Expert Systems With Applications 221 (2023) 119795

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Review

Data Science, Machine learning and big data in Digital Journalism: A survey of state-of-the-art, challenges and opportunities

Elizabeth Fernandes [a,*], Sérgio Moro [b], Paulo Cortez [c]

[a] ISCTE – Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Avenida das Forças Armadas, Edifício II, D615, 1649-026 LISBOA, Portugal
[b] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisboa, Portugal
[c] ALGORITMI Research Centre, University of Minho, Guimarães, Portugal

ARTICLE INFO

Keywords:
Data science
Digital journalism
Text mining
Systematic literature review
Media analytics
Machine Learning

ABSTRACT

Digital journalism has faced a dramatic change and media companies are challenged to use data science algorithms to be more competitive in a Big Data era. While this is a relatively new area of study in the media landscape, the use of machine learning and artificial intelligence has increased substantially over the last few years. In particular, the adoption of data science models for personalization and recommendation has attracted the attention of several media publishers. Following this trend, this paper presents a research literature analysis on the role of Data Science (DS) in Digital Journalism (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 integrating bibliometric search, text mining, and qualitative discussion, the relevant literature was identified and extensively analyzed. The review reveals an increasing use of DS methods in DJ, with almost 47% of the research being published in the last three years. An hierarchical clustering highlighted six main research domains focused on text mining, event extraction, online comment analysis, recommendation systems, automated journalism, and exploratory data analysis along with some machine learning approaches. Future research directions comprise developing models to improve personalization and engagement features, exploring recommendation algorithms, testing new automated journalism solutions, and improving paywall mechanisms.

1. Introduction

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 Expert Systems models have emerged (Davoudi et al., 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 Artificial Intelligence (AI) capabilities to the newsroom (Borges et al., 2021). Moreover, the adoption of Machine Learning (ML) methods is mentioned in the Reuters Digital Report as the new trend in media companies, especially for personalization and content recommendation (Newman et al., 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 sucessfully applied to diverse fields, such as Natural Language Processing, Social Network Analysis or business models development (Davoudi, 2018).

Motivated by the increase of interest in DS (including AI and ML) in Digital Journalism (DJ), this study presents a systematic, transparent and reproducible review process, through a Systematic Literature Review (SLR) (Abdelmageed & Zayed, 2020; Aria & Cuccurullo, 2017) integrating bibliometric search, text mining, and qualitative discussion of the literature. The time span time of study was 2010 and 2021. Only research literature about DS methods in DJ was considered. Our survey methodology presents four stages: study design, data collection and

* Corresponding author.
E-mail addresses: elisabeth.ferna@gmail.com, elizabeth.fernandes.data@gmail.com (E. Fernandes).

https://doi.org/10.1016/j.eswa.2023.119795
Received 14 September 2021; Received in revised form 2 March 2023; Accepted 2 March 2023
Available online 5 March 2023

selection, data analysis and findings, and discussion and results. Hence, the contribution of our research is threefold: (i) to identify and describe the state-of-the-art of existing approaches; (ii) to identify gaps, challenges, and opportunities on how to use DS to improve reader engagement in DJ; (iii) based on the identified gaps, to generate future recommendations and new research directions.

To summarize, the aim of this review is to present the most relevant works conducted in the field of DS in DJ by using a 4-step methodology. The remainder of the paper is structured as follows. First, Section 2 presents the background of Data Science in Digital Journalism and an introduction about literature analysis methods. Then, Section 3 describes the review methodology by describing the study design and the data collection steps. Followed by the descriptive analysis of the collection at Section 4. Then, the SLR main results are discussed in Section 5. Finally, Section 6 presents the main conclusions and research implications of this literature review.

2. Related work

2.1. Context and Motivation

Nowadays, media companies driven by economic pressures are investing in data and technological solutions to achieve business results. According to the International News Media Association (INMA) report (International News Media Association, 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.

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, no intensive survey on the role of DS in DJ has been recently published. Hence, by examining the existing research literature of the last decade, this paper surveys what has been done with DS methods in media. Moreover, one of the main contributions of this paper it is to present research gaps in the current literature and opportunities for future research.

2.2. Systematic literature review

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 paper assumes a Systematic Literature Review (SLR) (Abdelmageed & Zayed, 2020; Aria & Cuccurullo, 2017), which consists of a 4-step methodology. As presented at Table 1, 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 allow us to identify, evaluate and interpret the literature to answer relevant research questions (RQs) that are detailed at Section 3.

2.3. Contribution

Several studies present different techniques for gathering the state of the art on a research topic (Brous et al., 2020; Donthu et al., 2021). Table 2 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 clearly identified 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 while simultaneously avoiding human bias.

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, 2014), 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

Table 1
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 Protocol and Training	Study Design
	Data Collection	Search the literature Practical Screening Quality Assessment	Data Collection and Selection
Conducting the review	Data Analysis	Analysis and Findings	Data Analysis and Findings
	Data Visualization		
Reporting the review	Interpretation	Writing the review	Results and discussion

Table 2
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 communication 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.

knowledge, is innovative on a DJ survey (Zhou & Zhou, 2020).

3. Methodology

This section presents the proposed 4-step systematic method for reviewing the literature presented at Table 1. The SLR process begins, comprising study design, data collection and selection. Each stage encompasses several activities, as outlined in Fig. 1. The following subsection describes each stage of the SLR.

3.1. 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 paper, 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 preprocessing 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 above 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 (a head of digital media and analytics office at a wide audience journal; and two senior scholars in data science and analytics) analyzed all HLT decisions until a consensus was reached. It should be mentioned that one author is an analytics and audience insights manager in a

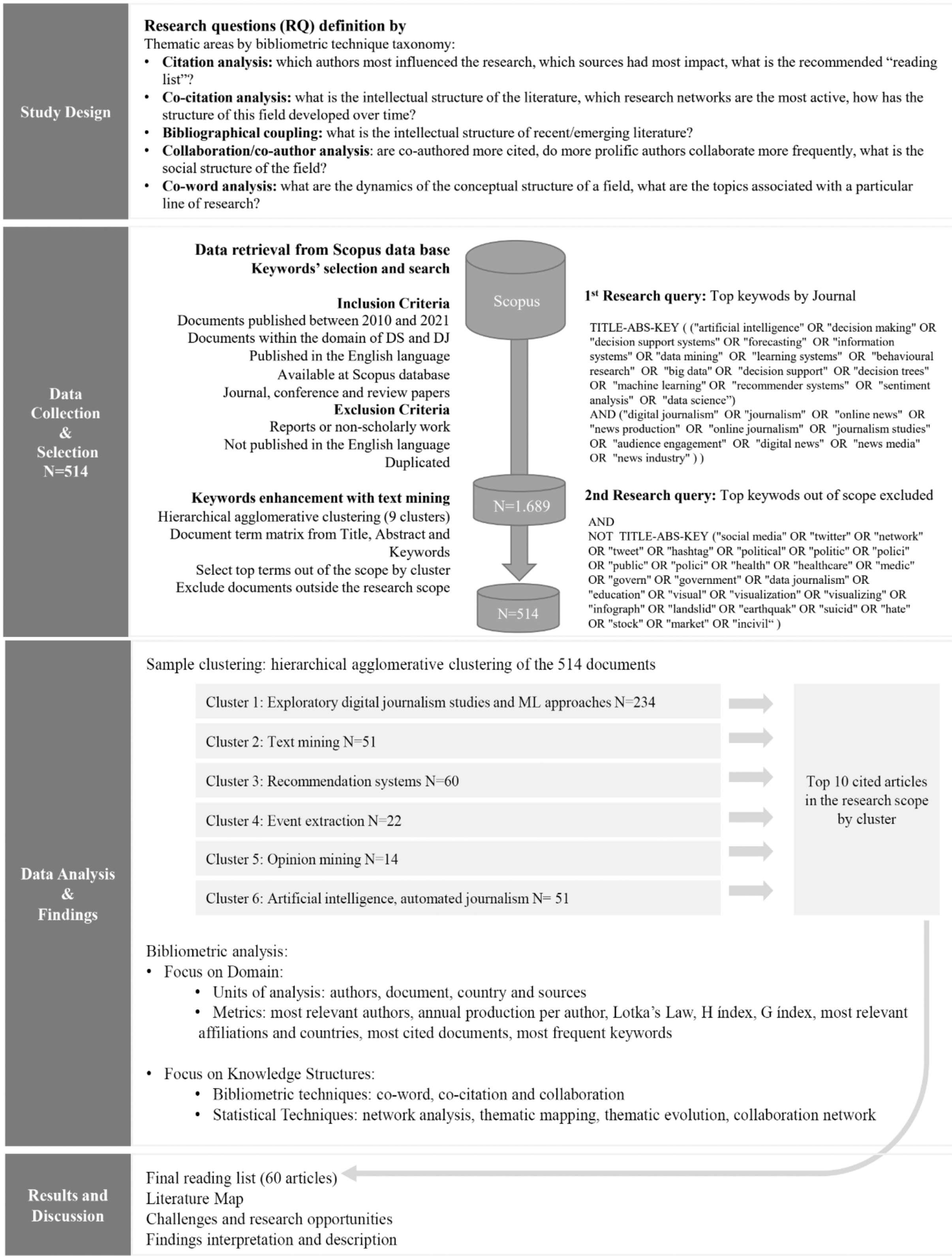


Fig. 1. Framework of the systematic literature review process.

national newspaper since 2015. Finally, the first query, with 25 keywords resulted in 1,689 documents, as presented at Fig. 1.

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 papers 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

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

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

R packages: stylo, factoextra

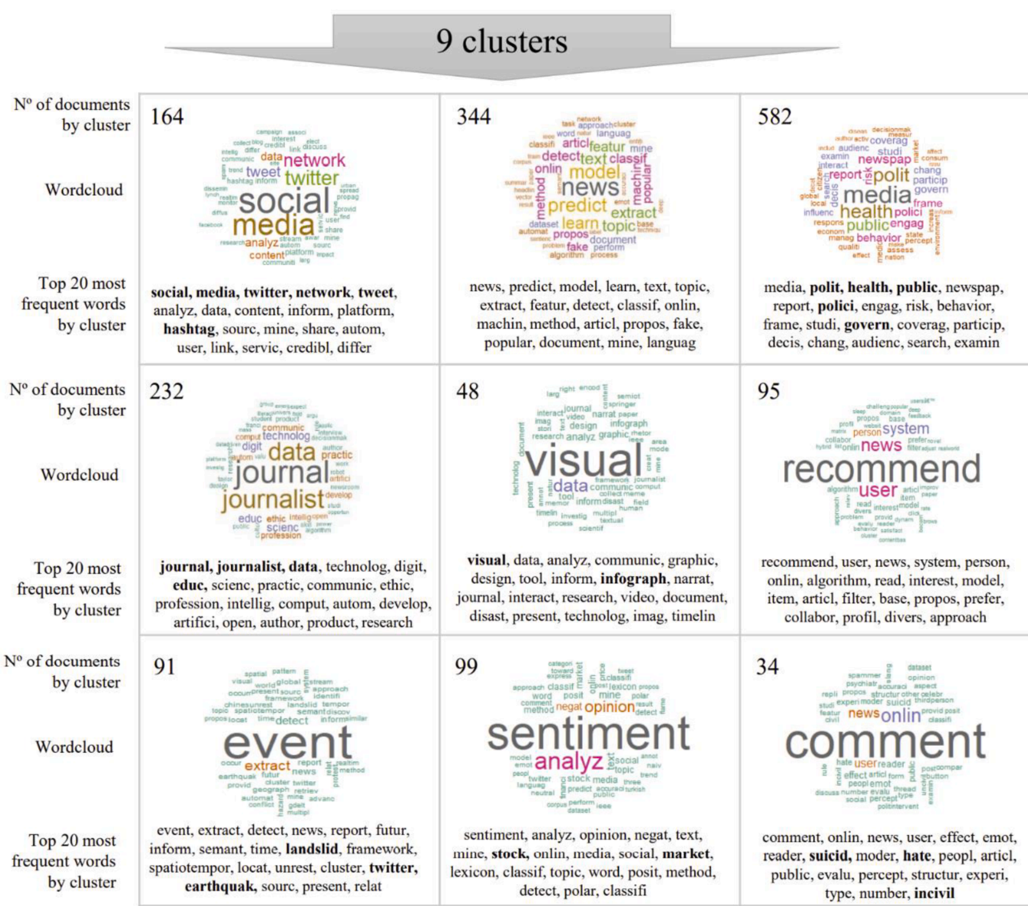


Fig. 2. Framework of the text mining process to find the keywords to exclude in the second search query.

research, nevertheless, we decided not to skim the article title and abstract to avoid a human bias.

3.2. 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 an example, the first cluster contains "social", "media", "twitter and "tweet" on the most frequent words. By reading the abstracts, we found research on social media platforms content and trends that were considered out of the scope (HTL). Hence, non-related documents were excluded from the Scopus search query by removing the words highlighted in bold (see Fig. 2).

4. Data analysis and findings

The present section aims to explore the thematic areas presented in Fig. 1. Hence, by performing citation, co-citation, collaboration and co-word analysis, complemented by a hierarchical clustering of the collection, the RQ1 presented in Section 3.1 can be answered. Then, in Section 5, a co-word analysis with keywords co-occurrences maps is also presented, which enables to answer both 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 papers (44%), 278 conference papers (54%) and 8 review papers (2%) (see Table 3). Furthermore, 47% of total sample was published between 2018 and 2020 (see Fig. 3). 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 Fig. 4 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 Fig. 4 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. Fig. 4 c) illustrates a bibliometric VOSviewer network visualization map of co-authorship (international collaboration) using country by average year

Table 3
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 paper	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

of publication and number of publications (Van 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 5 (thus answering to RQ1). In particular, Table 4 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 5 (Cole et al., 2011, 2015; Liu, Cole, 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).

In terms of researcher's impact measure, at Table 4 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. Fig. 5 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 4) and two of those are in the top most 20 cited articles of the sample (Cole et al., 2011; Liu, Cole, 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 5.

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, Fig. 6 presents the wordcloud of the top 50 author's keywords and highlights the most common keywords of the articles of the database

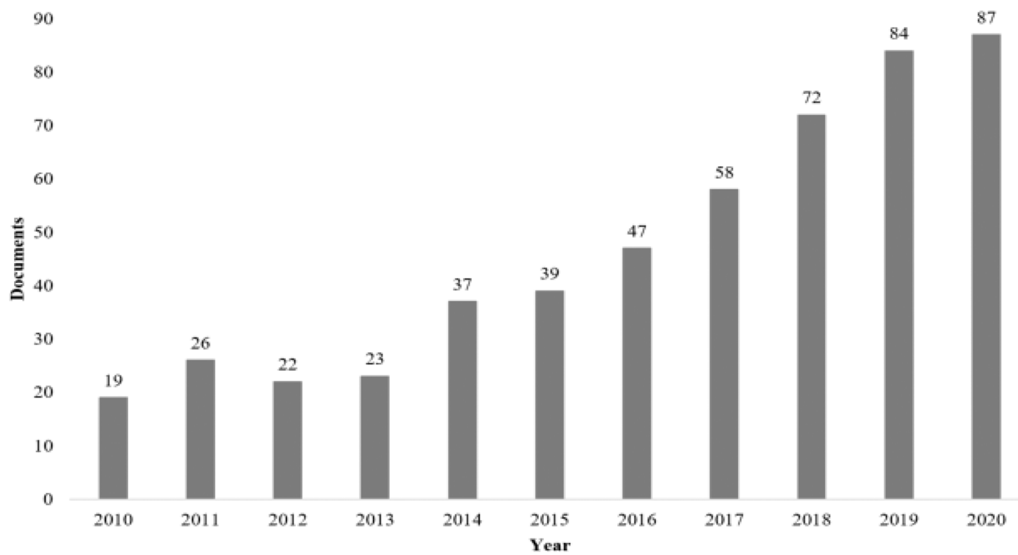


Fig. 3. Year-wise distribution of publications.

(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 and 13 times.

Moreover, Fig. 7 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 Fig. 8) to complement the data presented at Fig. 7, we can note that "text mining", "svm" (which stands for support vector machines, a supervised learning technique) 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 Fig. 7. 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 Fig. 9). 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 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.

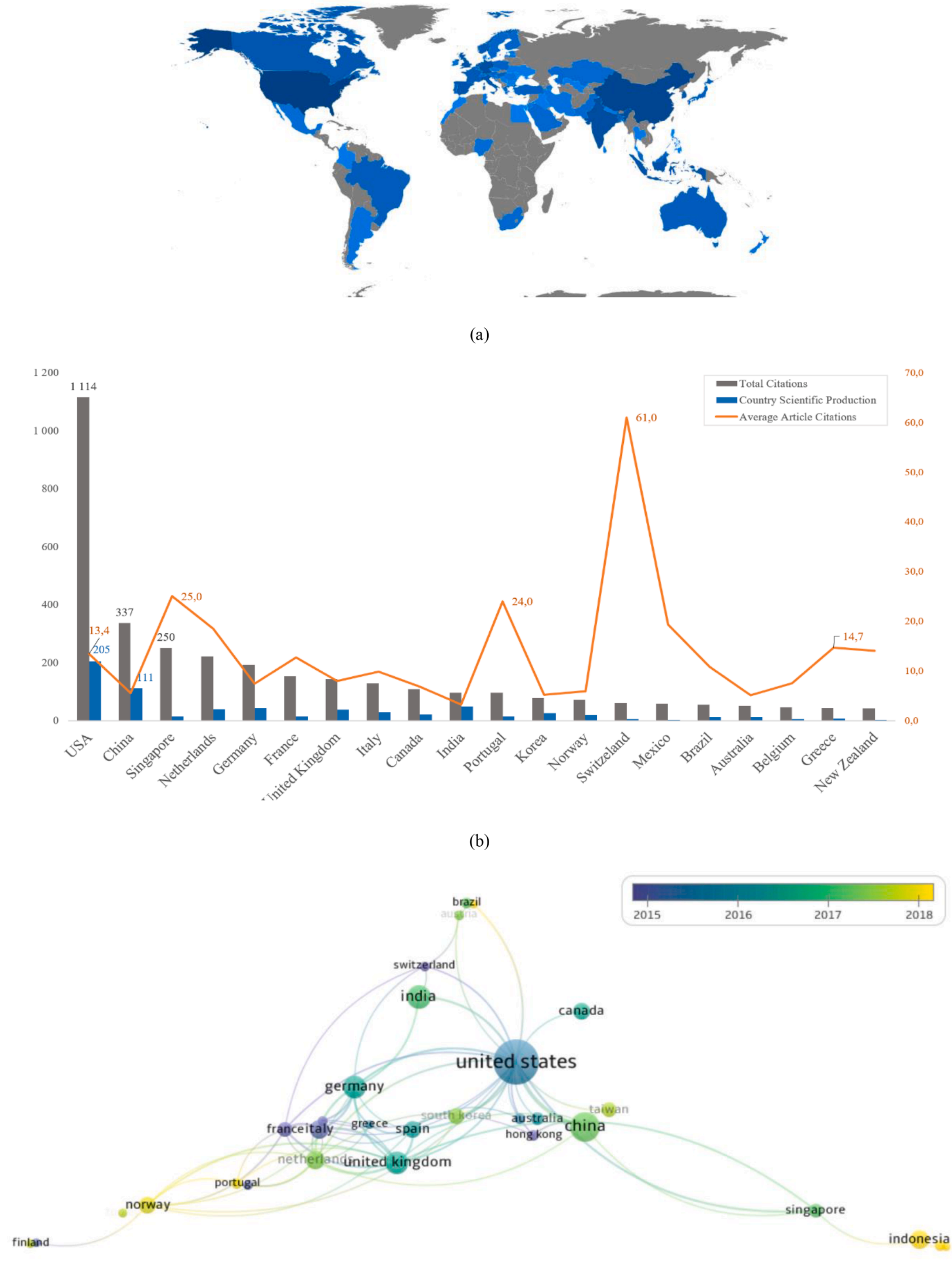
In order to define the final reading list, which presents the major topics of DS adoption in DJ (RQ1), this paper ranks the most cited articles by cluster (as presented in Table 5). Due to the number of documents, and to guarantee the quality of the selected publications, Table 5 was limited to the top 10 most cited articles related to the field under study by cluster. A content analysis was carried out by a meticulous abstracts' reading to guarantee that we selected contributions concerning the topic studied (HLT). Each document contains the number of total citations (TC), the TC per year and the normalized TC together with two journal's quality measures: SCImago Journal Rank (SJR) is a measure of the journal prestige calculated by considering the number of journal citations (Colledge et al., 2010); and the known Journal Citation Report (JCR) ISI impact factor quartiles, which reflect the quotient of a journal's rank in a scientific category and the total number of journals in the category (Garfield, 2006).

The top cited paper, which presents a news recommendation system, was published by (J. Liu, Dolan, 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 Fig. 8), 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; Lewis et al., 2019; Schonlau & Zou, 2020), related to "personalization", "journalism automation" and "statistical learning". However, the last two are not at Table 5, as their TC is lower than the top ten articles of their cluster. Nevertheless, both are mentioned in the literature map (see Fig. 12), as they present promising future research trends in journalism.

5. Discussion and challenges

In this section, the analysis conducted is based on the outcome from the procedure illustrated at Fig. 1. To answer the RQs, a deeper analysis across each cluster (see Fig. 9) allows 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 Fig. 12) that contains different levels of interactions, which are: the main domains found in the six clusters, DS topics of research in DJ and



(c)

Fig. 4. (a) Geographic distribution of published articles by country-based scientific production (b) total citations (left y-axis at grey colour), country scientific production (blue) and average article citations (right y-axis at orange) in the 20th most cited countries (c) VOSviewer network visualization map of country co-authorship by average year of publication and number of publications (documents weights). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
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%									

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, 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 Fig. 8) 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 5, 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 Fig. 12.

From the remaining five clusters, three are related to text analysis, the third main domain on the literature map (Fig. 12).

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 Fig. 10 indicating a line of research in DJ across the timespan (RQ1). Furthermore, 82% of the articles were published since 2015 (see Fig. 9) 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 (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 Fig. 12), defined as "event extraction" and "opinion mining" (see Fig. 9), 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 5). 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 (Wang, 2012; 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 Fig. 11) 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

Table 5
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	(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
	(I. Flaounas et al., 2010)	The structure of the EU mediasphere	23 (47 th)	1.9	0.6 (197th)	PLoS ONE	—	1.02
	(Rivera et al., 2014)	A text mining framework for advancing sustainability indicators	19 (57 th)	2.4	1.2 (129th)	Environmental Modelling and Software	4.807	1.9
	(Zhu et al., 2014)	Tracking the Evolution of Social Emotions: A Time-Aware Topic Modeling Perspective	18 (60 th)	2.3	1.1 (146th)	**Proceedings - IEEE International Conference on Data Mining, ICDM**	—	0.79
3 - Recommendation systems N = 60	(Liu 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)		32 (30th)	3.6	1.9 (85th)	Expert Systems with Applications	5.452	1.49

(continued on next page)

Table 5 (*continued*)

Cluster	Authors, Year	Title	TC (rank number)	TC per Year	Normalized TC (rank number)	Source (highlighted top 10 sources)	IF	SJR 2019
		Towards a journalist-based news recommendation system: The Wesomender approach						
	(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**	—	—
	(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
	(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
	(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	—	—
	(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	—	—
	(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
	(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
	(Chen & Ng, 2016)		23 (45th)	1.9	2.7 (197th)	**Computers in Human Behavior**	5.003	2.17

(*continued on next page*)

Table 5 (*continued*)

Cluster	Authors, Year	Title	TC (rank number)	TC per Year	Normalized TC (rank number)	Source (highlighted top 10 sources)	IF	SJR 2019
		Third-person perception of online comments: Civil ones persuade you more than me						
	(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, 2014).	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 & Rimestad, 2012)	Elimination of ideas and professional socialization: Lessons learned at newsroom meetings	12 (91st)	1.2	1.4 (125th)	**Journalism Practice**	1.542	1.26
	(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
	(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
	(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

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; Wu et al., 2019) (RQ3).

News recommendation systems development is a line of research that evolved from algorithms based on click behaviour (Liu, 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; Wang et al., 2017) or neural networks to solve the cold-start problem (Misztal-

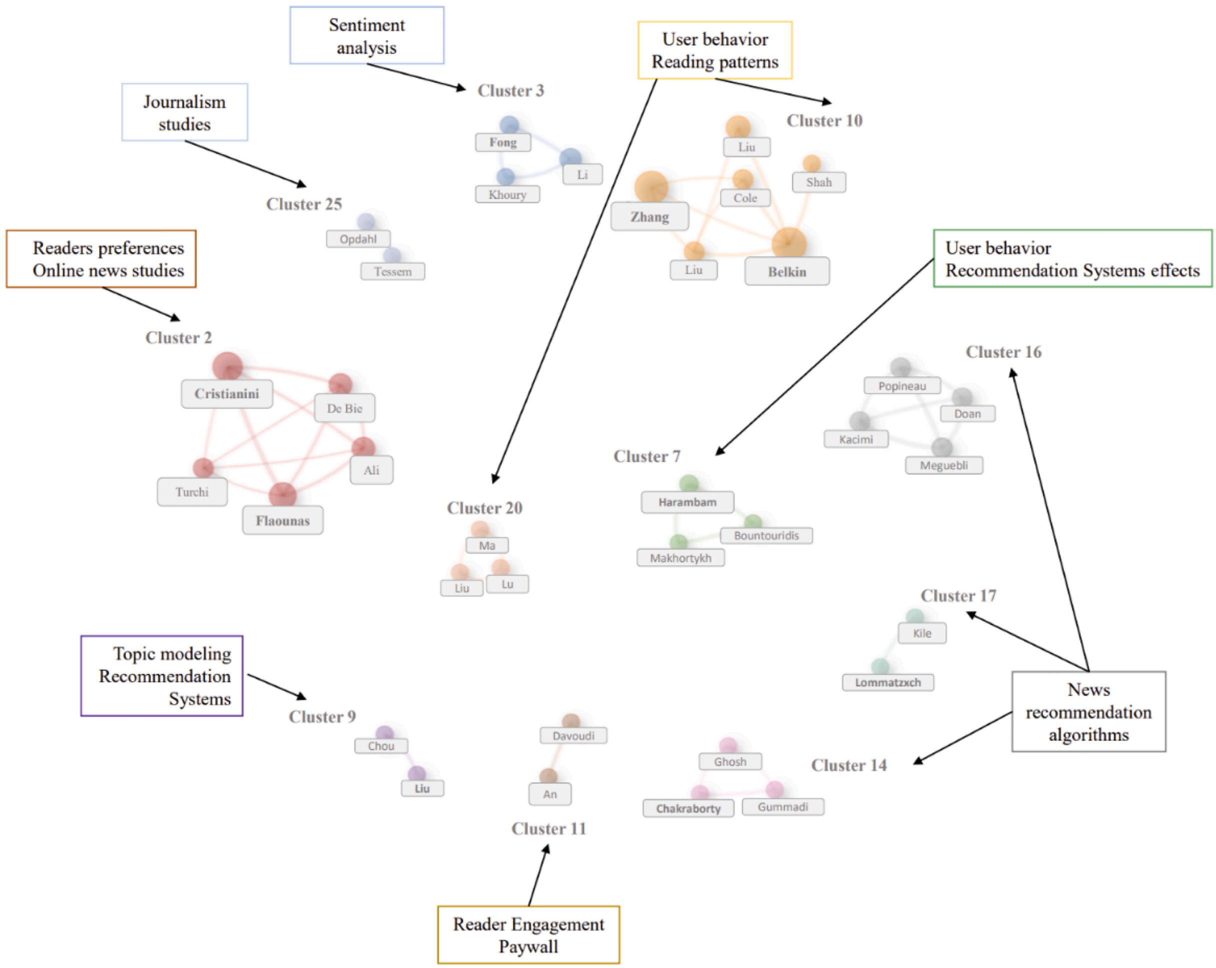


Fig. 5. *Bibliometrix* collaboration network map between authors from 11 of the 26 clusters of authors.

Fig. 6. Wordcloud of top 50 author's keywords (the word size depends on word occurrence).

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, 2014), 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 (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).

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 (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.

6. 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

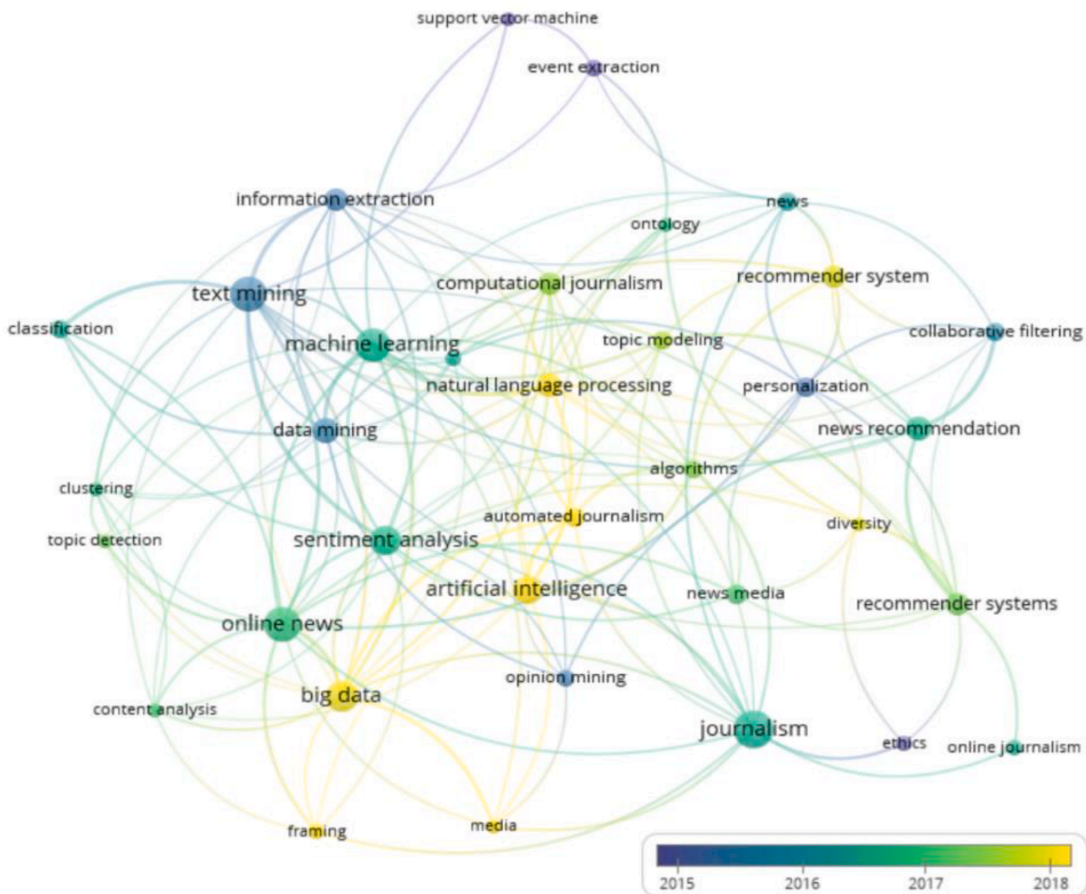


Fig. 7. VOSviewer co-occurrences map of keywords based on the full-counting method with a minimum number of occurrences of a keyword 5. The size of the nodes represent the relevance of the terms in the papers. The thickness of the lines means the bonding force between them. Finally, the colours indicate the average year of articles publication that mention those keywords.

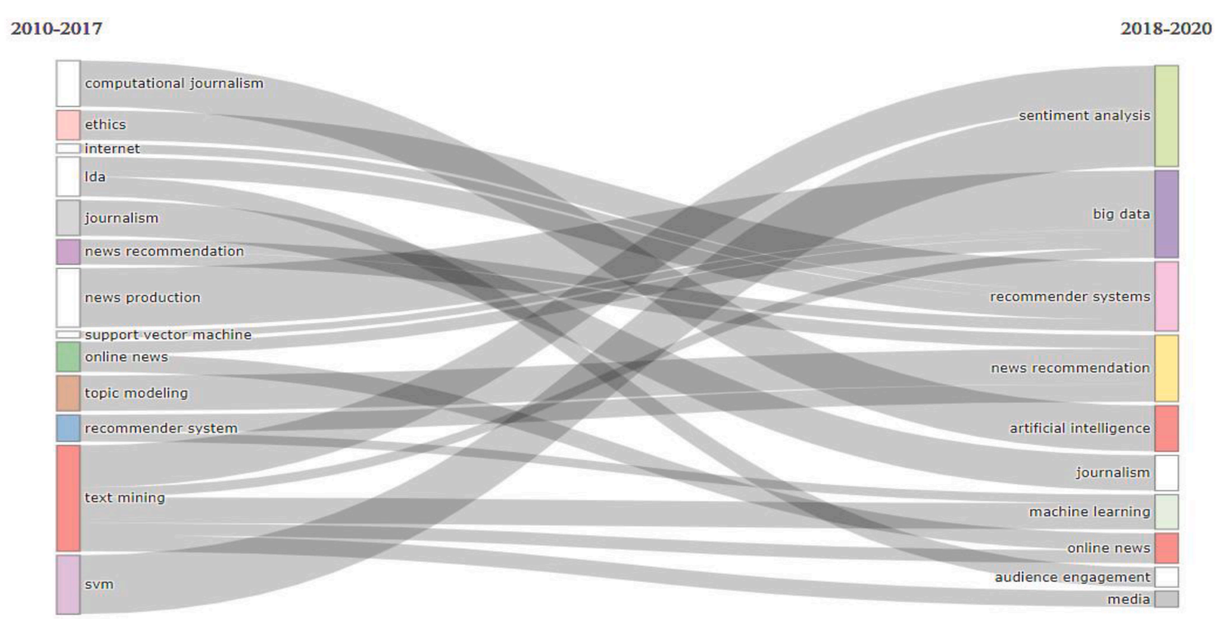


Fig. 8. *Bibliometrix* thematic evolution map that demonstrates the evolution of keywords in two different stages (2010–2017, 2018–2020).

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'

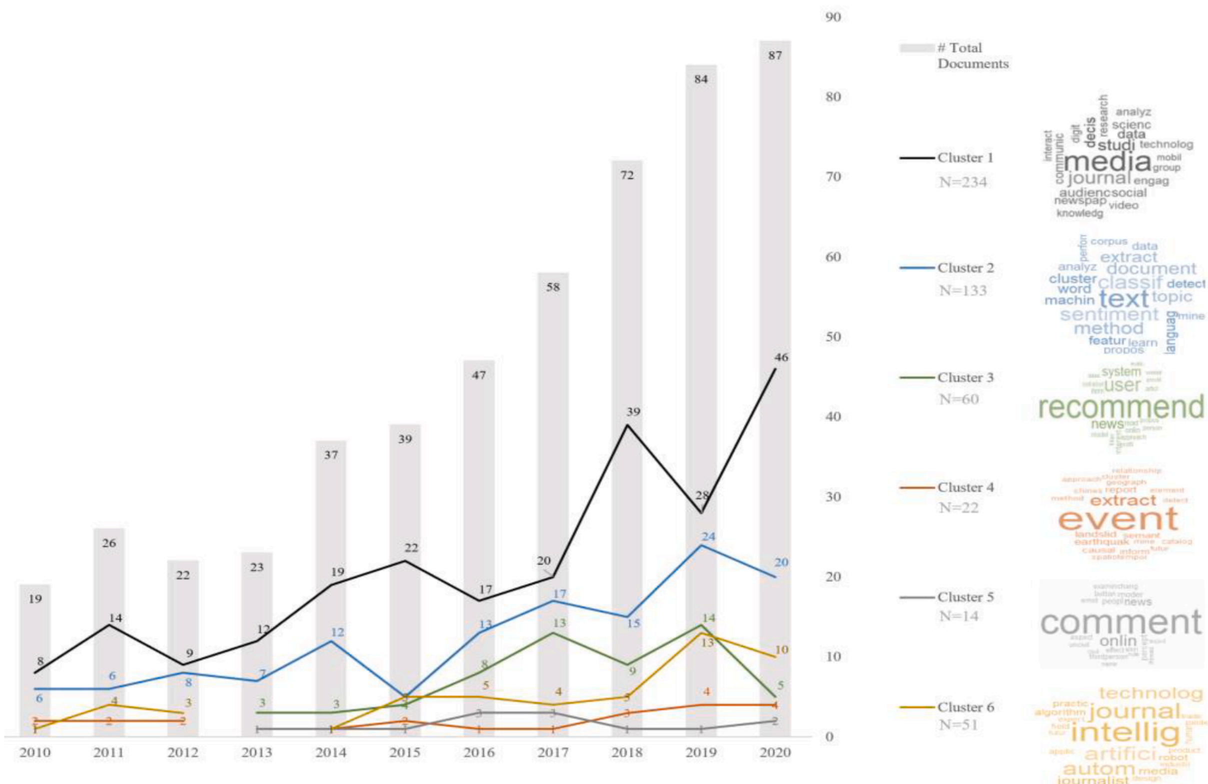


Fig. 9. Scientific Production by year and by cluster.

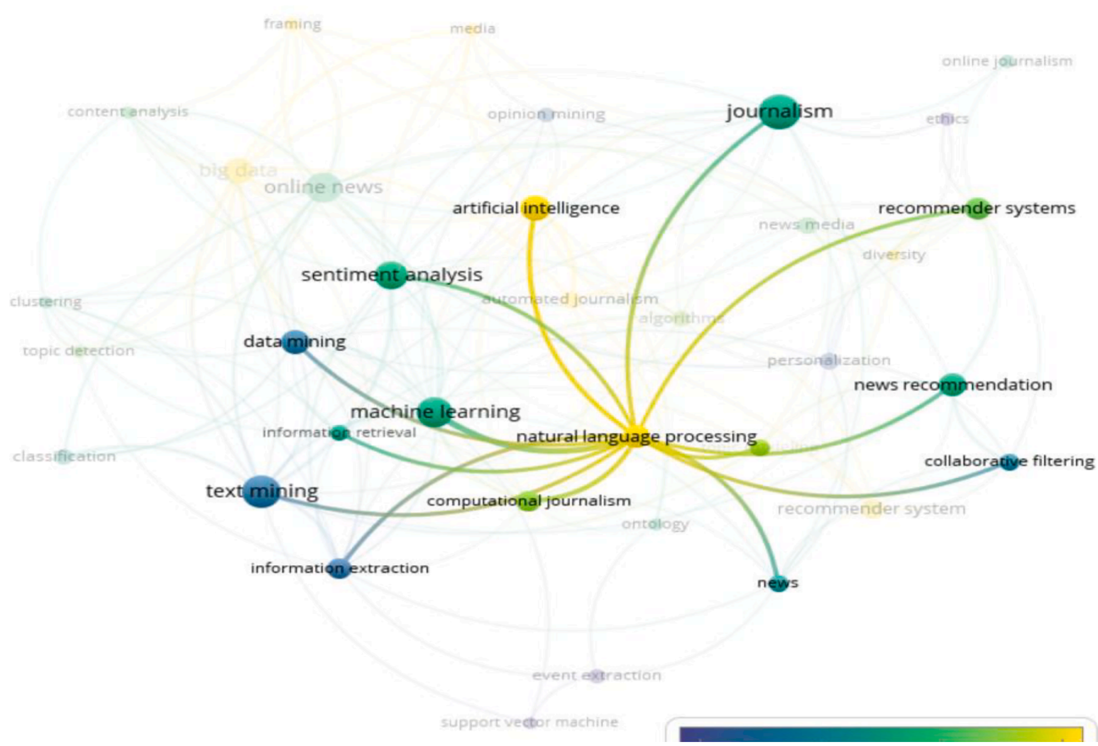


Fig. 10. VOSviewer keywords co-occurrences map based on the full-counting method (cluster 2 - "text mining"). The weight being visualized is the occurrence, thus when a keyword has a greater weight the label and bubble are bigger (Van Eck & Waltman, 2013).

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

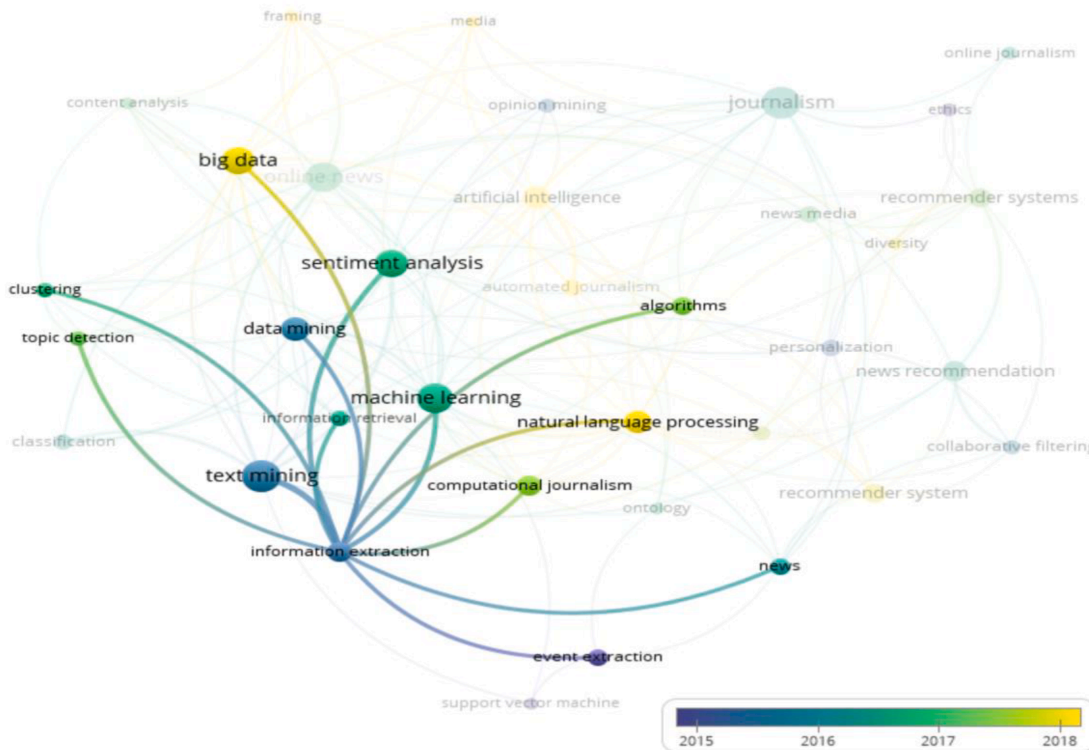


Fig. 11. VOSviewer keywords co-occurrences map based on the full-counting method (cluster 4 - "event extraction").

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, Dolan, 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 that there is a gap between journalism and ML communities.

7. Conclusion

In this paper, 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 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 (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 paper 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 papers and book chapters were neglected.

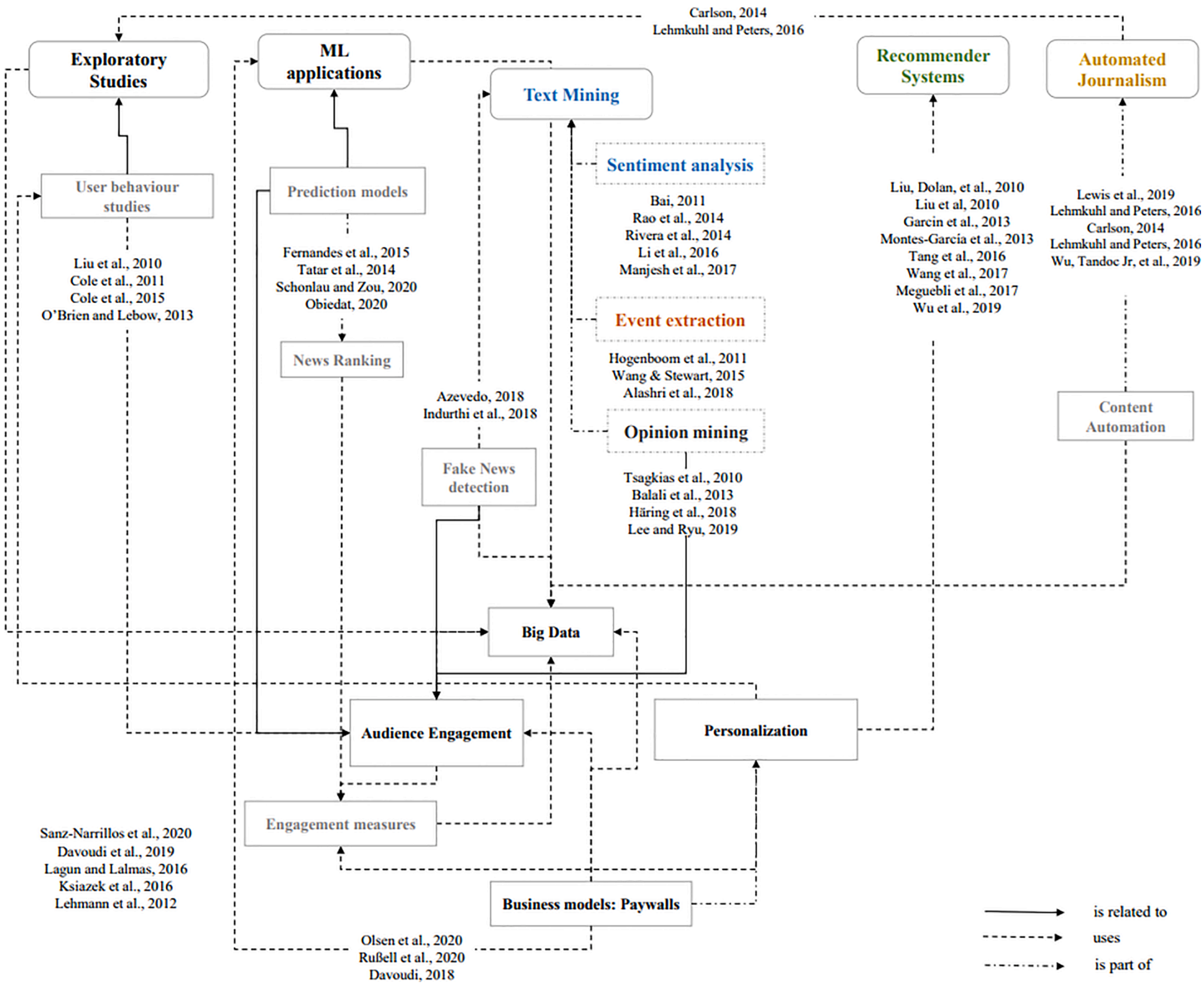


Fig. 12. Literature map.

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 papers 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.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the FCT - Fundação para a Ciência e Tecnologia, under the Projects: UIDB/04466/2020, UIDP/04466/2020, and UIDB/00319/2020.

References

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

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. 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.

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.

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, 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.

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.

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), Article 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), Article 101952. https://doi.org/10.1016/j.ijinfomgt.2019.05.008

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.

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

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

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

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.

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.

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

Colledge, L., Moya-Anegón, Guerrero-Bote, V., López-illescas, C., Aisati, M., & Moed, H. (2010). SJR and SNIP : two new journal metrics in Elsevier ' s Scopus. *Insights*, *23*(3), 215–221.

Cooper, H. (1998). *Synthesizing research*. SAGE.

Cortez, P. (2014). *Modern optimization with R*. Springer.

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

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.

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), Article 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.

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

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), Article 115555. https://doi.org/10.1016/j.eswa.2021.115555

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.

Flaounas, I., Ali, O., Lansdall-welfare, T., Bie, T. D., Lewis, J., Cristianini, N., … Bie, T. D. (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

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.

García-Avilés, J. A. (2014). 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.

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA—Journal of the American Medical Association, 295*(1), 90–93.

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), Article 104717. https://doi.org/10.1016/j.ssci.2020.104717

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

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

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

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), Article 101934. https://doi.org/10.1016/j.ijinfomgt.2019.04.007

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.

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

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.

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

International News Media Association, I. (2022). *The Benefits and Risks of Media Data Democratisation* (Issue January).

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. 10.1145/3377290.3377299.

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

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*. 10.1109/ICCUBEA47591.2019.9128691.

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. 10.1145/2835776.2835833.

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

Leetaru, K. (2011). *Culturomics 2.0: Forecasting Large-Scale human behavior using global news media tone in time and space*. First Monday.

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.

Lewis. (2015). Journalism In An Era Of Big Data Cases, concepts, and critiques. *Digital Journalism*, *3*(3), 321–330. 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.

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), Article 101965. https://doi.org/10.1016/j.techsoc.2022.101965

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.

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. 10.1145/3209978.3210007.

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)*.

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*, Article 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

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

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. 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.

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).

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

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

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.

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). 10.14569/IJACSA.2020.0110484.

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

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

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.

Reis, J., Olmo, P., Prates, R., Kwak, H., & An, J. (2015). Breaking the News : First Impressions Matter on Online News. *Proceedings of the International AAAI Conference on Web and Social Media, 9*(1), 357–366.

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 *In Proceedings of the 5th WSEAS International Conference on Computer Engineering and Applications* (pp. 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).

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

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.

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

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.

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), Article 115491. https://doi.org/10.1016/j.eswa.2021.115491

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

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

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

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

Tandoc, E. C., Jr (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. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 283–292).

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

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. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)* (pp. 95–98).
Tsagkias, M., Weerkamp, W., & De Rijke, M. (2010). News comments: Exploring, modeling, and online prediction. *European Conference on Information Retrieval*, 191–203.
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
Van Eck, N. J., & Waltman, L. (2013). VOSviewer manual. In Univeristeit Leiden (Issue February). http://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.1.pdf.
Viana, P., & Soares, M. (2016). A hybrid recommendation system for news in a mobile environment. In *6th International Conference on Web Intelligence, Mining and Semantics* (pp. 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).
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
Wang, H., Zhang, P., Lu, T., Gu, H., & Gu, N. (2017). Hybrid Recommendation Model Based on Incremental Collaborative Filtering and Content- based Algorithms. In *2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (pp. 337–342).
Wang, W. (2012). Chinese news event 5W1H semantic elements extraction for event ontology population. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 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., 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, Wei, Zhao, D., & Wang., D. (2012). Chinese news event 5W1H elements extraction using semantic role labeling. *2010 Third International Symposium on Information Processing*, 484–489. 10.1145/2187980.2188008.
Webster, J., & Watson, R. (2002). Analyzing the past to prepare for the future: writing a literature review. *MIS Quarterly*, *2*, xiii–xxiii. 10.1016/j.freeradbiomed.2005.02.032.
Welbers, K., Amsterdam, V. U., Atteveldt, W. V., 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
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, E. C., Jr, & 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.
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), Article 012135.
Yang, Y., Ma, X., & Fung, P. (2017). Perceived emotional intelligence in virtual agents. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2255–2262).
Yang, Y., Liu, Y., Lu, X., Xu, J., & Wang, F. (2020). A named entity topic model for news popularity prediction. *Knowledge-Based Systems, 208*, Article 106430. https://doi.org/10.1016/j.knosys.2020.106430
Yang, Z. (2020). Analysis of the Impact of Big Data Technology on News Ecology. *Journal of Physics: Conference Series, 1682*(1), Article 012084. https://doi.org/10.1088/1742-6596/1682/1/012084
Yeung, K. F., & Yang, Y. (2010). A proactive personalized mobile news recommendation system. In *Proceedings - 3rd International Conference on Developments in ESystems Engineering*. https://doi.org/10.1109/DeSE.2010.40
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. 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, 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. 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. *IEEE International Conference on Data Mining, 2014*, 697–706.
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. 10.1016/j.dss.2018.12.001.
Zupic, I., & Cater, T. (2015). Bibliometric Methods in Management and Organization. *Organizational Research Methods, 18*(3), 429–472. https://doi.org/10.1177/1094428114562629