017). By aiding awareness and distributing accurate information (Perreault & Perreault, 2021).

News organizations actively fought misinformation in Portugal by lifting paywalls for key COVID-19 related content (Newman, 2022). Furthermore, subscribers raised (Grzegorz, 2021) proving that customers pay for trustful journalism (Rußell et al., 2020). Therefore, SM platforms offer opportunities for reaching wider audiences (Nielsen & Ganter, 2018), increase customers' advocacy and induce subscriptions rise by developing a data-driven strategy across the reader conversion funnel (Kotler et al., 2016).

3.3.2. Instagram and visual journalism

Studies have examined user-generated content (UGC) to explore determinants of user engagement on social networks by focusing more on text analysis (Shahbaznezhad et al., 2021) and less on visual content (Overgaard, 2021). However, recent experiments suggest that including pictures in a post lead to greater levels of engagement (Li & Xie, 2020). Furthermore, visual solutions journalism provides comprehensive coverage of an issue, showing problems and responses (Midberry & Dahmen, 2020). According to (Aydin, 2020), "Vividness" is related to which a post simulates several senses by incorporating sound, movement and visual elements

to the post. Research have studied vividness and their effects on news coverage of violence and war (Overgaard, 2021).

Instagram is a visual social network and a multimedia sharing platform to use in mobile devices (Jarreau et al., 2019). The user can upload images and videos, and it can use a range of options to improve content presentation such as, filters, gifs, or captions (Jarreau et al., 2019). Instagram is the 5th preferred SM platform for sharing news (Newman, 2022). It's SM penetration in Portugal was 69% (Hootsuite, 2020). Furthermore, Instagram was the 5th SM platform most used by users aged 16 to 64 and verified the highest advertising reach compared to total population aged more than 13 (42%) (Hootsuite, 2020). In January 2020, 54% of Portuguese Instagram users were people aged 18 to 34. And, in 2021, this proportion was 50% (NapoleonCat Stats, 2021) indicating that half users are higher than 35 years, the target with higher propensity to buy an online newspaper subscription (Goyanes, 2020).

Instagram revealed a strong potential for brand awareness by spreading trustful journalism in a visual format (Vraga et al., 2020) that reduces readers' uncertainty, especially in a crisis context (Overgaard, 2021).

Hence, the main goal in the digitalization and platformization of Público journalism was to raise readers engagement along the reader funnel (Kotler et al., 2016) by distributing the right content at the right time at the right format.

3.3.3. Big data and analytics-driven Journalism

Editorial analytics represent an important element in news organizations (Moyo et al., 2019) to understand the media environment and to help audience-oriented editors to optimize decision making process (Hendrickx et al., 2021; Lamot & Paulussen, 2020). In fact, six main uses of analytics are identified by (Lamot & Paulussen, 2020): story placement, story packaging, story planning, story imitation, performance evaluation and audience conception.

Moreover, SMEs are intermediaries between audiences, newsrooms, and marketing departments (Neilson & Gibson, 2022). Thus, to maximize content mobile-first distribution, SMEs are allowed to observe internal and external data sources from historical to real-time data (Moyo et al., 2019) (see Table 8). Analytic tools provide web analytics metrics to measure website performance and the overall stories engagement (Moyo et al., 2019). SMEs can identify missed opportunities and improve their strategy in future decisions. Furthermore, SMEs have access to real-time metrics that help to decide the content to be promoted by platform according to content type and level of importance (Lamot & Paulussen, 2020; Neuberger et al., 2019).

To measure readers engagement, most authors use attention-based metrics such as, average time per session or viewport time (i.e., time spent at each part of the article) (Lagun & Lalmas, 2016). Furthermore, the literature seems to agree that reader retention is related to how often a user reads (regularity) (Davoudi, 2018). Thus, in this study, the number of sessions per user and pageviews per session were analysed (see Figure 21). Moreover, inspired on the successful business case of Financial Times (FTStrategies, 2022), the multi-dimensional RFV (recency, frequency and volume) engagement metric was implemented at Público in September 2020. The widely applied RFM (recency, frequency and monetary) model (Ernawati et al., 2021) was adapted to the media business context by replacing the monetary dimension by the number of articles visited, i.e., the Volume of content consumed. The RFV scoring process, for quantifying reader behaviour, uses the quintile method (Ernawati et al., 2021). Thus, a RFV value that goes from 111 to 555 is assigned to each reader. A total of 125 combinations (5 x 5 x 5) that divide the sample into 125 equal clusters. The readers who have the highest RFV scores are the most engaged, corresponding to readers that visited most recently the website (lowest recency), made more visits (highest frequency) and read many articles (highest volume) for 90 days.

Table 8. Data sources to optimize the decision-making process at Público.

Data Sources		**Historical Data**	**Real-time Data**
Internal	Website	Web analytics dashboards for general evaluations and comparisons on longer-term basis	Editorial Intelligence tool for real-time observational data about readers and website intra-day management
		Reader's profile, i.e., sociodemographic data, content consumption by day of the week, hour, author or tag.	
	Registered and Subscribers	Main registration or subscription articles, tags and authors	
		Register or Subscriber's profile	
External	Social Networks Competitors	Content discovery and social monitoring platform	

In order to measure users SM engagement in the platform (Manosevitch & Tenenboim, 2017), researchers' study two types of metrics (Li & Xie, 2020). One related to the direct response, or interactions, with the posts, such as number of likes or comments (Aydin, 2020). The second related to share and propagation of the published content that allows the user to recommend content to their followers. In this study, we will focus on interactions, such as number of likes or comments and**,** number of video visualizations.

As Instagram is a photo-sharing SM platform (Li & Xie, 2020), we aim to examine how posts' characteristics (e.g., content, interactivity, vividness) (Aydin, 2020) impacted user engagement and the relation between posts engagement and website traffic from Instagram users.

3.4. Methodology

To answer the previously raised RQ's (see Figure 20), our study consisted of five distinct steps: editorial strategy understanding, data gathering and preparation, results achieved through data analysis, results discussion and, finally, limitations and conclusions.

3.4.1. Editorial Strategy

In the daily routine, SMEs monitor audience engagement by measuring audience attentiveness and interactivity to continuously improve and create more engaging and relevant posts. A dynamic work of a multidisciplinary team composed of design, distribution, and analytics' elements that invested in testing innovative and creative approaches to engage Instagram users according to dashboards information and Key Performance Indicators (KPIs) results. Público analytics manager made available a central dashboards' page (see Figure 17) to allow SMEs to easily find the main information to drive the decision-making process (real-time and historical data, respectively shown in a blue and black colour). Dashboards were built into Google Data Studio (Kemp & White, 2020). The team analyses the content produced by journalists and decides which stories can be promoted on Instagram. SMEs do not influence the content production.

The current report presents results achieved by an iterative decision-making process where SMEs follow the next levels of priorities:

Firstly, to take real-time decisions SMEs combine information from the real-time analytics platform and the social monitoring platform to monitor articles performance across channels, devices or user types, as well as, to monitor and to discover misinformation spread in SM that can be clarified with Público content. The goal is to optimize content distribution by channel along with inform and clarify the misinformation that is over-performing in SM. As an example, the content presented at Figure 18 shows a carousel post and a story that clarifies readers about the hypothetic pandemic end.

Secondly, by identifying the content produced by the newsroom that needs to be posted on Instagram (i.e., content ideation), the team define the design approach (content creation) and select the right time for publishing. Furthermore, historical data presented in internal dashboards allows SMEs to analyse previous decisions, to decide the content to schedule and to promote on Instagram according to followers' profile and previous content performance. Dashboards combine information from social website traffic, registered and subscriber's website behaviour and conversion touchpoints. As an example, evergreen content, i.e., content that is consistently of readers interest (Y. Liao et al., 2019) is presented at the historical dashboard. By cross checking the maters that had good performance by hour or day of the week from Instagram, and the evergreen content for this target, SME decides the content to schedule on Instagram. As an example, at Figure 19 two types of content are presented, one related to the Alzheimer disease and the second about the bee's world. Both are multimedia formats produced by the newsroom that SME's posted because usually Instagram followers engage with this formats and content.

Finally, entertainment content, such as pictures or videos, are scheduled according to the best days and hours for this format. The team also focus on write informative text, in a constant tone (Denisova, 2022), and carefully select hashtags. Moreover, the team meet frequently with design team to improve content ideation, creation and planning always focused on readers' profile (i.e., reader's age, preferred tags, etc). For SMEs, it becomes clear the audiences to which they want to focus on and the goals to achieve.

Figure 17 Main data sources that support journalist decision-making process: real-time and historical data from website and social networks.

Figure 18. Example of an informative quiz, posted in a story and in a carousel, about group immunity and the possible pandemic end, posted at 19th March 2021.

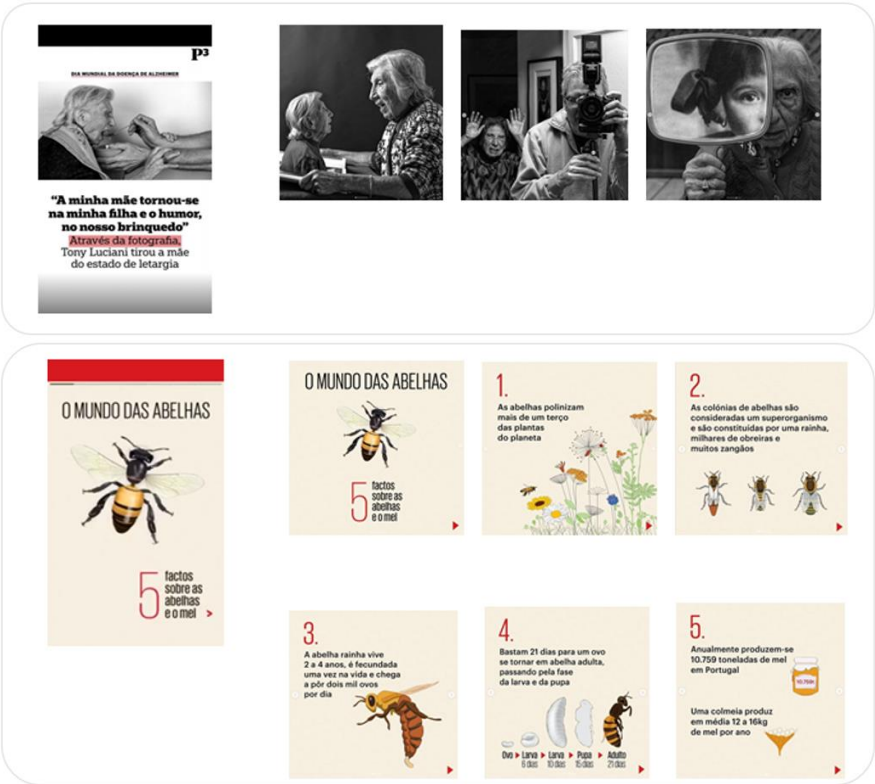


Figure 19. Example of content about Alzheimer disease posted at 21th September 2020 (a) and the "bee's world" posted at 6th June 2020 (b).

3.4.2. Data gathering and preparation

In order to analyse user's engagement and conversion, Público kindly provided the data that was collected from four main sources which are: website data from Google Analytics (Semerádová & Weinlich, 2020), internal data from BigQuery, a Google-managed data warehouse (Bisong, 2019), Instagram external data from InstLoadGram (Zimnitskiy, 2021), and Covid data from the *Our World Data* platform (Global Change Data Lab, 2021) (see Figure

21. (a) Monthly posts by post category (b) Website number of users from IG, sessions and respectively average RFV, plus Instagram followers and interactions during the timespan under study.Figure 20).

To analyse Instagram posts effectiveness, all 2,202 posts published at Público Instagram account from January 2020 to September 2021 were scrapped in November 2021 (see Table 9Table 9). Content shared may be in a variety of formats, such as, text, static or dynamic visuals, or a combination of all these (Aydin, 2020). Furthermore, each post can be published in a carousel horizontal format, i.e., by using multiple photos or videos that can be viewed by swiping or clicking left. Thus, inspired on the study by Li and Xie (2020) posts were manually categorized into seven types: only picture, picture with text, text and/or infographic, cartoon, quote, newspaper front-page and video (see Table 9).

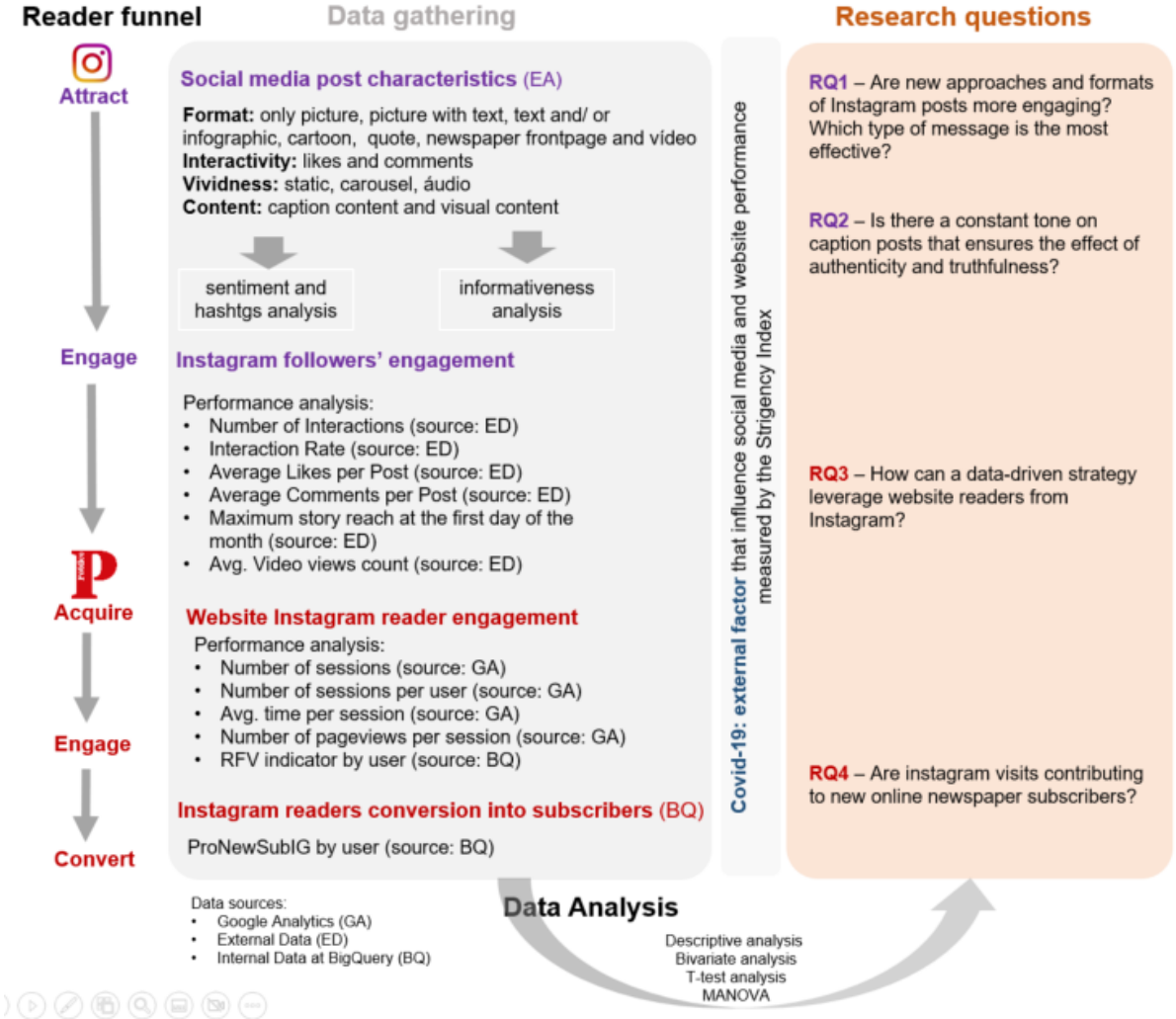


Figure 20. Research framework: variables under study across the reader funnel and RQ's by funnel level.

Prior research suggested that there are several factors affecting the popularity of posts and subsequent engagement, such as, the ease use of the technology, user motives and values, psychological factors and post characteristics (Aydin, 2020). Thus, the following post characteristics were collected (see Figure 20): Post' date, content (caption, hashtags), interactivity (total number of likes, total number of comments, and total number of views by post), vividness (is carousel, is video, has audio), and video duration in seconds. Then, the number of hashtags per post and number of words per caption was calculated. From Instagram we also saved user profile characteristics such as age, gender, and city, as well as competitor's public data such as number of followers.

Among the KPIs to analyse website user's engagement (see Figure 20), we highlight, the time per session, typically expressed in seconds, an important KPI for publishers (Hendrickx et al., 2021) that measures user attention. Furthermore, at the individual level, we monitor the multi-dimensional RFV, an indicator of reader's behaviour, implemented in September 2020. As well as the number of new subscribers that made at least one visit from Instagram during the subscription month. That allow us to calculate the ProNewSubIG that is the proportion of new subscribers with website visits from Instagram. These indicates how Instagram readers are converting into subscribers.

Besides, we added an external variable, the stringency index (see Figure 20) that "is a composite measure based on nine response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest)" (Global Change Data Lab, 2021). After the initial data collection, we perform a descriptive analysis presented as follows.

3.5. Results

3.5.1. Message type and format Analysis

In this section we will focus on post's characteristics by studying post's interactivity, format, vividness and content by post type. Along with, website user's engagement and conversion metrics. Message strategy and engagement effectiveness were examined. Microsoft Excel and the R statistical tool (Cortez, 2021) was chosen to conduct the analysis.

On the sample (see Table 9), less than a half of posts were Static (948), most of which was in a form of a simple picture (47%) or a picture with text (42%). From all the content shared, 44% were carousels, 12% were videos with audios, and only 1% were videos without sound. In terms of post format, the most frequent (54%) was only picture with 740 carousel posts and 449 static.

Table 9. Post statistics by format considering vividness levels and interactivity metrics.

				Format								
				Only Picture	Picture with Text	Text Infographic	Cartoon	Quote	Newspper Frontpage	Vídeo		Total
										No audio	Has audio	
			N	1198	556	93	61	11	15	19	259	**2,202**
Vividness	**Period**	**Interactivity**	%	54%	25%	4%	3%	0%	1%	1%	12%	**100%**
Static	2020		N	347	108	9	36	3	8			**511.0**
		Likes	Avg.	2,539.7	3,627.7	3,715.2	1,874.0	872.0	1,206.8			**2,712.8**
			StDev.	2,075.8	3,469.1	3,592.1	1,443.3	105.4	735.1			**2,474.2**
		Comments	Avg.	49.6	102.9	122.7	56.1	45.0	13.9			**62.0**
			StDev.	55.9	109.1	119.1	155.0	30.5	17.5			**84.0**
	Jan-Sept 2021		N	102	291	19	21	1	3			**437**
		Likes	Avg.	5,678.8	7,018.2	**11,789.6**	3,340.9	1,143.0	2,528.0			**6,692.0**
			StDev.	6,410.1	5,326.6	11,228.9	2,821.6		1,737.2			**6,012.4**
		Comments	Avg.	117.4	215.4	256.8	97.0	37.0	25.7			**186.9**
			StDev.	223.9	235.7	183.3	245.6		4.0			**235.0**
	Total		N	449	399	28	57	4	11			**948**
			%	47%	42%	3%	6%	0%	1%			**43%**
		Likes	Avg.	3,252.8	6,100.5	9,194.3	2,414.4	939.8	1,567.1			**4,547.1**
			StDev.	3,784.8	5,117.0	10,130.6	2,157.5	160.5	1,167.4			**4,886.5**
		Comments	Avg.	65.0	185.0	213.7	71.2	43.0	17.1			**119.6**
			StDev.	120.5	214.9	175.1	192.2	25.2	15.8			**182.0**
Carousel	2020		N	512	15	32	4	1	4			**568**
		Likes	Avg.	2,457.5	**2,292.1**	**2,354.4**	4,670.3	2,204.0	2,326.3			**2,461.6**
			StDev.	1,831.8	2,111.3	2,470.8	6,156.9		1,095.7			**1,926.4**
		Comments	Avg.	33,2	46.8	34.3	28.3	50.0	28.8			**33.6**
			StDev.	49,4	63.2	38.4	39.3		28.2			**49.0**
	Jan-Sept 2021		N	228	141	33		6				**408**
		Likes	Avg.	5,822.8	**8,625.2**	**6,809.2**		10,815.5				**6,944.5**
			StDev.	6,608.3	7,315.6	7,787.6		4,772.0				**7,051.2**
		Comments	Avg.	72.1	183.1	138.9		377.8				**120.4**
			StDev.	124.7	255.3	224.0		177.6				**197.7**
	Total		N	740	156	65	4	7	4			**976**
			%	76%	16%	7%	0%	1%	0%			**44%**
		Likes	Avg.	**3,495.4**	**8,016.2**	4,616.0	**4,670.3**	**9,585.3**	**2,326.3**			**4,335.6**
			StDev.	4,260.5	7,228.4	6,190.1	6,156.9	5,437.9	1,095.7			**5,273.2**
		Comments	Avg.	45,2	170.0	87.4	28.3	331.0	28.8			**69.8**
			StDev.	82,4	246.7	169.0	39.3	204.1	28.2			**139.8**
Audio	2020		N							14	158	**158**
		Likes	Avg.							869.9	1,351.8	**1,351.8**
			StDev.							852.8	1,439.1	**1,439.1**
		Comments	Avg.							8.9	39.2	**39.2**
			StDev.							9.8	55.2	**55.2**
		Count views	Avg.							16,842	23,969	**23,969**
			StDev.							12,171	18,164	**18,164**
	Jan-Sept 2021		N							5	101	**101**
		Likes	Avg.							648.6	2,281.7	**2,281.7**
			StDev.							318.7	3,224.2	**3,224.2**
		Comments	Avg.							10.0	76.8	**76.8**
			StDev.							9.3	155.4	**155.4**
		Count views	Avg.							14,351	39,095	**39,095**
			StDev.							6,998	39,584	**39,584**
	Total		N							19	259	**278**
		Likes	Avg.							811.6	1,714.4	**1,652.7**
			StDev.							746.9	2,344.4	**2,282.0**
		Comments	Avg.							9.2	53.9	**50.8**
			StDev.							9.4	107.5	**104.4**
		Count views	Avg.							16,187	29,867	**28.932.5**
			StDev.							10,915	29,372	**28,692.4**
Total		N		1,189	555	93	61	11	15	**278**		**2,202**
		%		54%	25%	4%	3%	0%	1%	**13%**		**100%**
		Likes	Avg.	3,383.6	6,625.3	5,994.4	2,562.4	6,441.5	1,769.5	**1,652.7**		**4,087.9**
			StDev.	4,078.8	5,848.4	7,825.2	2,560.7	6,064.3	1,162.5	**2,282.0**		**4,910.6**
		Comments	Avg.	52.4	180.2	125.4	68.4	226.3	20.2	**50.8**		**88.9**
			StDev.	98.6	224.2	179.6	186.2	215.2	19.4	**104.4**		**158.2**

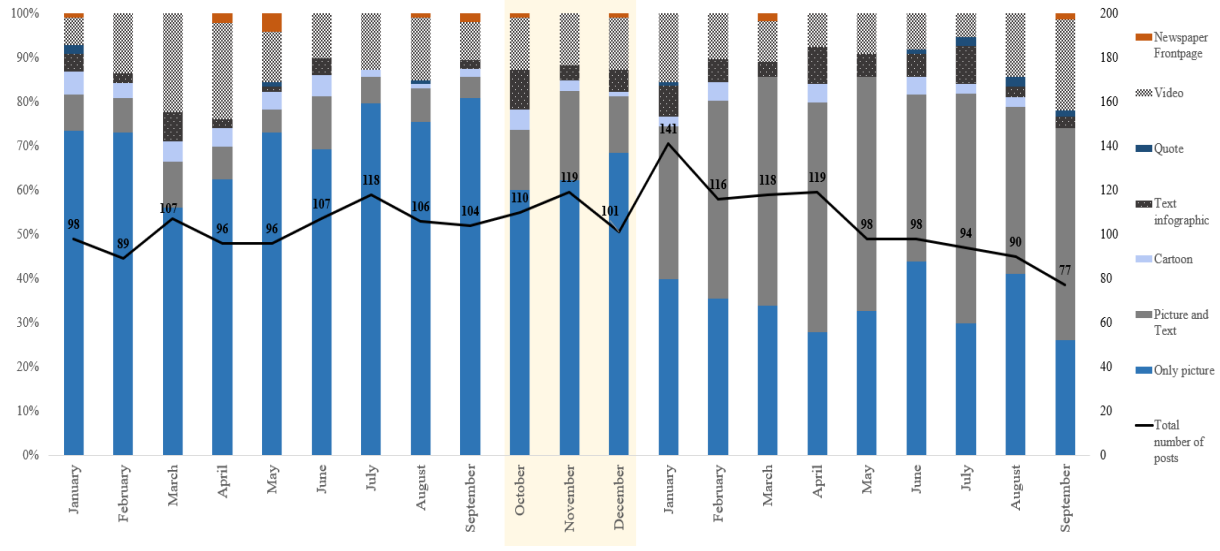
Despite the pandemic reality, at the first semester of 2020, Instagram engagement KPIs flatted (see Figure 21 b). Thus, in the last trimester 2020 (yellow box at Figure 21), a reader-centred approach was tested to address content readers' needs and preferences according to dashboards' information, coupled with design improvements. Depending on the content presentation, the level of interactivity differed. Thus, by experiments, the post's performance was continuously analysed to improve the next decision-making. As presented at Figure 21 a, more pictures with text and text infographics were posted. This resulted in a decrease of only picture posts (from 347 to 102, between 2020 and 2021, see Table 9).

In terms of interactivity, among the posts analysed, the mean number of likes was 4,088 with a standard deviation of 4,910 (see Table 9) as result of different types of post content and vividness. Picture with text posts verified highest average number of likes (6,625.3), followed by posts with author quotes (6,441). Interestingly, the increase on average likes between 2020 and 2021 on pictures with text in carousel format (from 2,292.1 to 8,625.2), as well as text and/or infographic (from 2,354 to 6,809) reflecting the importance of explaining the content in a visual and dynamic way. This finding is consisted with the work of (Midberry & Dahmen, 2020) that suggests that solutions-oriented photojournalism leads higher user engagement (Dahmen et al., 2021).

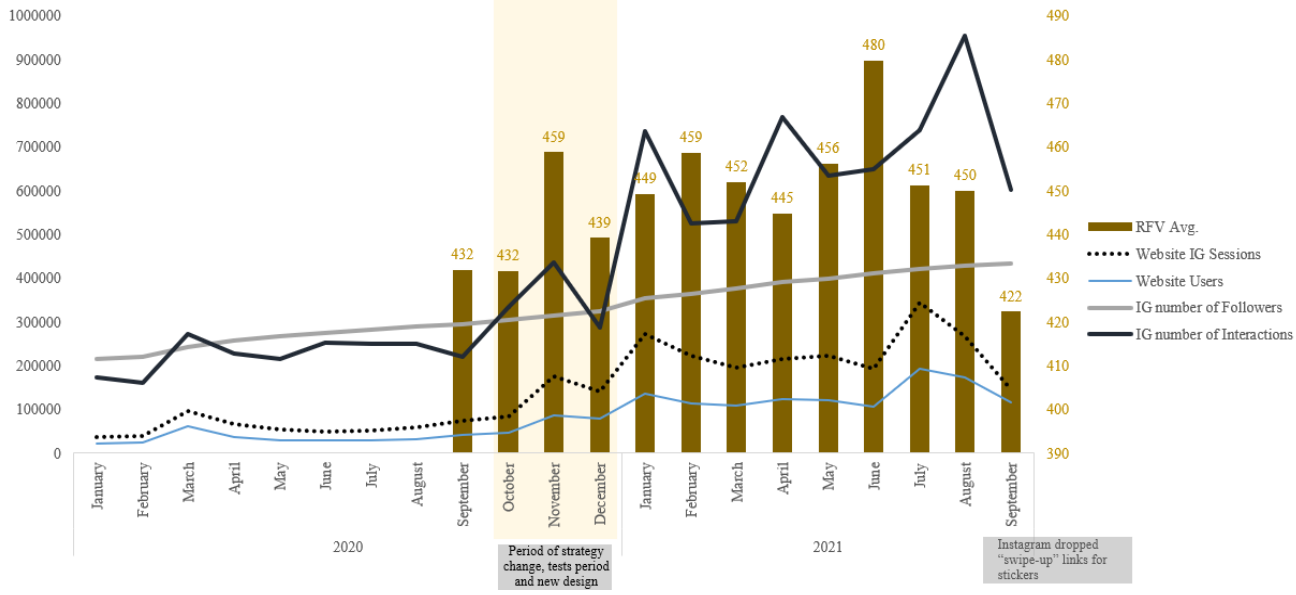
Unsurprising, carousel posts revealed high average number of likes by formats (see Table 9), except on infographic. Researchers argue that vividness features in posts lead to higher total interaction (Aydin, 2020), in our sample, static text infographic posts verified the higher average likes on 2021 (11,789.6), because those posts were related to pandemic news and closures clarifications.

In what concerns to comments, the average number of comments is higher on quotes (226.3), followed by pictures with text (180.2), proving that visual content with text promotes user reaction (Dahmen et al., 2021). Furthermore, posts with author quotes on carousel present an average of 331 comments, revealing higher reader involvement in a narrative content those results are consist with the work of (Dahmen et al., 2021).

To empirically examine the influence of type of image content on reader engagement, a t-test was applied to compare the average likes by year (see Table 10). There is a statistically significant difference on the followed formats: only picture, picture with text, cartoon, text infographic and quote, proving the positive impact on interactivity as result of the strategy followed.



(a)



(b)

Figure 21. (a) Monthly posts by post category (b) Website number of users from IG, sessions and respectively average RFV, plus Instagram followers and interactions during the timespan under study.

In order to analyse the effects of message type on the number of likes and comments. A MANOVA was conducted by using the post format as the independent variable. The overall MANOVA tests of Pillai's is significant ($p < 0.0001$), suggesting that the number of likes and comments varied across post format, supporting the hypothesis that different post formats generate different post effectiveness.

Table 10. t-Test: Two-Sample Assuming Unequal Variances.

t-Test: Two-Sample Assuming Unequal Variances	Only picture	Picture and Text	Cartoon	Text infographic	Video	Quote	Newspaper Frontpage
t Stat	-3,636	-6,356	-2,749	-4,924	-1,875	-3,128	-1,105
P(T<=t) two-tail	0,007	0,000	0,023	0,001	0,094	0,035	0,468
t Critical two-tail	2,306	2,145	2,262	2,262	2,262	2,776	12,706

Notes: Decision rule if $t\ Stat < -t\ Critical\ two-tail$ or if $t\ Stat > t\ Critical\ two-tail$ reject the null hypothesis

3.5.2. Text Analysis

In terms of text content features, we consider the influence of caption content and visual content. Therefore, on the caption, we study text sentiment and text dimension (number of words and number of hashtags), on the image, we analyse informativeness, i.e., if the image was edited by adding text or not.

The sentiment score was calculated through the sentimentR package using the open-source R statistical tool (Cortez, 2021). SentimentR was designed to calculate text polarity sentiment (TPS) at the sentence level and optionally aggregate by rows or grouping variable(s) (True, 2018). For corpus handling and text preprocessing we used the tm (Feinerer, 2019), the NLP (Hornik & Hornik, 2018), and the qdap packages (Jovi et al., 2015). Preprocessing involved punctuation, numbers, white spaces removal, as well as Portuguese stopwords such as “de” or “a”.

Descriptive statistics indicate that captions tend to be neutral ($TPS_{median} = 0.000$ and $TPS_{mean} = -0.0036$), ensuring a constant tone on text posts (RQ2) communication. The maximum TPS is 0.5897 related to a post about Marthin Luther King published on 16th January 2020, and the minimum TPS (-0.8215) is related to a cartoon about Donald Trump and democracy published on 11th January 2021.

Furthermore, by calculating a term frequency-inverse matrix, the words with more relative importance among captions emerge. As expected, on the top 10, we found *Portugal, Lisboa, covid, pandemia, mundo (world),* and *coronavirus*. Also, we note that words like *animais (animals), pessoas (people), ambiente (environment) and natureza (nature)* are also frequent revealing the diversity of content published. Moreover, text dimension was studied by calculating the number of words and hashtags. Those variables were grouped into 6 and 4 categories, respectively. The number of post words range between 6 and 284 words, with an average of 80. The number of hashtags vary between 0 and 20, with an average of 7 by post.

Among the 4,149 hashtags, the authors also examine hashtags frequency and the engagement KPIs registered on the respective hashtag's posts. On the most used hashtags, we found #covid (644), #portugal (346), #pandemia (300), #coronavirus (177), #pandemic (160) and #animais (153). Between the hashtags with frequency higher than 50 we found #ambiente, #natureza, #ciencia, or #sociedade. However, the most used hashtags are not necessarily related to the hashtags from posts with high interactivity. Therefore, we calculate an engagement score that sums the normalized average number of likes per hashtag and average number of comments per hashtag, given to each one the weight of 50%. A hashtag with high engagement score value indicates that, among the sample, this hashtag was used on posts that registered the highest average number of likes and average number of comments.

$$\text{engagement Instagram score} = \frac{X_{likes} - \bar{x}_{likes}}{s_{likes}} 0.5 + \frac{X_{comments} - \bar{x}_{comments}}{s_{comments}} 0.5 \quad (1)$$

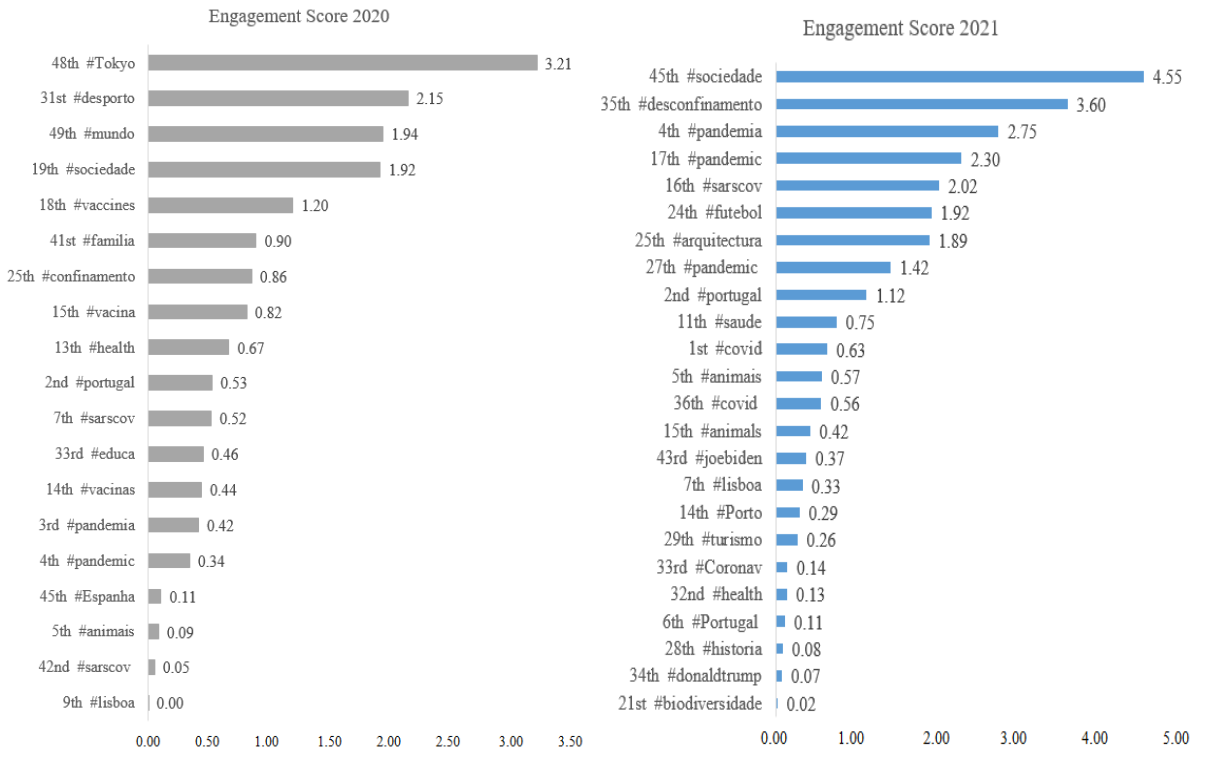


Figure 22. Hashtags with positive engagement score by year and respective rank position. For example, in 2020 the hashtag #portugal was the second hashtags most user with a score of 0.53.

For instance, #Tokyo presents the highest engagement score (3.21) in 2020, despite to be the 48th hashtag most used. Whilst, in 2021, #sociedade presented the highest score (4.55) and it was in the 45th place of the rank (see Figure 22). Furthermore, the year 2021 presents more hashtags with positive score (24 versus 19 on 2020) indicating more engagement in a wide variety of matters, such as, architecture, Joe Biden, tourism, history, or biodiversity.

In both years, sport posts conquer higher levels of engagement (#tokyo, #desporto, #futebol), as well as animals posts (#animais). As expected, hashtags related to the pandemic or related to the Portuguese actuality (#sociedade) are on the top of the most engaging posts and are also the most used. Interestingly, the hashtag #covid is the first one more written on both years, however, in 2020 the average number of likes and comments by post was significantly lower than the other hashtags, thus the engagement score was less than zero. That does not happen in 2021, indicating that the decisions made to improve content presentation on Instagram as well as to clarify pandemic information resulted in more user involvement.

Table 11. Posts' Instagram engagement metrics by caption number of words and number of hashtags.

		Post with text on the image (PwTextI) (more informativeness)				**Post without text on the image**				**Post with text versus without text**
Caption Nº of words		2020	2021	Total	YoY	2020	2021	Total	YoY	Total var.
[5,50]	**#**	16	15	31	-6%	174	68	242	-61%	-87%
	Avg. likes per post	2581	13042	**7643**	405%	2079	4177	2669	101%	186%
	Avg. Comments per post	76	264	167	246%	52	81	60	56%	179%
[51,100]	**#**	125	218	**343**	74%	674	190	**864**	-72%	-60%
	Avg. likes per post	3294	7821	**6172**	137%	2238	5278	2906	136%	112%
	Avg. Comments per post	86	191	153	123%	39	104	53	166%	188%
[101,150]	**#**	23	237	**260**	**930%**	228	186	**414**	-18%	-37%
	Avg. likes per post	3470	7023	**6708**	102%	2584	4840	3598	87%	86%
	Avg. Comments per post	87	202	192	131%	33	77	53	133%	261%
[151,200]	**#**	1	13	14	1200%	5	18	23	260%	-39%
	Avg. likes per post	2887	7838	**7485**	172%	1836	4992	4306	172%	74%
	Avg. Comments per post	68	247	234	263%	25	52	46	107%	407%
[201-250]	**#**		1	1		3	2	5	-33%	-80%
	Avg. likes per post		39078	**39078**		683	1329	941	95%	4051%
	Avg. Comments per post		1282	1282		6	14	9	133%	13835%
[251-350]	**#**		1	1		2	2	4	0%	-75%
	Avg. likes per post		1808	1808		1924	2049	1987	6%	-9%
	Avg. Comments per post		28	28		57	43	50	-25%	-43%
Caption Nº of hashtags										
[0,4]	**#**	33	91	124	176%	199	134	333	-33%	-63%
	Avg. likes per post	3461	8362	**7058**	142%	1997	3581	2634	79%	168%
	Avg. Comments per post	100	260	**217**	160%	36	80	54	123%	305%
[5,9]	**#**	107	360	**467**	236%	673	296	**969**	-56%	-52%
	Avg. likes per post	3243	7710	6687	138%	2366	5350	3277	126%	104%
	Avg. Comments per post	84	194	169	130%	42	88	56	109%	201%
[10,14]	**#**	25	34	59	36%	196	33	229	-83%	-74%
	Avg. likes per post	2982	5037	4166	69%	2288	5899	2808	158%	48%
	Avg. Comments per post	69	135	107	94%	36	111	47	211%	129%
[15,21]	**#**					18	3	21	-83%	-100%
	Avg. likes per post					2028	8510	2954	320%	-100%
	Avg. Comments per post					35	69	40	97%	**-100%**
Total	**#**	**165**	**485**	**650**	**194%**	**1086**	**466**	**1552**	-57%	**-58%**
	Avg. likes per post	**3247**	**7645**	**6529**	**135%**	**2278**	**4900**	**3066**	115%	**113%**
	Avg. Comments per post	**85**	**202**	**172**	**138%**	**40**	**87**	**54**	120%	**220%**

As presented at Table 11, the average number of comments on posts with edited text on the image (PwTextI) is three times higher than without text, and the average number of likes is more than double indicating that the presence of more information induces more user's reactions. Interestingly, PwTextI with edited text on the image increased (194%), especially on posts with more than 101 words and less than 150 words (930% YoY).

Furthermore, by comparing the average likes, by groups of caption number of words, posts without text increase the number of likes as the number of words increase in the caption (from 2,669 to 4,306). While PwTextI present values between 6,172 and 7,643 likes. In terms of hashtags, 467 of the 650 PwTextI present between 5 and 9 hashtags. However, posts with more likes and comments present less than 5 hashtags, indicating that the user reaction is more related to the information on the image than the number of hashtags.

Table 12. t-Test: Two-Sample Assuming Unequal Variances.

Variables under study	**Statistics**	**2020**	**2021**	**t Stat**	**P(T ≤ t) two-tail**	**t Critical two-tail**
Sessions	Avg.	34,922.33	167,715.44	-6.815	0.000	2.201
	StDev	26,668.02	53,700.67			
Visits/User	Avg.	1.76	1.75	0.048	0.963	2.201
	StDev	0.11	0.23			
Avg. Time Session (sec)	Avg.	122.50	108.89	1.691	0.115	2.160
	StDev	13.84	20.96			
PV/Session	Avg.	3.53	2.30	4.586	0.000	2.093
	StDev	0.74	0.48			
Interactions #	Avg.	257,175.42	682,528.33	-8.569	0.000	2.179
	StDev	73,541.59	134,600.04			
Interaction Rate	Avg.	0.89	1.66	-5.871	0.000	2.262
	StDev	0.12	0.38			
Avg. Likes per post	Avg.	2,099.87	6,001.57	-5.681	0.000	2.262
	StDev	622.97	1 988.49			
Avg. Comments per post	Avg.	47.74	150.72	-4.353	0.002	2.306
	StDev	14.22	69.90			
Maximum Stories Reach at first day of the month	Avg.	11,687.58	44,534.89	-3.424	0.009	2.306
	StDev	4,916.43	28,463.62			
ProNewSubIG	Avg.	0.01	0.03	-4.147	0.003	2.306
	StDev	0.00	0.02			

Notes: Decision rule if $t\ Stat < -t\ Critical\ two-tail$ or if $t\ Stat > t\ Critical\ two-tail$ reject the null hypothesis

3.5.3. Engagement and Conversion between Periods

To examine the impact of the new strategy followed on 2021, Table 12 presents the t-test on the metrics between 2020 and 2021. There is a statistically significant difference between the average number of sessions by month between periods, because of the sessions increase over 17% monthly in 2021. The monthly average of number of interactions, interaction rate, likes per post, comments per post and stories reach also presented a statistically difference between both years.

Furthermore, as presented at Figure 21 b, since September 2020, there was an increase of followers (3% monthly on average) and the average number of monthly website visits more than doubled.

The monthly number of interactions registered an increase over than 265% between 2020 and 2021. Moreover, we also found a significant increase on the proportion of new subscribers that came to the website from Instagram (3% in 2021). Besides, (see Figure 21) the monthly number of posts, the number of words and hashtags by post, on both years, was kept around 104 posts, 88 words and 7 hashtags.

Meanwhile, the variables Visits per User and Avg. Time Session do not present significant difference between years indicating that, despite the peak on sessions these users tend to keep the levels of engagement in the website visit. In fact, the RFV indicator since September 2020 presented an average value of 448 (see Figure 21 b) with a standard deviation of 14.7. Thus, the results suggested that the strategy followed attracted engaged audience.

3.5.4. Correlation analysis

According to (Kotler et al., 2016) across the five stages of the customer funnel, the goal is to convert the maximum number of anonymous users into high loyal customers. Low levels of conversion rates result in low customer attraction, low curiosity, low commitment and low affinity (Kotler et al., 2016). Thus, this section aims to analyse the relationship between variables under study. The correlation coefficient is a statistic that measures the strength of the relationship between two variables (Taylor, 1990). The Pearson correlation coefficient describes the proportion of the total variance in the observed data that can be explained by a linear model of the variables under study (Bermudez-Edo et al., 2018). It takes values between -1 and 1 and it is computed as:

$$\mathrm{corr}_{xy} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \qquad (1)$$

A positive correlation indicates a direct relationship between variables, i.e., when the first variable increases the second one will increase. However, to interpret those values labelling systems coupled with the statistical significance of the correlation coefficient allow to define the kind of relationship that exists between variables (Taylor, 1990). The p value indicates the probability that the data would be inconsistent with the hypothesis. Thus, for a confidence level of 95% when p value ≤ 0.05 the correlation is significant. When to variables are independent, the Pearson's correlation coefficient is 0 (Bermudez-Edo et al., 2018).

As presented at Table 13 the correlation matrix suggested that the variables – Sessions, Interactions, Interaction Rate, Average likes per post, Average comments per post, ProNewSubIG – were significantly positive correlated.

Furthermore, the number of followers increased along with the number of interactions, thus the interaction rate is significantly correlated with all the variables except Stringency Index. Therefore, across the conversion funnel as the number of Instagram followers increase the number of readers on the website increase and consequently the subscription conversions increase.

Table 13. Pearson correlation between variables

	KPI's	SpU	TpS	PpS	I	IR	LpP	CpP	Max. Story	PropNewSubIG	SI
Website KPIs	Sessions	0.099	-0.356	-0.730 (**)	**0.869 (**)**	0.798 (**)	**0.822 (**)**	**0.807 (**)**	0.464 (*)	0.727 (**)	0.269
	Session per User (SpU)	1	0.417	-0.028	0.000	-0.227	-0.208	-0.168	-0.520 (*)	-0.281	0.470 (*)
	Avg. Time per Session (TpS)		1	0.269	-0.449 (*)	-0.612 (**)	-0.575 (**)	-0.542 (*)	-0.391	-0.720 (**)	0.338
	Pageview per Session (PpS)			1	-0.747 (**)	-0.721 (**)	-0.752 (**)	-0.726 (**)	-0.527 (*)	-0.691 (**)	-0.416
Instagram KPIs	Interactions (I)				1	**0.949 (**)**	**0.945 (**)**	**0.863 (**)**	0.690 (*)	**0.879 (**)**	0.182
	Interaction Rate (IR)					1	**0.945 (**)**	**0.863 (**)**	0.690 (*)	**0.879 (**)**	0.18
	Avg. likes per post (LpP)						1	**0.948 (**)**	0.690 (*)	**0.928 (**)**	0.193
	Avg. Comments per post (Cp)							1	0.625 (*)	**0.904 (**)**	0.169
	Max. Story								1	0.614 (**)	0.132
	PropNewSubIG									1	0.111
	Stringency Index (SI)										1

Notes: (*) correlation is significant at the level 0.05 (**) correlation is significant at the level 0.01 (two-tailed)

3.5.5. Competitors analysis

To analyse Público market position, Table 14 presents a comparison between market players. In January 2020 Público already presented more followers than competitors. Through the strategy followed, Público increased 102% the number of followers and 251% the number of monthly interactions leading the Portuguese ranking in this platform.

Table 14. Follower's and monthly interactions comparison between Público and its competitors.

	January 2020		Comparison between September 2021 and January 2020	
Portuguese Publishers	Público number of followers versus competitors	Monthly interactions	Number of followers	Monthly Interactions
Público	----	**171,686**	**102%**	**251%**
Expresso	49%	72,458	83%	105%
JN	41%	95,822	73%	-39%
Observador	58%	57,054	65%	37%
Correio da Manhã	99%	21,287	44%	-74%
Diário de Notícias	287%	6,231	51%	949%

3.6. Discussion and research questions analysis

This study aims to answer four research questions (see Figure 20). Seeking to answer RQ1, previous analysis demonstrated that as the number of informative posts (i.e., PwTI) increases, the number of followers and interactions also increased. As argued by (Dahmen et al., 2021; Midberry & Dahmen, 2020) narrative engagement plays an important role involving audiences. Furthermore, the number of followers increased despite to keep the number of posts by month indicating the assertiveness in content selection. Moreover, the interaction rate and the maximum story reach at the first day of the month more than tripled. Also, the t-test showed significant positive difference between average likes and comments in both years. Therefore, to offer valuable information by introducing text on the visual format, attract reader's interest. These findings are consisted with the work of (Aydin, 2020).

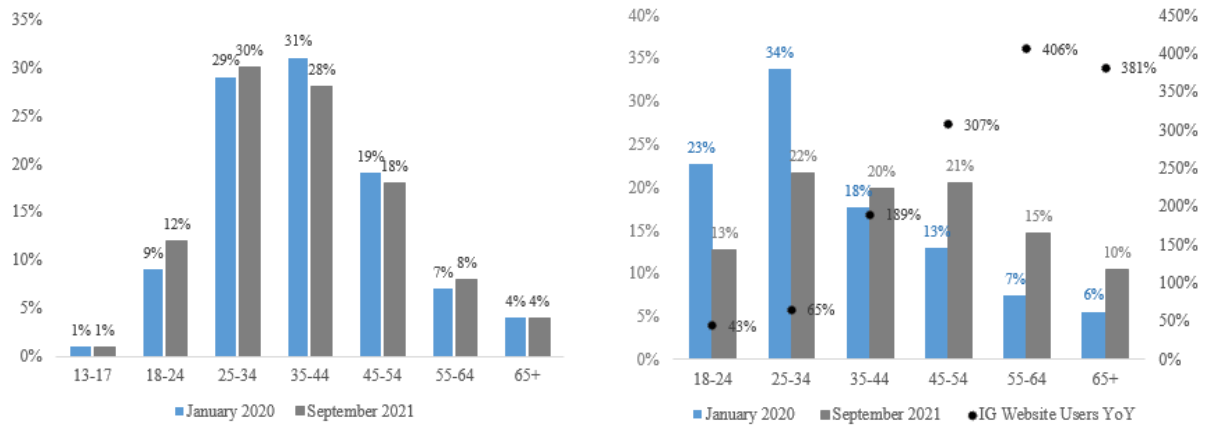
To answer RQ2 a text analysis was performed. The wide range of hashtags with high engagement score indicates that there was a daily understanding of the reader information needs. Furthermore, sentiment analysis demonstrated that the TPS tend to be neutral across the timespan. Thus, the engagement increased across the funnel always keeping the journalistic ethic perspective.

Authors defend that SM affects the way news is produced. However, Público SME's focus on selecting, from the content produced, the news that engage with Instagram readers (Cage et al., 2020) coupled with a need to fight against misinformation. Hence, the data-driven strategy guided the team to continuously identify the matters, produced by the newsroom, to post. Consequently, the number of sessions on the nine months of 2021 from Instagram readers surpassed the total visits of the previous year on 23%.

Seeking to investigate RQ3, a correlation analysis was performed. The low Pearson correlation between the Stringency Index and the number of sessions indicates that the increase on website sessions is not directly related to the strictness of lockdown. Thus, the website sessions increase was result of the efforts to optimize the SM strategy guided by data. Moreover, the strategic selection of matters, the careful choice of content formats (detailed at the Editorial Strategy understanding section) drove followers that kept engaged with posts, as the interaction rate increased and consequently drove readers to the website. Despite concerns about glorifying analytic tools (Moyo et al., 2019), the present research aims to provide an example on how a data-driven SM strategy can keep journalism principles and inform readers that consequently become subscribers.

On the other hand, in terms of acquiring readers with more propensity to subscribe (RQ4), an increase higher than 300% on readers aged 45 and more was verified, as result of a 156% growth on website sessions form Instagram. In January 2020, the majority (60%) Público Instagram followers fell in the age group 25-44 and 30% had more than 44 years (see Figure 23 a). When comparing January 2020 and September 2021, the distribution of the followers by age did not present significant differences. However, when we explore the users on the website that come from Instagram the distribution is quite different (see Figure 23 b), i.e., 46% of the readers had more than 44 years. Interestingly, the daily decision-making process increased the right target of readers to subscribe as the historical dashboards are prepared to understand content consumption by hour, day, age, etc. Besides, there is a strong positive correlation between PropNewSubIG, Sessions and Interactions, indicating that Público Instagram audience is reading more on the website and it is reading content that induces the subscription purchase.

Furthermore, Instagram readers on the website maintained the number of sessions per visit as the RFV engagement metric, revealing the assertiveness in content selection. Thus, and answering positively to RQ4, the proportion of new monthly subscribers that visit the website from Instagram increased, from 0.6% on January 2020 to 5.6% on August 2021. While some authors agree that SM contribution seems underwhelming (Ju et al., 2014), others agree that can boost website engagement (Manosevitch & Tenenboim, 2017). This research provides evidence that SME's can apply a subscription data-driven strategy in SM that increases reader engagement without affecting journalistic output.



(a) (b)

Figure 23. Instagram followers by age (a) and website users from Instagram by age (b).

3.7. Conclusions and limitations

The use of several channels in journalism reinforced the journalism move from a single channel to a multichannel communication strategy (Neuberger et al., 2019). While publishers handle their relations with digital intermediaries to not become too dependent (Nielsen & Ganter, 2018), little is known about the strategies followed and results achieved. Furthermore, studies focused on high reach channels, such as Facebook or Twitter (Aydin, 2020; Leung et al., 2017; Li & Xie, 2020), researchers have almost ignored Instagram (Vraga et al., 2020). Thus, the current research adds to recent efforts to analyse visual solutions journalism and understand its effects (Dahmen et al., 2021).

The empirical study analyses the impact of the strategy followed that focuses on readers' engagement. By attracting to the website users with the right content that present high RFV values that consequently subscribe. The findings demonstrate that users like and follow pages that provide them useful and timely information in a convenient way (Aydin, 2020). Furthermore, text analysis showed a neutral communication on posts proving that SME's avoided 'clickbait' strategies communication (Denisova, 2022). Moreover, the methodology of continuous test and results' analyses resulted in an increase of engagement metrics across de reader funnel.

Thus, this research shows how SME's deal with the digital challenge to deliver content in a SM platform, and how to continuously adapt to users' behaviour changes and Instagram updates.

Indeed, the data-driven strategy leverages data analytics in four main fronts: at Instagram, to determine optimal content and timing to post, since follower habits continuously change during the pandemic; at the website, to determine the best content to post and attract users to the website with high propensity to subscribe; at real-time decision-making to inform breaking news and update latest top stories, and finally, on misinformation monitor to determine the content that need to be clarified.

Despite this research identified several working tactics and successful patterns, the study was not immune to limitations. The biggest limitation is the generalizability. Indeed, this study only considered data from one Portuguese online newspaper between January 2020 and September 2021. Besides, other content information could be used to analyze post performance, such as colorfulness or the presence of human faces.

Furthermore, Instagram platform is always evolving by changing features and algorithms. Thus, it could be interesting analyse the evolution of the data-driven strategy to face Instagram changes and to keep growing engagement across the funnel.

Lastly, social media strategy is dynamic, must be an on-going task. The first step to implement a data-driven culture in the newsroom was well succeed. Results achieved proved the value delivered by data to help Público to conquer new readers and engage them. Thus the next step was to define a segmentation strategy to apply editorial and marketing actions, as presented in the following chapter.

CHAPTER 4

Segmenting online readers based on engagement features: a machine learning-based approach to increase digital revenue in a subscriptions business model

4.1. Research context

Digital subscription business models became popular in the last decade. To develop effective strategies, companies need to identify customers segments and the main drivers of purchase intention. Thus, this study aims to fell a literature gap by profiling online users to provide insights to support customer segmentation (as presented at Figure 1). Based on the interplay between several engagement' variables collected from a real-world online publisher, a clustering analysis is presented to identify online users' profiles. The combination of machine learning algorithms produced practical implications and business suggestions. Readers were segmented into eight clusters. The scroll down and the volume of premium content read are the most important drivers to subscribe. A twofold strategy is proposed to impact on acquisition, retention, and conversion of online readers across the audience funnel. Next sections present the third article mentioned at Table 2.

4.2. Article introduction

In the last decade, the decline of print advertising revenue, and the increase in digital subscription business models (DSBMs) changed the media business landscape (Arrese, 2016; Rios-Rodríguez et al., 2022). Across the media industry, DSBMs in the form of paywalls are documented in the academic literature mainly into three types: metered (consumption limited by a few articles), freemium (a selection of content is premium, i.e., only available for subscribers), and hard paywalls (all the content only accessible for subscribers) (Myllylahti, 2019). Furthermore, to improve user acquisition and engagement, to assure companies' sustainability, publishers need to build strong strategies (Davoudi, 2018). One way to personalize and optimize the product experience in DSBMs, it is to separate users into groups and expose each group to the product experience that most fits with the group's characteristics (Seufert, 2014).

Several studies argue that reader retention is related to how often a user reads (regularity) (Davoudi, 2018; Kim et al., 2021). Furthermore, researchers acknowledged the relationship between customer engagement and purchase intention in DSBMs (Clement Addo et al., 2021). Thus, marketing and editorial teams need to identify reader's segments across the conversion funnel to maximize reader's engagement (Lagun & Lalmas, 2016), retention and consequently conversion (Davoudi et al., 2018; Kotler et al., 2016; Villi & Picard, 2019).

Effective management and engagement analysis allows managers to understand reader behavior by providing insights to conduct data-driven strategies. Thus, based on the premise that the best predictor of future customer behavior is past customer behavior (Kamthania et al., 2018), the present research addresses the problem of identifying different reader profiles based on user behavior attributes collected from an online newspaper.

Different readers present more or less engagement and subscribe for different reasons (Pattabhiramaiah et al., 2019). By understanding reader patterns, marketing and editorial teams can customize user experience by group in a Big Data landscape. Although customer segmentations are popular across industries (Vinothini & Priya, 2018) none study was found in the particular case of publishers. Thus, authors aim to fill a gap in the research literature by presenting an innovative reader segmentation model to support retention and acquisition management for publishers. The results achieved can be helpful in targeting each group of readers to implement editorial and marketing tailored strategies (Kotler et al., 2016). Besides, the derived results are actionable, interpretable and experimented in the Portuguese newspaper Público that presents a reach higher that 4 million users and collect more than 200 million user events monthly.

4.3. Theoretical background

4.3.1. Segmenting online users

The concept of segmentation, firstly introduced by Smith (1956), consists in dividing a heterogeneous group into smaller homogeneous sub-groups in which customers share something in common (Fu et al., 2017; Smith, 1956). As customers are heterogeneous in their behaviors, a segmentation helps the company to develop customised strategies and suitable products for different segments (Fu et al., 2017; Y. Liu et al., 2021; Vinothini & Priya, 2018).

By searching in the Scopus database (Ballew, 2009) for the keywords "customer" and "segmentation" in titles, abstracts, and keywords, we found more than four thousands documents. In this sample, a wide range of successful applications across industries can be found. However, in the particular case of publishers, research is quite scarce. Among media research articles, there are investigation on segmenting readers according to similarities in their preference patterns (Chakrabarty et al., 2019), segmenting according to their perception of media website features such as, easy navigation, or content relevance or trustfullness (Cristobal-Fransi et al., 2017). Furthermore, to the best of the authors' knowledge, there is no approach that studies readers' segments by exploring user engagement metrics in online news. This is probably due to the complexity associated to the widely range of reader engagement levels, the high proportion of new monthly users (Jacob, 2021) and the high amount of data produced. However, behavioral segmentation is the most sophisticated method to segment that also has more potential to optimize user experience (Seufert, 2014).

In today's competitive Big Data environment, publishers need to segment and targert readership to meet revenue goals. Furthermore, DSBMs became quite popular in the last decade, in particular, across publishers as result of Paywalls Popularization (Arrese, 2016; Klopcic, A. L., Hojnik, J., Bojnec, S., & Papler, 2020).

One fundamental step in the application of cluster analysis is to choose the variables along which to group individuals (Ketchen & Shook, 1996). Thus, seeking to define the segmentation attributes, Table 15 summarizes different segmentation models and attributes used to identify customer' clusters in DSBMs. For each approach, the table mentions the industry where the method was applyed and respective attributes used. The first approach consists in segment online players by using engagement, performance, and social interaction features (Fu et al., 2017). Then, the second approach consider sociodemographic, user and behavior data (Tanuwijaya et al., 2021). Finally, the third approach segment customers by using user and behavior data (Smit et al., 2019).

The present approach focuses in perform a behavioral segmentation to influence future behavior (Seufert, 2014). Thus, the selection of attributes is based on the literature presented in Table 15, coupled with the main features of the reader engagement definitions found in the research literature, detailed further ahead.

Table 15 Literature review on customers' segmentations in DSBMs.

Author(s)	Business Domain	Attributes / Features	Method(s)
(Fu et al., 2017)	Games	**Engagement**: playtime, login **Performance**: level, mission, quest, coin frequency, coin **Social interaction**: guild status, guild role type, common point, guild frequency and friends	Fuzzy C-means clustering
(Tanuwijaya et al., 2021)	Streaming	**Sociodemographic**: age, gender **User**: smartphone brand. **Behavior**: traffic, duration, sessions	Naïve Bayes; Decision Tree; Random Forest; Logistics Regression; LGBM; XGBoost; Catboost
(Smit et al., 2019)	Education	**User**: date and time at which the recorded event happened, country, time zone corresponding the geo location of the user **Behavior**: sessions, URL, type of device, type of browser type of operating system	DBSCAN

4.3.2. Reader engagement measurement

The current highly competitive and nonlinear environment forces publishers to guarantee that they are offering useful, informative, compelling, gratifying content, and more than ever, engaging content (Ksiazek et al., 2016; Rios-Rodríguez et al., 2022). Customer engagement involves emotional attachment and rational loyalty (Clement Addo et al., 2021; Wenzel et al., 2022). Authors agree that reader engagement is a multidimensional phenomenon (Steensen et al., 2020) related to the level of attention and involvement (emotional, cognitive and behavioral) with media (Ksiazek et al., 2016; Mersey et al., 2010). Furthermore, reader engagement mesurement can be divided into three broad categories: self-report (such as surveys), physiological (such as observational methods), and web analytics methods (Davoudi, 2018).

To calculate user engagement, Peterson and Carrabis (2008) presented a mathematical linear function that combines some widely used metrics, such as, the number of clicks (C_i) or visit duration (D_i) (Peterson & Carrabis, 2008). However, the ability to act, interact, and co-create online are key characteristics of online media (Ksiazek et al., 2016). Thus, the involvement of the reader can be expressed as the interactivity, that can be measured by the number of interactions with article features like number of comments or likes (Ksiazek et al., 2016).

In 2016, the Financial Times marketing team adapted the renowned RFM (Recency, Frequency, Monetary) analysis (Stone, 1989) to the media sector by defining the new engagement metric RFV (Recency, Frequency and Volume) (Goad, 2016). RFM is a simple but effective method used in marketing to analyze customer purchasing behaviors (Coussement et al., 2014). Equivalently to RFM (Stone, 1989), the RFV relies on three dimensions: recency (R), frequency (F), and Volume (V). Thus, R measures how recently the customer have visited the website, F is the number of website visits within a time period and V indicates the total number of articles read in a period. Finally, the scores of all three variables are consolidated into a score (FTStrategies, 2022). As result of some successful cases around the media industry, the RFV became a strategic KPI across media companies to measure readers' engagement (Goad, 2016; Wenzel et al., 2022; Zontek, 2018). However, other approaches are emerging. The APV engagement score developed in The Independent (UK), combines active days (A), volume of premium content read (P) and number of articles read (V) in a period (Lajumoke et al., 2020). Furthermore, the Wall Street Journal team found a direct correlation between the number of user active days and churn that led them to define A as the engagement north star (Seale, 2021). Similarly, a churn model was computed with one year data of Público subscribers. Findings revealed that A is also a good predictor to predict Público subscriber's churn. It is also consisted with subscriptions studies in others industries, as an example, the telecom market where the usage affects churn and loyalty.

Other engagement metrics can be found across the literature, such as *dwell* time (i.e., the time spend on a resource) (Davoudi et al., 2019; Grinberg, 2018; Lehmann et al., 2012) or engaged time that measures the amount of time that users spend actively interacting with a page (Schwartz, 2013). However, each platform measures engaged time differently according to their definition of user active interaction. Furthermore, the same user can be interacting with two pages at same time. Moreover, dwell time only provides partial information about reader activity in the article (Grinberg, 2018). Thus, in the present research we decided to measure how far users scroll down the article page (Grinberg, 2018), i.e., the vertical scroll depth of 75%. It means that the reader visited the webpage at least at the end of the article text providing information about article relevance (Grinberg, 2018).

4.3.3. Big data and clustering

The explosive growth in available data from online businesses has induced a strong emphasis on big data analysis (BDA) (Mathew, 2021). Furthermore, BDA applications, such as algorithms in classification, clustering, and association, have been used in a disparate variety

industries like media, entertainment and communication (Vinothini & Priya, 2018). Moreover, BDA adds value to online businesses by transforming information into insights to solve business challenges leveraging the dynamics of people, processes, and technology (Mathew, 2021). For instance, Spotify or Netflix are among the companies with DSBMs that already uses BDA (Mathew, 2021; Omran et al., 2007), as well as some widely known publishers like The New York Times (Rußell et al., 2020) or Aftenposten (Sjøvaag, 2016).

Cluster analysis is an important machine learning (ML) modeling technique to discover natural grouping of the observed data to create customer profiles (Gonçalves et al., 2021; Silva et al., 2018; Vinothini & Priya, 2018). Two learning problems are addressed that are supervised, also known as classification, and unsupervised, known as clustering. The first referred to labeled data, the second one to unlabeled data (Gonçalves et al., 2021; Omran et al., 2007). One of the most popular and efficient clustering methods is the K-Means algorithm (Vergani & Binaghi, 2018). K-Means groups the data based on their closeness to each other. The cluster's centre point is the mean of that cluster, and the other points are the observations that are nearest to the mean value. Furthermore, the number of clusters is chosen in advance, and it uses an iterative procedure that minimize the squared error of the following objective function:

$$J = \sum_{i=1}^{D} \sum_{K=1}^{K} \left\| X_{i(K)} - A_K \right\|^2$$

where $\|...\|^2$ is a specific distance function, *D* are the data points, *K* are the clusters, $X_{i(K)}$ are the points in cluster *K*, A_K are the centroids of clusters *K*.

In a clustering analysis to determine the number of cluster (*K*) is crucial (Fu et al., 2017). However, no metric can guarantee optimal results. As result, many approaches are documented in the literature to determine a suitable number of clusters. As each method has limitations (Ketchen & Shook, 1996), the present approach uses a combination of three widely used metrics (Gustriansyah et al., 2019; Xiao et al., 2017) detailed as follows:

- *Distortion score* (Ketchen & Shook, 1996), also called Within Cluster Sum of Squares (WCSS), is the sum of squares of distance between points (*d*) and cluster centres (*C*). It is a convex monotonically decreasing curve. The widely known elbow method, results of select the value *k* were the curve reaches an elbow. By visual inspection it is possible to identify the range of reasonable values for *k* when the curve reaches a plateau, i.e., when we increase the number of clusters the new is very near to some of the exiting and the WCSS decreases slowly.

$$WCSS = \sum_{C_n}^{C_K} \left(\sum_{d_i in\ C_i}^{d_m} distance\ (d_i, C_K)^2 \right)$$

- *Calinski-Harabasz score* is an evaluation metric based on the degree of dispersion between clusters(Calinski & Harabasz, 1974). It consists in a ratio between inter-cluster covariance and intra-cluster divergence. The higher the ration is, the better the clustering effect is (Calinski & Harabasz, 1974).

$$CH(K) = \frac{(\sum_{k=1}^{K} a_k \|\overline{x_K} - \bar{x}\|^2)(n-K)}{\left(\sum_{K=1}^{K} \sum_{c(j)=K} \left\|x_j - \overline{x_K}\right\|^2\right)(K-1)}$$

- *Davies Bouldin Score* calculates the average of the similarity measures of each cluster with a cluster most similar to it (Davies & Bouldin, 1979; Vergani & Binaghi, 2018). Thus, the optimal number of *K* is where the average similarity is minimized.

$$DB\ (K) = \frac{1}{K} \sum_{i=1}^{k} max_{i \neq j} \left(\frac{s_i + s_j}{d_{ij}} \right),$$

where s_i is the cluster diameter and d_{ij} the distance between cluster centroids *i* and *j*.

4.4. Method

Our research assumes the workflow proposed in Figure 24 that is inpired on the known CRISP-DM methodology (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, & Rudiger, 2000; Moro et al., 2011). The following subsections detail the remainder framework steps to analyze the data that was kindly provided by Público.

4.4.1. Engagement window

Firstly, to collect Público reader's data, and preprocess the raw dataset, it was necessary to define the period to calculate the engagement metrics. Thus, an analysis was performed to define the "*engagement window*", i.e., the period of significant user content involvement prior to subscription. All the user events (Google, 2022) in the six months before subscribe, of Público readers that subscribed for the first time between 1st September 2021 and 15th June 2022, were analyzed. Thus, a total 12,721,603 events were considered, from 10,294 new Público subscribers. An example of a frequent event is "*Ler Mais – click*" indicates that the

reader clicked at the “Ler Mais” box, that is a recirculation element at the article page. The average number of events by day was computed to enable a time series analysis that shows the evolution of readers events on the website before subscribe. The goal was to identify changes in the time series. Thus, to solve this changepoint detection problem (CDP), the standard deviation was used as a cost function (Katser et al., 2021). The reader behavior pattern has a cycle of a week, thus the standard deviation was calculated for 7 and 14 days. A sliding window through the six-month time series, starting at the subscription day, presented maximums between the 34th and 39th day before subscribe (as presented at Figure 25). Both changes in mean and variance indicate a reader's behavior change through the increase in the number of events by day prior to subscription. Thus, the engagement window considered to calculate engagement attributes (EA) was 30 days that is also a frequent period of reader analysis at the media industry (Blazejewski, 2019; Jacob, 2021).

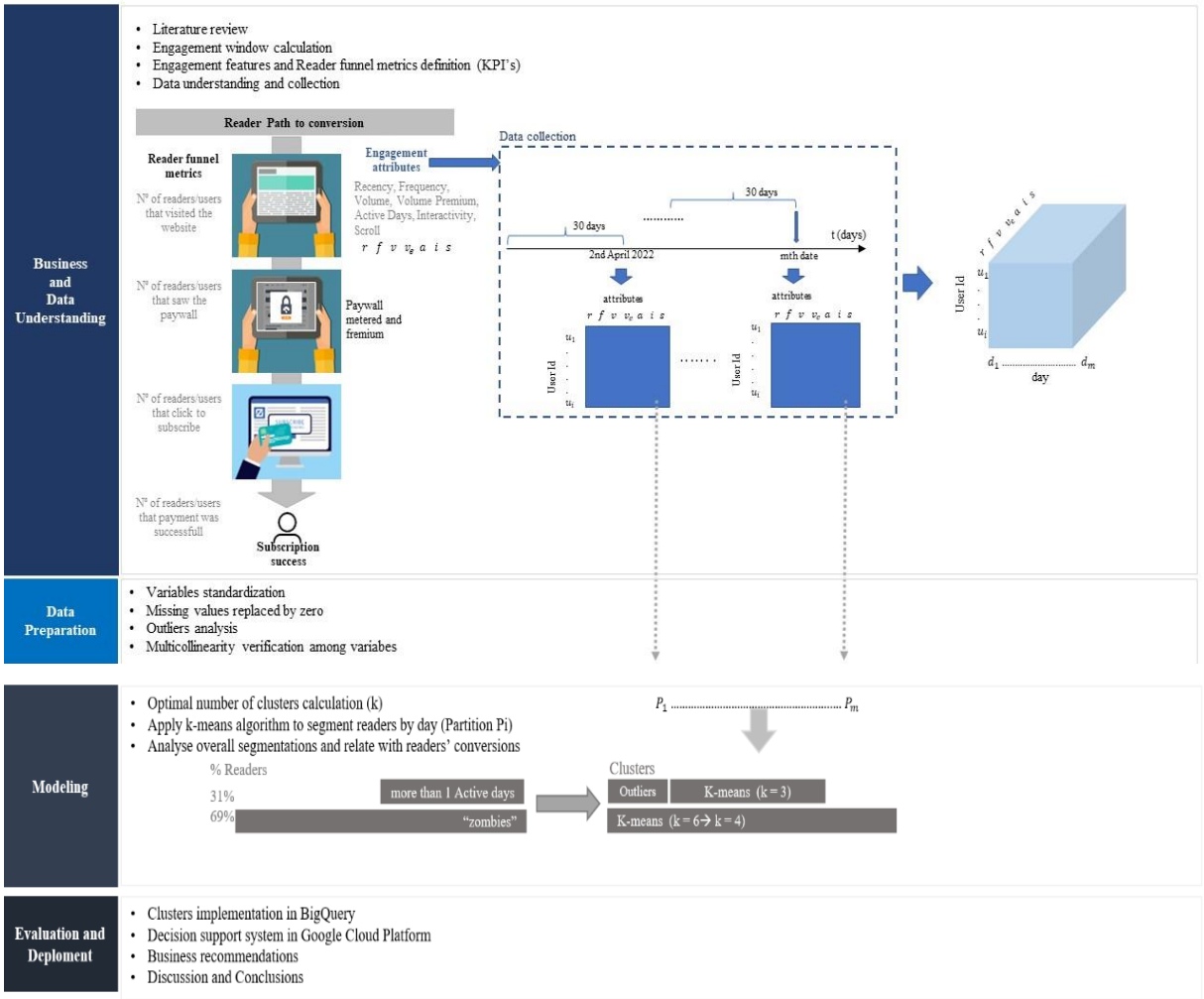


Figure 24 The proposed framework.

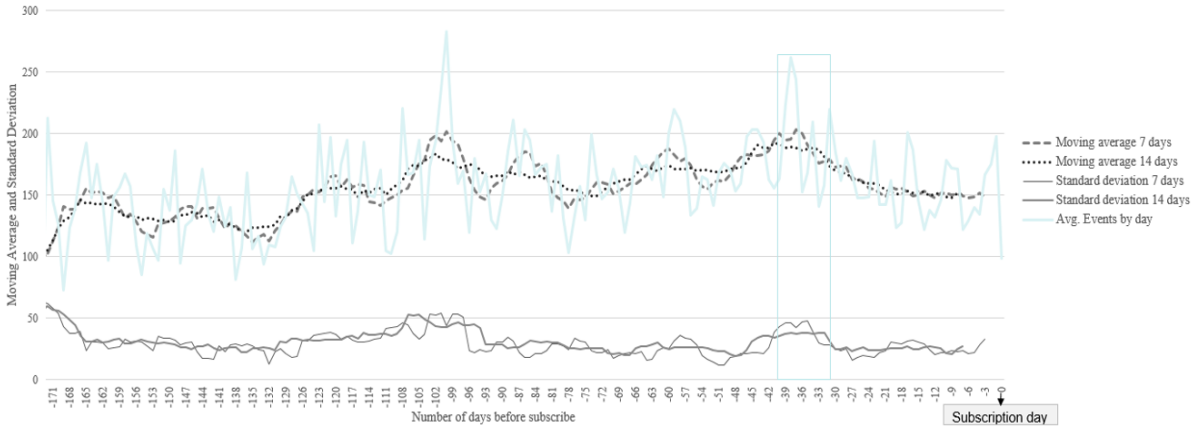


Figure 25. Changepoint detection across the time series.

4.4.2. Data understanding and data preparation

A real dynamic dataset was used in this research that is stored at Bigquery, a Google-managed data warehouse (Bisong, 2019; Mucchetti, 2020). Data was accessed through Colaboratory (see Figure 26), a product from Google Research that is a hosted Jupiter notebook service with access to Google hardware (Nelson & Hoover, 2020). BDA was performed by programming Python scripts in Colaboratory.

The dataset contains the *date*, *user id* and the engagement *features* calculated for a 30-days timeframe (see Table 16). Data values were stored on a daily basis, encompassing a period ranging from 2nd April to 1st June 2022. Missing values were replaced by zero as zero indicates that no user visit was registered. The zero-mean normalization method was used as attributes present different scales. Furthermore, as clustering methods are sensitive to outliers, those were considered as an individual cluster. For each column of the dataset the absolute Z-score was calculated and readers with values greater than 3 were grouped into a cluster (i.e., data points that fall outside of 3 standard deviations were classified as outliers).

In the news domain, two cold start problems are frequently studied: the *item cold start,* i.e., when a new article is published; and, *user cold start,* i.e., when a new or unknown user visit de website (Delpisheh et al., 2016; Zihayat et al., 2019). This problem becomes more prominent when light users are dominant (B. Liu et al., 2022), approximately 70% of readers only visit the website once in 30 days (see Figure 24). Thus, the dataset was divided into two samples.

The first sample contains readers that visited the website once in a month, frequently called "zombies" at the industry literature (Jacob, 2021; Lynes, 2021). The second group contains readers than present more than one active day in 30 days (Sample 1). Those readers present an average of 4.4 active days, 5.7 articles read with 3.4 articles of premium content, 5.1 times that achieved 75% of the article page and 4.9 times that the readers share or comment an article in 30 days (see Table 16).

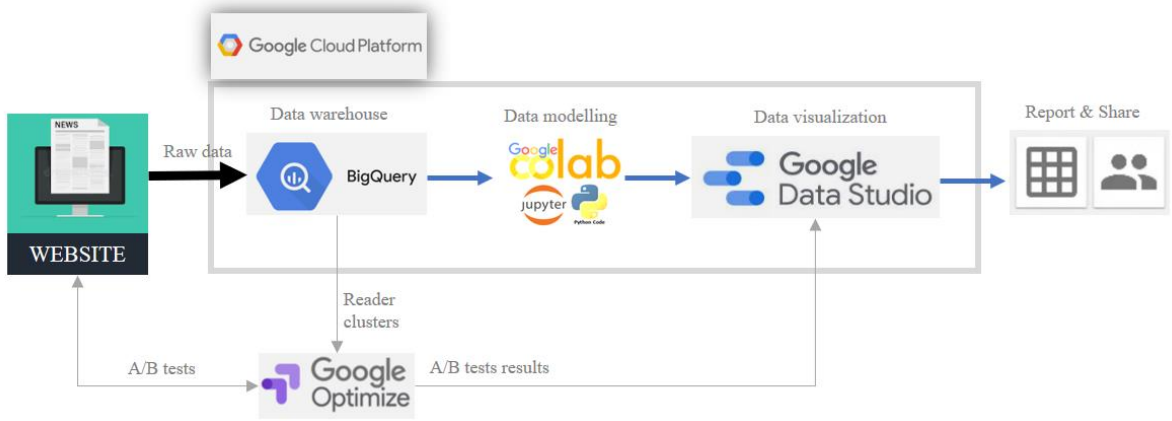


Figure 26. Decision support system designed in the Google Cloud Platform (GCP).

Table 16. Descriptive statistics of the EA by sample from 2nd April to 1st June 2022

EA	Definition	Descriptive Statistics Sample 2 "zombies" (n=10.326.306 users)				Descriptive Statistics Sample 1 " +1 active day" (n=3.379.611 users)			
		Avg.	Std.	Min.	Max.	Avg.	Std.	Min.	Max.
R	Number of days since the last visit (high value means that reader made a visit recently)	-15.6	8.5	-30.0	-2.0	-10.5	7.7	-30.0	2.0
F	Number of different hours with at least on visit	1.2	0.7	1.0	664.0	6.9	14.1	1.0	696.0
A	Number of different days with at least one visit	1	0	1.0	1.0	4.4	4.3	2.0	30.0
V	Number of articles read	1.2	0.8	1.0	239.0	5.7	12.5	1.0	2,484.0
VP	Number of premium articles read	1.1	0.6	1.0	55.0	3.4	7.2	1.0	735.0
S	Number of times that the reader achieved at least 75% of the webpage	1.4	2.5	1.0	768.0	5.1	31.3	1.0	13,935.0
I	Number of times that a reader shared an article by one social media options, or by email, or at least commnet the article.	1.6	1.8	1.0	139.0	4.9	29.9	1.0	3,467.0

In the news domain, two cold start problems are frequently studied: the *item cold start,* i.e., when a new article is published; and, *user cold start*, i.e., when a new or unknown user visit de website (Delpisheh et al., 2016; Zihayat et al., 2019). This problem becomes more prominent when light users are dominant (B. Liu et al., 2022), approximately 70% of readers only visit the website once in 30 days (see Figure 24). Thus, the dataset was divided into two samples.

The first sample contains readers that visited the website once in a month, frequently called "zombies" at the industry literature (Jacob, 2021; Lynes, 2021). The second group contains readers than present more than one active day in 30 days (Sample 1). Those readers present an average of 4.4 active days, 5.7 articles read with 3.4 articles of premium content, 5.1 times that achieved 75% of the article page and 4.9 times that the readers share or comment an article in 30 days (see Table 16).

4.4.3. Clustering analysis

In order to define a segmentation of 30-day users (see Figure 24) at each day of the timespan, a dataframe was collected and a segmentation was performed. The main goal was to define the optimal number of clusters that daily need to be calculated to implement actionable actions according to clusters' characteristics. For both samples, we stored the data values daily, encompassing a period ranging from 2nd April to 1st June 2022 (61 days).

Next steps present a one-day users' segmentation of readers from Sample 1, at the 1st of June, that represents approximately 1.4 million users. Furthermore, Table 17 presents Pearson correlations (Hauke & Kossowski, 2011) among engagement variables. As expected, V and VP present a very high correlation. Despite, high correlation among clustering variables can be problematic (Ketchen & Shook, 1996), we decided to keep both because in this business context the content type is important to define editorial strategies. Besides, Table 17 also indicates that an increase in A will increase F, as well as V and VP. Moreover, I, R and S does not present strong linear correlations with other variables.

Then, to define the optimum value of clusters *k* (i.e., when inter cluster dissimilarity and intra cluster similarity were maximized) the three metrics were calculated. The number of clusters is optimal when the marginal gain of adding a cluster drops dramatically in the distortion score (DS) ("elbow criterion"), and when the CH Score is maximum, and the DB Score is minimum. At 1st of June, as presented at Figure 27 the optimal number of cluster (*k*) for sample 1, is three.

Table 17. Pearson correlation between variables after standardization

	R	F	V	VP	A	I	S
R	1.00	0.16	0.14	0.12	0.29	0.01	0.04
F		1.00	0.65	0.61	0.74	0.13	0.36
V			1.00	0.95	0.62	0.23	0.30
VP				1.00	0.57	0.24	0.28
A					1.00	0.07	0.20
I						1.00	0.07
S							1.00

Table 18. Pearson correlation between variables after standardization (sample 2)

	R	F	V	VP	I	S
R	1.00	-0.004	0.017	0.014	0.000	0.002
F		1.00	0.181	0.116	0.048	0.149
V			1.00	**0.577**	0.142	0.240
VP				1.00	0.105	0.136
I					1.00	0.033
S						1.00

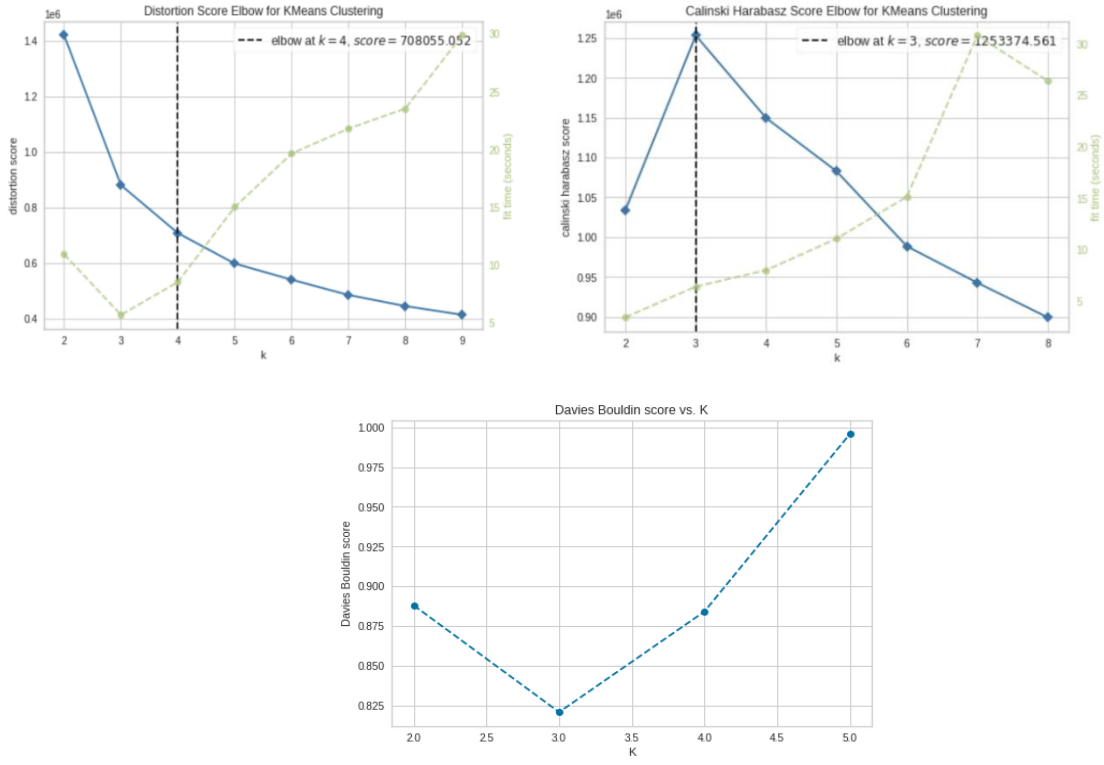


Figure 27. Evaluation metrics to find the optimal number of clusters for users with more than one active days at 1st June.

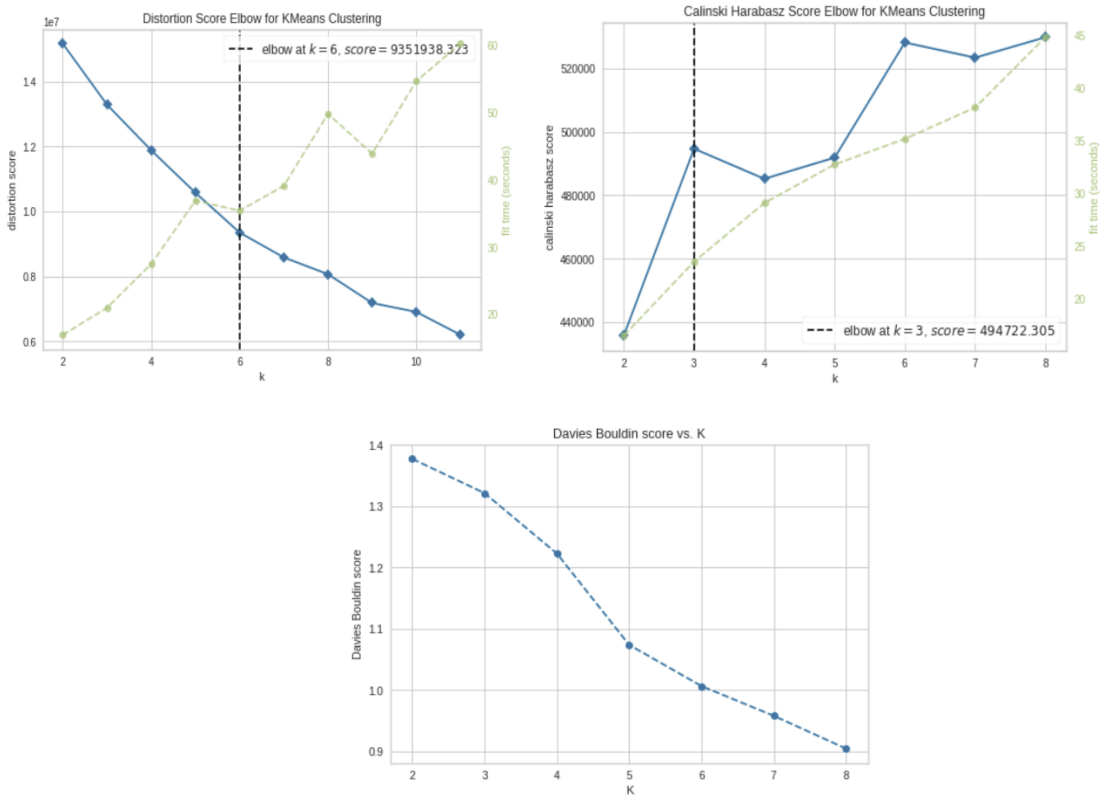


Figure 28. Evaluation metrics to find the optimal number of clusters for the "zombies" (sample 2).

Oppositely to sample 1, Pearson correlation is low between all EA except between V and VP that presents a moderate correlation (Table 18). Furthermore, Figure 28 presents the optimal number of clusters by metric. The DS score indicates six as the best number of clusters. However, CH Score is maximum at 6 and 8, whereas DB Score does not present an absolute minimum (Vergani & Binaghi, 2018). Thus, by combining the three measures, 6 seems to be the optimal number to partitionate "zombies".

4.5. Results, general discussion and implications

4.5.1. Group profiling

In the period under analysis, all readers were impacted with the same marketing campaigns and communications. Thus, the click action happened under the same marketing *stimulus*, despite different levels of content involvement. By using the EA $[R; F; V; VP; A; I; S]$, readers of Sample 1 were grouped into four clusters that are referred to $Cluster_0$, $Cluster_1$, $Cluster_2$ and $Cluster_3$. As presented at Table 19, larger values indicate high engagement levels. The characteristics of the detected clusters are summarized as follows:

- $Cluster_0$ (**active**) contains approximately 11% of the sample where around 30.43% visit the website in a day. From those readers, 61% see at least once the paywall and then 0.58% click to subscribe. These are recent visitors, with high number of A and consequently high values of F, V and VP. They are valuable readers with strong content involvement that increases ads revenue. However, they do not present the highest click and conversion rates.
- $Cluster_1$ (**moderate**) corresponds to the readers that did not visit recently the website, they present less A and also less S, revealing low engagement with the content. In fact, only 8.27% of the cluster users return to the website, and from those, 51.04% see the paywall at least once. Only 0.6% click, but after click, they are more propense to convert than $Cluster_0$ and $Cluster_2$ (5.41%). Those results indicate that there is opportunity to increase conversion rate in readers that present higher engagement levels.
- $Cluster_2$ (**need attention**) constitutes the highest cluster (55%). They reflect a segment of readers that visited the website recently. But, despite they are not active readers, after seeing the paywall they present the highest click rate. After manual inspection we verified that such readers were attracted by a premium content (from push notifications) and articles that induced button subscription click (van Damme et al., 2020).
- $Outliers$ (**super**) is the smaller cluster. However, 87.15% return to the website, indicating that the last visit was recent ($R = -2$) and A is the highest ($A = 6.77$). As result, those readers present high values of F, V and S. Nonetheless, I presents low value.

In what concerns to "*zombies*, clustering was performed considering the remaining six EA since A is constant, equal to 1.0. Furthermore, on average, those readers present low values of F, V, I and S (see Table 16) indicating there is a need of recirculation strategies (Lioudis, 2019). Also, they present an average of 1.4% website return that indicates an opportunity to a multichannel strategy improvement (Hullar, 2020; Loni et al., 2019). Moreover, the click rate is low (as presented at Table 20) indicating that low levels of engagement are related to a low propensity to subscribe. However, those readers play an important role on advertising revenue (Arrese, 2016), as they represent 20% of website pageviews. The characteristics of the clusters are detailed as follows:

- $Zombies\ 0$ (**casual**) consists of the first three clusters presented at Table 20 that represent less than 2% of the sample. They were grouped because they do not present

significant click rate and conversion rates. Furthermore, the last visit was around 16 days ago and between all the zombies, they present better content consumption ($V >$ 4).

- $Zombies\ 1$ (**sleepers**) accounts for 23% of total "zombies". This cluster contains readers that on average made the last visit 15 days ago. Furthermore, only 1.62% return to the website, from those 39.67% see the paywall where 0.46% click and 10% of those that clicked subscribe.
- $Zombies\ 2$ (**lost**) constitutes 35% of the sample. These readers do not perform well, the last visit was 23 days ago. Only 0.71% return to the website and only 0.3% that saw the paywall click to subscribe. However, 12.5% that click on the paywall subscribe, suggesting that there is an impulse related to breaking news.
- $Zombies\ 3$ (**recent visitors**) the larger cluster (41% of the sample) corresponding to readers that visited recently the website (9 days). However, they present lower V and only 1.57% return to the website.

Table 19. Engagement attributes average and reader funnel metrics by cluster (sample 1)

	EA	**Cluster 0** $\boldsymbol{n}$ = 141,361 11%	**Cluster 1** $\boldsymbol{n}$ =373,631 30%	**Cluster 2** $\boldsymbol{n}$ = 670,002 55%	**Outliers** $\boldsymbol{n}$ =49,356 4%
Engagement attributes	R	-6.01	**-20.36**	-6.51	**-2.0**0
	F	9.17	**4.39**	7.25	**9.71**
	V	**7.84**	**3.79**	6.21	6.32
	VP	3.51	**2.44**	**3.76**	2.62
	A	5.97	**2.97**	4.62	**6.77**
	I	4.48	3.98	**5.58**	**1.97**
	S	5.29	**3.62**	**5.53**	4.66
Reader funnel metrics	Proportion of users that visited the website (%)	30.43	8.27	23.63	87.15
	Proportion of users that saw the paywall (%)	60.88	51.04	52.95	21.80
	Users that click to subscribe (%)	0.58	0.60	0.96	0.81
	Users that converted (%)	4.17	5.41	4.70	90.66

Table 20. Engagement attributes average and reader funnel metrics by cluster (sample 2)

	Variables	**Zombies 0** $n = 276$	**Zombies 0** $n = 1{,}916$	**Zombies 0** $n = 38{,}185$	**Zombies 1** $n = 672{,}033$	**Zombies 2** $n = 1{,}041{,}994$	**Zombies 3** $n = 1{,}201{,}960$
		<0.5%	<0.5%	1%	23%	35%	41%
EA	R	-16.36	-15.68	-16.04	-15.43	**-23.67**	**-9.05**
	F	**9.13**	2.04	3.13	1.12	**1.10**	**1.10**
	V	**7.90**	4.56	4.92	1.21	**1.06**	**1.06**
	VP	**3.73**	3.03	2.60	1.05	**0.00**	**0.00**
	I	1.00	**5.05**	1.37	1.16	**1.13**	**1.13**
	S	**86.86**	2.70	4.17	1.29	1.19	1.14
Reader funnel metrics	Proportion of users that visited the website (%)	2.54	0.63	1.36	1.62	0.71	1.57
	Proportion of users that saw the paywall (%)	42.86	41.67	36.68	39.67	36.09	34.70
	Users that click to subscribe (%)			0.53	0.46	**0.30**	0.50
	Users that converted (%)				10.00	12.50	12.12

4.5.2. Managerial implications

In terms of managerial implications, a two-fold approach is proposed. Readers of Sample 1 have more intention to subscribe. Thus, by target marketing strategies the team can develop customized plans and schemes (Kotler et al., 2016). Furthermore, authors explored new paywall mechanisms (Davoudi et al., 2018) and the impact of paywall design (Aral & Dhillon, 2021) that could be adapted for tailoring strategies by cluster. Moreover, recommendation algorithms could be used to increase engagement, for example, in newsletters to provide a more personalized experience.

Meanwhile, at Sample 2, the main challenge is to increase visits to the website, i.e., invest in multichannel strategy. Authors agree that push notifications drive more engagement than newsletters and it is a fast and effective way to inform users (Budiman & Akhlis, 2021; Gao et al., 2018; Loni et al., 2019). Thus, it could be an opportunity to develop a specific experience for "zombies" to induce them to receive notifications (van Damme et al., 2020). Despite, too many notifications can create a poor experience, researchers developed optimization systems for notifications (Gao et al., 2018) and personalized systems for this business context (Loni et al., 2019).

4.5.3. The relation between engagement metrics and intention to subscribe

Another point of view, it is to analyze causal relationships between EA and reader's intention to subscribe. Hence, we considered a dataset with 189 thousand users from Sample1 that viewed the paywall. To start the payment flow, each reader may click or not click in the paywall, value 1 and 0, (see reader flow at Figure 24). Three ML classifiers were chosen given their popularity and results achieved (Coussement et al., 2014), namely: Logistic Regression (LR), K-Nearest Neighbor (KNN), and e**X**treme **G**radient **B**oosting (XGBoost) (T. Chen & Guestin, 2016). The obtained results are calculated and evaluated in terms of accuracy (Labatut & Cherifi, 2012), F1 score (Lipton et al., 2014), **A**rea **U**nder the receiver operating characteristic **C**urve (AUC) (Gonçalves et al., 2014), and **R**oot **M**ean **S**quare **E**rror (RMSE) (see Table 21). LR and KNN were widely applied in DSBMs, such as churn modeling (de Caigny et al., 2018; Shahraki et al., 2017). However, as XGBoost algorithm is an implementation of gradient boosted decision trees designed for speed and performance (Brownlee, 2016), it has achieved superior results in several ML challenges (Brownlee, 2016; T. Chen & Guestin, 2016; Wieland et al., 2021).

In this research, we split the database into the training and testing subsets on a ratio of 0.75/0.25. The training set is used for establishing the models, while the testing set evaluates the prediction. XGboost optimizes the four performance metrics (see Table 21). The percentage of features used per tree was set to 50%, in order to avoid overfitting. Moreover, we oversample the minority class to obtain a more balanced data distribution by using Synthetic Monority Oversampling Technique (SMOTE) (Chawla et al., 2022), six neighbors were found that maximize the mean AUC. As XGBoost is a tree-based method, it does not require standardization (Wieland et al., 2021). Furthermore, to explain how much each EA contributed to the model's prediction when compared to the mean prediction **SH**apley **A**dditive ex**P**lanations (SHAP) values were calculated (Lundberg & Lee, 2017). In the particular case of a binary model, SHAP values give the difference between the predicted log odds and average predicted log odds (Wieland et al., 2021).

Table 21. Performance of classification methods (best values are highlighted by using a boldface font).

Classification method	Accuracy	F1 score	AUC	RMSE
LR	0.915	0.019	0.676	0.292
KNN	0.987	0.099	0.696	0.113
XGBoost	**0.993**	**0.126**	**0.752**	**0.088**

At Figure 29, each point in the cloud represents a row from the original dataset. The colour code denotes high (red) to low (blue) feature values (Lundberg & Lee, 2017). The ranking indicates that S is the most important driver for the click action thus experiments on design can improve the user experience to increase the scroll down. Furthermore, VP was identified as the second major driver followed by A. Click on subscribe is more likely to happen through increasing S. In contrast, high values of A have a high negative contribution on subscribe. An explanation for this derived from the high proportion of readers that only read the first paragraph of the article. Furthermore, V also present a negative contribution, indicating that when a reader is impacted by the paywall, he is less likely to subscribe. These finding is in line with the study of (Davoudi et al., 2018) that argues the lack of assertively in the metered and freemium models. Thus, there is an opportunity to improve the user experience and innovate the business paywall model to increase propensity to subscribe. The effect of R comes almost at the bottom, indicating that, in the sample under study, R seems to have little effect on the intention to subscribe. Surprisingly, I also has low effect, despite to be a widely study engagement feature. Moreover, the engagement drivers to subscribe could be daily monitored to understand the action's impact and analyze the SHAP ranking fluctuations. The fact that the segmentation will be running daily into big query (see Figure 25) will help marketing team to monitor engagement and funnel metrics by cluster.

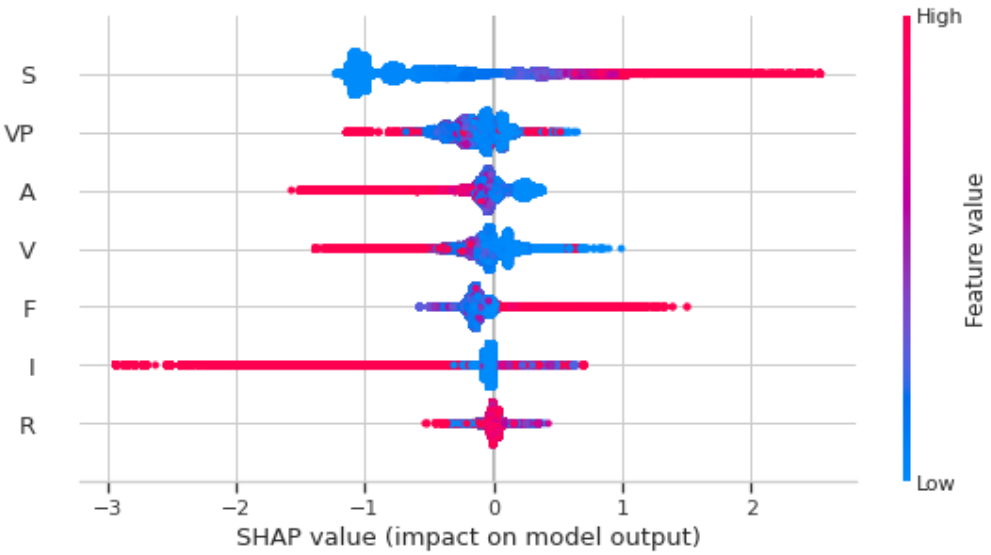


Figure 29 Engagement features ranking by SHAP values of the model.

4.6. Conclusions and limitations

Each reader presents different reading interests and pattern consumptions. Thus, in a Big Data context, it is impractical to apply the same strategy to engage all users. Reader segmentation is a challenge that combined with conversion data provides useful information to improve acquisition and retention strategies. Hence, the proposed method consists in attributes retrieval, clustering analysis, and business recommendations. Furthermore, the presented approach can be adapted to other online subscription businesses. Results obtained in previous sections can be studied in two complementary ways. Firstly, we can see the characteristics of each cluster in terms of engagement to implement multichannel (Hullar, 2020) and recirculation strategies (Lioudis, 2019). Then we can study how engagement is related to intention to purchase to develop different strategies.

In the present research, readers are divided into eight clusters, labelled as *super, active, moderate, need attention, casual, sleepers, lost*, and *recent visitors*. The new model identifies behavior patterns by cluster. From an operational perspective, such knowledge allows editorial and marketing team to define business strategies.

Furthermore, the classification algorithm provides information about the main subscription drives in Sample 1. The derived results reveal that, the number of times that the reader scroll depth (Grinberg, 2018) in the article page is a strong driver to increase the propensity to subscribe, followed by the volume and active days.

A two-fold strategy is proposed to guarantee engagement increase and conversion increase. Hence, a multichannel strategy improvement is recommended to motivate readers to subscribe newsletters or notifications to increase reach and engagement by device. Furthermore, the use of recommendation algorithms can play a pivotal role to increase website recirculation and consequently increase V, I and S. Meanwhile, segmented marketing and design strategies can induce a S increase and click rate increase. Besides, the advertising strategy could be segmented to increase this source of revenue according to cluster goals.

This proposal is of interest of the publisher, as it presents a useful and understandable model based on EA that are actionable, reliable and readable. Moreover, the model was built in the Google ecosystem allowing the teams to easily monitor values, implements website tests, and measure daily results impact. Thus, the publisher can ultimately improve the user experience, the content offer and the multichannel content distribution to increase reader loyalty, engagement and consequently maximize revenue.

Despite its benefits and contributions, there are still limitations. This study models and analyzes the behavior of different readers clusters in one publisher. K-means was selected as it is a widely used method. However, other segmentation algorithm or attributes could be applied. Despite the period of metrics calculation of 30 days is aligned to the most frequent practices across the industry, future work could be exploring the impact of engagement window variation.

The scope of future work lies in two main goals: to study recommendation algorithms to improve recirculation, and to improve multichannel strategy to face the user cold start problem. Thus, next section aims to present an experience implemented at Público to improve reader's engagement in a particular channel, the newsletters. As an important channel for personalization and engagement, the author provided a solution to increase subscribers engagement and decrease propensity to churn to impact in another level of the reader funnel (see Figure 1).

CHAPTER 5

Content personalization in editorial newsletters

In the big data era, recommendations systems (RS) play a pivotal role to overcome information overload. Furthermore, newsletters emerged as an important conversion channel to engage readers as they provide a personalized experience by building habit. However, the lack of resources and the need to more content assertiveness per reader lead to publishers to search for analytical solutions. We address this problem by proposing a recommendation algorithm inspired in the *table d'hôte* approach (Abdollahpouri et al., 2021). This approach, inspired on a table d'hôte meal, considers that a bundle of content is like a meal with a sequence of courses that create a balanced and enjoyable dining experience. Thus, the reader receives a personalized newsletter where he can discover informative and surprizing content. The goal is to offer a self-contained package that offers value and retain readers. Readers can control how and when to read the content, increasing loyalty and consequently the propensity to subscribe. We study the effects of content recommendations on the behaviour of Público newsletter subscribers.

According to Suárez (2020) churn is more important than acquisition. A small improvement on that will have a huge impact on company results (Suárez, 2020). Thus, to provide a data-driven solution to Público, this section presents an experiment on a particular segment of readers, i.e., readers that present high propensity to churn that are readers with less than 42% active days in 60 days (pattern detected in the churn model developed at Público). Our focus is to improve reader engagement from the newsletter channel to guarantee assertiveness and engagement increase.

Despite newsletters only represent 2% of monthly users of the website, readers from newsletters stay longer in the website (more than 6.5 minutes), see more than 4 pages by session, and visit the website approximately 4 times per month. While the overall average of the website users see 2.4 pages by session, stays 3 minutes per visit, and do less than 2 website visits per month. Furthermore, research shows that micro-segments are efficient acquisition channels (Suárez, 2020).

5.1. Introduction and background

Newsletters provide to the reader a personalized experience. The reader can control how and when to read the content, by creating habit there is a consequent loyalty increase (Mcmullan, 2018). Furthermore, as mentioned in the work of (Suárez, 2020), habits matter more than

content, and newsletter is a habit channel. Despite the newsletter's predictions about their demise, the medium continues to grow (Hendrickx et al., 2020). Moreover, newsletter represent an important channel to acquire subscribers (Mcmullan, 2018).

Some successful examples have been published in the industry literature, at The Globe they proved that a reader is 10 times more likely to become online subscriber after subscribing the editorial newsletter (Newman et al., 2019). As an example, in Belgium nearly a quarter of the population indicated newsletters as their primary source of news (Hendrickx et al., 2020). Hence, 70% of the publishers say that they will put more resources into email newsletters (Newman, 2022).

Researchers argue that news fatigue is becoming more recognized as a serious concern (Fitzpatrick, 2022). As mentioned in the annual Reuters Institute Digital News Report most readers "*remain engaged and use the news regularly*" but there are readers "*that also increasingly choose to ration or limit their exposure to it- or at least to certain types of news*" (Newman et al., 2022). As news fatigue is slowing down subscriptions sales, how to engage readers amid the news fatigue?

Furthermore, news personalization can help people managing information overload (Fayyaz et al., 2020) while publisher draw audiences to the website and keep them engaged (Lafrance, 2017). An example of a successful personalization feature is myFT at the Financial Times. This feature allows the user to personalize their news services and has proven results on engagement KPI's (Mcmullan, 2018).

To build an automated newsletter, authors were inspired by the *table d'hôte* approach presented by Abdollahpouri (2021) coupled with the concept of 'personalized diversity' presented by (Joris et al., 2021). News personalization makes individuals' news diet unique (Lafrance, 2017), thus the goal it is to make news consumption more diverse and unique. The goal is to satisfy the news-diet needs (Abdollahpouri et al., 2021) considering a channel with some particular characteristics. The reader can open the newsletter at any time after receiving the content, which induces a need of content with diverse lifetimes. Thus, a recommendation set can cover different needs of a reader. In one hand, can contain important stories to be informed, on the other hand, it can contain unexpected surprises. Furthermore, as algorithms emerge, the notion of serendipity gains scholarly attention (van Damme et al., 2020). Serendipity refers to incidental news consumption (van Damme et al., 2020), in RS refers to suggestions that are attractive and unexpected (Abdollahpouri et al., 2021). In this chapter, we aim to offer a self-contained package that offers value and retain readers by developing an automated newsletter for individual users.

In this exploratory research we aim to address the following research questions (RQs):

- RQ1: How to cover reader needs with a NL?
- RQ2: Can a newsletter impact positively the engagement KPI's?

We did two live controlled experiments with real readers in which we measured the impact on reader's behaviour and engagement. Then a more complex approach was proposed to be tested in the future.

5.2. Recommendation systems and newsletters

Recommendation systems (RS) are widely used in retail and e-commerce to meet users' and business' needs (Fayyaz et al., 2020). RS were built based on the following items: users, items, preference (Omar et al., 2020). Furthermore, other industries have proven the positive impact of RSs such as healthcare, transportation, agriculture, culture, or media (Fayyaz et al., 2020). A wide range of applications can be found across de literature (D. R. Liu et al., 2018; Nasir et al., 2021).

In the online news media context, RS usually focus on recommending stories in the front-page (Chakraborty et al., 2017) or across the recirculation elements of the website (Zihayat et al., 2019). According to (Zihayat et al., 2019), built an effective and efficient RS for the news domain is more challenging than other domains (Bangari et al., 2021). Chapter 2 already presents at Table 7 the ten most cited articles in RS until the end of 2020. Seeking to define the RS for this study, Table 22 summarizes research on RS in the news domain published after 2020. Furthermore, (Fayyaz et al., 2020) provides the landscape of RS research and identifies directions by presenting types of RS, challenges, limitations, and business adoptions.

We further note that only one article (Abdollahpouri et al., 2021) at Table 22 is focused on newsletters in media, as the overall algorithms developed are usually adapted to newsletters. Furthermore, more research has been made in the e-commerce context to improve newsletters performance (J. Bai et al., 2019; Zhu, Harrington, et al., 2014).

According to (Mcmullan, 2018), there are two kinds of newsletters created by publishers: those that drive content back to a news website and those promoting content specifically created just for the newsletter. In the present approach, the goal is to drive the reader to the website by providing a bundle of content. A set of recommendations is a bundle with items interacting each other (Zhu, Harrington, et al., 2014), a high-quality bundle boosts the user-experience and can increase transaction volume (J. Bai et al., 2019).

Across the literature the most common news selection mechanisms are content-based similarity, collaborative similarity, content-based diversity (Joris et al., 2021), association rule-based, utility-based, knowledge-based (Feng et al., 2021), demographic-based, and hybrid-based (Fayyaz et al., 2020). RS are studied in (Joris et al., 2021) from their reader preference point of view. According to (Joris et al., 2021), people prefer content-based similarity over collaborative similarity and content-based diversity.

Table 22 Literature review on RS in news

Author(s)	Title	Methods and Findings
(Fayyaz et al., 2020)	Recommendation systems: Algorithms, challenges, metrics, and business opportunities	The work provides the landscape of RS research and identifies directions by presenting types of RS, challenges, limitations, and business adoptions.
(Y.-S. Liao et al., 2019)	News Recommendation based on Collaborative semantic Topic Models and Recommendation Adjustment	This research propose a Collaborative Semantic Topic Model and an ensemble model to predict reader preferences. The algorithm improved the recoomendation quality of articles.
(Bangari et al., 2021)	A Review on Reinforcement Learning based News Recommendation Systems and its challenges	The article reviews the different Reinforcement algorithms to develop the news RS and also mentioned the challenges by those algorithms.
(Qin & Zhang, 2021)	Research on News Recommendation Algorithm Based on User Interest and Timeliness Modeling	This work presents a model based on user interest and timeliness. Authors show that this model is superior to the traditional news RS in terms of accuracy and recall rate, also presents higher recommendation performance.
(Yi et al., 2021)	DebiasRec: Bias-aware User Modeling and Click Prediction for Personalized News Recommendation	The research presents a bias-aware personalized RS named DebiasRec to solve the problem of CTR variation influenced by position or size in the online news platform. The method includes a bias representation module, a bias-aware user modeling module, and a bias-aware click prediction module. The authors reduced the bias effect on user interest inference and model training.
(Joris et al., 2021)	Appreciating News Algorithms: Examining Audiences' Perceptions to Different News Selection Mechanisms	The study gives insights into audience's perception to news RS and news selection mechanisms. People prefer content-based similarity over collaborative similarity and content-based diversity. Concerns about information overload and missing viewpoints New concept 'personalized diversity' introduced to slowly guide audiences into a more diverse news diet
(Abdollahpouri et al., 2021)	Toward the Next Generation of News Recommender Systems	*Table d'hôte* approach: set of articles that together can fulfill reader needs for information and joy.

Furthermore, RS present some challenges that need to be detailed and considered when we aim to present a RS for NL. Feng et al. (2021) argue that RS can become a "warped mirror" as result of two types of bottlenecks: technical and moral (Feng et al., 2021).

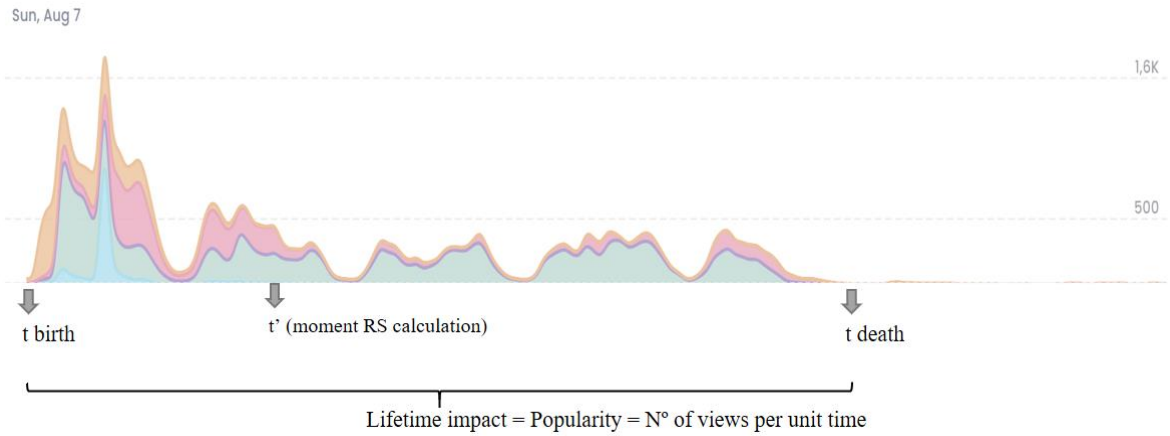
- In a technical perspective, the problem of data sparsity and the cold starts have been addresses by researchers by presenting sparse algorithms (Fayyaz et al., 2020; Huang et al., 2004; Y.-S. Liao et al., 2019; Sun et al., 2009), utility-based news RS (Zihayat et al., 2019), and cold start explicit and implicit solutions (Gope & Jain, 2017; Misztal-Radecka et al., 2021; Rahmat et al., 2020).
- In a moral perspective, some problems are detailed in the literature, the information imbalance or diversity issues, habituation effect (Fayyaz et al., 2020), and information manipulation (Feng et al., 2021). To avoid lack of diversity, surprisal and personalization measures have been proposed (Zhou et al., 2010). Recently the concept of 'personalized diversity' was introduced by (Joris et al., 2021) to slowly guide audiences into a more diverse news diet by using personalization features.

Email newsletters offer to the publishers the chance to maintain a strong direct relationship with readers (Jack, 2016). They provide user personality, provide context and value, content focused, consistency by using automation, despite to be personal it is also an important editorial marketing channel. Thus, Público aims to start automating newsletters to create personalized experiences and activate habits in their digital subscribers. The articles of the recommendation package must be dissimilar as well as, complement each other for a certain purpose (Abdollahpouri et al., 2021). News articles must be recommended on what make a user read in a diverse way, rather than recommend considering clicks (Zihayat et al., 2019) and likes the content has (Joris et al., 2021). Furthermore, the assumption that a click behaviour can indicate user interest can be erroneous as a click could be consequence of other factors, such as the bias of news presentation (Yi et al., 2021). Moreover, the *table d'hôte* approach presented by (Abdollahpouri et al., 2021) propose a combination of articles that the read needs to read (survellant content) and articles that are not related to the reader desire of information but bring him joy and surprise (serendipitous content). Similarly to how a chef put together a set of items in a balanced and enjoyable dining experience, the order of the content is also an important factor to give the user the best experience (Abdollahpouri et al., 2021).

In the surveillance perspective, the popularity of news is often measured by considering the number of interactions in the web social network (Fernandes et al., 2015; Obiedat, 2020). However, other algorithms can be implemented to predict the reader needs, as argued by (Abdollahpouri et al., 2021), this is a research challenge. In the serendipity perspective, content brings surprise, diversion or relaxation can be selected by a wide range of criterions (Mizgajski & Morzy, 2019) or automations (Chakraborty et al., 2017; Garcin et al., 2013). In this chapter we present an approach to address this issue that could be improved in future research.

To increase the propensity to click in the newsletter, we aim to provide a selection of articles that are useful to the reader by providing a saving-time solution. The reader receives a personalized newsletter where he can discover surprizing and informative content. The criteria selection to automatically recommend a story could be influenced or not by past reading history, personalized or non-personalized (Chakraborty et al., 2019). However, in the case on low engaged users, i.e., in a cold-start situation low historical browsing data are available (Misztal-Radecka et al., 2021).

Another important challenge in RS for newsletters is related to how quickly content can become irrelevant due to new developments (Mizgajski & Morzy, 2019). Thus, it is important to consider the lifecycle of a news story to develop a RS in a channel that can be accessed some minutes or hours later. The lifetime of a story is the interval between the moment when the story is published on the website and the instant when the lifecycle is over, as presented at Figure 30. Furthermore, the lifetime-impact is the number of views that the story gets in its entire lifetime (Chakraborty et al., 2019), i.e., is the total area under the popularity curve see Figure 30



Source: Real-time software Compass analytics www.marfeel.com/audience

Figure 30. Popularity lifecycle of a news story.

5.3. Methodology

According to (Suárez, 2020) churn is more important than acquisition. A small improvement on that will have a huge impact on company results (Suárez, 2020). Thus, to provide a data-driven solution to Público, this section presents an ongoing experiment on a particular segment of readers, i.e., readers that present high propensity to churn. Our focus is to improve reader engagement from the newsletter channel to guarantee assertiveness and reader engagement increase in the website.

Despite newsletters only represent 2% of monthly users of the website, readers from newsletters stay longer on the website (more than 6.5 minutes by visit), they see more than 4 pages by session, and visit the website approximately 4 times per month. While the overall average of the website users see 2.4 pages by session, stays 3 minutes per visit, and do less than 2 website visits per month. Furthermore, research shows that micro-segments are efficient acquisition channels (Suárez, 2020).

5.3.1. Research motivation, continuous experimentation, and improvement

The present experience started after running a subscriptions churn model at Público. Findings revealed that, when a Público subscriber presented less than 42% of active days (AD) for 60 days, the propensity to churn strongly increases. Thus, the team decided to send an automatic weekly newsletter to those subscribers that verify the pattern, i.e., subscribers with low engagement.

The *first experiment* consisted in deliver a bundle of articles based on a simple news selection mechanism (SNSM), i.e., select N articles from the most read by all subscribers, that have high lifecycle (evergreen content published in the last 3 days) that those subscribers didn't visit yet. Those decisions were based into premises: one related to the engagement definition presented at section 4.3.2 (Ksiazek et al., 2016; Mersey et al., 2010). Thus, content that involved many subscribers' will have high propensity to be of other subscribers interest. Futhermore, content that catch the attention of the subscriber, i.e., long articles mostly evergreen have preference to be recommended. Thus, as the goal was to provide serendipitous content, unexpected and diverse (Kotkov et al., 2016), the content was selected from the sub brands Fugas, íPsilon, P3, Culto and Azul. Respectively, soft news journalism related to travel, culture, wellbeing, environment, and opinion, that represents 35% to 40% of total content produced.

The NL template was designed together with the design team (see Figure 31) to guarantee a good user experience, and the right content balance, on desktop and mobile, by giving more emphasis to the Público sub brands logo, article main picture, title and subtitle. Furthermore, to avoid news bias regarding boxes size (Yi et al., 2021), all the content is presented in the same format size and structure.

Results of 16 weekly newsletters sent to an average of 9,905 low engaged subscribers, across the first semester of 2022, showed an average of 45% open rate (OR) and 6% of click through rate (CTR). A better performance than the average values registered at Público newsletters (25% OR and 5% CTR), and better than the benchmark values reported by (Mailchimp, 2022) for the media industry (22% OR and 5% CTR).

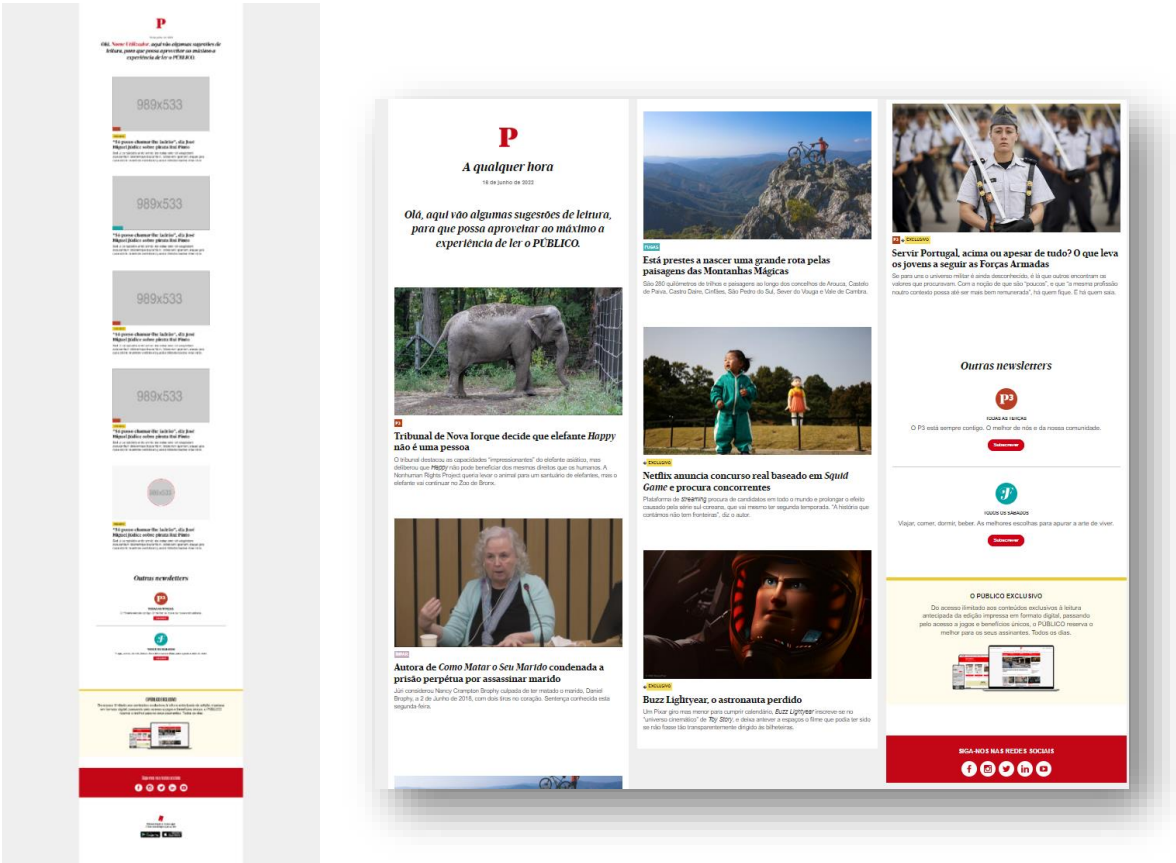


Figure 31. Engagement newsletter template, an example can be found here.

Furthermore, 55% of the subscribers that clicked on the newsletter increased AD, 18% kept the same AD, while 27% decreased AD in the next 60 days. Thus, subscribers were induced to read more. Moreover, from those subscribers that had a subscription end in the next 60 days after open at least one newsletter, 70% renewed while those that did not open the newsletters 66% renewed. Findings revealed that act on those low engaged subscribers with a content email induces an engagement recover.

Furthermore, to improve CTR and consequent website visits, a *second experiment* was performed for continuous improvement. The marketing team decided to ask a provider for a ML model to compete with our SNSM approach. The experiment consisted of applying a A/B test in four newsletters sent in a weekly basis, between 27th July and 17th August 2022. The reader's sample was a group of register readers that made only one website visit in the last 30 days. The provider ML model considers all the content published by Público to build the bundle of articles.

Results revealed no statistically significant difference between both CTR after 24 hours, as the ML model CTR was 6.88%, 7.07%, 6.58%, 6.19% while the SNSM CTR was 6.99%, 8.06%, 6.89%, and 7.77%. However, the SNSM after one week achieves more 1p.p. in the CTR on average. Furthermore, after two days of the newsletter delivery, the CTR of the newsletter with the SNSM approach still increasing, while CTR of the newsletter with ML decreases. This indicates that the content recommended by SNSM is more engaging across more days. Thus, the lifetime of the content delivered is an important factor as a newsletter is not a real-time channel that could be accessed some days after send. Therefore, those results drove the authors to propose a RS inspired on the table d'hôte RS suggested by (Abdollahpouri et al., 2021).

The methodology followed was inspired by the CRISP-DM methodology (Moro et al., 2011). Data analysis and modelling was performed through Colaboratory a product from Google Research that is a hosted Jupiter notebook service with access to Google hardware (Nelson & Hoover, 2020).

5.3.2. Data collection and data preparation

The data used in this research was collected from Público database. When a logged user reads an article, it is tracked on the website, and then is recorded as a hit that represents a row in the data warehouse (e.g. BigQuery by Google). Furthermore, a hit contains information like *date, time, user id, user environment variables, article url,* and e*vents* (Google, 2022) (such as, social share, click in recirculation boxes, and etc). However, clickstream data collection contains unrelated information and some noisy data. Thus, data processing and cleaning is a pivotal step before data modelling. The list of attributes collected for each article is presented at Table 23. Three main tables of the database were combined: table of the articles and their main characteristics, table of website hits, and table of tags. By article we also added the new variable *homepage post* that indicates if journalists decided to post the article in the homepage, it is an indicator of journalistic relevance level of the content.

5.3.3. Data modelling and future work

Create a *table d'hôte* RS is a multi-objective process (Abdollahpouri et al., 2021). The list of recommended news results of two main elements combination: **Surveillance** and **Serendipity**. As presented at Figure 32, the first three articles are information that reader needs, and the last three are stories offer a surprise to the reader. Calculations are detailed as follows.

Firstly, to define the three articles that each reader needs, we calculate the expected value of views for new articles. From the articles expected to have high number of views, we excluded those that the reader already reed, the top three are sent to the reader.

Table 23. List of attributes

Feature group	Feature	Type	Description
Article characteristics	**Title characters**	number	Number of characters in the title
	Tags	number	Number of tags
	Date day	number	Day of the week published
	Hour	number	Hour published
	Antiquity	number	Nº of article days
	Autor ID	number	Autor ID
	Premium	number	Article Premium (y/n)
	News type ID	number	Article type defined by the journalist from the list: Noticia, Listicle, Reportagem, Perguntaserespostas, Imagem_semana, Foto_legenda, Shopping, Depoimento, Opiniao, Entrevista, Cronica, Analise, Listicle _Number, Perfil, Cronica_de_jogo, Newsletter, Prepublicacao, Critica, Comentario, Editorial, Investigacao, Ensaio, Ficha
	Section	number	Editorial section: Cultura, Mundo, Economia, Azul, Sociedade, Política, Culto, Ciências, Fugas, p3, Local, Publico Desporto, Tecnologia, Terroir, Newsletters, Iniciativas Público, Opinião, Público na Escola
Traffic performance metrics	Users	number	Number of Users
	Views	number	Number of Views
	Logged Users	number	Number of logged users, users that are registered into the website
Article content relevance	**Homepage post**	number	Article posted at the Homepage (y/n)
	Views of the main tag	number	Average pageviews of the main tag

To predict the views by article, we apply daily the XGBoost algorithm to model the number of views considering the features highlighted at Table 23. Then, for articles published in the last N hours the number of views is predicted. Some experiments could be performed to better define the better N, also that depends on the hour of the day that the newsletter is send. It is recommended to run some experiments across the day to find the best hour to send.

Secondly, to define the three surprising articles by user we use the SNSM approach. Then, by reader, we concatenate the respective articles (see Figure 32). As presented at Table 24, two models were built. However, as a result of technological challenges, at the moment that we wrote this thesis, it was not possible to have ready the live controlled experiment with real readers. Future work will be to test this approach and measure the impact of the newsletter on reader's reading behaviour.

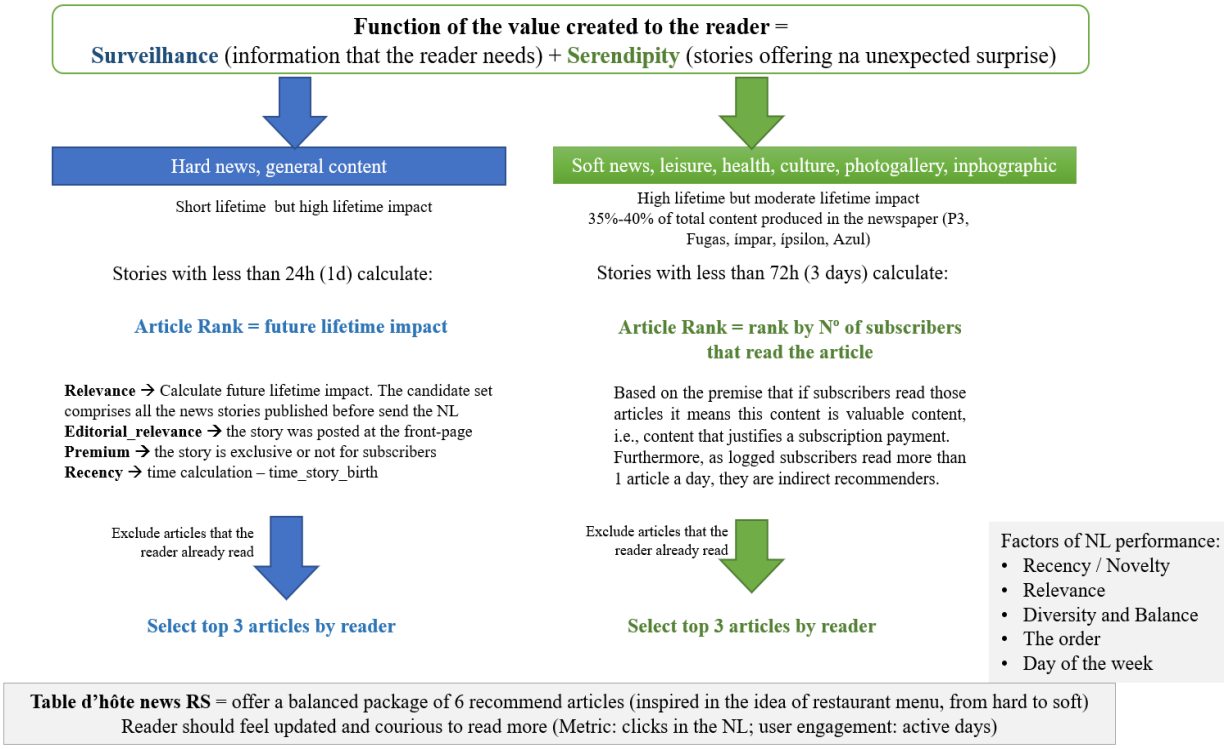


Figure 32. Multi-objective process to build a news recommendation list by reader.

Table 24. XGBoost model evaluation

	Mean absolute error	Mean squared error	Mean squared log error	Median absolute error	R2 score
Model 1	4,868	210,534,291	2.77	1,911	0.485
Model 2	**4,306**	**133,022,393**	**2.58**	**1,584**	**0.648**

5.4. Discussion and future research opportunities

The *second experiment* already demonstrated that long lifecycle articles with serendipitous content induce higher CTR. As the reader can open the newsletter some hours and days after the sending date, content with short lifecycle could decrease reader newsletter loyalty. This finding is consistent with the work of (Kotkov et al., 2016). Serendipitous content is relevant while novel content could be irrelevant (Kotkov et al., 2016). Thus, the newsletter should be a

combination of two strategies to provide the news-diet that user needs with diversity and balance, as argued by (Abdollahpouri et al., 2021).

The engagement newsletter is sent weekly to low engaged subscribers (around 12 thousand) since January 2022 automatically. On average, 256 users click at least once in the newsletter and by newsletter 2,825 pageviews are generated. This increases the number of active days on low loyal subscribers and the perceived value (Hsiao & Chen, 2017).

After, more that forty weeks, there was a slightly decrease in performance to 43.4% open rate (OR) and 5.5% of click through rate (CTR). Thus, future work will be focused on improve the bundle generation and the newsletter subject that it is always the same. Some text mining approach could help to improve subject to increase open rate. Furthermore, the algorithm to predict article views could be improved, and other features could be added to increase CTR and open rate. Also, data from the real-time software could be added.

5.5. Conclusions

In this section, we present some experiments around newsletters content automation to increase reader engagement and to developed a save-time solution for the newsroom. The main goal it is to improve the present approach by using the *table d'hôte* and *personalized diversity* concepts adapted to the publisher context (that allow us to answer RQ1). In one hand, we aim to provide surveillant content that are the news that must be read. On another hand, readers expect serendipitous content that bring them surprising news. We believe that the present approach will increase reader engagement as the newsletter that only contains serendipitous content already shown an increase of active days from subscribers more propense to churn (this will allow us to answer RQ2). However, future work should improve the assertiveness of the content selected by analyzing which are the serendipitous news that are more joyful, a text mining approach with tags and keywords study can provide new directions. Furthermore, we should consider an analysis of the best day and hour of the week to send the newsletter. Moreover, in the experiment were sent six articles in the newsletters, however other bundle sizing can be tested considering desktop and mobile browser.

Thus, future work will be focused on improve the bundle generation and the newsletter subject that it was always the same. Some text mining approach could help to improve subject to increase open rate. Furthermore, the algorithm to predict article views could be improved, and other features could be added to increase CTR and open rate. Also, data from the real-time software could be added.

CHAPTER 6

Conclusions

6.1. Overview

Online publishers face enormous challenges to survive and become sustainable in a competitive digital market. The technology development and the constant reader change behavior require efficient editorial and marketing digital actions (Suárez, 2020). Furthermore, define successful user engagement, acquisition and retention strategies as become fundamental to increase profitability. Thus, data analytics plays a pivotal role by presenting solutions to optimize engagement, acquisition, and conversion KPI's.

There was a hype of DS in DJ research since 2017. As presented at Chapter 1, the main motivations and major topics to adopt DS in DJ (MRQ1) vary between exploratory studies, web analytics research, to more advanced analytics, such as TM methods or RS. Furthermore, authors agree that more can be done to improve personalization and investigate innovative business models. Thus, to answer MRQ1, the present research analyzed the literature published at Scopus between 2010 and 2021. Results allowed us to guide our research and present some data-driven actions to Público.

One research gap found was the lack of research around micro-segments that are efficient acquisition channels (Suárez, 2020). As an example, Instagram research on its strategic use is rather then scarce. Thus, Chapter 3 presents a data-driven strategy to engage those readers (MRQ2). Interactive content and vivid content involve readers and proved to increase reader engagement., Furthermore, the data-driven strategy with effective information delivery and informative dashboards, allowed to improve the decision making of the social media managers. Thus, this experiment made an important contribution on the data culture across the newsroom.

Besides, to develop effective strategies, we identified readers segments and the main drivers of purchase intention (MRQ3). From the literature, engagement variables were defined. Furthermore, eight clusters were identified. Then, editorial and marketing actions were proposed to increase engagement and conversion. Finally, to answer MRQ4 and considering the churn drivers, we tested an automated newsletter than increased the active days of subscribers.

As mentioned at Section 1.3, the Data Science lifecycle management (DSLM) (Abonyi et al., 2022) is a framework that supports the development of ML algorithms to develop, maintain and manage the whole lifecycle. Hence, one of the main goals in this research is to measure the impact of the actions applied across the reader funnel, at each chapter results were presented and proved engagement increase.

In addition, the Economic Success Criteria mentioned by (Abonyi et al., 2022) it is a criteria in the form of a KPI to the project. Economical measures for the relevance of the applications proposed can be defined as "cost saving with the segmentation from an external provider", "cost saving with real-time software for social media", "cost saving with newsletter algorithm from an external provider", "revenue from Instagram ads" as Público is national leader, "revenue from personalized newsletter" or brand awareness increase.

Despite Público saved those costs, the value delivered by the present research can be indirectly inferred by the equation presented at Section 1.5. Across the years of this research, the number of use cases and utility increased as we started from no DS approach to a newsroom that asks for reports and KPi's. The dashboard' users increased as result of a data-driven culture implemented. Moreover, as published by Público on 30th November 2022, Público is the Portuguese publisher with more digital subscriptions verifying a consistent increase across the years (Público, 2022). According to Netaudience, the national system to measure and compare audiences, Público was the Portuguese news media platform with more visits in 2020 and 2021 (Marketest, 2022).

6.2. Discussion

According to Abraham (2019), Data Governance is the exercise of authority and control over the management of data, that implements a corporate-wide data agenda to maximize the value from data and manages data-related risks (Abraham et al., 2019; DAMA, 2017; Koltay, 2016). Furthermore, the success of DS methods depends on the maturity level of each section of the Data Governance framework (see Figure 34) (DAMA, 2017; Koltay, 2016). Moreover, data will continue to grow, various forms of data from diverse sources may come and create the need for a common understanding of data across the company (Abraham et al., 2019). Therefore, a good data management and data strategy is needed to use the data more effectively and keep delivering value from data (Brous, Janssen, & Krans, 2020).

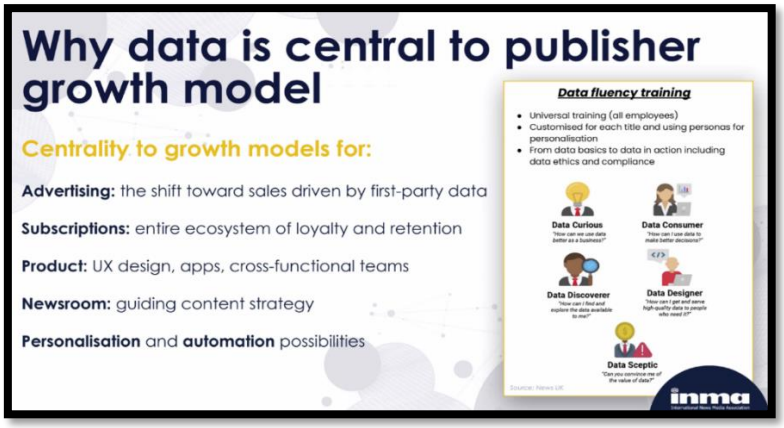
Despite that Público achieved positive results, the use of DS faced some resistance and challenges related to the data culture, data strategy, and data governance. The implementation of DS gets blocked by the lack of improvement on the main areas of the Data Governance framework. As an example, if data quality is monitored and controlled, DS will be more likely to have useful outcomes (Brous, Janssen, & Krans, 2020).

Media researchers agree that data is central to publishers. As presented at Figure 33a, data induces the development of important sources of revenue, such as advertising and subscriptions by allowing to improve product, guiding the newsroom strategy, and promoting personalization. Meanwhile, some missed opportunities have been reported (see Figure 33a), such as, reader engagement, reader revenue, advertising revenue, and technology (Schmidt, 2022).

The present research focuses on Público reader engagement actions by applying DS methods that consequently increases revenue. However, in a big data environment, Público has some improvement opportunities on data governance that impact on DS success.

To measure the maturity level of the data strategy at Público, and to do a diagnosis, we applied a methodology developed by a well-known group of Spanish data leaders (CDO Club Spain & Latam, 2022). The *dataMat* is a tool that consists of a questionnaire to measure 21 ambits of four main areas: data governance, data strategy, advanced analytics, and data visualization or reporting (see Figure 36 (CDO Club Spain & Latam, 2022)). For each ambit, the tool calculates a score that varies from 1 to 5. Low values indicate the main areas to improve.

The questionnaire was answered by the Business Analyst, Data Engineer, and Head of Analytics and Audience of Público. Results allow us to understand the main ambits to improve (see Figure 36a) and how to prioritize those elements (see Figure 36b). As an example, at Figure 36, it is possible to observe that Data Monetization and Data Ethics verify the lowest score (1). Oppositely, Change Management (3.6), and Self-service and Data Discovery (3.5) present the highest values. Another important insight is that, to improve the Advanced Analytics area (at yellow in the diagram) more effort must be done in the Data Governance area (green box). The maximum score at the Data Governance area is 2.8 related to Information Security. As consequence, the evolution of advanced analytics depends on the improvement of Data Governance actions.



(a)

(b)

Figure 33 Data Culture at publishers (a) "Why data is central to publishers' growth model" from INMA; (b) Data Maturity for Publishers (Schmidt, 2022).

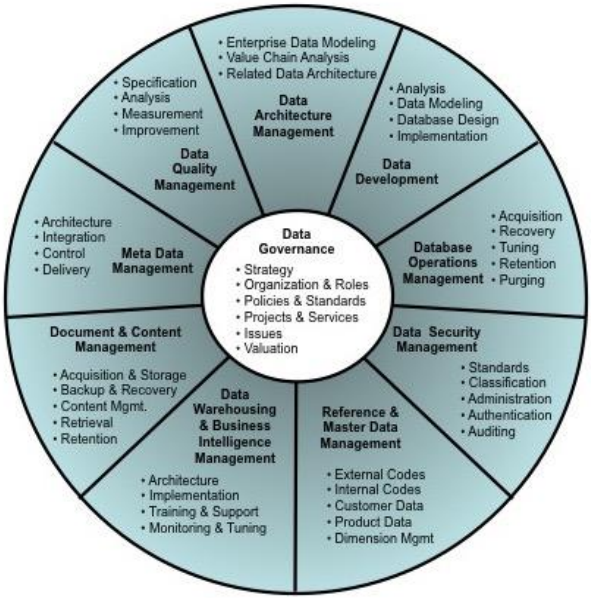


Figure 34 Data Governance Framework DAMA Internacional.

Furthermore, the Self-service and Data Discovery presents the highest score as result of the data democratization strategy implemented and designed by the PhD researcher. The main goal was to increase the value delivered by data, as presented at the equation of Section 1.5, we aimed to increase the number of users and the strategic value of the Data Science work developed. Thus, a landing page with the structure of the reader funnel (Kotler et al., 2016) was provided to the users (see Figure 35), that was also mentioned at Figure 17. This design allowed the users to find the dashboards available to increase reach, engagement, and conversion. Also, the user understands where his work is impacting on the funnel. For instance, the dashboards developed to implement the data driven strategy presented at Chapter 3, the dashboard with the segmentation of readers presented at Chapter 4, or the report to monitor experiment results with the newsletter presented at Chapter 5, where allocated in the Landing page according to the main contribution in the funnel, respectively, increase website reach (Aware), increase user conversion (Act), and increase subscriber engagement (Advocate) (see Figure 35).

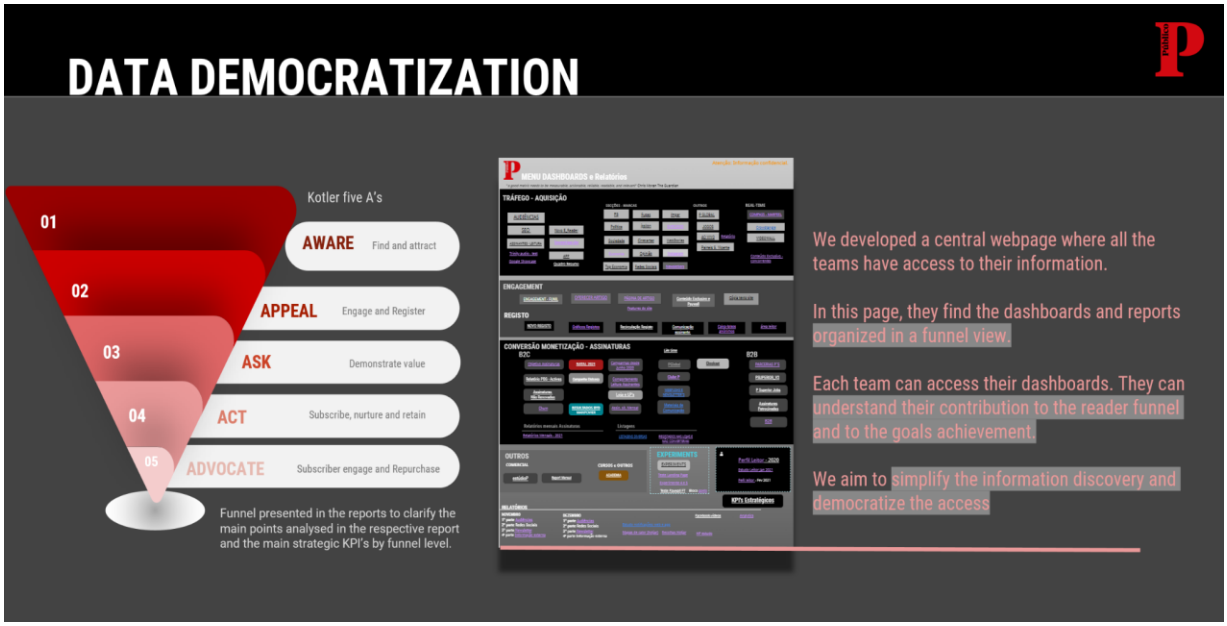
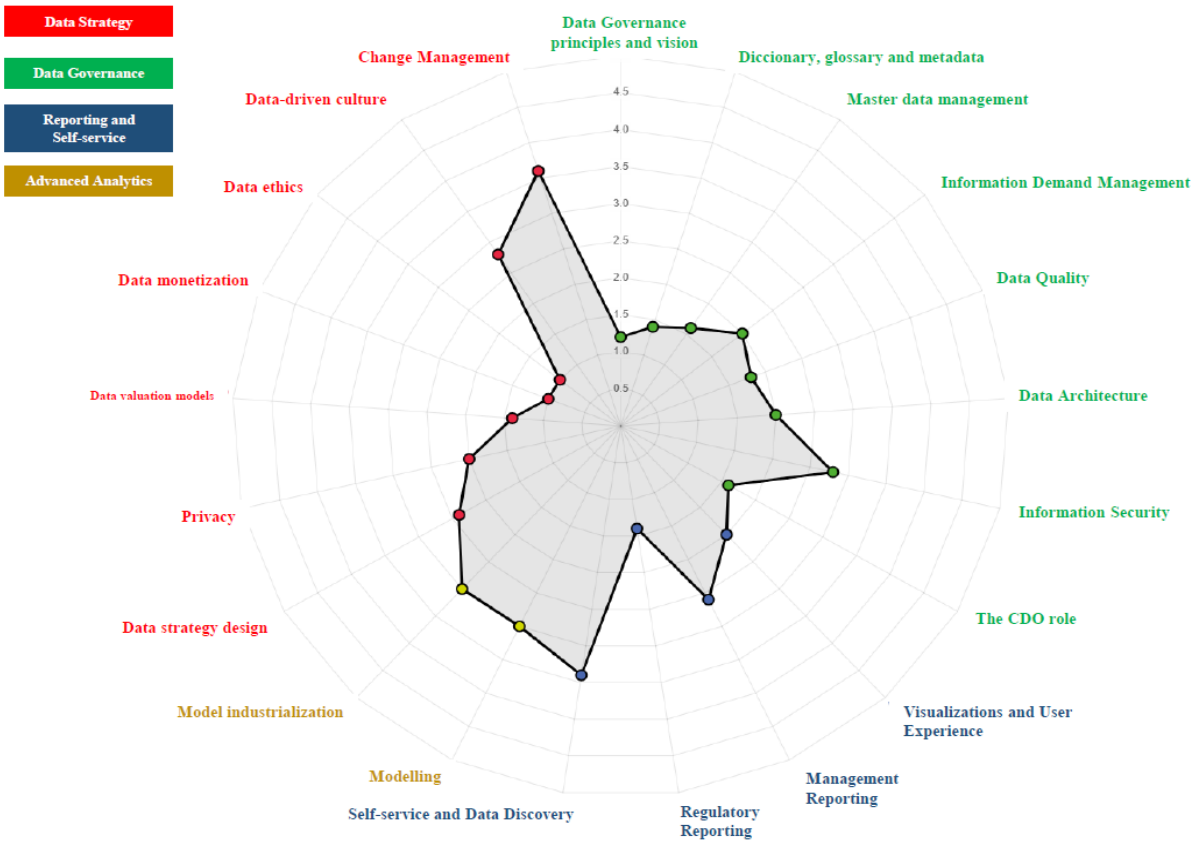
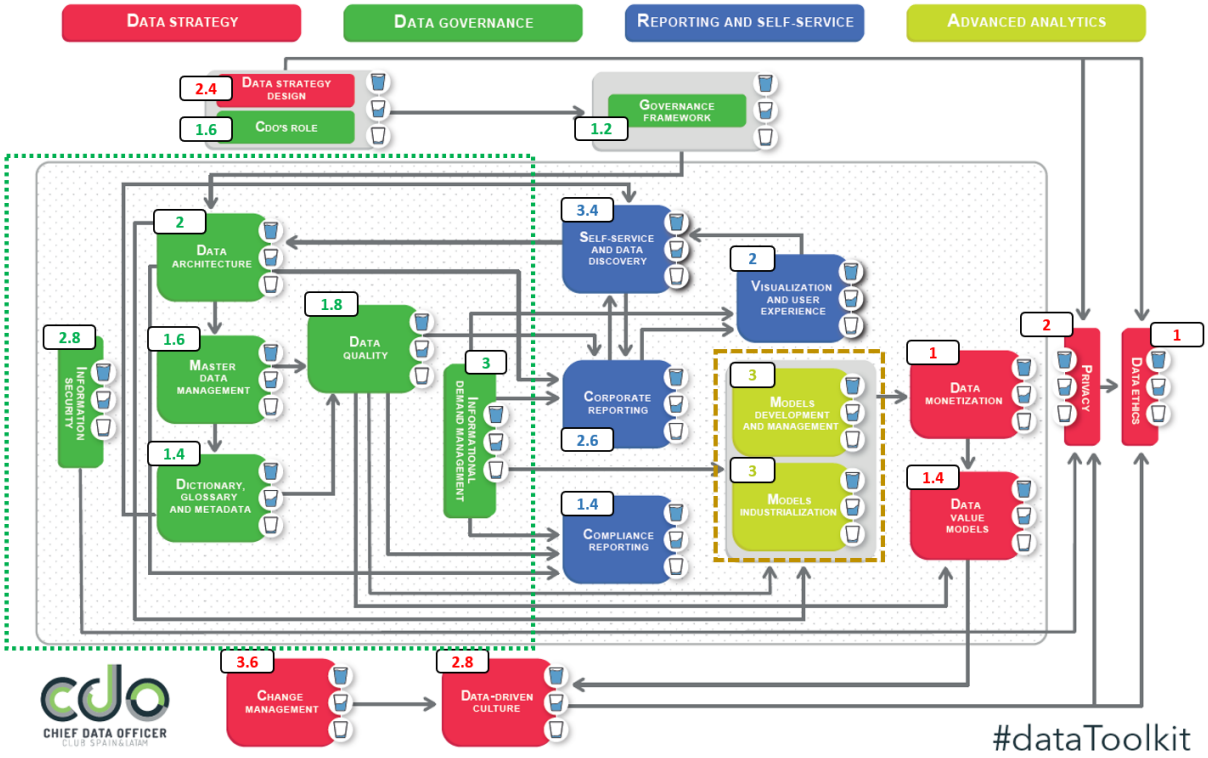


Figure 35. Global Público landing page to increase data culture and data-driven decision making.

Regarding the detection of the main areas of improvement, future work relies on implement Data Governance and Data strategy projects that increase the maturity level of the overall Data framework to lead Público to achieve another level in Advanced Analytics and DS innovation. Furthermore, investment on how to monetize data will introduce new sources of revenue that can be potentiated by Público.



(a)



(b)

Figure 36 Data Management, Governance and Strategy Diagnosis; (a) radar graph with the main areas under analysis; (b) Map with the main areas and calculated score.

6.3. Contributions and implications

The present research provides important contributions to the literature. The literature review performed is replicable and helps researchers to find research opportunities in a wide range of industries. Furthermore, DJ researchers can be inspired on the present literature review to focus on new research lines and develop real business solutions to contribute on the sustainability challenge of publishers. Moreover, the literature review combines TM methods, science mapping analysis (SMA) by using VOSviewer (Donthu et al., 2021; Van-Eck & Waltman, 2010) and *bibliometrix,* the R-tool for comprehensive science mapping analysis (Aria & Cuccurullo, 2017), that results in a novelty approach.

The main areas of research opportunities comprise recommend systems, Big Data, personalization, content automation, fact-checking, engagement metrics, and new paywall mechanisms.

Despite that the present research was driven by Público business needs, data sources available, and limited technical conditions, the researcher focused on some of these topics. Furthermore, in a real context, the researcher also had to face Público daily challenges, provide insights to help on company strategy definition, and adapt the analytical approach to be understandable and useful.

By looking at the key topics that raised from the literature review, the researcher presented a list of engagement metrics used to segment readers at Chapter 4. Furthermore, personalization and content automation experiments in newsletters were presented at Chapter 5. Moreover, all the experiments were implemented in a big data context for real-time decision making. As presented at Chapter 3, internal and external data was combined to perform the analysis.

Meanwhile, with limited analytical and technological resources it was possible to develop a framework that mobilized editorial and marketing teams to be more data-driven and open to experimentation in a fast-paced digital context.

Unsurprising, the literature presents more studies on text and less in visual content for SM. However, as IG is a visual SM platform, the researcher explored the advantages of vividness content to reach wider audience, increase reader engagement and consequently convert them into subscribers by keeping the journalistic values. The analysis performed at Chapter 3 comprises statistical tests, sentiment analysis and hashtag analysis. Important contributions raised that are aligned with journalistic principals, such as, the important role of the narrative to engage readers, and the possibility to fight misinformation that increases bran trust and brand awareness.

Additionally, at Chapter 4 an effective reader segmentation approach was developed by considering the business context, with strong focus on retention and engagement management, as well as the need of an algorithm that is understandable by the marketing team. The approach presents eight clusters that are daily updated on Big Query (BQ). From an operational perspective, it is practical and efficient to design customised strategies. Furthermore, this allows the team to monitor the number of readers by cluster, their levels of engagement, and conversions rates to improve marketing and editorial actions. Moreover, the main engagement drivers found, i.e., the volume of premium content and scroll down level, drove the team to test new paywall designs and communications. Also, the possibility to run the XGBoost ML model weekly or monthly and see the results on the Data Studio dashboard (see Figure 26), could potentially help the team to understand reader's change behaviours and improve conversion strategies.

6.4. Limitations and future research

Despite this research present useful DS methods to deliver value in a real publisher, the study was not immune to limitations. The high level of exigence to deliver information daily to face Público needs limited the resources available to perform analytical experiments.

From the main areas of research opportunities detailed at the literature review, improving paywall mechanisms and fact-checking are two examples that could be developed for future work. Furthermore, Table 25 presents the main findings and future research proposed. The scope of future work lies into enhancing and developing algorithms for content automation in NL, improving content distribution by using algorithms to distribute content on SM, also fostering paywall mechanisms to increase assertiveness.

Although the present research was of Público interest, as it presents analytical solutions to improve business results that showed proved results, more can be done. By looking at the scale of maturity level of subscriptions analytics (see Figure 37) presented by Greg Piechota at the institute for media strategies in 2018, future research opportunities are detailed to achieve level 4. Público could invest on advanced analytics to increase automation to better perform at each channel and device. Experiments should be improved and potentiated. Also, to be more competitive, algorithms to present different prices and offers could be developed. Moreover, improved communication by using AI could increase engagement and brand awareness.

Table 25. Generalizable findings and future research

Generalizable Findings	Future Research
The main areas of research are recommendation systems, personalization, content automation, Big Data, fact-checking, engagement metrics, and new paywall mechanisms	• Improve and study new paywall mechanisms • Develop novelty algorithms for fact-checking • Invest on personalization algorithms (as an example, differentiate between desktop and mobile) • Improve content recommendation systems
Lack of Data governance can compromise operational efficiency, revenue increase and costs control (Brous, Janssen, & Krans, 2020)	• Explore how are publishers embracing the data governance main areas to guarantee good data management
Low research has been done to mobile traffic regarding incidental news (van Damme et al., 2020)	• Improve UX and RS across the website for mobile users
Low social media research in DJ with the use or implementation of DS methods	• Explore ML or AI algorithms to better distribute content on SM platforms to increase reach and engagement
Kmeans was an effective algorithm to segment readers based on engagement and retention management. Also, it is easy to understand regarding the data maturity level of the company.	• Other segmentation algorithms could be developed and tested
Despite that newsletter emerged as a big source of incoming online traffic (Hendrickx et al., 2020) there is a lack of research on RS adapted to this channel (Abdollahpouri et al., 2021).	• Research on RS for publishers NL

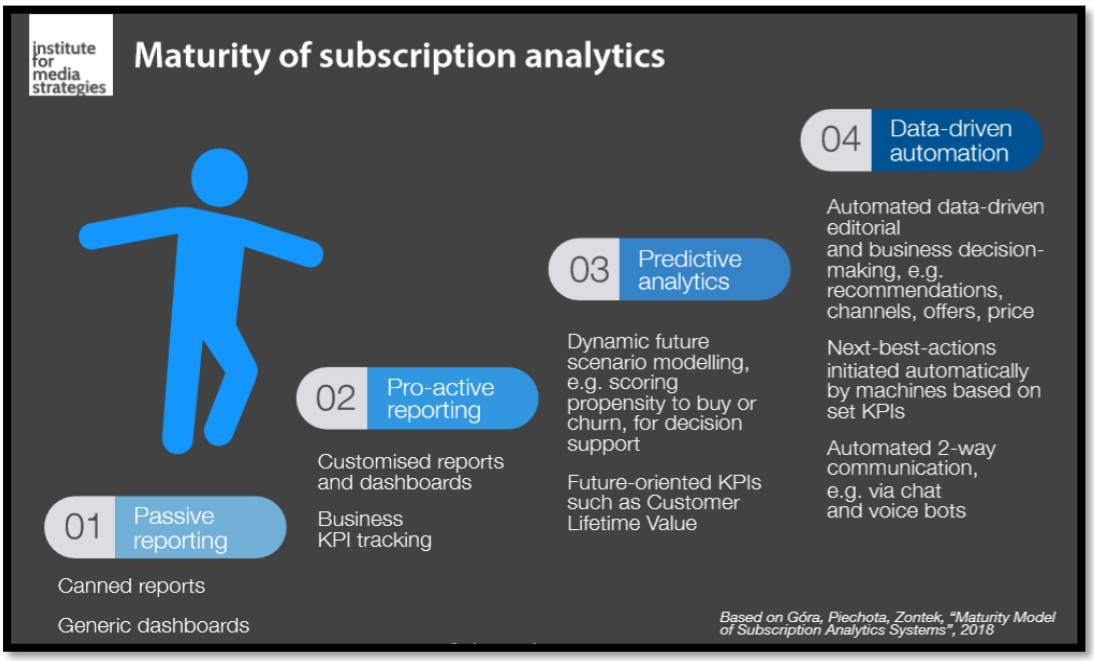


Figure 37 Maturity level of subscriptions analytics.

Concerning the maturity level of subscriptions analytics, it is interesting to note that, at the beginning of this research Público was on level 1 and now it is on level 3. Part of this path, of continuous data culture increase, data analytics development and advanced analytics research is presented in this document. The present research can be used by other publishers that are starting their digital analytics transformation. Our findings can guide other researchers or analytics professionals in the media industry and can be adapted to other digital industries.

Despite the high importance and contribution of the explored advanced analytics methods, data visualization played a pivotal role to improve data culture and data-driven decision making. data storytelling and the analysis of data visualizations performance it is also an important area of future work.

References

Abdelmageed, S., & Zayed, T. (2020). A study of literature in modular integrated construction - Critical review and future directions. *Journal of Cleaner Production*, *277*, 124044. https://doi.org/10.1016/j.jclepro.2020.124044

Abdollahpouri, H., Malthouse, E. C., Konstan, J. A., Mobasher, B., & Gilbert, J. (2021). Toward the Next Generation of News Recommender Systems. *The Web Conference 2021 - Companion of the World Wide Web Conference, WWW 2021*, 402–406. https://doi.org/10.1145/3442442.3452327

Abonyi, J., Kummer, A., & Hanzelik, P. (2022). Edge-Computing and Machine-Learning-Based Framework for Software Sensor Development. *Sensors 22*, *11*(4268).

Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. In *International Journal of Information Management* (Vol. 49, pp. 424–438). Elsevier Ltd. https://doi.org/10.1016/j.ijinfomgt.2019.07.008

Alashri, S., Tsai, J. Y., Koppela, A. R., & Davulcu, H. (2018). Snowball: Extracting causal chains from climate change text corpora. *Proceedings - 2018 1st International Conference on Data Intelligence and Security, ICDIS 2018*, 234–241. https://doi.org/10.1109/ICDIS.2018.00045

Amado, A., Cortez, P., Rita, P., & Moro, S. (2018). Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. *European Research on Management and Business Economics*, *24*(1), 1–7. https://doi.org/10.1016/j.iedeen.2017.06.002

António, N., Almeida, A. de, & Nunes, L. (2018). Predictive models for hotel booking cancellation: A semiautomated analysis of the literature. *Tourism & Management Studies International Conference TMS Algarve*.

Antoun, W., Baly, F., Achour, R., Hussein, A., & Hajj, H. (2020). State of the Art Models for Fake News Detection Tasks. *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, 519–524.

Aral, S., & Dhillon, P. S. (2021). Digital paywall design: Implications for content demand and subscriptions. *Management Science*, *67*(4), 2381–2402. https://doi.org/10.1287/mnsc.2020.3650

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Arrese, Á. (2016). From Gratis to Paywalls: A brief history of a retro-innovation in the press's business. *Journalism Studies*, *17*(8), 1051–1067. https://doi.org/10.1080/1461670X.2015.1027788

Attfield, S., Kazai, G., & Lalmas, M. (2011). Towards a science of user engagement (Position Paper). *WSDM Workshop on User Modelling for Web Applications*. http://www.dcs.gla.ac.uk/~mounia/Papers/engagement.pdf

Aydin, G. (2020). Social media engagement and organic post effectiveness: A roadmap for increasing the effectiveness of social media use in hospitality industry. *Journal of Hospitality Marketing and Management*, *29*(1), 1–21. https://doi.org/10.1080/19368623.2019.1588824

Azevedo, L. (2018). Truth or Lie: Automatically Fact Checking News. *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*, 807–811. https://doi.org/10.1145/3184558.3186567

Babanejad, N., Agrawal, A., Davoudi, H., An, A., & Papagelis, M. (2020). Leveraging emotion features in news recommendations. *INRA@ RecSys*, *2554*, 70–78.

Bai, J., Zhou, C., Song, J., Qu, X., An, W., Li, Z., & Gao, J. (2019). Personalized Bundle List Recommendation. *The World Wide Web Conference*, 60–71.

Bai, X. (2011). Predicting consumer sentiments from online text. *Decision Support Systems*, *50*(4), 732–742. https://doi.org/10.1016/j.dss.2010.08.024

Balali, A., Faili, H., Asadpour, M., & Dehghani, M. (2013). A supervised approach for reconstructing thread structure in comments on blogs and online news agencies. *Computación y Sistemas*, *17*(2), 207–217.

Ballew, B. (2009). Elsevier's Scopus® Database. *Journal of Electronic Resources in Medical Libraries*, *6*(3), 245–252.

Bangari, S., Nayak, S., Patel, L., & Rashmi, K. T. (2021). A Review on Reinforcement Learning based News Recommendation Systems and its challenges. *Proceedings - International Conference on Artificial Intelligence and Smart Systems, ICAIS 2021*, 260–265. https://doi.org/10.1109/ICAIS50930.2021.9395812

Barriuso, A. L., de La Prieta, F., Murciego, Á. L., Hernández, D., & Herrero, J. R. (2016). An Intelligent Agent-Based Journalism Platform. *International Conference on Practical Applications of Agents and Multi-Agent System*, 322–332.

Bermudez-Edo, M., Barnaghi, P., & Moessner, K. (2018). Analysing real world data streams with spatio-temporal correlations: Entropy vs. Pearson correlation. *Automation in Construction*, *88*(January 2021), 87–100. https://doi.org/10.1016/j.autcon.2017.12.036

Bisong, E. (2019). Google bigquery. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform* (pp. 485–517).

Blazejewski, P. (2019). *User Engagement drives subscriptions —RFV user engagement scores from Deep.BI*. https://medium.com/deep-bi/user-engagement-drives-subscriptions-rfv-user-engagement-scores-from-deep-bi-8aa1ed23a923

Borges, A. F. S., Laurindo, F. J. B., Spínola, M. M., Gonçalves, R. F., & Mattos, C. A. (2021). The strategic use of artificial intelligence in the digital era : Systematic literature review and future research directions. *International Journal of Information Management*, *57*(September 2020), 102225. https://doi.org/10.1016/j.ijinfomgt.2020.102225

Brous, P., Janssen, M., & Herder, P. (2020). The dual effects of the Internet of Things (IoT): A systematic review of the benefits and risks of IoT adoption by organizations. *International Journal of Information Management*, *51*(September 2018), 101952. https://doi.org/10.1016/j.ijinfomgt.2019.05.008

Brous, P., Janssen, M., & Krans, R. (2020). Data Governance as Success Factor for Data Science. In *Responsible Design, Implementation and Use of Information and Communication Technology: 19th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society* (Vol. 12066, pp. 431–442). Springer International Publishing.

Brownlee, J. (2016). XGBoost With Python: Gradient Boosted Trees With XGBoost and scikit-learn. In *Machine Learning Mastery*. Machine Learning Mastery.

Budiman, K., & Akhlis, I. (2021). Changing user needs and motivation to visit a website through ad experience: A case study of a university website. *Journal of Physics: Conference Series*, *1918*(4), 042008.

Burggraaff, C., & Trilling, D. (2017). Through a different gate: An automated content analysis of how online news and print news differ. *Journalism*, *21*(1), 112–129.

Burrows, S., Potthast, M., & Stein, B. (2013). Paraphrase Acquisition via Crowdsourcing and Machine Learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *4*(3), 1–21.

Cage, J., Herve, N., & Mazoyer, B. (2020). Social Media and Newsroom Production Decisions. *SSRN Electronic Journal*, *3663899*. https://doi.org/10.2139/ssrn.3663899

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, *3*(1), 1–27.

Campos, J., Teixeira, A., Ferreira, T., Cozman, F., & Pagano, A. (2020). Towards Fully Automated News Reporting in Brazilian Portuguese. *Anais Do XVII Encontro Nacional de Inteligência Artificial e Computacional*, 543–554. https://doi.org/10.5753/eniac.2020.12158

Campos, & Rodríguez, J. (2020). *El valor del dato: La brújula para gestionar tus datos como un activo* (1st editio). Valor del Dato.

Canito, J., Ramos, P., Moro, S., & Rita, P. (2018). Unfolding the relations between companies and technologies under the Big Data umbrella. *Computers in IIdustry*, *99*, 1–8.

Carlson, M. (2014). The Robotic Reporter Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. *Diital Journalism*, *3*(3), 416–431.

Carlson, M. (2017). *Journalistic authority: Legitimating news in the digital era* (C. U. Press, Ed.).

CDO Club Spain & Latam. (2022). *dataMat*. DataMat by CDO Club Spain & Latam. https://clubcdo.com/en/datamat/

Chakrabarty, N., Rana, S., Chowdhury, S., & Maitra, R. (2019). RBM Based Joke Recommendation System and Joke Reader Segmentation. In *International Conference on Pattern Recognition and Machine Intelligence* (Issue January 2020). Springer International Publishing. https://doi.org/10.1007/978-3-030-34872-4_26

Chakraborty, A., Ghosh, S., Ganguly, N., & Gummadi, K. P. (2017). Optimizing the recency-relevancy trade-off in online news recommendations. *26th International World Wide Web Conference, WWW 2017*, *i*, 837–846. https://doi.org/10.1145/3038912.3052656

Chakraborty, A., Ghosh, S., Ganguly, N., & Gummadi, K. P. (2019). Optimizing the recency - relevance - diversity trade - offs in non - personalized news recommendations. *Information Retrieval Journal*, *22*(5), 447–475. https://doi.org/10.1007/s10791-019-09351-2

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Rudiger, W. (2000). Crisp-Dm 1.0. *CRISP-DM Consortium*, *3*, 76.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. *CRISP-DM Consortium*.

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2022). SMOTE: Synthetic Minority Over-sampling Technique. *Journal Of Artificial Intelligence Research*, *16*, 321–357.

Checkland, P., & Holwell, S. (1998). Action Research: Its Nature and Validity. *Systemic Practice and Action Research*, *11*(2), 9–21.

Chen, G. M., & Ng, Y. M. M. (2016). Third-person perception of online comments: Civil ones persuade you more than me. *Computers in Human Behavior*, *55*, 736–742.

Chen, G. M., & Ng, Y. M. M. (2017). Nasty online comments anger you more than me, but nice ones make me as happy as you. *Computers in Human Behavior*, *71*, 181–188.

Chen, T., & Guestin, C. (2016). XGBoost: A Scalable Tree Boosting System. In A. Press (Ed.), *Proceedings of the 22nd 611 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (pp. 785–794).

Christin, A. (2017). Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*, *4*(2), 2053951717718855.

Chung, M., Munno, G. J., & Moritz, B. (2015). Triggering participation: Exploring the effects of third-person and hostile media perceptions on online participation. *Computers in Human Behavior*, *53*, 452–461.

Clement Addo, P., Fang, J., Asare, A. O., & Kulbo, N. B. (2021). Customer engagement and purchase intention in live-streaming digital marketing platforms. *Service Industries Journal*, *0*(0), 1–20. https://doi.org/10.1080/02642069.2021.1905798

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology*, *62*(7), 1382–1402. https://doi.org/10.1002/asi.21525

Cole, M. J., Gwizdka, J., Liu, C., Bierig, R., Belkin, N. J., & Zhang, X. (2011). Task and user effects on reading patterns in information search. *Interacting with Computers*, *23*(4), 346–362. https://doi.org/10.1016/j.intcom.2011.04.007

Cole, M. J., Hendahewa, C., Belkin, N. J., & Shah, C. (2015). User activity patterns during information search. *ACM Transactions on Information Systems*, *33*(1). https://doi.org/10.1145/2699656

Cooper, H. (1998). *Synthesizing research* (C. Thousand Oaks, Ed.). SAGE.

Cortez, P. (2021). *Modern optimization with R* (2nd Edition). Springer. https://pcortez.dsi.uminho.pt/mor-book#h.7ak9plkw1435

Coussement, K., van den Bossche, F. A. M., & de Bock, K. W. (2014). Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees. *Journal of Business Research*, *67*(1), 2751–2758. https://doi.org/10.1016/j.jbusres.2012.09.024

Cristobal-Fransi, E., Hernández-Soriano, F., & Daries-Ramon, N. (2017). Nuevos lectores para nuevos medios: Segmentación de los e-lectores de un cibermedio. *Espacios*, *38*(39).

Dahmen, N. S., Thier, K., & Walth, B. (2021). Creating engagement with solutions visuals: testing the effects of problem-oriented versus solution-oriented photojournalism. *Visual Communication*, *20*(2), 271–288. https://doi.org/10.1177/1470357219838601

DAMA. (2017). *DAMA-DMBOK: Data Management Body of Knowledge* (2nd edition). Technics Publications.

Danzon-Chambaud, S. (2021). A systematic review of automated journalism scholarship: guidelines and suggestions for future research. *Open Research Europe*, *1*(May), 4. https://doi.org/10.12688/openreseurope.13096.1

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227.

Davoudi, H. (2018). *User Acquisition and engagement in digital News Media* (Issue December).

Davoudi, H., An, A., & Edall, G. (2019). Content-based Dwell Time Engagement Prediction Model for News Articles. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *2*, 226–233.

Davoudi, H., An, A., Zihayat, M., & Edall, G. (2018). Adaptive Paywall Mechanism for Digital News Media Heidar. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2018) 205-214*, 205–214.

Davoudi, H., & Edall, G. (2018). *Adaptive Paywall Mechanism for Digital News Media*. 205–214.

de Caigny, A., Coussement, K., & de Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, *269*(2), 760–772. https://doi.org/10.1016/j.ejor.2018.02.009

Delpisheh, E., Davoudi, H., Boroujerdi, E. G., & An, A. (2016). Time aware topic based recommender system. *Big Data and Information Analytics*, *1*(2/3), 261–274. https://doi.org/10.3934/bdia.2016008

Denisova, A. (2022). Viral journalism. Strategy, tactics and limitations of the fast spread of content on social media: Case study of the United Kingdom quality publications. *Journalism*, 146488492210777. https://doi.org/10.1177/14648849221077749

Donthu, N., Kumar, S., Pandey, N., & Gupta, P. (2021). Forty years of the International Journal of Information Management: A bibliometric analysis. *International Journal of Information Management*, *57*(December 2020), 102307. https://doi.org/10.1016/j.ijinfomgt.2020.102307

Du, N., Farajtabar, M., Ahmed, A., Smola, A. J., & Song, L. (2015). Dirichlet-Hawkes Processes with Applications to Clustering Continuous-Time Document Streams Categories and Subject Descriptors. *21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 219–228.

Eck, N. J. van, & Waltman, L. (2013). VOSviewer manual. In *Univeristeit Leiden* (Issue February). http://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.1.pdf

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*, 131–152.

Engelke, K. M. (2019). Online participatory journalism: A systematic literature review. *Media and Communication*, *7*(4), 31–44. https://doi.org/10.17645/mac.v7i4.2250

Ernawati, E., Baharin, S. S. K., & Kasmin, F. (2021). A review of data mining methods in RFM-based customer segmentation. *Journal of Physics: Conference Series*, *1869*(1). https://doi.org/10.1088/1742-6596/1869/1/012085

Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., & Kashef, R. (2020). Recommendation systems: Algorithms, challenges, metrics, and business opportunities. *Applied Sciences (Switzerland)*, *10*(21), 1–20. https://doi.org/10.3390/app10217748

Feinerer, I. (2019). Introduction to the tm Package Text Mining in R. *Retrieved March*, *1*.

Feng, S., Meng, J., & Zhang, J. (2021). News recommendation systems in the Era of information overload. *Journal of Web Engineering*, *20*(2), 459–470. https://doi.org/10.13052/jwe1540-9589.20210

Fernandes, K., Vinagre, P., & Cortez, P. (2015). A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. *Portuguese Conference on Artificial Intelligence*, 535–546.

Ficel, H., Haddad, M. R., & Baazaoui Zghal, H. (2021). A graph-based recommendation approach for highly interactive platforms. *Expert Systems with Applications*, *185*(May), 115555. https://doi.org/10.1016/j.eswa.2021.115555

Flaounas, I., Ali, O., Lansdall-welfare, T., Bie, T. De, Lewis, J., Cristianini, N., Flaounas, I., Ali, O., Lansdall-welfare, T., & Bie, T. De. (2013). Research methods in the age of digital journalism: Massive-scale automated analysis of news-content—topics, style and gender. *Digital Journalism*, *1*(1), 102–116. https://doi.org/10.1080/21670811.2012.714928

Flaounas, I., Turchi, M., Ali, O., Fyson, N., De Bie, T., Mosdell, N., & Cristianini, N. (2010). The structure of the EU mediasphere. *PloS One*, *5*(12), 14243.

FTStrategies. (2022). *How the Financial Times brought data into the Newsroom*. FTStrategies. https://www.ftstrategies.com/en-gb/insights/how-the-financial-times-brought-data-into-the-newsroom/

Fu, X., Chen, X., Shi, Y. T., Bose, I., & Cai, S. (2017). User segmentation for retention management in online social games. *Decision Support Systems*, *101*, 51–68. https://doi.org/10.1016/j.dss.2017.05.015

Fu, X., Lee, J., Yan, C., & Gao, L. (2019). Mining newsworthy events in the traffic accident domain from Chinese microblog. *International Journal of Information Technology & Decision Making*, 717–742.

Galily, Y. (2018). Artificial intelligence and sports journalism: Is it a sweeping change? *Technology in Society*, *54*, 47–51.

Gao, Y., Gupta, V., Yan, J., Shi, C., Tao, Z., Xiao, P. J., Wang, C., Yu, S., Rosales, R., Muralidharan, A., & Chatterjee, S. (2018). Near real-time optimization of activity-based notifications. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 283–292. https://doi.org/10.1145/3219819.3219880

García-Avilés, J. A. (2014a). Online Newsrooms as Communities of Practice: Exploring Digital Journalists' Applied Ethics. *Journal of Mass Media Ethics*, *29*(4), 258–272.

García-Avilés, J. A. (2014b). Online Newsrooms as Communities of Practice: Exploring Digital Journalists' Applied Ethics. *Journal of Mass Media Ethics*, *29*(4), 258–272.

Garcin, F., Dimitrakakis, C., & Faltings, B. (2013). Personalized News Recommendation with Context Trees. *Proceedings of the 7th ACM Conference on Recommender Systems*, 105–112.

Gil, M., Wróbel, K., Montewka, J., & Goerlandt, F. (2020). A bibliometric analysis and systematic review of shipboard Decision Support Systems for accident prevention. *Safety Science*, *128*(March), 104717. https://doi.org/10.1016/j.ssci.2020.104717

Global Change Data Lab. (2021). *Our World Data*. https://ourworldindata.org/

Goad, R. (2016). Transforming a Media Organisation with Big Data. *EBU Big Data Conference*.

Goldani, M. H., Safabakhsh, R., & Momtazi, S. (2021). Convolutional neural network with margin loss for fake news detection. *Information Processing and Management*, *58*(1), 102418. https://doi.org/10.1016/j.ipm.2020.102418

Gonçalves, Cortez, P., & Carvalho, M. S. (2021). K -means clustering combined with principal component analysis for material profiling in automotive supply chains. *European Journal of Industrial Engineering*, *15*(2), 273–294.

Gonçalves, Subtil, A., Rosário Oliveira, M., & de Zea Bermudez, P. (2014). ROC curve estimation: An overview. *Revstat Statistical Journal*, *12*(1), 1–20.

Gonzalez Camacho, L. A., & Alves-Souza, S. N. (2018). Social network data to alleviate cold-start in recommender system: A systematic review. *Information Processing and Management*, *54*(4), 529–544. https://doi.org/10.1016/j.ipm.2018.03.004

Google. (2022). *Event Measurement*. https://developers.google.com/analytics/devguides/collection/analyticsjs/events

Gope, J., & Jain, S. K. (2017). A survey on solving cold start problem in recommender systems. *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 133–138.

Gordon, A. D. (1999). *Cassification 2nd Edition*. Chapman & Hall/CRC.

Goyanes, M. (2020). Why Do Citizens Pay for Online Political News and Public Affairs ? Socio-psychological Antecedents of Local News Paying Behaviour Why Do Citizens Pay for Online Political News and Public Paying Behaviour. *Journalism Studies*, *21*(4), 547–563. https://doi.org/10.1080/1461670X.2019.1694429

Gravengaard, G. Rimestad, L. (2012). Elimination of ideas and professional socialisation: Lessons learned at newsroom meetings. *Journalism Practice*, *6*(4), 465–481.

Greco, F., & Polli, A. (2020). Emotional Text Mining: Customer profiling in brand management. *International Journal of Information Management*, *51*(April 2019), 101934. https://doi.org/10.1016/j.ijinfomgt.2019.04.007

Grinberg, N. (2018). Identifying modes of user engagement with online news and their relationship to information gain in text. *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*, 1745–1754. https://doi.org/10.1145/3178876.3186180

Grzegorz, P. (2021). News Subscriptions In the Age of Coronavirus. *INMA Readers First Iniciative*.

Gustriansyah, R., Suhandi, N., & Antony, F. (2019). Clustering optimization in RFM analysis based on k-means. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*(1), 470–477. https://doi.org/10.11591/ijeecs.v18.i1.pp470-477

Haim, M., Graefe, A., Brosius, H., Haim, M., Graefe, A., & Brosius, H. (2018). Burst of the Filter Bubble ? Effects of personalization on the diversity of Google News. *Digital Journalism*, *6*(3), 330–343. https://doi.org/10.1080/21670811.2017.1338145

Häring, M., Loosen, W., & Maalej, W. (2018). Who is addressed in this comment? Automatically classifying meta-comments in news comments. *Proceedings of the ACM on Human-Computer Interaction*, 1–20.

Hauke, J., & Kossowski, T. (2011). Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae*, *30*(2), 87–93. https://doi.org/10.2478/v10117-011-0021-1

Hazrati, N., & Elahi, M. (2021). Addressing the New Item problem in video recommender systems by incorporation of visual features with restricted Boltzmann machines. *Expert Systems*, *38*(3), 1–20. https://doi.org/10.1111/exsy.12645

Hendrickx, J., Karen, D., & Picone, I. (2020). Innovating journalism by going back in time? The curious case of newsletters as a news source in Belgium. In *Journalistic metamorphosis* (pp. 57–68). Springer.

Hendrickx, J., Montero, E., Ranaivoson, H., & Ballon, P. (2021). Becoming the Data-Informed Newsroom? The Promotion of Audience Metrics in the Newsroom and Journalists' Interactions with Them. *Digital Journalism*, *9*(4), 427–442. https://doi.org/10.1080/21670811.2021.1890170

Hermida, A., & Mellado, C. (2020). Dimensions of Social Media Logics : Mapping Forms of Journalistic Norms and Practices on Twitter and Instagram Dimensions of Social Media Logics : Mapping Forms of Journalistic Norms and Practices on Twitter. *Digital Journalism*, *8*(7), 864–884. https://doi.org/10.1080/21670811.2020.1805779

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572.

Ho, S. S., Lieberman, M., Wang, P., & Samet, H. (2012). Mining future spatiotemporal events and their sentiment from online news articles for location-aware recommendation system. *Proceedings of the First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems*, 25–32.

Hogenboom, F., Frasincar, F., Kaymak, U., & De Jong, F. (2011). An overview of event extraction from text. *DeRiVE@ ISWC*, 48–57.

Hootsuite. (2020). *Digital 2020: Portugal, All the data, trends, and insights you need to help you understand how people use the internet, mobile, social media, and ecommerce*.

Hornik, K., & Hornik, M. K. (2018). *Package ' NLP .'* http://cran.salud.gob.sv/web/packages/NLP/

Hsiao, K. L., & Chen, C. C. (2017). Value-based adoption of e-book subscription services: The roles of environmental concerns and reading habits. *Telematics and Informatics*, *34*(5), 434–448. https://doi.org/10.1016/j.tele.2016.09.004

Huang, Z., Chen, H., & Zeng, D. (2004). Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *Daniel* , *22*(1), 116–142.

Hullar, K. (2020). *3 social media tips from Chartbeat that will enhance your presence across channels*. Chartbeat Blog. https://blog.chartbeat.com/2020/04/01/3-social-media-tips-across-channels/

Indurthi, V., Oota, S. R., Gupta, M., & Varma, V. (2018). Believe it or not! Identifying bizarre news in online news media. *ACM International Conference Proceeding Series*, 257–264. https://doi.org/10.1145/3152494.3152524

INMA. (2022). *The Benefits and Risks of Media Data Democratisation* (Issue January). https://www.inma.org/report/the-benefits-and-risks-of-media-data-democratisation

Jääskeläinen, A., Taimela, E., & Heiskanen, T. (2020). Predicting the success of news: Using an ML-based language model in predicting the performance of news articles before publishing. *Proceedings of the 23rd International Conference on Academic Mindtrek*, 27–36. https://doi.org/10.1145/3377290.3377299

Jack, A. (2016). Editorial Email newsletters: The Medium Is Not the Only Message. In *Reuters Institute for the Study of Journalism*. https://www.researchgate.net/publication/269107473_What_is_governance/link/548173090cf22525dcb61443/download%0Ahttp://www.econ.upf.edu/~reynal/Civil wars_12December2010.pdf%0Ahttps://think-asia.org/handle/11540/8282%0Ahttps://www.jstor.org/stable/41857625

Jacob, M. (2021). *Nearly Half of Digital Subscribers Are 'Zombies,' Medill Analysis Finds*. Northwestern Local News Initiative. https://localnewsinitiative.northwestern.edu/posts/2021/03/01/zombies/index.html

Jarreau, P., Dahmen, N., & Jones, E. (2019). Instagram and the science museum: a missed opportunity for public engagement. *Journal of Science Communication*, *8*(5), 55.

Jin, R., Gao, S., Cheshmehzangi, A., & Aboagye-Nimo, E. (2019). A holistic review of public-private partnership literature published between 2008 and 2018. *Journal of Cleaner Production*, *202*, 1202–1219. https://doi.org/10.1155/2019/7094653

Joris, G., Grove, F. De, Van Damme, K., & De Marez, L. (2021). Appreciating News Algorithms: Examining Audiences' Perceptions to Different News Selection Mechanisms. *Digital Journalism*, *9*(5), 589–618. https://doi.org/10.1080/21670811.2021.1912626

Jovi, A., Brki, K., & Bogunovi, N. (2015). An overview of free software tools for general data mining. *37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1112–1117. https://doi.org/10.1109/MIPRO.2014.6859735

Ju, A., Jeong, S. H., & Chyi, H. I. (2014). Will Social Media Save Newspapers? *Journalism Practice*, *8*(1), 1–17. https://doi.org/10.1080/17512786.2013.794022

Kamthania, D., Pahwa, 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. https://doi.org/10.20532/cit.2018.1003863

Katser, I., Kozitsin, V., Lobachev, V., & Maksimov, I. (2021). Unsupervised offline changepoint detection ensembles. *Applied Sciences (Switzerland)*, *11*(9), 1–19. https://doi.org/10.3390/app11094280

Kemp, G., & White, G. (2020). *Google Data Studio for Beginners - Start Making Your Data Actionable* (Apress, Ed.; 1st Editio). Apress.

Ketchen, D., & Shook, C. (1996). The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal*, *17*(6), 441–458.

Kim, S. J., Zhou, Y., Malthouse, E. C., & Kamyab Hessary, Y. (2021). In Search for an Audience-Supported Business Model for Local Newspapers: Findings from Clickstream and Subscriber Data. *Digital Journalism*, 1–21. https://doi.org/10.1080/21670811.2021.1948347

Klopcic, A. L., Hojnik, J., Bojnec, S., & Papler, D. (2020). Global Transition to the Subscription Economy: Literature Review on Business Model Changes in the Media Landscape. *International Research Journal*, *18*(4), 323–348.

Koltay, T. (2016). Data governance, data literacy and the management of data quality. *IFLA Journal*, *42*(4), 303–312. https://doi.org/10.1177/0340035216672238

Kothari, C. R. (2004). *Research Methodology: Methods and Techniques*. New Age International (P) Ltd.

Kotkov, D., Wang, S., & Veijalainen, J. (2016). A survey of serendipity in recommender systems. *Knowledge-Based Systems*, *111*, 180–192. https://doi.org/10.1016/j.knosys.2016.08.014

Kotler, P., Kartajaya, H., & Setiawan, I. (2016). *Marketing 4.0: Moving from traditional to digital*. John Wiley & Sons.

Ksiazek, T. B., Peer, L., & Lessard, K. (2016). User engagement with online news: Conceptualizing interactivity and exploring the relationship between online news videos and user comments. *New Media and Society*, *18*(3), 502–520. https://doi.org/10.1177/1461444814545073

Kulkarni, H., Joshi, T., Sanap, N., Kalyanpur, R., & Marathe, M. (2019). Personalized newspaper based on emotional traits using machine learning. *Proceedings - 2019 5th International Conference on Computing, Communication Control and Automation, ICCUBEA 2019*. https://doi.org/10.1109/ICCUBEA47591.2019.9128691

Labatut, V., & Cherifi, H. (2012). *Accuracy Measures for the Comparison of Classifiers*.

Lafrance, A. (2017). *The Power of Personalization*. Nieman Reports. https://niemanreports.org/articles/the-power-of-personalization/

Lagun, D., & Lalmas, M. (2016). Understanding and measuring user engagement and attention in online news reading. *WSDM 2016 - Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, *22–25*, 113–122. https://doi.org/10.1145/2835776.2835833

Lajumoke, T., Betts, T., Gautier, L., Part, T., Patel, U., & Meirinhos, L. (2020). *Towards your North Star - Report on the outcomes of the european GNI Subscriptions LAB 2020*.

Lamot, K., & Paulussen, S. (2020). Six Uses of Analytics: Digital Editors' Perceptions of Audience Analytics in the Newsroom. *Journalism Practice*, *14*(3), 358–373. https://doi.org/10.1080/17512786.2019.1617043

Larsson, A. O. (2018). The News User on Social Media A comparative study of interacting with media organizations on Facebook and Instagram. *Journalism Studies*, *19*(15), 2225–2242. https://doi.org/10.1080/1461670X.2017.1332957

Lee, S. Y., & Ryu, M. H. (2019). Exploring characteristics of online news comments and commenters with machine learning approaches. *Telematics and Informatics*, *43*(101249).

Lehmann, J., Lalmas, M., Yom-Tov, E., & Dupret, G. (2012). Models of user engagement. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 164–175). Springer. https://doi.org/10.1007/978-3-642-31454-4_14

Lehmkuhl, M., & Peters, H. P. (2016). Constructing (un-) certainty: An exploration of journalistic decision-making in the reporting of neuroscience. *Public Understanding of Science*, *25*(8), 909–926.

Leung, X. Y., Bai, B., & Erdem, M. (2017). Hotel social media marketing : a study on message strategy and its effectiveness Journal of Hospitality and Tourism Technology Article information : *Journal of Hospitality and Tourism Technology*, *8*(2), 239–255. https://doi.org/10.1108/JHTT-02-2017-0012

Lewis. (2015). Journalism In An Era Of Big Data Cases , concepts , and critiques. *Digital Journalism*, *3*(3), 321–330. https://doi.org/10.1080/21670811.2014.976399

Lewis, S. C., Sanders, A. K., & Carmody, C. (2019). Libel by Algorithm? Automated Journalism and the Threat of Legal Liability. *Journalism & Mass Communication Quarterly*, *96*(1), 60–81.

Lewis, S., Guzman, A., & Schmidt, T. (2019). Automation, Journalism, and Human–Machine Communication: Rethinking Roles and Relationships of Humans and Machines in News. *Digital Journalism*, *7*(4), 409–427.

Li, J., Fong, S., Zhuang, Y., & Khoury, R. (2016). Hierarchical classification in text mining for sentiment analysis of online news. *Soft Computing*, *20*(9), 3411–3420.

Li, & Xie. (2020). Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement. *Journal of Marketing Research*, *57*(1), 1–19. https://doi.org/10.1177/0022243719881113

Liao, Y., Wang, S., Han, E. H., Lee, J., & Lee, D. (2019). Characterization and Early Detection of Evergreen News Articles. *In ECML/PKDD*, *3*, 552–568. https://doi.org/10.1007/978-3-030-46133-1_33

Liao, Y.-S., Lu, J.-Y., & Liu, D.-R. (2019). NEWS RECOMMENDATION BASED ON COLLABORATIVE SEMANTIC TOPIC MODELS AND RECOMMENDATION ADJUSTMENT. In *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*.

Lim, J. S., & Zhang, J. (2022). Adoption of AI-driven personalization in digital news platforms: An integrative model of technology acceptance and perceived contingency. *Technology in Society*, *69*(February), 101965. https://doi.org/10.1016/j.techsoc.2022.101965

Lioudis, N. (2019). *How Recirculation builds engagement, supports reader acquisition efforts*. Chartbeat Blog. https://blog.chartbeat.com/2019/08/21/recirculation-data-reader-acquisition/

Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). *Thresholding Classifiers to Maximize F1 Score*.

Liu, B., Bai, B., Xie, W., Guo, Y., & Chen, H. (2022). Task-optimized User Clustering based on Mobile App Usage for Cold-start Recommendations. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2022)*. https://doi.org/10.1145/3534678.3539105

Liu, Cole, M., Liu, C., Bierig, R., Gwizdka, J., & Belkin, N. (2010). Search Behaviors in Different Task Types. *Proceedings of the 10th Annual Joint Conference on Digital Libraries.*, 69–78.

Liu, D. R., Chen, K. Y., Chou, Y. C., & Lee, J. H. (2018). Online recommendations based on dynamic adjustment of recommendation lists. *Knowledge-Based Systems*, *161*, 375–389. https://doi.org/10.1016/j.knosys.2018.07.038

Liu, J., Dolan, P., & Pedersen, E. R. (2010). Personalized news recommendation based on click behavior. *Proceedings of the 15th International Conference on Intelligent User Interfaces*, 31–40.

Liu, Y., Hsiao, A., & Ma, E. (2021). Segmenting Tourism Markets Based on Demand Growth Patterns: A Longitudinal Profile Analysis Approach. In *Journal of Hospitality and Tourism Research* (Vol. 45, Issue 6). https://doi.org/10.1177/1096348020962564

Loni, B., Schuth, A., Nl, A. S., De Haas, L., Jansze, J., & Visser, V. (2019). Personalized Push Notifications for News Recommendation. *Proceedings of Machine Learning Research*, *109*, 36–45.

Lu, H., Zhang, M., & Ma, S. (2018). Between clicks and satisfaction: Study on multi-phase user preferences and satisfaction for online news reading. *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, 435–444. https://doi.org/10.1145/3209978.3210007

Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. *In Advances in Neural Information Processing Systems*, 4765–4774. https://doi.org/10.1016/j.ophtha.2018.11.016

Lynes, M. (2021). *3 types of newsreaders and strategies to engage them*. Twipe. https://www.twipemobile.com/3-types-of-newsreaders-and-strategies-to-engage-them/

Mailchimp. (2022). *2022 Email Marketing Statistics & Benchmarks - Mailchimp*. Https://Mailchimp.Com/Pt-Br/Resources/Email-Marketing-Benchmarks/.

Makridakis, S., Hyndman, R. J., & Petropoulos, F. (2020). Forecasting in social settings: The state of the art. *International Journal of Forecasting*, *36*(1), 15–28. https://doi.org/10.1016/j.ijforecast.2019.05.011

Manjesh, S., Kanakagiri, T., Vaishak, P., Chettiar, V., & Shobha, G. . (2017). Clickbait pattern detection and classification of news headlines using natural language processing. *2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*.

Manosevitch, I., & Tenenboim, O. (2017). The Multifaceted Role of User-Generated Content in News Websites: An analytical framework. *Digital Journalism*, *5*(6), 731–752. https://doi.org/10.1080/21670811.2016.1189840

Marketest. (2022). netAudience. In *https://www.marktest.com/wap/a/grp/p~109.aspx*. https://www.marktest.com/wap/a/grp/p~109.aspx

Mathew, A. (2021). *Role of Big Data Analysis and Machine Learning in Ecommerce - Customer Segmentation*. *3*(1).

Mcmullan, D. (2018). *How Newsletters Are Redefining Media Subscriptions INMA*. https://www.inma.org/report/how-newsletters-are-redefining-media-subscriptions

Meel, P., & Vishwakarma, D. K. (2020). Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, *153*, 112986. https://doi.org/10.1016/j.eswa.2019.112986

Meguebli, Y., Kacimi, M., Doan, B. L., & Popineau, F. (2017). Towards better news article recommendation. *World Wide Web*, *20*(6), 1293–1312.

Melki, J. P., & Mallat, S. E. (2016). Block Her Entry, Keep Her Down and Push Her Out: Gender discrimination and women journalists in the Arab world. *Journalism Studies*, *17*(1), 57–79.

Mersey, R. D., Malthouse, E. C., & Calder, B. J. (2010). Engagement with online media. *Journal of Media Business Studies*, *7*(2), 39–56. https://doi.org/10.1080/16522354.2010.11073506

Midberry, J., & Dahmen, N. S. (2020). Visual Solutions Journalism: A Theoretical Framework. *Journalism Practice*, *14*(10), 1159–1178. https://doi.org/10.1080/17512786.2019.1689371

Misztal-Radecka, J., Indurkhya, B., & Smywiński-Pohl, A. (2021). Meta-User2Vec model for addressing the user and item cold-start problem in recommender systems. *User Modeling and User-Adapted Interaction*, *31*(2), 261–286.

Mizgajski, J., & Morzy, M. (2019). Affective recommender systems in online news industry : how emotions influence reading choices. *User Modeling and User-Adapted Interaction*, *29*(2), 345–379. https://doi.org/10.1007/s11257-018-9213-x

Montes-García, A., Álvarez-Rodríguez, J. M., Labra-Gayo, J. E., & Martínez-Merino, M. (2013). Towards a journalist-based news recommendation system: The Wesomender approach. *Expert Systems with Applications*, *40*(17), 6735–6741. https://doi.org/10.1016/j.eswa.2013.06.032

Moro, S., Cortez, P., & Rita, P. (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Systems with Applications*, *42*(3), 1314–1324. https://doi.org/10.1016/j.eswa.2014.09.024

Moro, S., Laureano, R. M. S., & Cortez, P. (2011). Using Data Mining for Bank Direct Marketing : An Application of the CRISP-DM Methodology. *Proceedings of European Simulation and Modelling Conference-ESM'2011*, 117–121.

Moyo, D., Mare, A., & Matsilele, T. (2019). Analytics-Driven Journalism? Editorial Metrics and the Reconfiguration of Online News Production Practices in African Newsrooms". *Digital Journalism*, *7*(4), 490–506. https://doi.org/10.1080/21670811.2018.1533788

Mucchetti, M. (2020). Google Data Studio. In *BigQuery for Data Warehousing* (pp. 401–416).

Muralidhar, N., Rangwala, H., & Han, E.-H. S. (2015). Recommending Temporally Relevant News Content from Implicit Feedback Data. *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, 689–696. https://doi.org/10.1109/ICTAI.2015.104

Myllylahti, M. (2017). We need to talk about metrics. In *Themes and debates in contemporary journalism.* (pp. 87–103). Cambridge: Cambridge Scholar Publishing.

Myllylahti, M. (2019). Paywalls. *The International Encyclopedia of Journalism Studies*, 1–6.

NapoleonCat Stats, W. (2021). *Instagram users in Portugal*. https://napoleoncat.com/stats/instagram-users-in-portugal, last accessed 2021/11/07

Napoles, C., Pappu, A., & Tetreault, J. (2017). Automatically identifying good conversations online (yes, they do exist!). *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1).

Nasir, M., Ezeife, C. I., & Gidado, A. (2021). Improving e-commerce product recommendation using semantic context and sequential historical purchases. *Social Network Analysis and Mining*, *11*(1). https://doi.org/10.1007/s13278-021-00784-6

Neilson, T., & Gibson, T. A. (2022). Social Media Editors and the Audience Funnel: Tensions between Commercial Pressures and Professional Norms in the Data-Saturated Newsroom. *Digital Journalism*, *10*(4), 556–578. https://doi.org/10.1080/21670811.2021.2004553

Nelson, M. J., & Hoover, A. K. (2020). Notes on Using Google Colaboratory in AI Education. *Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE*, 533–534. https://doi.org/10.1145/3341525.3393997

Neuberger, C., Nuernbergk, C., & Langenohl, S. (2019). Journalism as Multichannel Communication: A newsroom survey on the multiple uses of social media. *Journalism Studies*, *20*(9), 1260–1280. https://doi.org/10.1080/1461670X.2018.1507685

Newman. (2019). *Journalism, Media, and Technology Trends and Predictions 2019*.

Newman, N. (2022). Digital News Report: Journalism, Media, and Technology Trends and Predictions 2022. *Reuters Institute*, 46.

Newman, N., Fletcher, R., Kalogeropoulos, A., & Nielsen, R. K. (2019). *Reuters Institute Digital News Report 2019*.

Newman, N., Fletcher, R., Schulz, A., Andi, S., Robertson, C., & Nielsen, R. (2022). Digital News Report 2021 10th Edition. In *Reuters Institute*. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021

Nielsen, R. K., & Ganter, S. A. (2018). Dealing with digital intermediaries: A case study of the relations between publishers and platforms. *New Media and Society*, *20*(4), 1600–1617. https://doi.org/10.1177/1461444817701318

Obiedat, R. (2020). Predicting the popularity of online news using classification methods with feature filtering techniques. *Journal of Theoretical and Applied Information Technology*, *98*(8), 1163–1172.

O'Brien, D., Wellbrock, C. M., & Kleer, N. (2020). Content for Free? Drivers of Past Payment, Paying Intent and Willingness to Pay for Digital Journalism–A Systematic Literature Review. *Digital Journalism*, *8*(5), 643–672. https://doi.org/10.1080/21670811.2020.1770112

O'Brien, H. L., & Lebow, M. (2013). Mixed-Methods Approach to Measuring User Experience in Online News Interactions. *Journal of the American Society for Information Science and Technology*, *64*(8), 1543–1556. https://doi.org/10.1002/asi

Olsen, R. K., Kammer, A., Solvoll, M. K., & Olsen, R. K. (2020). *Paywalls ' Impact on Local News Websites ' Traffic and Their Civic and Business Implications*. *9699*. https://doi.org/10.1080/1461670X.2019.1633946

Omar, N., Omar, Y. M. K., & Maghraby, F. A. (2020). Machine Learning Model for Personalizing Online Arabic Journalism. *Machine Learning*, *11*(4). https://doi.org/10.14569/IJACSA.2020.0110484

Omran, M. G. H., Engelbrecht, A. P., & Salman, A. (2007). An overview of clustering methods. *Intelligent Data Analysis*, *11*(6), 583–605. https://doi.org/10.3233/ida-2007-11602

Overgaard, C. S. B. (2021). Constructive Journalism in the Face of a Crisis: The Effects of Social Media News Updates About COVID-19. *Journalism Studies*, *22*(14), 1875–1893. https://doi.org/10.1080/1461670X.2021.1971107

Pattabhiramaiah, A., Sriram, S., & Manchanda, P. (2019). Paywalls: Monetizing online content. *Journal of Marketing*, *83*(2), 19–36. https://doi.org/10.1177/0022242918815163

Perreault, M. F., & Perreault, G. P. (2021). Journalists on COVID-19 Journalism : Communication Ecology of Pandemic Reporting. *American Behavioral Scientist*, *65*(7), 976–991. https://doi.org/10.1177/0002764221992813

Peterson, E. T., & Carrabis, J. (2008). Measuring the Immeasurable: Visitor Engagement. *Web Analytics Demystified*, *14*(16).

Piccinelli, S., Moro, S., & Rita, P. (2021). Air-travelers' concerns emerging from online comments during the COVID-19 outbreak. *Tourism Management*, *85*(104313).

Piechota, G. (2020). COVID-19 tactics driving engagement across the funnel. *INMA World Congress of News Media 2020*.

Público. (2022, November 30). *PÚBLICO aumentou circulação e é o jornal diário com mais assinaturas digitais*. Https://Www.Publico.Pt/2022/11/30/Sociedade/Noticia/Publico-Aumentou-Circulacao-Jornal-Diario-Assinaturas-Digitais-2029829.

Qin, Z., & Zhang, M. (2021). Research on News Recommendation Algorithm Based on User Interest and Timeliness Modeling. *ACM International Conference Proceeding Series*, *PartF16898*. https://doi.org/10.1145/3448734.3450933

Rahmat, R. F., Lini, T. Z., Purnawati, S., Mulki, R., Sukarja, D., & Lubis. (2020). Sparse matrix implementation on personal news recommendation for anonymous user. *4rd International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM)*, 140–145.

Rao, Y., Lei, J., Wenyin, L., Li, Q., & Chen, M. (2014). Building emotional dictionary for sentiment analysis of online news. *World Wide Web*, *17*(4), 723–742.

Rendón, E., Abundez, I. M., Gutierrez, C., Díaz, S., Arizmendi, A., Quiroz, E. M., & H, E. A. (2011). A comparison of internal and external cluster validation indexes. *In Proceedings of the 5th WSEAS International Conference on Computer Engineering and Applications*, 158–163.

Renó, D., & Renó, L. (2015). The newsroom, Big Data and social media as information sources. *Estudios Sobre El Mensaje Periodistico*, *21*(21), 131–142. https://doi.org/10.5209/rev_ESMP.2015.v21.51135

Riedl, M. J., Masullo, G. M., & Whipple, K. N. (2020). The downsides of digital labor: Exploring the toll incivility takes on online comment moderators. *Computers in Human Behavior*, *107*(106262).

Rios-Rodríguez, R., Fernández-López, S., Dios-Vicente, A., & Rodeiro-Pazos, D. (2022). Reconversion in a declining market: the return to profitability of the print newspaper industry. *Journal of Media Business Studies*. https://doi.org/10.1080/16522354.2022.2104556

Rivera, S. J., Minsker, B. S., Work, D. B., & Roth, D. (2014). A text mining framework for advancing sustainability indicators. *Environmental Modelling and Software*, *62*, 128–138. https://doi.org/10.1016/j.envsoft.2014.08.016

Romero, L., & Portillo-Salido, E. (2019). Trends in sigma-1 receptor research: A 25-year bibliometric analysis. *Frontiers in Pharmacology*, *10*(MAY). https://doi.org/10.3389/fphar.2019.00564

Rosenberg, H., Syed, S., & Rezaie, S. (2020). The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic. *Canadian Journal of Emergency Medicine*, *22*(4), 418–421. https://doi.org/10.1017/cem.2020.361

Rußell, R., Berger, B., Stich, L., Hess, T., & Spann, M. (2020). Monetizing Online Content : Digital Paywall Design and Configuration. *Business & Information Systems Engineering*, 1–8. https://doi.org/10.1007/s12599-020-00632-5

Sanz-Narrillos, M., Masneri, S., & Zorrilla, M. (2020). Combining video and wireless signals for enhanced audience analysis. *International Conference on Agents and Artificial Intelligence*, 151–161.

Sapian, A., & Vyshnevska, M. (2019). The marketing funnel as an effective way of a business strategy. *ΛΌΓΟΣ. The Art of Scientific Mind*, *4*, 16–18.

Saranya, K. G., & Sadasivam, G. S. (2017). Personalized news article recommendation with novelty using collaborative filtering based rough set theory. *Mobile Networks and Applications*, *22*(4), 719–729.

Schmidt, S. (2022, August 11). *3 data drivers are key to understanding news audiences*. International News Media Association INMA .

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, *20*(1), 3–29.

Schwartz, J. (2013). *Using Engaged Time to understand your audience*. Chartbeat. https://blog.chartbeat.com/2013/03/18/using-engaged-time-to-understand-your-audience/

Seale, S. (2021). *How Wall Street Journal uses metrics and engagement to drive digital subscriptions*. INMA International News Media Association. https://www.inma.org/blogs/conference/post.cfm/how-wall-street-journal-uses-metrics-and-enagagement-to-drive-digital-subscriptions

Semerádová, T., & Weinlich, P. (2020). Using Google Analytics to Examine the Website Traffic. In *Website Quality and Shopping Behaviour Quantitative and Qualitative Evidence* (pp. 91–112).

Seufert, E. B. (2014). Freemium Economics: Leveraging Analytics and User Segmentation to Drive Revenue. In *Freemium Economics: Leveraging Analytics and User Segmentation to Drive Revenue*. https://doi.org/10.1016/C2013-0-00599-3

Shahbaznezhad, H., Dolan, R., & Rashidirad, M. (2021). The Role of Social Media Content Format and Platform in Users' Engagement Behavior. *Journal of Interactive Marketing*, *53*, 47–65. https://doi.org/10.1016/j.intmar.2020.05.001

Shahraki, H., Pourahmad, S., & Zare, N. (2017). K Important Neighbors: A Novel Approach to Binary Classification in High Dimensional Data. *BioMed Research International*.

Shim, J. S., Lee, Y., & Ahn, H. (2021). A link2vec-based fake news detection model using web search results. *Expert Systems with Applications*, *184*(June), 115491. https://doi.org/10.1016/j.eswa.2021.115491

Silge, J., & Robinson, D. (2019). *Text Mining with R - A Tidy Approach*. O'Reilly.

Silva, S., Cortez, P., Mendes, R., Pereira, P. J., Matos, L. M., & Garcia, L. (2018). A Categorical Clustering of Publishers for Mobile Performance Marketing. *The 13th International Conference on Soft Computing Models in Industrial and Environmental Applications*, 145.154.

Simon, A. F. M., & Graves, L. (2019). *Pay Models for Online News in the US and Europe : 2019 Update*. *May*, 1–16.

Sjøvaag, H. (2016). Introducing the paywall: A case study of content changes in three online newspapers. *Journalism Practice*, *10*(3), 304–322. https://doi.org/10.1080/17512786.2015.1017595

Smit, G., Fahland, D., Dongen, B. F., & Farzami, T. (2019). *Customer Segmentation using Clickstream*. Eindhoven University of Technology.

Smith, W. R. (1956). Product Differentiation and Market Segmentation as Alternative Marketing Strategies. *Journal of Marketing*, *21*(1), 3–8.

Souza Freire, P. M., Matias da Silva, F. R., & Goldschmidt, R. R. (2021). Fake news detection based on explicit and implicit signals of a hybrid crowd: An approach inspired in meta-learning. *Expert Systems with Applications*, *183*(February). https://doi.org/10.1016/j.eswa.2021.115414

Steensen, S., Ferrer-Conill, R., & Peters, C. (2020). (Against a) Theory of Audience Engagement with News. *Journalism Studies*, *20*, 1–19. https://doi.org/10.1080/1461670X.2020.1788414

Steinberger, R. (2012). A survey of methods to ease the development of highly multilingual text mining applications. *Language Resources and Evaluation*, *46*(2), 155–176. https://doi.org/10.1007/s10579-011-9165-9

Stone, B. (1989). *Successful Direct Marketing Methods: The Bob Stone direct marketing book* (3dr Editio). NTC Business Books.

Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Mueller, K.-R. (2021). *Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology*. http://arxiv.org/abs/2003.05155

Suárez, E. (2020). *How to build a successful subscription news business: lessons from Britain and Spain* (Issue February).

Sun, D., Zhou, T., Liu, J. G., Liu, R. R., Jia, C. X., & Wang, B. H. (2009). Information filtering based on transferring similarity. *Information Filtering Based on Transferring Similarity. Physical Review*, *80*(1).

Tandoc Jr, E. C. (2014). Journalism is twerking ? How web analytics is changing the process of gatekeeping. *New Media & Society*, *16*(4), 559–575. https://doi.org/10.1177/1461444814530541

Tang, L., Long, B., Chen, B. C., & Agarwal, D. (2016). An empirical study on recommendation with multiple types of feedback. *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 283–292.

Tanuwijaya, S., Alamsyah, A., & Ariyanti, M. (2021). Mobile Customer Behaviour Predictive Analysis for Targeting Netflix Potential Customer. *9th International Conference on Information and Communication Technology (ICoICT)*, 348–352. https://doi.org/10.1109/ICoICT52021.2021.9527487

Tatar, A., Antoniadis, P., De Amorim, M. D., & Fdida, S. (2014). From popularity prediction to ranking online news. *Social Network Analysis and Mining*, *4*(1), 174. https://doi.org/10.1007/s13278-014-0174-8

Taylor, R. (1990). Interpretaion of the Correlation Coefficient: A Basic Review. *Journal of Diagnostic Medical Sonography*, *6*(1), 35–39.

Tessem, B., & Opdahl, A. L. (2019). Supporting journalistic news angles with models and analogies. *Proceedings - International Conference on Research Challenges in Information Science*, 1–7. https://doi.org/10.1109/RCIS.2019.8877058

Tewari, A. S., Yadav, N., & Barman, A. G. (2016). Efficient tag based personalised collaborative movie reccommendation system. *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, 95–98.

True, L. (2018). *Package ' sentimentr .'* https://cran.r-project.org/web/packages/sentimentr/

Tsagkias, M., Weerkamp, W., & De Rijke, M. (2010). News comments: Exploring, modeling, and online prediction. *European Conference on Information Retrieval*, 191–203.

van Damme, K., Martens, M., van Leuven, S., Vanden, A., & Marez, L. (2020). Mapping the Mobile DNA of News. Understanding Incidental and Serendipitous Mobile News Consumption. *Digital Journalism* , *8*(1), 4968.

Van Der Aalst, W. (2011). *Process mining: discovery, conformance and enhancement of business processes*. Springer.

Van-Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

Vergani, A. A., & Binaghi, E. (2018). A Soft Davies-Bouldin Separation Measure. *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*.

Viana, P., & Soares, M. (2016). A hybrid recommendation system for news in a mobile environment. *6th International Conference on Web Intelligence, Mining and Semantics*, 1–9.

Villi, M., & Picard, R. G. (2019). Transformation and Innovation of Media Business Models. In *Making Media: production, Practices, and Professions* (pp. 121–132).

Vinothini, A., & Priya, S. B. (2018). Survey of machine learning methods for big data applications. *ICCIDS 2017 - International Conference on Computational Intelligence in Data Science, Proceedings*, *2018-Janua*, 1–5. https://doi.org/10.1109/ICCIDS.2017.8272638

Volkmer, I. (2021). *Social media & COVID-19: A global study of digital crisis interaction among Gen Z and Millennials*. https://arts.unimelb.edu.au/__data/assets/pdf_file/0007/3958684/Volkmer-Social-Media-and-COVID.pdf

Von Bloh, J., Broekel, T., Özgun, B., & Sternberg, R. (2020). New(s) data for entrepreneurship research? An innovative approach to use Big Data on media coverage. *Small Business Economics*, *55*(3), 673–694. https://doi.org/10.1007/s11187-019-00209-x

Vraga, E. K., Kim, S. C., Cook, J., & Bode, L. (2020). Testing the Effectiveness of Correction Placement and Type on Instagram. *The International Journal of Press/Politics*, *25*(4), 632–652. https://doi.org/10.1177/1940161220919082

Wang, H., Zhang, P., Lu, T., Gu, H., & Gu, N. (2017). Hybrid Recommendation Model Based on Incremental Collaborative Filtering and Content- based Algorithms. *2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 337–342.

Wang, W. (2012). Chinese news event 5W1H semantic elements extraction for event ontology population. *Proceedings of the 21st International Conference on World Wide Web*, 197–202.

Wang, W., & Stewart, K. (2015). Spatiotemporal and semantic information extraction from Web news reports about natural hazards. *Computers, Environment and Urban Systems*, *50*, 30–40.

Wang, W., Zhao, D., & Wang., D. (2012). Chinese news event 5W1H elements extraction using semantic role labeling. *2010 Third International Symposium on Information Processing*, 484–489. https://doi.org/10.1145/2187980.2188008

Wang, W., Zhao, D., Zou, L., Wang, D., & Zheng, W. E. (2010). Extracting 5W1H event semantic elements from Chinese online news. *International Conference on Web-Age Information Management*, 644–655.

Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019). Systematic literature review on the spread of health-related misinformation on social media. *Social Science & Medicine*, *240*(112552).

Webster, J., & Watson, R. (2002). Analyzing the past to prepare for the future: writing a literature review. *MIS Quarterly*, *2*, xiii–xxiii. https://doi.org/10.1016/j.freeradbiomed.2005.02.032

Welbers, K., Amsterdam, V. U., Atteveldt, W. Van, Amsterdam, V. U., & Benoit, K. (2017). Text Analysis in R. *Communication Methods and Measures*, *11*(4), 245–265. https://doi.org/10.1080/19312458.2017.1387238

Wenzel, S., Kleer, N., & Kunz, R. E. (2022). Customer engagement behavior in the media and technology industry: a quantitative content analysis of content types and COVID-19 context. *Journal of Media Business Studies*. https://doi.org/10.1080/16522354.2022.2139997

Wieland, R., Lakes, T., & Nendel, C. (2021). Using SHAP to interpret XGBoost predictions of grassland degradation in Xilingol, China. *Geoscientific Model Development*, *14*(3), 1493–1510.

Wu, C., Wu, F., An, M., Huang, J., Huang, Y., & Xie, X. (2019). Neural News Recommendation with Attentive Multi-View Learning. *IJCAI International Joint Conference on Artificial Intelligence*, 1907.05576.

Wu, S., Tandoc Jr, E. C., & Salmon, C. T. (2019). When journalism and automation intersect: Assessing the influence of the technological field on contemporary newsrooms. *Journalism Practice*, *13*(10), 1238–1254.

Xiao, J., Lu, J., & Li, X. (2017). Davies Bouldin Index based hierarchical initialization K-means. *Intelligent Data Analysis*, *21*(6), 1327–1338.

Yang, J. A. (2016). Effects of popularity-based news recommendations ("most-viewed") on users' exposure to online news. *Media Psychology*, *19*(2), 243–271. https://doi.org/10.1080/15213269.2015.1006333

Yang, W. (2020). Ux Design of Artificial Intelligence News Robot. *IOP Conference Series: Materials Science and Engineering*, *740*(1), 012135.

Yang, Y., Liu, Y., Lu, X., Xu, J., & Wang, F. (2020). A named entity topic model for news popularity prediction. *Knowledge-Based Systems*, *208*, 106430. https://doi.org/10.1016/j.knosys.2020.106430

Yang, Y., Ma, X., & Fung, P. (2017). Perceived emotional intelligence in virtual agents. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2255–2262.

Yang, Z. (2020). Analysis of the Impact of Big Data Technology on News Ecology. *Journal of Physics: Conference Series*, *1682*(1), 012084. https://doi.org/10.1088/1742-6596/1682/1/012084

Yeung, K. F., & Yang, Y. (2010). A proactive personalized mobile news recommendation system. *Proceedings - 3rd International Conference on Developments in ESystems Engineering, DeSE 2010*, 207–212. https://doi.org/10.1109/DeSE.2010.40

Yi, J., Wu, F., Wu, C., Li, Q., Sun, G., & Xie, X. (2021). DebiasRec: Bias-aware User Modeling and Click Prediction for Personalized News Recommendation. In *Proceedings of ACM Conference (Conference'17)* (Vol. 1, Issue 1). Association for Computing Machinery. http://arxiv.org/abs/2104.07360

Yu, X., Yuan, C., Kim, J., & Wang, S. (2021). A new form of brand experience in online social networks: An empirical analysis. *Journal of Business Research*, *130*(February 2020), 426–435. https://doi.org/10.1016/j.jbusres.2020.02.011

Zhang, C., Wang, H., Wang, W., & Xu, F. (2015). RCFGED: Retrospective Coarse and Fine-Grained Event Detection from Online News. *Proceedings - 2015 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2015*, 139–144. https://doi.org/10.1109/SMC.2015.37

Zheng, Y., Zhong, B., & Yang, F. (2018). When algorithms meet journalism: The user perception to automated news in a cross-cultural context. *Computers in Human Behavior*, *86*, 266–275.

Zhou, Kuscsik, Z., Liu, J. G., Medo, M., Wakeling, J. R., & Zhang, Y. C. (2010). Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences* , *107*, 4511–4515.

Zhou, Y., & Liao, H.-T. (2020). A Bibliometric Analysis of Communication Research on Artificial Intelligence and Big Data. *6th International Conference on Humanities and Social Science Research*, *435*, 456–459. https://doi.org/10.2991/assehr.k.200428.097

Zhou, Y., & Zhou, Z. (2020). Towards a Responsible Intelligent HCI for Journalism: A Systematic Review of Digital Journalism. *International Conference on Intelligent Human Computer Interaction*, 488–498.

Zhu, C., Zhu, H., Ge, Y., Chen, E., & Liu, Q. (2014). Tracking the evolution of social emotions: A time-aware topic modeling perspective. *2014 IEEE International Conference on Data Mining*, 697–706.

Zhu, Harrington, P., Li, J., & Tang, L. (2014). Bundle Recommendation in eCommerce. *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 657–666.

Zihayat, M., Ayanso, A., Zhao, X., Davoudi, H., An, A., Rogers, T., & Technology, I. (2019). A utility-based news recommendation system. *Decision Support Systems*, *117*(December 2018), 14–27. https://doi.org/10.1016/j.dss.2018.12.001

Zimnitskiy, I. (2021). *InstLoadGram*. Date Accessed 31st November 2021. https://www.instaloadgram.com/

Zontek, S. (2018). *User Engagement Drives Subscriptions. New RFV Engagement Scores from Deep.BI*. Deep.BI. https://www.deep.bi/blog/user-engagement-drives-subscriptions-new-rfv-engagement-scores-from-deep-bi

Zupic, I., & Cater, T. (2015). Bibliometric Methods in Management and Organization. *Organizational Research Methods*, *18*(3), 429–472. https://doi.org/10.1177/1094428114562629