




Universitat Autònoma de Barcelona

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  http://cat.creativecommons.org/?page_id=184

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <http://es.creativecommons.org/blog/licencias/>

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>

UNIVERSITAT AUTÒNOMA DE BARCELONA

DOCTORAL THESIS

Discovering Twitter through Computational Social Science Methods

Author:

Jingyuan YU

Supervisor:

Dr. Juan MUÑOZ JUSTICIA

*A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Psychology
in the*

Doctoral Program: Person and Society in the Contemporary World

Department of Social Psychology

December 3, 2020

Declaration of Authorship

I, Jingyuan YU, declare that this thesis titled, “Discovering Twitter through Computational Social Science Methods” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UNIVERSITAT AUTÒNOMA DE BARCELONA

Abstract

Faculty of Psychology
Department of Social Psychology

Doctor of Psychology

Discovering Twitter through Computational Social Science Methods

by Jingyuan YU

As Twitter has covered up people's daily life, it has become one of the most important information exchange platforms, and quickly attracted scientists' attention. Researchers around the world have highly focused on social science and internet studies with Twitter data as a real world sample, and numerous analytics tools and algorithms have been designed in the last decade. The present doctoral thesis consists of three researches, first, given the 14 years (until 2020) of history since the foundation of Twitter, an explosion of related scientific publications have been witnessed, but the current research landscape on this social media platform remained unknown, to fill this research gap, we did a bibliometric analysis on Twitter-related studies to analyze how the Twitter studies evolved over time, and to provide a general description of the Twitter research academic environment from a macro level. Second, since there are many analytic software tools that are currently available for Twitter research, a practical question for junior researchers is how to choose the most appropriate software for their own research project, to solve this problem, we did a software review for some of the integrated frameworks that are considered most relevant for social science research, given that junior social science researchers may face possible financial constraints, we narrowed our scope to solely focus on the free and low-cost software. Third, given the current public health crisis, we have noticed that social media are one of the most accessed information and news sources for the public. During a pandemic, how health issues and diseases are framed in the news release impacts public's understanding of the current epidemic outbreak and their attitudes and behaviors. Hence, we decided to use Twitter as an easy-access news source to analyze the evolution of the Spanish news frames during the COVID-19 pandemic. Overall, the three researches have closely associated with the application of computational methods, including online data collection, text mining, complex network and data visualization. And this doctoral project has discovered how people study and use Twitter from three different levels: the academic level, the practical level and the empirical level.

Acknowledgements

Writing a PhD thesis is a long and lonely journey, there were countless failures and frustrations I met during the last three years, it was my greatest luck that I could have companions during the hardest time. I would first express my gratitude to my PhD supervisor, Dr. Juan Muñoz-Justicia, I couldn't have finished my doctoral study without his generous help and tutorial. I would say during my study in UAB, I learned so many things far beyond academic research: the philosophy of open science, open access publication, free software etc. I started to use Linux, L^AT_EX, open source programming languages because I know how important open science is for researchers. And all these lessons came from him. Thanks Juan, words and letters can never express the same feelings I have right now, but you are one of the most important people in my life.

I would also give my sincere appreciation to all the professors and staffs in the department, which include, but not limit to Dr. Lupicinio Íñiguez-Rueda, Dr. Susana Pallarès-Parejo, Dr. Miguel Torregrossa, Dr. Miquel Domènech-Argemí, Dr. Jesús Rojas-Arredondo, Dr. Isabel Pellicer, Mrs. Cristina Prats-Vilarós etc. Our department is like a warm family, I wish I could never leave your side.

To my dear PhD colleagues, including all the Laicos members, Dani Baltrán, Renata Guerda, Carmen Rojo, Jinfang Yang etc. It is always my pleasure to be able to share my details with you, and I won't forget the beautiful moments which I enjoyed the most with you.

To my dear friends, Anna Ventura, Meiying Zhu, Tuo Liu and Yanyan Zhou. Thanks for giving me your constant care and support, and most importantly, the encouragement to help me to finish my doctoral study.

To my family members, my parents Jianhua Xing and Jianhua Yu, and my little sister Jinghang Yu, it's been more than 10 years since the first day I left home and started to live on my own. I know you are my closest ties in this world, and thank you for all your supports whenever and whatever I need during so many years.

Finally, to Pei Li, always the most special one in my life, my best friend, Alt+QQ. You got me.

Contents

Declaration of Authorship	iii
Abstract	v
Acknowledgements	vii
List of Figures	xi
List of Tables	xiii
1 Introduction	1
1.1 Social networking our contemporary world	2
1.2 Synthesizing Twitter-related studies	4
1.2.1 Common methods and techniques in literature review	4
1.2.2 Current research tendencies on Twitter-related studies	5
1.3 Twitter research tools in general	7
1.4 Health communication, framing and information sharing on Twitter	8
1.4.1 Framing: from traditional media to social media	8
1.4.2 COVID-19 communication studies in Spain and other countries	10
1.4.3 Computational social science and automated methods on news frame detection	11
1.5 Research Aim and Objective	13
1.5.1 General Objective	13
1.5.2 Specific Objectives	13
2 First Publication: A Bibliometric Overview of Twitter-Related Studies Indexed in Web of Science	15
3 Second Publication: Free and Low-Cost Twitter Research Software Tools for Social Science	35
4 Third Publication: Analyzing Spanish News Frames on Twitter during COVID-19-A Network Study of El País and El Mundo	63
5 Conclusion	77
Bibliography	81

List of Figures

1.1	Process Model of Framing Research	9
1.2	Spanish news media tweet heatmap	11

List of Tables

1.1	Previous bibliometric/literature reviews on Twitter-related studies . . .	6
1.2	Spanish news media tweet dataset: detailed information	11

List of Abbreviations

COVID-19	CO rona VI rus Disease 2019
API	A pplication P rogramming Interface
LDA	L atent D irichlet A llocation
BA	B ibliometric A nalysis
R&D	R esearch & D evelopment
WHO	W orld H ealth O rganization
CSS	C omputational S ocial S cience
CCS	C omputational C ommunication S cience

Chapter 1

Introduction

The present doctoral thesis is carried out by the mode of compendium, under the framework of the doctoral program Person and Society in the Contemporary World, Department of Social Psychology, Universitat Autònoma de Barcelona. The structure of this thesis is introduction, peer-reviewed publications, and conclusion. Three published scientific articles have been included: First, a bibliometric analysis which aimed to describe the current academic research tendencies about Twitter and to analyze the evolution of the main research themes of Twitter-related studies. Second, a software review which aimed to provide a Twitter research guideline for social science research beginners from the perspective of methodology and affordability. Third, an empirical study which employed computational methods and focused on analyzing the evolution of Spanish news frames during the COVID-19 pandemic.

Corresponding to the three included papers, which are the core components of this thesis, my PhD project conducts Twitter research from three different dimensions, the dimension of academic research, by using the metadata of the Twitter-related publications, my paper illustrated the general research environment of Twitter studies during the last 14 years (since 2006), it is considered as a synthesis of all the relevant scientific literatures, providing an overview of the previous research hotspots. The dimension of utility, by analyzing the advantages and disadvantages of 9 free and low cost Twitter research software tools, my research is not only able to provide a guideline for choosing the most appropriate software tool for Twitter research, but also offers an easy-understanding explanation of the frequently adopted research methods. Based on the two fore-mentioned researches, we have already solved two questions: what did the researchers study about Twitter? and what kind of research can we do by using the currently available tools? Therefore, the third dimension of this PhD thesis emphasized on the application of these knowledge. Given the ongoing public health crisis and the explosion of news coverage on health issues, we did an empirical study to analyze the reaction of Spanish news media (El País and El Mundo) before, during and after the announcement of Spanish national lockdown, our data was retrieved from the news media's Twitter account and a set of computational methods (topic modeling and semantic network) were adopted in this study.

The implication of this PhD thesis is also threefold. First, it provided an on-time update for the previous bibliometric and systematic literature reviews, these studies were published at least 4 years ago, in today's terms, they are a bit dated due to the rapid development of computer science. Second, it benefited the junior social science researchers who are interested in Twitter research, unfamiliar with open source programming languages and facing financial constraints. Third, taking the advantage of short-text information, my thesis enhanced our understanding of the news bias in different stages of the public health crisis. It is a fact that this thesis may contain some limitations, for example, the research data of the bibliometric analysis were

retrieved from Web of Science, other prestigious scientific databases (e.g. Scopus) were neglected. In the software review we only discussed integrated frameworks, but the widely used open source software (e.g. R, Python) packages were not within our scope. In the empirical research, as the short-text news posts may be different from lengthy news articles, and our study solely focused on the Twitter data, it may not fully equivalent to the news media themselves because of the gatekeeping effect. But given the previously mentioned implications, I believe that my PhD thesis has brought sufficient findings and innovations to fulfill the requirement of our doctoral program.

1.1 Social networking our contemporary world

Social media have covered up people's daily life, becoming one of the most important information exchange platforms. They are considered powerful tools and medium in many fields, which include, but not limit to the politics (Rainie & Smith, 2012), health (Mansfield et al., 2011), business (Qualman, 2012), sports (Filo, Lock, & Karg, 2015) etc.

According to Pew Research Center, in the United States, about 70% of the people are social media users, and the majority of them visit these sites everyday (Perrin & Anderson, 2019). In the case of Spain, the statistics of the *Instituto Nacional de Estadística* (INE) have shown that about 65% of the Spanish people are social media users, and more than 90% of the total population use Instant Messaging Service (INE, 2019). The high penetration of internet and daily usage of social media has provided people more possibilities to engage and interact with each other, even with professionals from different areas (e.g. politics, health, sport etc.) in a more direct way, thus, get involved in social events (e.g. (Filo et al., 2015; Guidry, Jin, Orr, Messner, & Meganck, 2017; Hand & Ching, 2011)).

There are different kinds of social media, some of them are highly popular around the world, for example, Facebook, Instagram, Snapchat, and Twitter (Smith & Anderson, 2018). Among them, Twitter has the highest bridging social capital (Phua, Jin, & Kim, 2017), the bridging social capital refers to weak and distant relationships between individuals that make available opportunities for information sharing (Putnam et al., 2000), in other words, information on Twitter have a high possibility to be shared across the boundaries of social groups (e.g. race, class, religion etc.). This unique characteristic of Twitter makes this online platform a valuable research object for social scientists. Because Twitter users may have a bigger opportunity to engage and participate in social events, and exchange information with the "unknown" individuals in their real-life (Jin & Phua, 2014; D. Williams, 2006).

An excellent example can be made by analyzing Twitter use in the politics. It has been demonstrated that the use of social media has a positive relation with citizen engagement (Skoric, Zhu, Goh, & Pang, 2016) and political participation (Boulianne, 2015), and Twitter in obviously not an exception, it has shown its power in political elections (Jungherr, 2016) and social movements (Buettner & Buettner, 2016; Murthy, 2012; Zhu, 2017). This microblogging service was seen as a key tool in leading Obama and Trump's victory in the US presidential elections (Enli, 2017; Francia, 2018; Tumasjan, Sprenger, Sandner, & Welpel, 2011), and the Twitter campaign strategy is widely adopted by politician all over the world (e.g. (Bajaj, 2017; Kruikeimeier, 2014; López-Meri, Marcos-Garciá, & Casero-Ripollés, 2017)).

Contemporary social movements has been tightly associated with the use of social media, some of them were even called as "Twitter revolution" (Christensen,

2011), Lotan et al (2011) argued that Twitter plays a key role in amplifying and spreading timely information during the 2011 Tunisian and Egyptian revolutions. Also, by sharing images on Twitter, domestic grass-root users are more associated with activists and foreign users, the cyber-movement went to a global scale thanks to the Twitter engagement (Kharroub & Bas, 2016). From another perspective, this online platform provide journalists and/or online opinion leaders a boarder space to disseminate their information, and these "elite" users may be more influential in the internet era than in the past, serving as catalysts for different types of online movements (Barnard, 2018; Carter Olson, 2016; Isa & Himelboim, 2018).

In addition to socio-political issues, Twitter also shows its power on health information and communication. "Twitter discussion could remove boundaries between scientists, health professionals, and policy makers, creating a new diverse community that gives everyone a voice and an opportunity to contribute" (Lancet, 2014, p. 1641). The fast information exchange could facilitate the coordination among health workers, especially when facing an epidemic outbreak. On the other hand, the huge amount of Twitter data have also provided public health practitioners with "a quantitative indicator of anxiety, anger or negative emotions in the general public" (p. 2207) and this indicator could help to alleviate anxiety and correctly communicate the risk associated with the public health crisis (Fung, Tse, Cheung, Miu, & Fu, 2014). What's more, given Twitter's broad reach and high bridging social capital, it also has the potential to make public health campaigns (Wehner et al., 2014), and to be used for medical education (Goff et al., 2019).

At the time of writing this doctoral thesis (2020), the world is suffering from the pandemic of COVID-19 (coronavirus disease 2019), a worldwide epidemic outbreak first identified in December 2019 in Wuhan, China. Until September 2020, a total number of 28.3 million confirmed cases have been reported, and the virus has caused 913 thousand deaths. The unprecedented public health crisis triggered a huge amount of online discussion in Twitter, and relevant studies have already become one of the hottest research field in 2020 (Belli, Mugnaini, Baltà, & Abadal, 2020; Chahrour et al., 2020). And scientists have retrieved massive twitter datasets for COVID-19 studies in social science and computer science (Shuja, Alanazi, Alasmay, & Alashaikh, 2020). Such data provides researchers the opportunity to track online behaviors and reactions regarding the ongoing pandemic, for example, Das and Dutta (2020) used sentiment analysis and topic modeling methods to analyze the COVID-19 related tweets in India, examining the evolution of public attitude toward this health crisis in the Southern Asian country. Similarly, Kruspe et al (2020) did a cross-language sentiment analysis of European Twitter messages during the pandemic, by correlating the temporal development with events in European countries, the authors explained the effect of the health situation on people's moods. From another perspective, as news media have widely adopted Twitter as a primary and timely tool to release news, this microblogging service has become one of the main information source for the public, especially during the health crisis (Masip et al., 2020). Researchers have also used tweet dataset from news media accounts to analyze the evolution of news focus during the different development stages of the pandemic (Yu, 2020; Yu, Lu, & Muñoz-Justicia, 2020).

Twitter is certainly one of the most studied social media platforms, for example, it is the primary data source for most of the social media and disease surveillance studies (Charles-Smith et al., 2015), the second most focused research object for social media marketing studies (Alves, Fernandes, & Raposo, 2016), and misinformation (Y. Wang, McKee, Torbica, & Stuckler, 2019). Despite its value on bridging social capital, the easy-access of Twitter data (with a built-in API) and short-text form of

message (with a maximum of 280 characters) provide researchers more availability to manage and analyze Twitter information. However, doing social science research with Twitter data may contain certain limitations, for example, Twitter users tend to be younger, wealthier and more educated (Blank, 2017), "the unrepresentative characteristics of Twitter users suggest that Twitter data are not suitable for research where representativeness is important" (P. 679). Besides, due to the strict word limit of a single tweet, the online narratives in 280 characters (or less) may provide limited value for semantic analysis (Veil, Buehner, & Palenchar, 2011). Researchers should pay attention to the potential limitations while dealing with Twitter-related studies.

1.2 Synthesizing Twitter-related studies

1.2.1 Common methods and techniques in literature review

There are many ways to illustrate and analyze the landscape of a research field, for example, systematic literature review, meta-analysis and bibliometric analysis. A systematic literature review is a review of "a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research and to collect and analyze data from the studies that are included in the review" (Moher, Liberati, Tetzlaff, Altman, Group, et al., 2009, p. 1). Systematic literature review is different from traditional (ordinary) literature review in a journal publication or a thesis/dissertation, in other words, a systematic literature review is a research study, addressing research questions and using the literature as data to be coded, analyzed and synthesized to make results or conclusions (Ridley, 2012).

Meta-analysis is a statistical procedure that integrates the results of several independent studies from a specific research domain, in a general sense, it is considered to be a special type of literature review (Egger, Smith, & Phillips, 1997; Ridley, 2012). "It determines the direction and size of the effect and whether it is consistent or different across studies. This synthesis gives more strength to the assessment of the effectiveness of an intervention because it is a combined result from a number of different investigations and thus gives a more precise measure" (Ridley, 2012, p. 193).

However, the two fore-mentioned review methods may have certain limitations. For the systematic reviews, they are usually carried out from qualitative approach, in which the manual selection, coding and analysis are ordinary procedures, and it may make the review article opportunistic and biased, because it report only the literature that is pertinent to the purpose or point of view of the author (Galvagno, 2017). And because of these limitations, the reproducibility of the review results is sometimes questionable. As for the meta-analysis, it is only suited for quantitative studies, and is sometimes biased on the statistical methods (Schmidt, 2008). At the same time, to make sure the study objects (the reviewed publications) are relevant enough for the research scope of the author, doing meta-analysis also requires a strict manual selection and coding process (Borenstein, Hedges, Higgins, & Rothstein, 2011), which may lead to the same limitation as the one systematic review has. Last but not least, due to the manual selection/coding procedure, they are highly time consuming, and both of the two methods may more suitable for a very narrow research field/scope.

Bibliometric analysis (BA) uses the metadata of the scientific publications (e.g. authors, citations, keywords etc) to measure the scientific impact, influence and relationships of the academic works in a certain research field (Van Raan, 2003). Amidst an explosion of academic publications, manually analyzing and studying the current research tendencies of a subject became more and more complicated and

difficult. Generally speaking, BA contains two main procedures, performance analysis and science mapping (Gutiérrez-Salcedo, Martínez, Moral-Muñoz, Herrera-Viedma, & Cobo, 2018; Tang, Liao, Wan, Herrera-Viedma, & Rosen, 2018). "Performance analysis aims at evaluating groups of scientific actors (countries, universities, departments, researchers) and the impact of their activity on the basis of bibliographic data." (Gutiérrez-Salcedo et al., 2018, p. 1275). Science mapping uses network and/or computational methods to visualize and extract knowledge from the intellectual, social or conceptual structure of a research field (Gutiérrez-Salcedo et al., 2018).

One of the biggest advantages of BA is that it is able to summarize and synthesize the landscape of a (relatively broader) research domain, and it is suitable for any kind of research (for social science, mainly qualitative and quantitative study), because it uses the bibliographic information of a publication as its data, in this sense, it is more comprehensive than meta-analysis, which, as we have previously mentioned, is only for analyzing quantitative researches. Considering the main research object in this doctoral project – Twitter, it is a well studied social media platform, a huge amount of scientific publications can be found on academic databases (Yu & Muñoz-Justicia, 2020), different kinds of research methods have been applied (Weller, 2014; Shirley A Williams, Terras, & Warwick, 2013; Shirley Ann Williams, Terras, & Warwick, 2013), BA may be a better option for illustrating the landscape the Twitter research.

1.2.2 Current research tendencies on Twitter-related studies

Researchers have tried to review Twitter-related studies from systematic review and BA approaches. Williams et al (2013) manually coded, reviewed and analyzed 1161 Twitter-related publications, they classified these works according to three types of categories (methods, subject and approach), they found that the majority of the publications focus on the study of messages and users. Four main methodological approaches have been highlighted, they are analytic, design and development, examination, and knowledge discovery, varying across different research domains. In another of their publication (Shirley Ann Williams et al., 2013), by retrieving and analyzing 134 Twitter-related scientific literatures from PubMed, they argued that the early Twitter studies on medical science are mainly for introducing the topic and highlighting the potential of the social media platform, but later scholars started to use knowledge discovery methods and data mining techniques to analyze vast datasets, and the study of Twitter is becoming quantitative research. Zimmer and Proferes (2014) put their focus on Twitter related research disciplines, methods and especially, ethics. They did a content analysis on 382 Twitter literatures, computer science and information science are the two main research disciplines, content analysis is the predominant research method in Twitter studies. The authors highlighted their concerns on big data research ethics, along with the emerging of data mining methods and techniques in Twitter studies, very few scientific works (16 out of 382) made mention of ethical issues (data collection, public information, user privacy etc.), which should be paid a larger attention for future studies.

To the best of my knowledge, the study of Weller (2014) is the first to adopt BA methods to analyze Twitter-related studies, the author retrieved the metadata of 370 Twitter-related studies in social science, by executing a set of performance analysis (e.g. yearly output, citation rank etc.) and manual categorization (method, domain and dataset), the author has listed the most important researches of Twitter in the

early years, and argued that both experimental and analytical approaches are significant in social science Twitter studies, the dataset size varies across different kinds of studies. Kang and Lee (2014) collected 539 articles about Twitter study from 2009 to 2014, they used a co-word analysis method, and successfully visualized 53 disciplines on a two-dimensional semantic network, with journalism, business, computer science and political issues to be the core research fields. The most recent bibliometric research on Twitter-related studies was conducted by Gupta et al (2016), compared to the fore-mentioned studies, their research used a bigger dataset (with 4709 publications), their study was solely based on performance analysis (e.g. annual growth rate, average citation, top countries, top institutions etc.), their study has presented a basic description of the Twitter research landscape until 2015. Table 1.1 summarised the detailed information of all the fore-mentioned reviews.

TABLE 1.1: Previous bibliometric/literature reviews on Twitter-related studies

Item	Author	Year	Sample Size	Research Domain	Publication Title
1	Williams et al (a)	2013	1161	General	What do people study when they study Twitter? Classifying Twitter related academic papers
2	Williams et al (b)	2013	134	Medical science	How Twitter is studied in the medical professions: A classification of Twitter papers indexed in PubMed
3	Bruns et al	2014	382	General	A topology of Twitter research: disciplines, methods, and ethics
4	Weller	2014	370	Social science	What do we get from Twitter—and what not? a close look at Twitter research in the social sciences
5	Kang and Lee	2014	539	General	A bibliometric analysis on Twitter research
6	Gupta et al	2016	4709	General	A Bibliometric Assessment of Global Literature on " Twitter" during 2008-15

All the listed publications are excellent reviews about Twitter studies, but several important limitations should be mentioned. First, all the studies are carried out several years ago, even the most recent one, was published in 2016, with the rapid growth of Twitter-related publications and the application of new computational methods (e.g. machine learning, deep learning etc.), the current research landscape on this social media platform is still unknown. Second, the research methods used in the fore-mentioned reviews are simple, either by manual categorization or by performance analysis, no in-depth analysis results were presented, the only work which provides insights with mapping techniques (Kang & Lee, 2014) applied a very limited sample size. Third, none of the above-mentioned studies are able to make a longitudinal analysis, the evolution of the research trends on this subject remained in blank. Hence, a more timely and comprehensive research to illustrate the current landscape of Twitter study should be prepared and drafted.

1.3 Twitter research tools in general

As Twitter-related studies are becoming quantitative (Weller, 2014) and computational research (Yu & Muñoz-Justicia, 2020), more and more studies rely on data-driven analysis, numerous computer software and analytic tools were developed (Ahmed, Bath, & Demartini, 2017). They have demonstrated a high efficiency of dealing with large-scale online data.

For example, the study of Vicari (2020) used the tool "Mozdeh" to collect and analyze the Twitter data about the discussion of breast cancer gene mutation, the author explored how and to what extent personal stories shape health content on the online platform. Ahmed et al (2020) used the tool "NodeXL" to collect and analyze the 5G conspiracy during the COVID-19 pandemic, their study demonstrated that although the topic attracted high volume on Twitter, only a small group of users genuinely believed the conspiracy. Burnap et al (2016) used the tool "COSMOS" to collect and analyze Twitter data and to predict the results of UK 2015 General Election. Poell and Rajagopalan (2015) used the tool "DMI-TCAT" (Digital Methods Initiative - Twitter Capture and Analysis Toolset) to collect and analyze the tweets about the gang rape incident in New Delhi in 2012, they argued that while traditional media moved their focus to other issues several days after the incident, online media like Twitter has taken the incident more vitality, keeping the discussion consistently on the front burner. Park, Kim and Ok (S. " Park, Kim, & Ok, 2018) used the tool "TAGS" (Twitter Archiving Google Sheet) to collect the geo-tagged tweets at Disneyland, by running an emotion analysis, they identified three hot spots in the part where pleasant tweets were posted.

Besides the fore-mentioned integrated frameworks (which allow both data access and data analysis) (Antonakaki, Fragopoulou, & Ioannidis, 2020), there are also plenty of semi-integrated software packages that provides more flexibility for researchers, for example R packages (e.g. Rtweet, twitteR etc.) and/or Python modules (e.g. tweepy), and interested researchers can also use these packages together with other types of software/package/module (e.g. packages for text mining, network analysis) to make an advanced study. Despite the fact that they are all high-quality software for Twitter research, it should be mentioned that in order to fully master these software, an extensive knowledge of programming is required. Given the education and training background of the majority of the social science early-stage researchers, it is recommended that the beginners shall start to learn from the basics of data collection and execution mechanism. For this purpose, the integrated frameworks are excellent examples. Apart from the above-mentioned Ahmed's software list (Ahmed et al., 2017), there are also social media software lists like Social Media Data Stewardship - Social Media Research Toolkit ¹, which gives researchers a wide selection of their ideal tools. However, none of these lists are able to present the details of the software tools, in other words, by browsing these lists, it is not easy for the Twitter research beginners to choose the right one for their project. To solve this problem, I plan to make a roadmap for the early-stage researchers, pointing out the advantages and disadvantages of these tools, reducing the unnecessary waste of time.

Due to the pandemic, scientists around the world are suffering from financial constraints, European Union has greatly reduced the research budget for the next 7 years (2021-2027) (Wallace, 2020). On the other hand, in the case of Spain, the R&D funding from Spanish government has not fully recovered since the 2008 economic

¹<https://socialmediadata.org/social-media-research-toolkit/>

recession, and it dropped by 25% between 2009 and 2016 (Rehm, 2018). According to another statistics, the Spanish public funding in R&D was reduced by 9.8% between 2010 and 2018, and in 2018, only 51% of the budget was executed ². Going back to our discussion on Twitter research tools, some of them require a costly purchase, which may be not a wise investment for social science researchers under the current financial situation, but there are indeed some free and low-cost software tools that may bring interested scientists the hope to conduct big data study. Our study on the Twitter research software shall put the focus on these tools.

1.4 Health communication, framing and information sharing on Twitter

1.4.1 Framing: from traditional media to social media

Social media have become one of the primary information sources for the public in the time of health crisis (Masip et al., 2020). Like we have argued in the first chapter, health communication is an important research area in Twitter study. Many traditional news media (newspapers, television etc.) registered their official Twitter account, and use this social media platform to post news updates in the first time. However, there are indeed some significant differences between the Twitter news updates and full-length news articles. Throughout an automatic topic analysis for both Twitter updates and news articles, Zhao et al (2011) believe that Twitter is a good source of entity-oriented topics that have low coverage in traditional news media. Regarding health issues, Twitter messages not only play the role of spreading information from credible sources, but also serve as a source of opinions and experiences, which is beneficial for health authorities, because this allows the professionals to respond to public concerns (Chew & Eysenbach, 2010). In another comparative study, Zhang, Bie and Billings (2017) compared the newspaper articles and Twitter posts in the 2014 Ebola outbreak. While newspapers fulfilled traditional media responsibilities, the Twitter updates reflected more public concern during the epidemic escalation. Besides, newspaper articles trend to use more alarming and reassuring tones than Twitter posts, for the latter case, contained mainly neutral tone.

An important research field in communication and media study is framing. "Framing essentially involves selection and salience. To frame is to select some aspects of a perceived reality and make them more salient in a communication text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described" (Entman, 1993, p. 52). Frames highlight part of the information about a specific topic or theme, and thereby elevating them in salience (Entman, 1993). Scheufele (1999) developed a process model of framing research (see Figure 1.1), in which the relation of four main processes (frame building, frame setting, individual-level effects of framing, and journalists as audiences) have been explained. And they are considered fundamental elements in constructing the framing as a theory of media effects.

A common understanding of frame building is how the professional and elite actors (e.g. journalists, politicians etc) can impact the framing of news content. And the framed content can thus transmit the attribute salience to the public, shaping the audience's perception and frames, known as the framing setting. This effect may eventually influence the individual's behavior, attitude and cognition (individual-level effects of framing). And finally, the journalists, like their audiences, are equally

²<http://informecotec.es/>

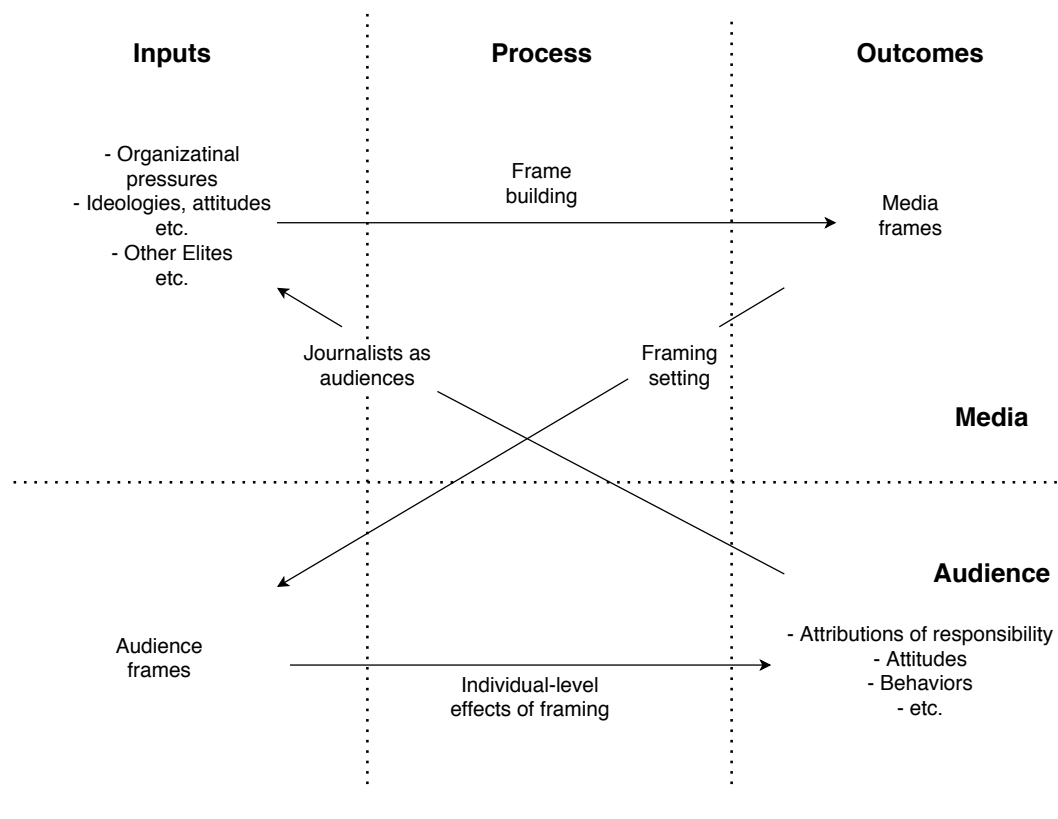


FIGURE 1.1: Process model of framing research (Scheufele, 1999)

susceptible to the frames that they used to describe events and issues (journalists as audience), making the loop go back to the beginning point (Scheufele, 1999).

A huge number of health communication publications have focused on the social media framing study during sanitary crisis. Liu and Kim (2011) studied how organizations framed 2009 H1N1 pandemic via social and traditional media, they classified the content into four frames, general crisis, disaster, health crisis and general health issues, arguing that the organizations tended to use social media as much as traditional media regarding the general crisis frame, while for other cases, the organizations relied more on traditional media than social media. Kilgo, Yoo and Johnson (2018) focused on media and news frames, they examined the news shared on Reddit during the 2014 Ebola crisis, they classified the news frames into six categories, blame, praise, general risk, risk in the United States, solution and speculation. They argued that the Reddit news amplified panic and uncertainty surrounding the epidemic outbreak. The study of Wang and Guo (2018) is grounded in the framing setting process, they investigate how the online news and Twitter framed the discussion during the Zika crisis, throughout a combination of manual coding and machine learning methods, they classified the text data into two general categories, benefit-oriented (includes health, cost-effectiveness, economic frames) and risk-oriented (includes health, environmental, ethical, experimental, cost-effectiveness frames). They found that Twitter discussion was more benefit-oriented, while the online news were more balanced. At the same time, by doing a time series analysis, they argued that intermedia frame setting may change its focus and orientation over time.

1.4.2 COVID-19 communication studies in Spain and other countries

The global health situation in 2020 is complicated, with regarding the worldwide outbreak of COVID-19, humankind may facing the most severe public health crisis in the 21st century. Following World Health Organization's (WHO) announcement of pandemic on March 11, 2020³. Spain stepped into a strict national lockdown from March 15, 2020. People were forced to temporally work from home. There are already a great number of social science studies focusing on the pandemic in Spain, especially in information and health communication areas. At the level of government information, the communication strategy of the Spanish government mainly focused on the impact of the pandemic, policymaking, hygiene standards and social behavior (Castillo-Esparcia, Fernández-Souto, & Puentes-Rivera, 2020). The survey of Moreno, Fuentes-Lara and Navarro (2020) analyzed the effect of information-seeking behavior and message reception in public's evaluation during the Spanish lockdown, the results showed that people who relied more on mainstream news media tended to express positive opinion toward government's communication strategy. On the contrary, people who are less able to make correct attributions of governmental information tended to be the most critical individuals toward the official response. At the level of news consuming, since the very beginning of the national lockdown, there is a significant increase of news reports focused on the health problem, digital media play an important role in the news dissemination (Lázaro-Rodríguez & Herrera-Viedma, 2020), and the public is more informed in the lockdown period, but maintaining a critical attitude toward the news coverage, conditioned by media ideology (Masip et al., 2020). Tejedor et al (2020) did an content analysis on Spanish and Italian newspaper's front-pages, their results showed that the front-page images tend to foster humanization through an emotional representation of the pandemic, and politicians are the most represented actors, meaning a high degree of politicization of the crisis.

There are also several studies focused on the framing study about the pandemic. For example, the study of Basch, Kecojevic and Wagner (2020) showed that financial impact of COVID-19, stories of affected individuals, death and death rates, precaution recommendation for public and quarantine are the five most common news frames for highly circulated US newspapers, they also argued that news media play a vital role in enhancing the public understanding of the current pandemic. Poirier et al (2020) adopted an automatic content analysis to analyze the Canadian news frames on the COVID-19 pandemic, they found a significant difference between francophone and anglophone media regarding the use of health crisis, social impact and chinese outbreak frames. They also pointed out that computational methods (topic modeling) is a useful approach for frame analysis. More importantly, Park, Park and Chong (2020) used Twitter as an information source, analyzed the COVID-19 news frames of South Korean media, they manually organized six news frames (conflict, human interest, attribution of responsibility, morality, medical, entertainment), and found that tweets containing medically framed news articles were more popular than those with non-medical frames. Their study has demonstrated the potential of using Twitter information to detect news frames.

Since social media is one of the main information source for the public, and Twitter is one of the most important social media platforms, I built an open source institutional and news media Twitter dataset from the very beginning of the pandemic

³<https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020>

(Yu, 2020), which include the tweets from more than one hundred international organizations, governments and news media. Among them, six Spanish news media are selected as the most representative ones. Table 1.2 gives the detailed information of the selected Spanish media (updated on September 23, 2020). And Figure 1.2 presents the COVID-19 tweeting frequency heatmap by Spanish news media, related tweets are selected by keywords "covid" and "coronavirus". It is easy to observe that at the beginning of the pandemic, and the first days of the Spanish national lockdown, the media attention on COVID-19 are much higher than other normal days. Considering the fore-mentioned scientific literatures, it would be an interesting idea to adopt our news media tweet dataset to analyze the evolution of Spanish news frames across time.

TABLE 1.2: Spanish news media tweet dataset: detailed information

News media	Twitter account	Timespan	Tweet amount
ABC	@abc_es	2020/02/24-2020/09/23	42652
El País	@el_pais	2020/02/25-2020/09/23	41458
El Mundo	@elmundoes	2020/02/19-2020/09/23	35708
El Periodico	@elperiodico	2020/02/24-2020/09/23	31817
La Vanguardia	@LaVanguardia	2020/02/25-2020/09/23	37758
RTVE	@rtve	2020/02/17-2020/09/23	31949

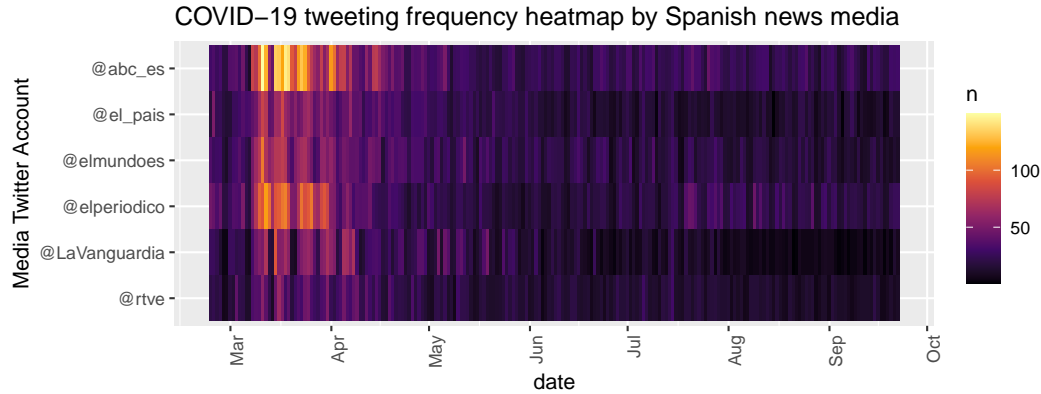


FIGURE 1.2: Spanish news media tweet heatmap

1.4.3 Computational social science and automated methods on news frame detection

Computational social science (CSS) is a rapidly developing interdisciplinary research domain, emerged in the last decades (Lazer et al., 2009), it aims to employ the cutting-edge computational methods (e.g. large-scale networks, natural language processing etc) to solve social science questions, leveraging "the capacity to collect and analyze data with an unprecedented breadth and depth and scale" (p. 722). One of the most important research objects in CSS is the Internet, because "the Internet offers an entirely different channel for understanding what people are saying, and

how they are connecting" (p. 722), naturally, CSS shows a great power in social media research.

Following the development of CSS, computational communication science (CCS) is becoming an important discipline in communication science (Domahidi, Yang, Niemann-Lenz, & Reinecke, 2019), similar to CSS, CCS "offers an opportunity to accelerate the scope and pace of discovery in communication research" (Van Atteveldt, Strycharz, Trilling, & Welbers, 2019, p. 3935). The terms of CSS and CCS seem intuitive, but the concepts are not easy to define, computational researches highly rely on computer hardware and software, but "a method is executed on a computer does not make it a computational method" (Van Atteveldt & Peng, 2018, p. 82). In turn, researchers have provided a clear criteria to conceptualize CCS: "(1) large and complex data sets; (2) consisting of digital traces and other 'naturally occurring' data; (3) requiring algorithmic solutions to analyze; and (4) allowing the study of human communication by applying and testing communication theory" (Van Atteveldt & Peng, 2018, p. 82). Computational methods are "revolutionary", because they allow us to analyze and investigate social behavior in ways that were impossible before.

Computational methods have furthered the application of many communication theories, which include, but not limit to agenda-setting, two-step flow of information, selective exposure, and of course, framing (Nicholls & Culpepper, 2020; Van Atteveldt & Peng, 2018). For different types of communication research, there are also more than one possible (computational) solutions, taking framing studies as an example, clustering algorithm (e.g. K-means) (Burscher, Vliegenthart, & Vreese, 2016), factor analysis (e.g. Evolutionary Factor Analysis) (Motta & Baden, 2013), and topic modeling (e.g. Latent Dirichlet Allocation, Structural Topic Models) (Poirier et al., 2020; Roberts et al., 2014) were widely applied to answer related research questions. According to the newest research outcomes, topic modeling is considered a better way to conduct computational framing study (Nicholls & Culpepper, 2020; Poirier et al., 2020).

In natural language processing, a topic model is a statistical model that aims to discover the latent topics from vast text data sets, apart from the fore-mentioned topic models (LDA, STM), there are also other types of models such as Latent Semantic Analysis (LSA), Non-negative Matrix Factorization (NMF) etc. But LDA is undoubtedly one of the most used topic models in CCS (Maier et al., 2018; Walter & Ophir, 2019). It is "a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document" (Blei, Ng, & Jordan, 2003, p. 993).

In addition to the previously mentioned Poirier's study (2020) in identifying the Canadian COVID-19 news frames. LDA was widely adopted in other empirical researches, the study of Li et al (2020) used the data from Chinese microblogging platform (Sina Weibo), identified and analyzed the frames adopted in the Chinese #MeToo movement, their study demonstrated the effectiveness of using LDA on short-text data. The study of Xu, Ellis and Laffidy (2020) analyzed the flu vaccine related frames in US newspapers, their research has shown the power of using LDA in health news frame detection. Therefore, based on our Twitter news dataset, we believe that LDA is an appropriate method to analyze the Spanish news frames during the COVID-19 pandemic.

1.5 Research Aim and Objective

1.5.1 General Objective

To fill our previously proposed research gaps: the current Twitter research environment remain uncertain, the availability of the free and low-cost Twitter research software tools should be further investigated, and the needs to conduct framing study on Spanish news media tweets regarding enfacing the pandemic. The main objective of this doctoral thesis is to investigate, through computational methods, the current environment of how people study and use this microblogging service.

1.5.2 Specific Objectives



- To identify the main research themes of the Twitter-related studies, and how the Twitter studies evolved over time.
- To study the academic environment of Twitter-related study from a macro level, by using bibliometric methods.
- To identify the characteristics of Twitter analytic tools and their use in the field of social sciences.
- To analyze the evolution of the Spanish news frames regarding their COVID-19 Twitter updates.

Chapter 2

First Publication: A Bibliometric Overview of Twitter-Related Studies Indexed in Web of Science

Article

A Bibliometric Overview of Twitter-Related Studies Indexed in Web of Science

Jingyuan Yu *  and Juan Muñoz-Justicia 

Department of Social Psychology, Universitat Autònoma de Barcelona, 08193 Barcelona, Spain;
Juan.Munoz@uab.cat

* Correspondence: jingyuan.yu@e-campus.uab.cat

Received: 23 April 2020; Accepted: 19 May 2020; Published: 20 May 2020



Abstract: Twitter has been one of the most popular social network sites for academic research; the main objective of this study was to update the current knowledge boundary surrounding Twitter-related investigations and, further, identify the major research topics and analyze their evolution across time. A bibliometric analysis has been applied in this article: we retrieved 19,205 Twitter-related academic articles from Web of Science after several steps of data cleaning and preparation. The R package “Bibliometrix” was mainly used in analyzing this content. Our study has two sections, and performance analysis contains 5 categories (Annual Scientific Production, Most Relevant Sources, Most Productive Authors, Most Cited Publications, Most Relevant Keywords.). The science mapping included country collaboration analysis and thematic analysis. We highlight our thematic analysis by splitting the whole bibliographic dataset into three temporal periods, thus a thematic evolution across time has been presented. This study is one of the most comprehensive bibliometric overview in analyzing Twitter-related studies by far. We proceed to explain how the results will benefit the understanding of current academic research interests on the social media giant.

Keywords: twitter; bibliometric analysis; science mapping; bibliometrix

1. Introduction

With more than ten years of prosperity and development, Twitter possesses 330 million monthly active users that send about 500 million tweets per day [1]. Previous reports [2,3] indicated that Twitter was losing its users, but statistics show that the trend of active users in this social network platform is still relatively positive [4].

Data from diverse social network platforms is being used by researchers to develop “a better understanding of how people are using social media in specific circumstances” [5]. Under the global tendency of using Twitter as a daily communication and information tool [6], scientific research about this social network platform has maintained a high growth rate year by year [7]. Twitter data, compared with other digital platforms (e.g., Facebook, Instagram, Snapchat, etc.), is more accessible and can contain valuable resources for academic research; besides, the wide range of data-retrieving method options makes Twitter one of the most studied objects in the social sciences [5,8].

Figuring out the focus of scholars when they study Twitter became a realistic problem in understating such a rapidly developing research field. There are some academic works focusing on this issue; for example, Williams, Terras and Warwick [9] qualitatively reviewed the title and abstract of 1161 Twitter-related articles, they classified these remaining academic works across three dimensions: aspect, method and domain, they found that the majority of the publications relating to Twitter concentrates on messages sent and details of the users. Kang and Lee [10] applied a co-word analysis to a limited bibliographic data of the Korea Citation Index, revealing 53 different disciplines in Twitter scientific literatures. Gupta et al. [7] quantitatively ranked 4709 Twitter-related studies by

various categories, including annual global publication, geographic distribution, subject distribution, top keywords, top productive institutions, top authors etc.

Above-mentioned studies have successfully argued the current research environment about Twitter-related studies, but important limitations were also included: First, as the study of Gupta et al. revealed, the total number of academic output of Twitter study is growing rapidly; thus, their study may lose accuracy and representability in today's view. Second, none of the listed academic publications systematically analyzed the common characteristics of the Twitter scientific literatures, the current Twitter studies' community structure remains in blank. Third, fore-mentioned studies were mainly descriptive, no analytic insights were explicitly discussed or concluded regarding to how do the related study hotspots or domains were evolved across time.

In this paper, we aim to update the current knowledge boundary in Twitter-related studies by amplifying the research sample, and provide a longitudinal analysis to discuss our proposed research gap.

2. Literature Review

2.1. Twitter and Its Research Lines

One of the most discussed research field of Twitter was its implication on political issues [10], recent years, scholars have argued the influence of using Twitter in sociopolitical movements [11–13], in political elections and campaigns [14–16]. Despite the fact that how much influence Twitter has in such events remains under discussion, scholars' enthusiasm toward Twitter in politics seems increasing. Along with the development of computer science and artificial intelligence, using Twitter as a social, political and economic monitor and predictor becomes a new subject for debate in both engineering and social sciences subjects. For example, scholars used Twitter data to monitor natural disaster social dynamics [17], to detect traffic events [18], to predict general election results [19], to make stock market predictions [20] etc. Table 1 presents a summary table of the aforementioned articles, which provides the researchers easy access to these studies.

Such research domains and examples are too numerous to list here; there are also several academic works that provided a panorama for this subject. Williams et al., [9] qualitatively classified more than 1000 Twitter-related academic works, they categorized them into 13 domains, which were Business, Classification, Communication, Education, Emergency, Geography, Health, Libraries, Linguistics, Search, Security, Technical, Other. Zimmer and Proferes [21] analyzed the content of 382 Twitter-related academic publications from 2006 to 2012, they classified 17 different domains and 9 categories of research methods regarding to their analyzed papers. On the other hand, they found that the publications related to emerging innovative research methods such as data-driven analysis were developed more rapidly than other types of publication, at the same time, the demand for tweet content as research raw data is also increasing. Hence, they argued that more studies must be updated with the continued growth of Twitter-based research.

Weller [22] analyzed Twitter-related scientific literature within social science disciplines, with a focus on the most highly cited articles. The common patterns inside these publications have been found, they fit new methods and research designs into classical methodological backgrounds in both qualitative and quantitative approaches. Meanwhile, she argued that studies about Twitter should not solely rely on single datasets and methods, and that the combination of newly emerged methods and classical methods and the connection of Twitter data with other online or offline data sources would positively improve future studies. Researchers have also studied 134 Twitter-related scientific articles indexed in PubMed [23]: they found the early Twitter-focused publications introduced the topic and highlighted its potential, but without any form of data analysis. However, data analytic techniques were mainstream methods in most of the later publications. Despite the fact that the size of the dataset in these papers varies significantly, they argued that the study of Twitter is becoming quantitative research.

Table 1. Summary table of the reviewed scientific literature.

Title	Author	Year	Domain and Research Focus	Reference Pointer
A bibliometric analysis on Twitter Research	Kang, B.; Lee, J. Y.	2014	Bibliometric study. Argued that political issues are one of the core subjects in Twitter research.	[10]
Spanish Indignados and the evolution of the 15 M movement on Twitter: towards networked para-institutions	Pena-Lopez, I.; Congosto, M.; Aragon, P.	2014	Social dynamics. Using Twitter as a communication tool in regional social movements.	[11]
A social networks approach to online social movement: social mediators and mediated content in #FreeAJStaff Twitter network	Isa, D.; Himelboim, I.	2018	Social dynamics. Twitter as a mediator in news freedom online movements.	[12]
Movember: Twitter conversations of a hairy social movement	Jacobson, J.; Mascaro, C.	2016	Social dynamics. Twitter as a platform to engage individuals in social campaigns and sociotechnical social movements.	[13]
Communication dynamics in Twitter during political campaigns	Aragon, P.; Kappler, K. E. et al.	2013	Politics. Political elites use Twitter as a campaign platform in general elections	[14]
E-campaigning on Twitter: The effectiveness of distributive promises and negative campaign in the 2013 Italian election.	Ceron, A.; d'Adda, G.	2016	Politics. Using Twitter content to evaluate the impact of different electoral strategies in political elections	[15]
The 13th General Elections: Changes in Malaysian Political Culture And Barisan Nasional's Crisis of Moral Legitimacy	Jaharudin, M.H.	2014	Politics. The role and importance that Twitter and other social media played in political elections.	[16]
Using Twitter data to monitor natural disaster social dynamics: A recurrent neural network approach with word embeddings and kernel density estimation.	Hernandez-Suarez, A.; Sanchez-Perez, G; et al.	2019	Geographical information system and disaster management. Using Twitter data to monitor natural disasters and to evaluate the post-effect of such catastrophe	[17]
Twitter mining for traffic events detection.	Gutierrez, C.; Figuerias, P et al.	2015	Traffic and management. Twitter as a monitor to detect traffic events	[18]
Prediction of the 2017 French Election Based on Twitter Data Analysis	Wang, L.; Gan, J.Q.	2017	Politics. Using Twitter content to predict political event	[19]
Twitter mood predicts the stock market.	Bollen, J.; Mao, H.; Zeng, X.	2001	Economics. Using Twitter content to predict stock market	[20]

2.2. Methodological Background

For fully completing our research aim, an in-depth bibliometric analysis is going to be applied. Bibliometric analysis is a useful method for measuring the scientific impact, influence and relationships of the published academic works in a certain research framework [24]. Due to the huge amount of scientific literature, manually organizing results within a specific subject under a giant database becomes unfeasible; hence, scientific measurement technique was considered a viable approach for obtaining a detailed overview of a large bibliographic information [25,26].

In bibliometric studies, two main procedures are contained: *performance analysis* and *science mapping* [27,28]. Performance analysis enables the evaluation of scientific publication and citation structures on the basis of bibliographic data such as author(s), author affiliation(s) (university, department), academic journal, conference and country, etc., as well as the impact of their activities on the basis of those data [29,30]. Science mapping displays structural and dynamic aspects of scientific research, which can be generated by the visualization function of digital bibliometric tools [27,31]. Corresponding to our objectives, performance analysis serves for describing the current environment of Twitter studies (e.g., annual scientific production, most productive authors etc.) Science mapping will allow us to illustrate the collaboration structure between countries, the main themes of Twitter-related studies and their evolution over time.

There are different ways to analyze and visualize the research topics of an academic subject; one of them is thematic map. It was first proposed by Callon, Courtial and Laville [32], and is a coordinate system consisting of centrality (x-axis) and density (y-axis). According to them [32] “centrality measures for a given cluster the intensity of its links with other clusters, the more numerous and stronger are these links, the more this cluster designates a set of research problems considered crucial by the scientific or technological community” (p. 164), while “density characterizes the strength of the links that tie the words making up the cluster together. The stronger these links are, the more the research problems corresponding to the cluster constitute a coherent and integrated whole” (p. 165). Thus, a research subject could be classified in 4 quadrants by these two values, each representing a specific theme module, and it would be displayed by a relevant (author) keyword of the bibliographic data, analyzing where the keyword (research theme) lies on is the essential method to interpret the thematic map, thus, the research topics.

Figure 1 shows a thematic map strategic diagram [32]. In the last ten years, researchers have also interpreted this diagram in a more easily understandable way. Cobo et al. [33] take the first quadrant (central and developed) as the space of motor themes, the second quadrant (Central and undeveloped) as the space of basic and transversal themes, the third quadrant (Peripheral and developed) as the space of highly developed and isolated themes, and the fourth quadrant (Peripheral and undeveloped) as the space of emerging or declining themes.

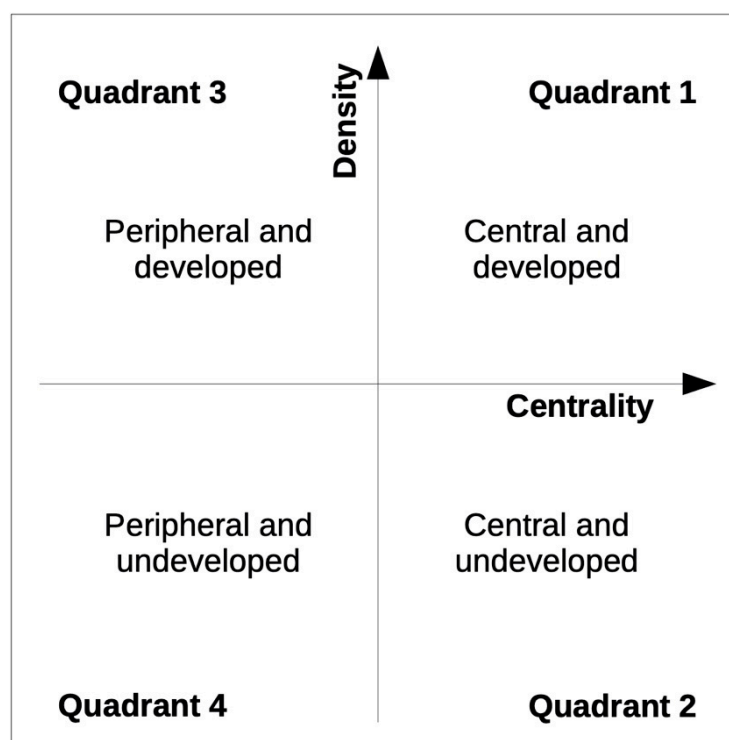


Figure 1. Thematic map strategic diagram with 4 quadrants.

3. Methods

3.1. Data Collection and Preparation

We retrieved our original data from Web of Science (Core Collection) with the keyword (topic) ‘Twitter’, during the period from January 2006 to April 2020. Searched documents (articles, conference proceedings, books, book chapters) are saved with full records and cited references.

The data preparation phase contained two parts. First, a keyword data depuration step was performed. For this purpose, we built a de-pluralization corpus with the help of SciMAT word manager function [34], such function provides an automatic procedure to generate de-pluralization list of the existing keywords (e.g., tweets - tweet), as a result, a total number of 1864 terms were set for this phase. Second, since “Twitter” was the term used for the selection of data, apparently it is the most common keyword in our data, and appears in every document, it might be too impactful to best present our results. Inspired by Leopold, May and Paaß [35], we eliminated it from the set of keywords to improve the quality of our results.

3.2. Bibliometric Analysis Strategies

In the *performance analysis* phase, by using R package “Bibliometrix” [26], basic analysis results about Twitter-related research were calculated and reported in 5 categories: Annual Scientific Production, Most Relevant Sources, Most Productive Authors, Most Cited Publications and Most Relevant Keywords.

In the *science mapping* phase, a country collaboration network based on association strength normalization [36] will be plotted. This network is made by using bibliometric analysis tool Vosviewer [37] with its own clustering algorithm [38]. For studying the research topics and their temporal evolution, we will split our bibliographic dataset according to the Annual Scientific Production, three main research periods will be sliced: initial research period, developing research period, and advanced research period. Bibliometrix provides the possibility to plot thematic map for each of the period based on co-word networks and clustering [26,32].

4. Results and Discussion

4.1. Performance Analysis

A total number of 19,205 academic publications were collected according to our searching strategy. There were 7033 different sources (journals, books etc.) for the publication of all the retrieved bibliographic data, including 37,455 authors. The number of average citations per article was 9.06, and the number of authors per article was 1.95. A total number of 73,178 Author Keywords (AK, keywords provided by the original authors) and 39,747 Keywords Plus (KP, keywords extracted from the titles of the cited references by Thomson Reuters) have been collected, among them, there were 27,179 unique AK, and 7066 unique KP. After applying the de-pluralization corpus, the number of AK has reduced to 25,686, and the number of KP was 6565.

Wang and Chai has introduced the concept of indicator K to quantitatively describe the discipline’s development stages [39], it is measured by the ratio between the unique AK number and the overall AK number. The indicator K of Twitter-related scientific literature is 0.35, which means Twitter research is currently on its normal science stage. This stage means a long-period development of the subject, with further establishment of mature concepts; this stage is expected to step into the post-normal stage with less scientific innovation and vitality [39].

4.1.1. Annual Scientific Production

The annual scientific production (Figure 2) consists of four parts, productions by year, relative growth rate (RGR), doubling time (DT) and average citation rate (ACR). As we retrieved our bibliographic data in April 2020, the total number of scientific publications of 2020 is not complete, hence, we did not include the data of 2020 in this analysis. RGR represents the increase in the

cumulative number of publications per unit of time (year), while DT refers to the required time for publications to become double the existing amount [40,41], and the ACR represents the normalized number of citations per document. It should be mentioned that in this section, only bibliographic data with year information can be calculated, in our retrieved dataset, there are 297 documents have no such information, so the total number of calculated documents in this section is 18,474 (with publications of the year 2020 excluded).

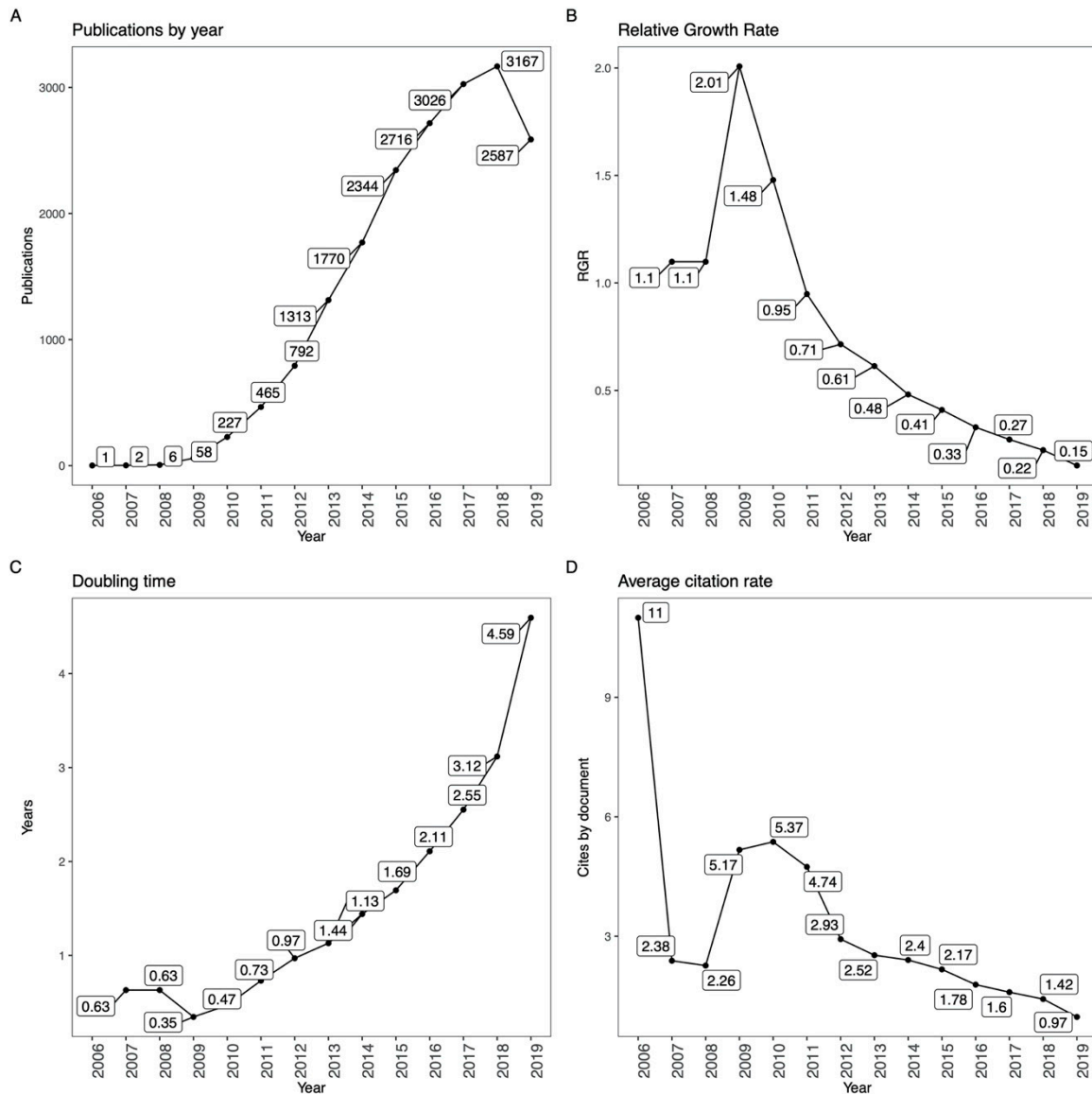


Figure 2. Annual Scientific Production, $RGR = (\ln c_2 - \ln c_1) / (t_2 - t_1)$, \ln = natural logarithm, c_1 = cumulative number of publications in period one, c_2 = cumulative number of publications in period two. $DT = ((t_2 - t_1) * \ln 2) / (\ln c_2 - \ln c_1)$. $ACR = citations / documents / years_since_publication$. (a) Publications by year, (b) relative growth rate, (c) doubling time, (d) average citation rate.

In general, the production of academic research kept increasing year by year, however, the number of Twitter-related publication of 2019 is less than 2018. The RGR and DT demonstrated that although the quantity of related research keeps growing, their growth rate and speed have been largely turned down in recent years. As for ACR, due to the very limited number of publications in the first three years, the ACR index in those years is considered meaningless, in general, the ACR presents a negative

growth trending, it is understandable, because older articles tend to be more cited than new published articles [42].

4.1.2. Most Relevant Sources

PLOS ONE is the most popular journal in publishing academic works for studies on Twitter. A total number of 251 articles were published on this scientific journal. In addition to PLOS ONE, there are 7 journals (*Computers in Human Behavior*, *Journal of Medical Internet Research*, *Information, Communication & Society*, *New Media & Society*, *Social Network Analysis and Mining*, *International Journal of Communication and Social Media + Society*) that have published more than 100 articles with the theme ‘Twitter’. Table 2 shows our results in detail; the column ‘Subject’ refers to the journals’ domain according to the classification information of Web of Science.

Table 2. Most Relevant Sources.

Rank	Sources	Subject	Articles
1	PLOS One	Multidisciplinary Sciences	251
2	International Conference on Advances in Social Networks Analysis and Mining ¹	Computer Science, Computer Networks and Communications, Information Systems	239
3	Computers in Human Behavior	Psychology, Experimental; Psychology, Multidisciplinary	176
4	IEEE International Conference on Big Data ²	Computer Science, Software	145
5	Journal of Medical Internet Research	Health Care Sciences & Services; Medical Informatics	142
6	Information Communication & Society	Communication; Sociology	141
7	New Media & Society	Communication	118
8	Social Network Analysis and Mining	Computer Science; Information Systems	118
9	International Journal of Communication	Communication	108
10	Social Media + Society	Computer Science Applications, Communication, Cultural Studies	107

¹ Different editions (years) have been grouped together; ² Different editions (years) have been grouped together.

Corresponding to the most relevant sources of academic publication, most of them belong to the subjects of communication and computer science. The rest of the subjects are mostly related to social sciences and informational science. Only a few journals dedicated to psychology and medical information. Figure 3 presents a year-by year evolution line chart of the fore-mentioned subjects: x-axis represents the year and the y-axis represents the number of publications under a certain subject. This line chart has proved our previous argument, that communication and computer science are the two main subjects in Twitter-related researches—both of the two disciplines have been largely developed since 2012. Twitter studies published in social science and information science journals are slightly more numerous than those in psychology and medical journals. All the four minor disciplines kept a relatively low increase rate.

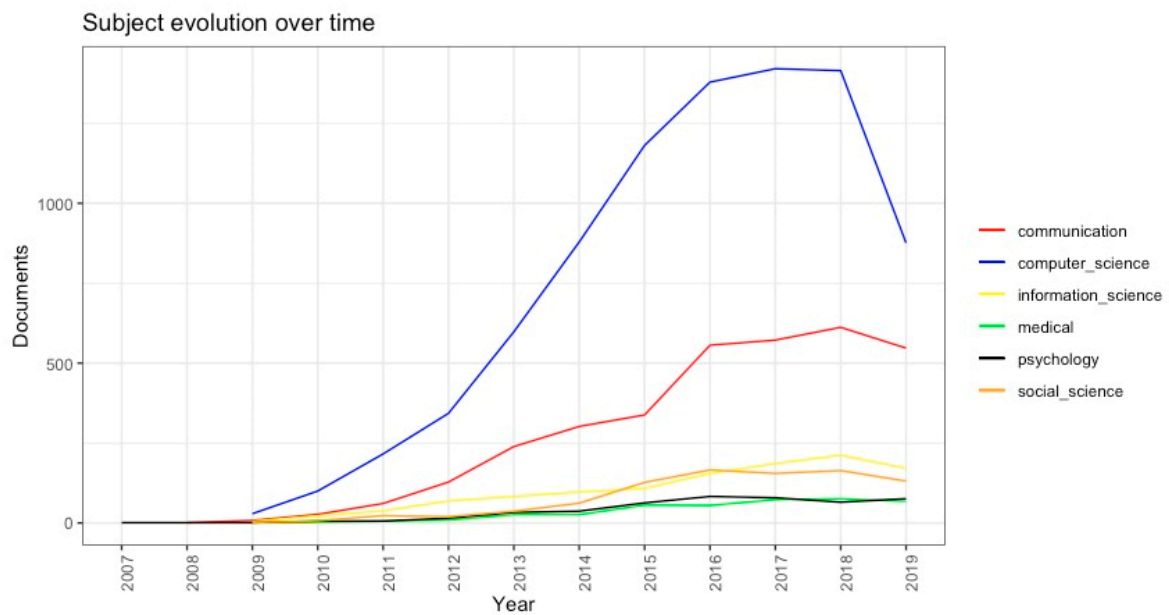


Figure 3. Subject evolution over time.

4.1.3. Author Statistics and Most Cited Publications

Table 3 shows the most productive authors and most cited publications (ranked by total citation) in Twitter-related studies. Different from previous results of most relevant sources, we find three highly cited papers were published in the journal *Business Horizon*: this proves the study of Twitter may have a high interdisciplinary impact. However, as row citation counts are not useful for comparison purpose because older articles tend to be more cited [42], here we are not going to further discuss about this ranking, the table of most cited publications is only intended to help researchers master the information in its entirety.

However, the table of top 10 most cited publications would be slightly changed if we rank the publications by their annual citation rate, another 4 papers would appear on this table, they are “Vosoughi S, 2018, *Science*” (218), “Isola P, 2017, *Proc CVPR IEEE*” (138), “Stephens ZD, 2015, *Plos Biol*” (77), “Huang JD, 2019, *Tob Control*” (76). The numbers inside the parenthesis are their average citation number per year.

Table 3. Most productive authors and most cited publications.

Rank	Most Productive Authors			Most Cited Publications			
	Name	N. Articles	Corresponding Author	Year	Journal	Total Citation	Citation per Year
1	Wang Y	55	Kaplan AM	2010	Bus Horizons	4169	417
2	Kim J	45	Boyd D	2012	Inform Commun Soc	1624	203
3	Kim Y	44	Bollen J	2011	J Comput Sci-Neth	1414	157
4	Zhang Y	44	Kietzmann JH	2011	Bus Horizons	1248	139
5	Liu H	43	Marwick AE	2011	New Media Soc	1126	125
6	Liu Y	42	Jansen BJ	2009	J Am Soc Inf Sci Tec	828	75
7	Wang D	36	Casler K	2013	Comput Hum Behav	577	82
8	Park HW	35	O’Keefe GS	2011	Pediatrics	549	61
9	Lee J	34	Chew C	2010	Plos One	504	50
10	Bruns A	33	Hanna R	2011	Bus Horizons	492	55

Figure 4 presents a line chart of the average number of authors per year per document; for example, in 2019, there were 3.29 authors per publication in Twitter-related researches. Given the very limited number of publications in the year 2006(1), 2007(2) and 2008(6), the mean number of authors in these years is considered meaningless. From the year 2009, the average number of authors per document kept increasing, this implies that scholars are becoming more and more cooperative with each other in Twitter-related studies.

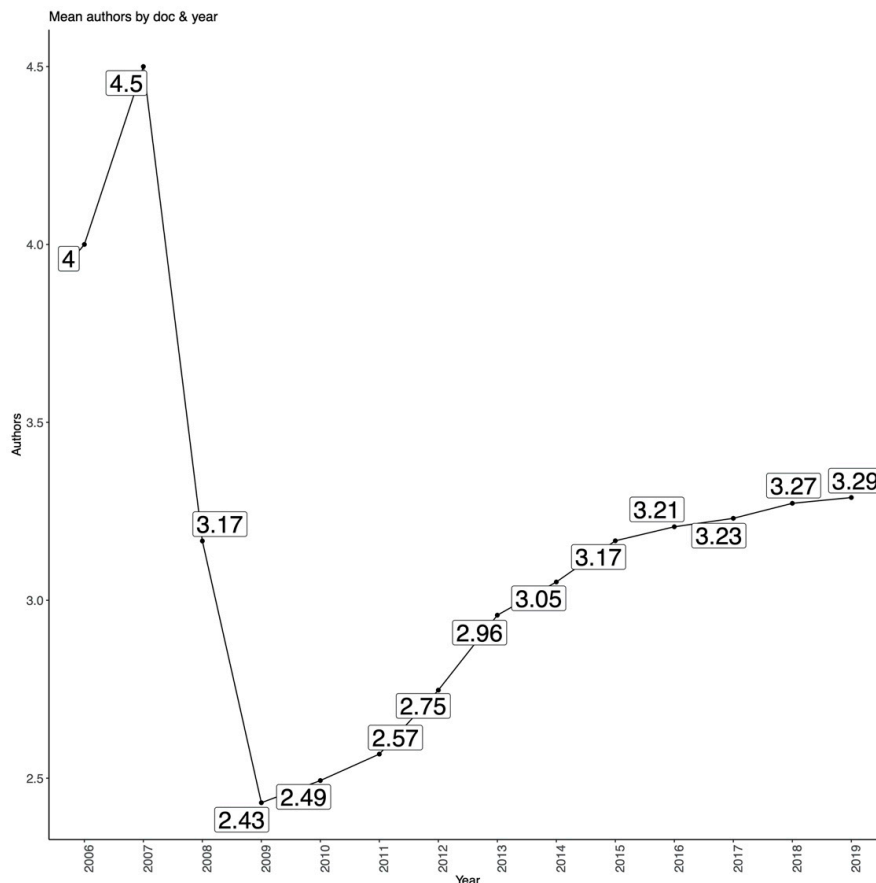


Figure 4. Average number of authors per document.

4.1.4. Most Relevant Keywords

Table 4 shows the most relevant author keywords and keyword plus. Both of the two kinds of keywords are mostly related to computer science and communication. On the whole, Author Keywords and Keywords Plus revealed similar research trends; both of the two types of keywords described equally the focus of Twitter-related studies. However, small differences can still be observed.

As presented, Author Keywords emphasized research methods and techniques, for example, there are terms like “sentiment analysis”, “machine learning”, “social network analysis”, “text mining”, whereas Keyword Plus tended to focus on specific research objects, like “media”, “news” etc. As Keywords Plus are words or phrases that frequently appear in the titles of the articles’ references [43], here we agree with the argument of Zhang et al, that Keywords Plus is less comprehensive in representing an article’s content [44].

Table 4. Most relevant keywords.

Rank	Author Keywords	Documents	Keyword Plus	Documents
1	Social media	4699	Social media	1408
2	Sentiment analysis	1148	Media	776
3	Social networks	1015	Communication	680
4	Facebook	753	Facebook	672
5	Machine learning	508	Internet	613
6	Big data	482	Impact	540
7	Social network	428	Online	534
8	Social network analysis	390	News	444
9	Internet	353	Networks	412
10	Text mining	327	Model	405

4.2. Science Mapping

4.2.1. Country Collaboration Network

Vosviewer presents the country collaboration network based on co-occurrence frequencies. By default, the association strength is employed to normalize the network [45], this method has also been proved as one of the best [36]. The clustering algorithm is based on a weighted and parameterized variant of the well-known modularity function of Newman and Girvan [46].

Figure 5 shows the top 40 country collaboration network of our retrieved bibliographic data, it is able to reflect the degree of communication between countries as well as the influential countries in this field [47]. Three major communities (with different node colors) can be found from the network. The size of the nodes represents the impact of the country on Twitter-related studies (based on the number of publications). The edges between nodes represent strength of the cooperative relationships between countries.

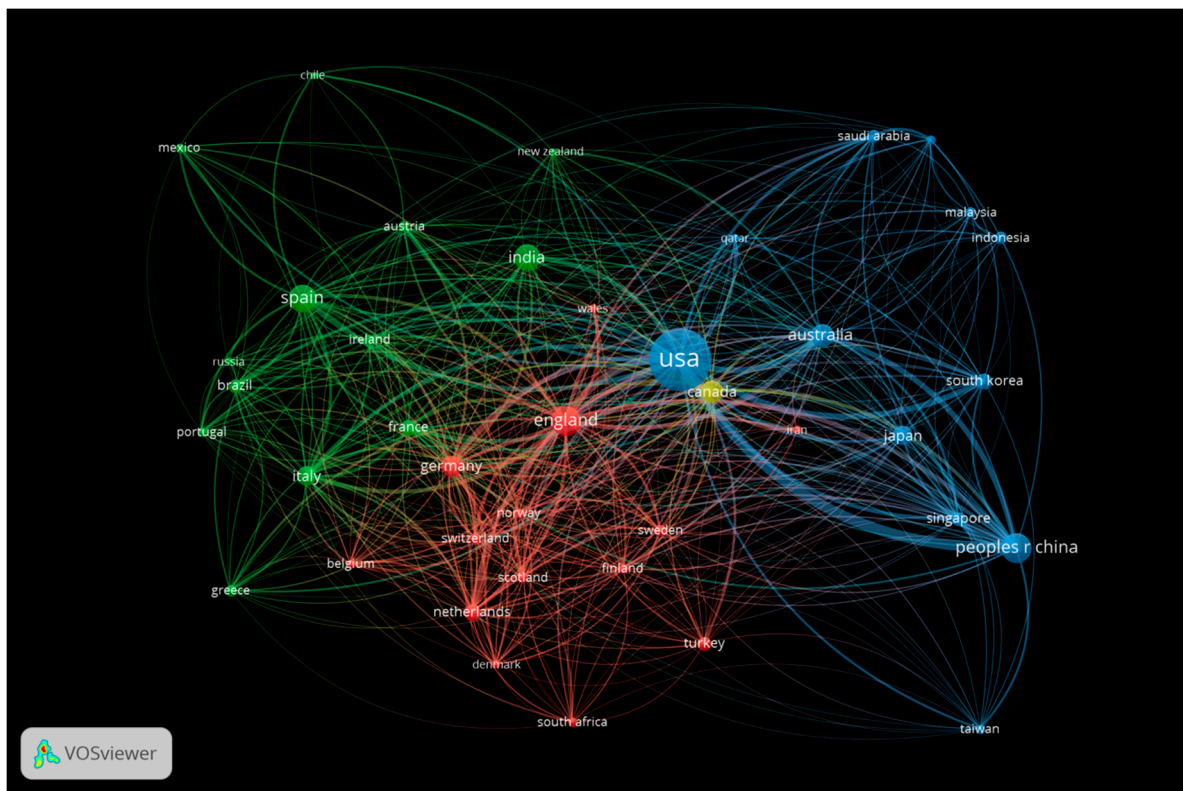


Figure 5. Country collaboration network.

It can be easily observed that European countries has a highly internal collaboration ties, while for Asian-Pacific countries, North American countries are their most frequent collaboration partners. However, for USA and Canada, they have strong ties with both European and Asian-Pacific countries. There are also close relations between Iberian countries and Latin American countries, naturally, we believe the common language usage among these countries are the main reason of their close ties.

Table 5 gives the detailed information about the top 10 most productive countries of Twitter-related studies, SCP is the abbreviation of Single Country Publications, and MCP is Multiple Country Publications, MCP Ratio is MCP as a proportion of total publication number. European countries like the UK, Spain, Germany and Italy share a relatively high degree of international collaboration. Despite the fact that China has the highest index, other Asian countries (India and Japan) hold the lowest ratio. From another perspective, English-speaking countries (USA, UK, Australia, Canada) hold a relatively high degree of international collaboration than other countries.

Table 5. Top 10 most productive countries.

Country	Publications	SCP	MCP	MCP Ratio
USA	5340	4626	714	13.37%
United Kingdom	1300	997	303	23.31%
China	1251	820	431	34.45%
Spain	1098	934	164	14.94%
India	1086	1001	85	7.83%
Australia	707	523	184	26.03%
Canada	620	448	172	27.74%
Japan	610	547	63	10.33%
Germany	518	372	146	28.19%
Italy	510	381	129	25.29%

4.2.2. Thematic Analysis

For the analysis of topic evolution across time, a set of time slices is made. According to the Annual Scientific Production, we take three periods to segment the whole Twitter-related scientific development process into three phases: Initial period is from 2006 to 2012: in this period, the publication number is not so much as later years, but RGR is relatively high, DT kept steadily with mild changes. The developing period is from 2013 to 2016; in this period the number of publications increased rapidly, RGR slowed down while DT started to slightly grow. The advanced period is from 2017 to 2020; in this period the number of publications arrived peak, while RGR kept turning down, DT grew immensely.

Figure 6 presents the thematic maps of the three periods, each of the circles represents a cluster and the size of the circle represents the size of the cluster (the number of included terms/keywords). There are fewer clusters in developing and advanced period than the initial period, which implies that there are fewer research topics in last years than the first years.

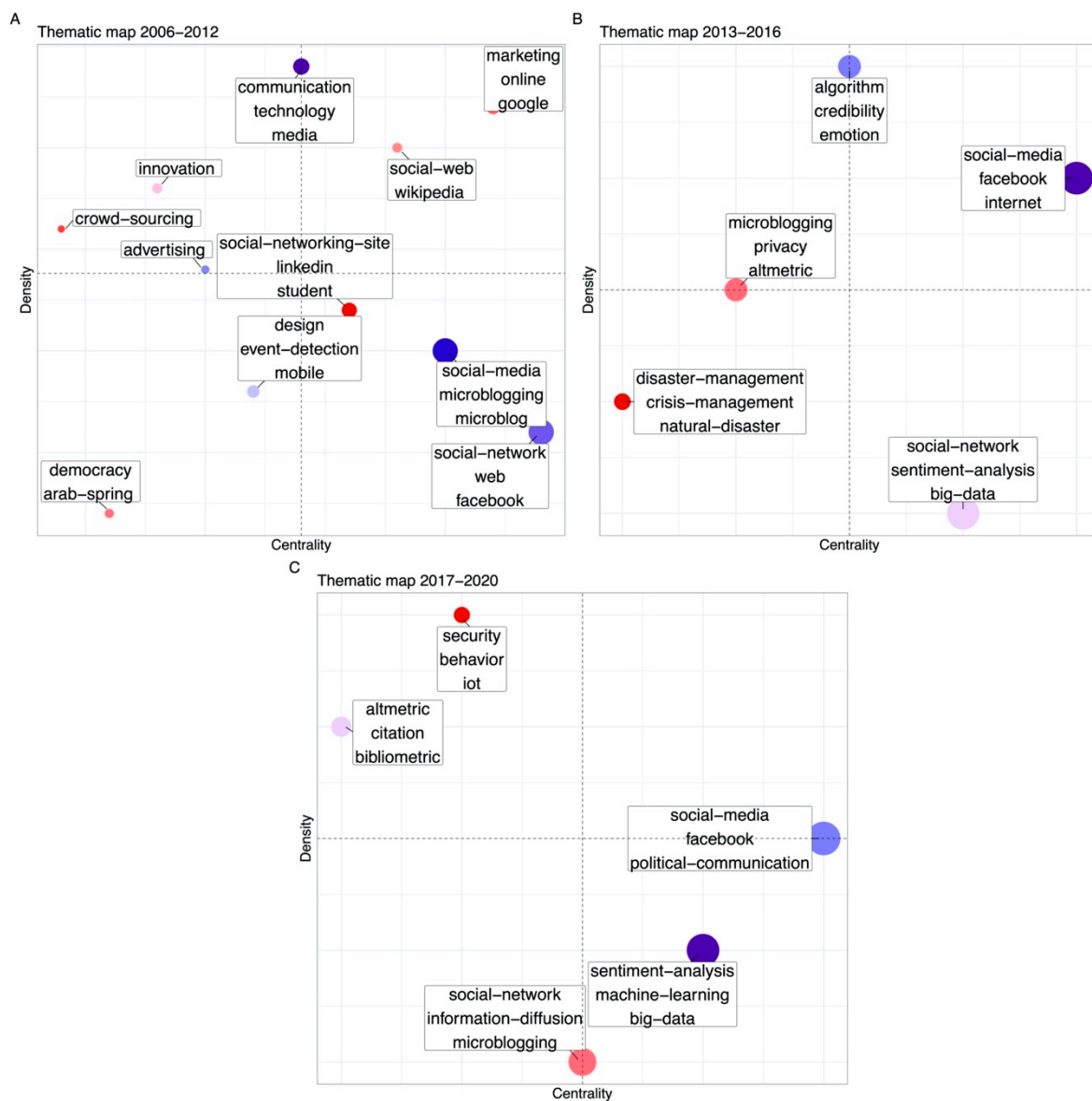


Figure 6. Thematic maps of the three periods. (A) Initial period, (B) developing period and (C) advanced period.

For the initial period (2006–2012), there are two clusters on the first quadrant with high centrality and density, “marketing, online, google” and “social-web, wikipedia”, these clusters focused on Twitter and other well-known website and marketing, are the motor research themes of this period. The third quadrant mainly consists of three clusters, “innovation”, “crowd-sourcing” and “advertising”, all these three clusters can be considered as specific research topics for business subject, they are the highly developed and isolated themes of 2006–2012. While Twitter was a newly emerged social media in that time, business related topics revealed a high centrality in the initial period, they have been hugely developed in the first years since the foundation of Twitter.

“Democracy, arab-spring” and “design, event-detection, mobile” are the emerging or declining themes, they are independent from each other, “democracy, arab-spring” corresponds to 2010 arab-spring revolution, “design, event-detection, mobile” might related to the studies about smartphone and mobile application, such new electronic device and software also appeared after 2010, there are publications such as “Tweeting with the telly on! Mobile phones as second screen for TV”, “Mobile apps: innovative technology for globalization and inclusion of developing countries” can prove our assumption. It is more reasonable to classify these two clusters as emerging themes, compared to the foundation of Twitter (2006), from 2006 to 2012, such political events and technological innovation occurred in 2010 was even newer.

“Social-networking-site, linkedin, student”, “social-media, microblogging, microblog”, “social-network, web, facebook” are the three clusters that belong to basic and transversal themes; they are mainly focused on other virtual social networks, comparative studies about Twitter and other similar platforms are another important research line in the initial period. However, based on the previous argument, the “social-networking-site, linkedin, student” cluster may also refer to the studies of human resources, online employment and education, there are publications like “Using facebook, linkedin and Twitter for your career”, “Friend or foe? The promise and pitfalls of using social networking sites for HR decisions”, “Comparative survey of students’ behavior on social networks (in Czech perspective)” can prove our assumption.

For the developing period (2013–2016), in general, topics related to business, mobile and arab-spring disappeared from the map, contrarily, computer science related nouns emerged in this period (e.g., algorithm, sentiment-analysis). Cross-platform comparative studies (“social-media, facebook, internet” cluster) moved from basic and transversal themes to motor themes. “Algorithm, credibility, emotion” cluster locates between the first and second quadrant with a very high density, this cluster refers to using computational methods to detect online emotion, and is highly developed within this period. “Microblogging, privacy, altmetric” cluster locates between the third and fourth quadrant, as big data is gaining attention and popularity among researchers in this period, the usage of big data starts to be important, which have also caused people’s awareness about privacy. This cluster may contain two research lines, using Twitter metrics as a tool to measure research impact [48,49], and the privacy caution of using microblog service [50].

Disaster-management, crisis-management, natural-disaster” cluster is the emerging and declining theme of the developing period, apparently, this cluster refers to studies about crisis management and crisis communication during severe disasters, for example, earthquakes [51], tsunami [52], and epidemic crisis [53] etc. The last cluster of this period is “social-network, sentiment-analysis, big-data”—this cluster belongs to basic and transversal theme, data-driven sentiment analysis becomes a popular research method for social media studies in this period.

For the advanced period (2017–2020), there is no absolute motor theme, “social-media, facebook, political-communication” locates between the first and the second quadrant with a high centrality, this cluster refers to the study of political communication with social media. Two clusters are on the second quadrant, “security, behavior, iot (internet of things)” and “altmetric, citation, bibliometric”; they are highly developed and isolated research themes, and independent from each other. Alongside the rapid development of social network sites, the integration of social media and internet of things has formed a new concept, social internet of things (siot) [54], meanwhile, social network-based recommendation

system emerges as a new research topic, for example, researchers used Twitter data to personalize movie recommendation system [55], but such advanced technologies also contain considerable security risk. We believe the cluster “security, behavior, iot” refers to use Twitter as an iot medium to study user’s online behavior and the potential cybersecurity concerns of iot. The cluster “altmetric, citation, bibliometric” is easier to interpret—it refers to Twitter-based scientometric studies, compared to the “altmetric” cluster in developing period, the study of scientometrics during 2017 to 2020 becomes an independent and developed research theme.

“Sentiment-analysis, machine-learning, big-data” was the only basic and transversal research theme, this implies computational methods and techniques are widely used in Twitter research from 2017 to 2020. The cluster “social-network, information-diffusion, microblogging” locates between the third and the fourth quadrant, with a low density, this means that although the study of information diffusion on Twitter and microblogs emerged in recent years, yet not fully developed.

Figure 7 presents the alluvial diagram of research thematic evolution across the three previously segmented periods; it provides us a global view of the changes. Each of the nodes represents a cluster, and is labeled by the first three words of the clusters, the edges are their temporal evolution track, generated by keyword co-occurrence of the topics between two time slices [33].

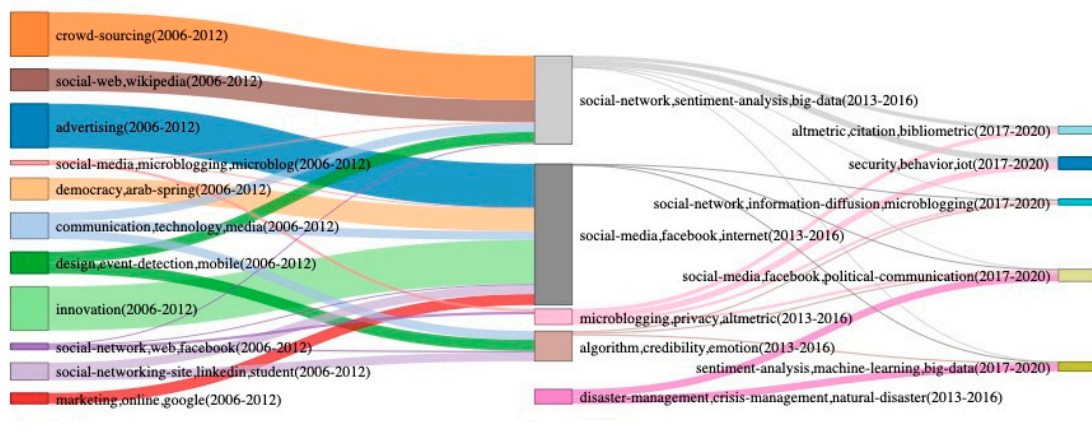


Figure 7. Alluvial diagram of thematic evolution.

Overall, research topics in the initial period were more than in later periods; business-related research lines took an important place in that time. There are two major research topics in the developing period, “social-network” (social-network, sentiment-analysis, big-data) and “social-media” (social-media, facebook, internet). As we have discussed, they imply different research lines, the former represents Twitter study with computational methods, the latter represents cross-platform comparative studies. Most of the research themes of the initial period were lumped together under these two large topics. Furthermore, “disaster-management” (“disaster-management, crisis-management, natural-disaster”) emerged in the developing period, and it evolved to be an important component for the clusters with information diffusion (“social-network, information-diffusion, microblogging”) and big data (“sentiment-analysis, machine-learning, big-data”) in the advanced period. Scientometric study (“altmetric, citation, bibliometric”) was an important research topic in recent years—naturally, it is strongly associated with clusters containing altmetric (microblogging, privacy, altmetric) and big data (social-network, sentiment-analysis, big-data). Such clusters were also evolution sources for the cluster “security, behavior, iot”.

5. Conclusions

A general approach to analyze and visualize the basic status of Twitter-related studies has been presented in this paper. Compared to previous studies [9,56], our research has largely expanded the number of bibliographic data. With the general description of our bibliographic data, we have

successfully illustrated the current twitter study environment. In a nutshell, Twitter is still a research hotspot for both social science and computer science scholars. 2019 was the first year with negative growth, this might be a signal that Twitter-related studies have surpassed the advanced period, but this assumption should be further confirmed by future research. Other descriptive results, for example, the most relevant sources and most relevant keywords have also revealed some of the main research interests regarding Twitter-related scientific literature.

In the science mapping section, we first presented a country collaboration network, in which a set of country collaboration patterns have been identified, Asian-Pacific countries are closely linked to North American countries, while European countries refer to collaborate within themselves, the 40 most important countries in Twitter research are presented as nodes on the network. The detailed information of the top 10 most productive countries has been further presented. Among them, European countries and English speaking countries have a relatively high international collaboration degree.

For the thematic analysis, we have successfully identified the most important research topics, they are mainly related to business (including marketing, advertising etc.), communication (including political communication, new media studies etc.), disaster management, scientometrics and computer science (including sentiment analysis, machine learning etc.). Although the research lines seem to become more homogenous over time, new research topics in Twitter-related studies emerged in recent years: while studies in the subject of business took an important place in the first years, individual research focuses like marketing, advertising and crowd-sourcing disappeared from the thematic map in later periods, they have been involved into larger interdisciplinary clusters.

Twitter research is highly associated with a real world timeline; the 2010 Arab spring revolution has been shown to be an emerging topic in the thematic map. While in the developing period (2013–2016), disaster management and crisis communication appeared to be an important research focus, as discussed, they have a strong tie with the natural disaster and epidemic crisis in those years. At last, computational methods (e.g., machine learning, sentiment analysis, etc.) were developed rapidly in later years; the above-mentioned research topics have shown a strong association with these new techniques. As Williams et al. [23] once indicated, Twitter-related studies are becoming quantitative research and we agree with their argument; however, quantitative research is a broad concept—it involves both traditional and new methods, and we would like to say Twitter-related studies are becoming computational research.

Author Contributions: Conceptualization, J.Y. and J.M.-J.; methodology, J.Y. and J.M.-J.; validation, J.Y. and J.M.-J.; formal analysis, J.Y.; investigation, J.Y.; data curation, J.M.-J.; writing—original draft preparation, J.Y.; writing—review and editing, J.Y. and J.M.-J.; visualization, J.Y.; supervision, J.M.-J. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by the Department of Social Psychology, Universitat Autònoma de Barcelona.

Acknowledgments: This work belongs to the framework of the doctoral programme in Person and Society in the Contemporary World of the Autonomous University of Barcelona.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Twitter. *Twitter Annual Report 2018*; Twitter: San Francisco, CA, USA, 2018.
2. Fiegerman, S. Twitter Now Losing Users in the U.S. Available online: <https://money.cnn.com/2017/07/27/technology/business/twitter-earnings/index.html?iid> (accessed on 27 July 2018).
3. Haque, U. The Reason Twitter's Losing Active Users. Available online: <https://hbr.org/2016/02/the-reason-twiters-losing-active-users> (accessed on 27 July 2018).
4. Statista Twitter: Number of Active Users 2010–2018|Statista. Available online: <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/> (accessed on 27 July 2018).
5. Ahmed, W.; Bath, P.A.; Demartini, G. *Chapter 4: Using Twitter as a Data Source: An Overview of Ethical, Legal, and Methodological Challenges*; Emerald Publishing Limited: Bingley, UK, 2017; pp. 79–107.

6. Kwak, H.; Lee, C.; Park, H.; Moon, S. What is Twitter, a social network or a news media? In Proceedings of the 19th International Conference on World Wide Web—WWW '10, Raleigh, CA, USA, 26–30 April 2010; ACM Press: New York, NY, USA, 2010; p. 591.
7. Gupta, B.M.; Kumar, A.; Gupta, R.; Dhawan, S.M. A bibliometric assessment of Global Literature on “Twitter” during 2008–15. *Int. J. Inf. Dissem. Technol.* **2016**, *6*, 199–206.
8. Yu, J.; Muñoz-Justicia, J. Free and Low-Cost Twitter Research Software Tools for Social Science. *Soc. Sci. Comput. Rev.* **2020**. [[CrossRef](#)]
9. Williams, S.A.; Terras, M.; Warwick, C. What do people study when they study Twitter? Classifying Twitter related academic papers. *J. Doc.* **2013**, *69*, 384–410. [[CrossRef](#)]
10. Kang, B.; Lee, J.Y. A Bibliometric Analysis on Twitter Research. *J. Korean Soc. Inf. Manag.* **2014**, *31*, 293–311. [[CrossRef](#)]
11. Peña-López, I.; Congosto, M.; Aragón, P. Spanish Indignados and the evolution of the 15M movement on Twitter: Towards networked para-institutions. *J. Span. Cult. Stud.* **2014**, *15*, 189–216. [[CrossRef](#)]
12. Isa, D.; Himelboim, I. A Social Networks Approach to Online Social Movement: Social Mediators and Mediated Content in #FreeAJStaff Twitter Network. *Soc. Media Soc.* **2018**, *4*, 4. [[CrossRef](#)]
13. Jacobson, J.; Mascaro, C. Movember: Twitter Conversations of a Hairy Social Movement. *Soc. Media Soc.* **2016**, *2*. [[CrossRef](#)]
14. Aragón, P.; Kappler, K.E.; Kaltenbrunner, A.; Laniado, D.; Volkovich, Y. Communication dynamics in twitter during political campaigns: The case of the 2011 Spanish national election. *Policy Internet* **2013**, *5*, 183–206. [[CrossRef](#)]
15. Ceron, A.; D’Adda, G. E-campaigning on Twitter: The effectiveness of distributive promises and negative campaign in the 2013 Italian election. *New Media Soc.* **2016**, *18*, 1935–1955. [[CrossRef](#)]
16. Jaharudin, M.H. The 13th General Elections: Changes in Malaysian Political Culture and Barsian Nasional’s Crisis of Moral Legitimacy. *Kaji Malays.* **2014**, *32*, 149–169.
17. Hernandez-Suarez, A.; Sanchez-Perez, G.; Toscano-Medina, L.K.; Perez-Meana, H.M.; Portillo-Portillo, J.; Villalba, L.J.G.; Villalba, L.J.G. Using Twitter Data to Monitor Natural Disaster Social Dynamics: A Recurrent Neural Network Approach with Word Embeddings and Kernel Density Estimation. *Sensors* **2019**, *19*, 1746. [[CrossRef](#)] [[PubMed](#)]
18. Gutierrez, C.; Figuerias, P.; Oliveira, P.; Costa, R.; Jardim-Goncalves, R. Twitter mining for traffic events detection. In Proceedings of the 2015 Science and Information Conference, SAI, London, UK, 28–30 July 2015; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2015; pp. 371–378.
19. Wang, L.; Gan, J.Q. Prediction of the 2017 French Election Based on Twitter Data Analysis. In Proceedings of the 2017 9th Computer Science and Electronic Engineering (CEEC), Colchester, UK, 27–29 September 2017.
20. Bollen, J.; Mao, H.; Zeng, X.-J. Twitter mood predicts the stock market. *J. Comput. Sci.* **2011**, *2*, 1–8. [[CrossRef](#)]
21. Zimmer, M.; Proferes, N. A topology of Twitter research: Disciplines, methods, and ethics. *Aslib J. Inf. Manag.* **2014**, *66*, 250–261. [[CrossRef](#)]
22. Weller, K. What do we get from Twitter—and What Not? A Close Look at Twitter Research in the Social Sciences. *Knowl. Organ.* **2014**, *41*, 238–248. [[CrossRef](#)]
23. Williams, S.A.; Terras, M.; Warwick, C.; McGowan, B.; Pedrana, A. How Twitter Is Studied in the Medical Professions: A Classification of Twitter Papers Indexed in PubMed. *Med. 2.0* **2013**, *2*, e2. [[CrossRef](#)]
24. Van Raan, A.F.J. The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments. *Tech. Theor. Prax* **2003**, *1*, 20–29. [[CrossRef](#)]
25. Broadus, R.N. Toward a definition of “bibliometrics”. *Scientometrics* **1987**, *12*, 373–379. [[CrossRef](#)]
26. Aria, M.; Cuccurullo, C. bibliometrix: An R-tool for comprehensive science mapping analysis. *J. Inf.* **2017**, *11*, 959–975. [[CrossRef](#)]
27. Noyons, E.C.M.; Moed, H.F.; Luwel, M. Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. *J. Am. Soc. Inf. Sci.* **1999**, *50*, 115–131. [[CrossRef](#)]
28. van Raan, A.F.J. Measuring Science. In *Handbook of Quantitative Science and Technology Research*; Springer: Dordrecht, The Netherlands, 2005; pp. 19–50.
29. Van Raan, A.F.J. Measurement of Central Aspects of Scientific Research: Performance, Interdisciplinarity, Structure. *Meas. Interdiscip. Res. Perspect.* **2005**, *3*, 1–19. [[CrossRef](#)]
30. Gutierrez-Salcedo, M.; Martínez, M.Á.; Moral-Munoz, J.A.; Herrera, F.; Cobo, M.J. Some bibliometric procedures for analyzing and evaluating research fields. *Appl. Intell.* **2017**, *48*, 1275–1287. [[CrossRef](#)]

31. Börner, K.; Chen, C.; Boyack, K. Visualizing knowledge domains. *Annu. Rev. Inf. Sci. Technol.* **2005**, *37*, 179–255. [[CrossRef](#)]
32. Callon, M.; Courtial, J.P.; Laville, F. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics* **1991**, *22*, 155–205. [[CrossRef](#)]
33. Cobo, M.J.; López-Herrera, A.G.; Herrera-Viedma, E.; Herrera, F. An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. *J. Inf.* **2011**, *5*, 146–166. [[CrossRef](#)]
34. Cobo, M.J.; Herrera-Viedma, E.; Herrera, F.; López-Herrera, A. SciMAT: A new science mapping analysis software tool. *J. Am. Soc. Inf. Sci. Technol.* **2012**, *63*, 1609–1630. [[CrossRef](#)]
35. Leopold, E.; May, M.; Paaß, G. Data Mining and Text Mining for Science & Technology Research. In *Handbook of Quantitative Science and Technology Research*; Springer: Dordrecht, The Netherlands, 2004; pp. 187–213.
36. Van Eck, N.J.; Waltman, L. How to normalize cooccurrence data? An analysis of some well-known similarity measures. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 1635–1651. [[CrossRef](#)]
37. Van Eck, N.J.; Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* **2009**, *84*, 523–538. [[CrossRef](#)]
38. Waltman, L.; Van Eck, N.J.; Noyons, E. A unified approach to mapping and clustering of bibliometric networks. *J. Inf.* **2010**, *4*, 629–635. [[CrossRef](#)]
39. Wang, M.; Chai, L. Three new bibliometric indicators/approaches derived from keyword analysis. *Scientometrics* **2018**, *116*, 721–750. [[CrossRef](#)]
40. Waila, P.; Singh, V.K.; Singh, M.K. A Scientometric Analysis of Research in Recommender Systems. *J. Sci. Res.* **2016**, *5*, 71–84. [[CrossRef](#)]
41. Sweileh, W.; Al-Jabi, S.W.; AbuTaha, A.S.; Zyoud, S.; Anayah, F.M.A.; Sawalha, A.F. Bibliometric analysis of worldwide scientific literature in mobile - health: 2006–2016. *BMC Med. Inform. Decis. Mak.* **2017**, *17*, 72. [[CrossRef](#)] [[PubMed](#)]
42. Thelwall, M. Author gender differences in psychology citation impact 1996–2018. *Int. J. Psychol.* **2019**, 12633. [[CrossRef](#)] [[PubMed](#)]
43. Clarivate Analytics KeyWords Plus Generation, Creation, and Changes. Available online: https://support.clarivate.com/ScientificandAcademicResearch/s/article/KeyWords-Plus-generation-creation-and-changes?language=en_US (accessed on 12 May 2020).
44. Zhang, J.; Yu, Q.; Zheng, F.; Long, C.; Lu, Z.; Duan, Z. Comparing keywords plus of WOS and author keywords: A case study of patient adherence research. *J. Assoc. Inf. Sci. Technol.* **2015**, *67*, 967–972. [[CrossRef](#)]
45. Van Eck, N.J.; Waltman, L. BIBLIOMETRIC MAPPING OF THE COMPUTATIONAL INTELLIGENCE FIELD. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* **2007**, *15*, 625–645. [[CrossRef](#)]
46. Newman, M.E.J.; Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **2004**, *69*, 026113. [[CrossRef](#)] [[PubMed](#)]
47. Liao, H.; Tang, M.; Luo, L.; Li, C.; Chiclana, F.; Zeng, X.-J. A Bibliometric Analysis and Visualization of Medical Big Data Research. *Sustainability* **2018**, *10*, 166. [[CrossRef](#)]
48. Holmberg, K.; Thelwall, M. Disciplinary differences in Twitter scholarly communication. *Scientometrics* **2014**, *101*, 1027–1042. [[CrossRef](#)]
49. Thelwall, M.; Haustein, S.; Larivière, V.; Sugimoto, C.R. Do Altmetrics Work? Twitter and Ten Other Social Web Services. *PLoS ONE* **2013**, *8*, e64841. [[CrossRef](#)]
50. Buccafurri, F.; Lax, G.; Nicolazzo, S.; Nocera, A. Comparing Twitter and Facebook user behavior: Privacy and other aspects. *Comput. Hum. Behav.* **2015**, *52*, 87–95. [[CrossRef](#)]
51. Lu, X.; Brelsford, C. Network Structure and Community Evolution on Twitter: Human Behavior Change in Response to the 2011 Japanese Earthquake and Tsunami. *Sci. Rep.* **2014**, *4*, 6773. [[CrossRef](#)]
52. Chatfield, A.; Scholl, H.J.; Brajawidagda, U. Tsunami early warnings via Twitter in government: Net-savvy citizens' co-production of time-critical public information services. *Gov. Inf. Q.* **2013**, *30*, 377–386. [[CrossRef](#)]
53. Fung, I.C.-H.; Tse, Z.T.H.; Cheung, C.-N.; Miu, A.S.; Fu, K.-W. Ebola and the social media. *Lancet* **2014**, *384*, 2207. [[CrossRef](#)]
54. Atzori, L.; Iera, A.; Morabito, G.; Nitti, M. The Social Internet of Things (SIoT)—When social networks meet the Internet of Things: Concept, architecture and network characterization. *Comput. Netw.* **2012**, *56*, 3594–3608. [[CrossRef](#)]

55. Das, D.; Chidananda, H.T.; Sahoo, L. Personalized movie recommendation system using twitter data. In *Progress in Computing, Analytics and Networking. Advances in Intelligent Systems and Computing*; Pattnaik, P., Rautaray, S., Das, H., Nayak, J., Eds.; Springer: Singapore, 2018; Volume 710, pp. 339–347.
56. Fausto, S.; Aventurier, P. *Scientific Literature on Twitter as a Subject Research: Findings Based on Bibliometric Analysis*; Handbook Twitter For Research 2015–2016; EMLYON Press: Lyon, France, 2016; p. 242.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Chapter 3

Second Publication: Free and Low-Cost Twitter Research Software Tools for Social Science

Free and Low-Cost Twitter Research Software Tools for Social Science

Social Science Computer Review
1-26

© The Author(s) 2020

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0894439320904318

journals.sagepub.com/home/ssc



Jingyuan Yu¹ and Juan Muñoz-Justicia¹

Abstract

Twitter is an important object of research in social science studies, and the emergence of numerous Twitter software tools has provided researchers with the possibility of gaining insights into Twitter data; however, at the same time, early stage researchers may experience difficulties in selecting the most suitable tool for their own projects. Due to personal or institutional financial constraints, the purchase of commercial software is not a wise investment for all researchers. Hence, this article aims to provide a review of nine different free-of-charge and low-cost software tools for studying Twitter and highlights their advantages and disadvantages, in order to serve as a guide for social science scholars. This review is divided into two parts: background information and data retrieval features of the software tools, and their data analysis features. At the end of the review, several recommendations and suggestions are made for the use of these tools.

Keywords

Twitter, software, social science, free-of-charge software, low-cost, social network analysis

Twitter is an important academic data resource and object of research for social scientists (Burnap et al., 2015; Golder & Macy, 2012; McCormick et al., 2017). A single tweet can contain various types of information, including a username and biography, a hashtag, the content of the tweet, the time and location of posting, language, and so on. Based on these elements, several kinds of Twitter analyses have been developed in recent years. Roenneberg (2017) carried out a user analysis of the Twitter account @realdonaldtrump, analyzing the account owner's preferences regarding the use of an electronic device and the main user's chronotype, based on his tweeting time line. Barnard (2017) studied the hashtag #Ferguson using user network analysis and link analysis and examined how journalists and activists used Twitter to identify changes in field relations and practices. Bollen et al. (2011) analyzed Twitter text content using two types of sentiment analysis (positive vs. negative) and carried out mood measurement using six dimensions (calm, alert, sure, vital, kind, and happy).

¹ Universitat Autònoma de Barcelona, Spain

Corresponding Author:

Jingyuan Yu, Facultat de Psicologia, Departament de Psicologia Social, Universitat Autònoma de Barcelona, Edifici B, Departament B5-034, Campus de la UAB, Bellaterra (Cerdanyola del Valles), Barcelona 08193, Spain.

Email: jingyuan.yu@e-campus.uab.cat

These authors cross-validated the sentiment time series of Twitter content with the closing values of the Dow Jones Industrial Average and provided a potential method for predicting the stock market using the mood of Twitter content. Wilson and Dunn (2011) examined the language usage and user geolocation distribution of tweets containing the hashtag #jan25 during the 2011 Egyptian revolution and found that although the revolution occurred in an Arabic-speaking country, English dominated the language usage of the collected data set since the majority of powerful users were from Western countries. Their research showed that Twitter provided a transnational sphere for public dialogue that helped to enlarge the Egyptian revolution.

As in most empirical studies, one of the most significant elements during the research process is the selection of tools for both the data collection and data analysis steps. In other types of studies, there may be fewer options (or more standard tools); however, in Twitter analysis, we have a wide variety of possibilities when choosing a suitable software application. For example, Brooker et al. (2016) used the Chorus tool to provide a new methodology for Twitter data collection and analysis, allowing the user to explore the construction of developed phenomena through social media. Thelwall and Cugelman (2017), using the Mozdeh tool, proposed the “resonating topic method” to monitor an international organization’s Twitter strategy and to provide new propaganda strategies. Himelboim et al. (2017) used the NodeXL tool to classify Twitter topic networks into six different structures of information flow, which can be useful in evaluating Twitter activities and constructing strategies for Twitter and other social media spaces. Blaszkas et al. (2012) used DiscoverText to collect tweets related to #WorldSeries, and their study examined the use of this hashtag during the 2011 World Series.

A large number of these tools are commercial software (e.g., DiscoverText), and although these tools have been proved to be powerful options for social science research projects, some social science scholars are not able to purchase these software programs due to financial constraints. Thus, free, freeware, or low-cost (we define low cost as no more than US\$100) software tools are an indispensable option for these users. Different from free software, freeware refers to software that is available free of charge, while free software means that the users have the freedom to run, copy, distribute, study, change, and improve the software (“What is free software?,” 2018). On the other hand, Freeware and low-cost software tools have also shown good performance in retrieving and analyzing Twitter data (e.g., DMI-TCAT, Mozdeh) and in some cases, even better (Blaszkas et al., 2012). This article aims to provide an in-depth review of free-of-charge and low-cost Twitter software tools for social science research, highlighting their advantages and disadvantages in academic use and offering an easy method of software selection for researchers studying Twitter.

Ahmed (2019) has listed more than 20 different software tools that can be used for Twitter research. According to the author’s description, seven of them are completely free of charge (Chorus, COSMOS, Mozdeh, TAGS, Webometric Analyst, Gephi, and DMI-TCAT). In this review, in addition to the above-mentioned freeware, we select three low-cost software tools (NodeXL, Netlytic, and SocioViz) which we consider affordable for Twitter study beginners. Despite the fact that all these three software tools (including many other software tools which Ahmed mentioned) provide short-term free trial or limited free services (e.g., limited data samples, limited data retrieval time spans), the free trial version tools are considered inadequate for academic research. During our testing period, Chorus could not provide a stable performance on our computers, hence discussions about this software tool are excluded from this review. On the other side, although there are tools like R or Python packages that contain more flexible data analysis functions, all the software included in our review do not require programming skills.

General Twitter analysis consists of two steps: data retrieval and data analysis. In the following sections of this article, we first present the background and the data retrieval features of each selected tool, then we discuss and compare the data analysis functions of these tools, using several small-scale case study examples (data retrieval strategy based on the #MeToo hashtag).

Three important limitations should be mentioned. First, in addition to the software tools reviewed by Ahmed, there are still numerous Twitter analysis software applications (e.g., Social Media Data

Stewardship, 2018) that have been shown to be powerful for Twitter research and are free of charge. Due to the limited length of this article and institutional financial constraints, it is not possible to review all the free and low-cost Twitter study software tools at this time. Second, no guarantee can be made about the quality of our study objects, therefore using the reviewed software tools (and the collected Twitter data) should be at the users' own risk. Third, it will be helpful to do a deeper analysis about the types of social science research methods that the tools would support; although it is out of our actual scope, it would be interesting to be developed in the future.

Background Information and Data Retrieval Features of Software Tools

Application programming interface (API) and Twitter's data collection policy

All the abovementioned software tools require Twitter API (<https://apps.twitter.com>) access credentials or authorization before collecting Twitter data. Since a huge number of Twitter analytic tools require a manual authorization (not only our selected tools), scholars are strongly recommended to register at least one API before using these tools. There are two kinds of Twitter API: representational state transfer (REST) API and streaming API. REST API is mainly used to download historical Twitter data and user profile info. Streaming API is for real-time data collection.

Retrieving Twitter data should follow the guidelines and policy of the company; there are several official limitations, and here we present what we consider the most important (detailed information is available at <https://developer.twitter.com/>). The totally free-of-charge data collection option is called "standard search API" (consisting of both REST API and streaming API), which supports up to 7 days of historical data collection. Twitter data analyzers may collect tweets by both key word(s) and @username(s); in this last case, gathering tweets by certain @username(s) can only get the most recent 3,200 tweets of each Twitter user. In retrieving streaming tweets, a random 1% sample is allowed. In addition, Twitter has established data collection rate limits, which are divided into 15-min intervals (Twitter, n.d.). However, within the permitted data collection policy, there are still several differences among our selected software tools.

Windows tools

Mozdeh (<http://mozdeh.wlv.ac.uk/>; Thelwall, 2018b) and Webometric Analyst (n.d.; <http://lexiurl.wlv.ac.uk/>) are two desktop software tools that can be run only on Windows systems. Mozdeh is a free program for key word, issue, time series, sentiment, gender, and content analysis of social media texts (Thelwall, 2018a). Webometric Analyst was programmed by the same developer and institution as Mozdeh; it is a free program for gathering and analyzing web data and can be used for social web analysis, altmetrics, citation analysis, and link analysis (Thelwall, 2009, 2018a).

Regarding the data collection features of Mozdeh, this tool is able to retrieve historical tweets by key word(s) or Twitter username(s); both these data retrieval strategies allow to collect a maximum number of 72,000 tweets per hour. In the case of retrieving real-time data, a possible method is to keep Mozdeh running indefinitely; the retrieved data set is renewed every 15 min. Mozdeh is one of only two reviewed tools that can collect tweets by specific language (the other one being SocioViz). Boolean operators such as AND or OR can be added while formatting data retrieving queries. All retrieved data are saved automatically in local files into plain text format.

In the case of Webometric Analyst, as they are written by the same developer and institution, a data set retrieved using Mozdeh is compatible to this tool. Although developers suggest that it is better to use Mozdeh rather than Webometric Analyst to search for and download tweets ("Twitter—Webometric Analyst," n.d.), Webometric Analyst can be used to obtain Twitter user information from user IDs.

NodeXL (<https://www.smrfoundation.org/nodexl/>; Smith et al., 2010) is an add-in for Microsoft Excel (2007, 2010, 2013, 2016), available only for Windows Office (Mac MS Office users interested

in this software tool may use a Virtual Machine). NodeXL provides both a gratis version (with limited functions) “NodeXL Basic” and a paid version (with full functions) “NodeXL Pro.” NodeXL Pro offers a student/academic purchase plan. In the next sections, discussions about NodeXL refer to NodeXL Pro.

Comparing to Mozdeh, the data collection logic of NodeXL is different; given that NodeXL defines itself as a social network and content analysis tool (Smith et al., 2010), in addition to the Twitter content information, it also retrieves Twitter network information (relationships between users from replies and mentions). The main collection results are divided into two worksheet pages: “edges” and “vertices.” The edges consist of the relationship contained in a tweet, which includes mentions, replies to, retweet, and tweet. The vertices are the nodes in the network, which refer to the Twitter users involved in the relationship. The edge page also contains tweet content information (e.g., tweet text, tweet time, original tweet link), while in the vertices page, it provides Twitter user information (e.g., total number of tweets, total number of followers, user description).

NodeXL is able to collect up to 18,000 tweets in the same interval. By default, it downloads historical data; in the case of retrieving streaming data, the Connected Action Graph Server Importer should be installed to connect to the streaming API. Like Mozdeh, Boolean operators are also applicable in formatting the search query. The retrieved data set can be saved and exported in both spreadsheet formats and network structure formats (e.g., pajek, gexf).

Multi-Platform Tools

COSMOS (<http://socialdatalab.net/cosmos>; Burnap et al., 2015) and Gephi (<https://gephi.org/>; Bastian et al., 2009) are developed in Java. COSMOS (n.d.) is available at no cost to academic institutions and not-for-profit organizations; the developers have suggested that it runs best on Mac OS X and Linux Ubuntu, and installation in Windows is not recommended except in the last resort. Gephi is a free, open-source software for graph and network analysis. It can be used for Twitter research once the Gephi Twitter Streaming Importer plugin is installed.

Both of them can retrieve as much real-time data as the Twitter official policy allows. COSMOS allows formatting search queries by Boolean operators and provides an automatic data deuration function, while Gephi does not. COSMOS can export collected data in spreadsheet, plain text, and JavaScript Object Notation (json) format; Gephi can export data in spreadsheet formats (csv and tsv) and network structure formats (e.g., gexf, gml). While collecting data, COSMOS automatically generates tweet sentiment (range from -5 to $+5$) and tweeter gender information.

Similar to NodeXL, Gephi is mainly featured in network analysis. It is able to retrieve Twitter data by three categories: key word, username, and location. There are five types of networks that Gephi can collect: full Twitter network (represents all entities, including user, tweet, hashtags, URL, media, symbol, and so on), user network (that will allow to represent relations between users from mentions or retweets), hashtag network, emoji network, and Bernardamus projection (represents user network via hashtag present in tweets). Gephi retrieves edges and nodes; edges are relationships between the nodes; each of the nodes represents an element (e.g., user, hashtag, emoji) involved in the network.

DMI-TCAT (<https://github.com/digitalmethodsinitiative/dmi-tcat>; Borra & Rieder, 2014) is a software tool designed by researchers at the University of Amsterdam. It is written mostly in PHP and runs in a webserver (LAMP) environment; it is recommended to be installed on Linux distribution Ubuntu and Debian rather than on Windows and/or the Mac OS system.

DMI-TCAT is able to collect both historical and real-time Twitter data with Boolean operators. It collects not only tweets information but also the tweeter’s profile information. All retrieved data can be exported in csv and tsv format. There is no data deuration function in DMI-TCAT; however, different from other selected tools, it provides a wide range of data selection filters. Users are able to

choose the most suitable sample from the retrieved data set by various parameters such as startdate, enddate, tweet language, from user, exclude user, and so on.

Web-Based Tools

TAGS (<https://tags.hawksey.info/>) is a free Google Sheet template that allows the user to set up and run an automated collection of search results from Twitter (TAGS—Twitter Archiving Google Sheet, n.d.). Netlytic (<https://netlytic.org/>; Gruzd, 2016) and SocioViz (n.d.; <https://socioviz.net/>) are two commercial software tools; users must register their own account on the website and link it to Twitter before using these tools. Both of them provide free trial version with limited sample size; however, they also offer a low-cost advanced service for student/academic use.

TAGS can be used to retrieve historical data with Boolean operators; it can download up to 18,000 tweets. To retrieve real-time tweets, users can activate the “update archive by hour” function. There are three types of data collection strategies in TAGS: search and download historical tweets by key word, extract favorite tweets, and extract user time line or status updates by entering a screen name. TAGS also allows users to delete duplicate data. The retrieved data set can be exported into spreadsheet, pdf, and html formats.

There are three tiers of Netlytic account: The Tiers 1 and 2 accounts are free, and the Tier 3 account requires payment. Tier 1 users are able to freely upgrade to Tier 2 by filling out a simple form. The free account can save up to five data sets; each of them contains a maximum of 10,000 tweets. The Tier 3 account can save up to 300 data sets, and each of them contains a maximum of 100,000 tweets.

Netlytic can collect historical data with Boolean operators; in order to get real-time data, users should manually (free plan) or automatically (purchase plan) update their data set every 15 min. The retrieved data set can be filtered by four fields: link, pubdate, author, and title. It contains an automatic data cleaning function, which allows users to remove quoted text from all messages in the data set. All retrieved data are able to be exported in csv file.

SocioViz is able to collect both historical and real-time streaming data. Users can easily define the time range of the target data in the searching interface; three result types are available: most recent results, both popular and real-time results, and most popular results. However, neither data cleaning/depuration nor filter functions exist in SocioViz. Retrieved data can be exported in spreadsheet formats.

The information on downloaded data is slightly different among these three webpage-based tools. When searching and collecting data by key word(s), the tweeter’s profile information retrieved by TAGS and Netlytics is more detailed than in SocioViz. Data gathered by TAGS and SocioViz include geographic coordinate information (although there is only a small percentage of tweets that have such information), while in Netlytic, geographic information cannot be collected. In SocioViz and Netlytic, information about the language of the tweets is contained within the exported file, but in TAGS, there is no such variable.

Tables 1 and 2 are cheat sheets with details of the background information and the data retrieval features of our research objects.

Data Analysis Features of Software Tools

Data retrieval and preparation strategy. In order to achieve our research objectives and to enable a comprehensive comparison of the data analysis functions of the selected software tools, several similar small-scaled case studies are analyzed. We used each of the software to collect tweets based on the #MeToo hashtag (Date retrieved: October 17, 2019).

Since the number of recovered tweets is different in each case, we used different selection strategies and a filter to homogenize the size of each data set. As a result, we selected approximately 1,300 tweets with each of the selected tools; however, there are several exceptions: For NodeXL and

Table 1. Software Tools' Background Information Table.

Software	Latest Version	Latest Update Date	Initial Date	Author/Developer/Company	Software Property	Platform	Webpage
Mozdeh	V2	September 2018	—	Statistical Cybermetrics Research Group at the University of Wolverhampton	Freeware	Windows	http://mozdeh.wlv.ac.uk/
Webometric Analyst	V4.1	September 2018	April 2011	Statistical Cybermetrics Research Group at the University of Wolverhampton	Freeware	Windows	http://lexiurl.wlv.ac.uk/
NodeXL	V 1.0.1.413	June 2019	July 2008	Social Media Research Foundation	Commercial software	Windows Excel Add-in	https://www.smfoundation.org
COSMOS	V 1.5	—	2014	Social Data Science Laboratory at Cardiff University	Freeware	Multi-platform	http://socialdatalab.net/COSMOS
Gephi	V 0.9.2	September 2017	July 2008	Students of the University of Technology of Compiègne	Freeware	Multi-platform	https://gephi.org/
DMI-TCAT	—	Updated continuously	2014	Digital Method Initiative	Open-source software	Multi-platform	https://wiki.digitalmethods.net/Dmi/ToolDmiTcat
TAGS	V 6.1	May 2016	June 2010	Martin Hawksey	Freeware	Linux Ubuntu or Debian	https://tags.hawksey.info/
Netlytic	—	—	—	Social Media Laboratory at Ryerson University	Free software	Web-based	https://netlytic.org/home/
SocioViz	—	—	—	—	Commercial software	Web-based	https://socioviz.net/

Table 2. Software Tools' Data Retrieval Features.

Software	Twitter API		Real-Time Data Retrieval	Boolean Operators	Data Retrieval Quantity Limitation	Automatic Data Depuration or De-Duplication	Data Export Format	Programming Knowledge Requirement
	Access Credential or Authorization Request	Maximum Time Span for Historic Data Retrieval						
Mozdeh	Yes	Up to 7 days	Conditional (keep Mozdeh monitoring indefinitely)	Yes	Up to 72,000 tweets	Yes	Plain text	No
Webometric Analyst NodeXL	Yes	—	—	No	—	—	Plain text	No
COSMOS	Yes	Up to 7 days	Yes	Yes	Up to 18,000 tweets	Yes	Network structure format (Pajek, GEXF, GDF, etc.) .xls .csv, .json, .xlsx, plain text, etc.	No
Gephi DMI-TCAT	Yes Yes	— Up to 7 days	Yes Yes	No Yes	No limit Up to the maximum data quantity of the data sets	No No	.csv or .tsv .csv or .tsv	No No
TAGS	Yes	Up to 7 days	Conditional (data can be updated by hour)	Yes	Up to 18,000 tweets	Yes	.csv, .tsv, .xlsx, .ods, .pdf, etc.	No
Netlytic	Yes	Up to 7 days	Conditional (data can be manually or automatically updated)	Yes	Free account: 10,000 tweets per data set (max. five data sets) Paid account: 100,000 tweets per data set (maximum 300 data sets)	Yes	.csv	No
SocioViz	Yes	Up to 7 days	Yes	Yes	Free plan: up to 100 tweets Academic plan: up to 5,000 tweets Business plan: up to 50,000 tweets	No	.xls or .csv	No

Table 3. Mozdeh Word Association Analysis.

Word	Matches (%)	No Match (%)	Matches	Total	DiffPZ	χ^2	Significance (3,742 tests)
Harassment	62.3	1.9	38	61	21.6	453.1	***
Damaged	9.8	0.0	6	6	11.0	100.1	***
Trivial	9.8	0.0	6	6	11.0	100.1	***
Conflating	9.8	0.0	6	6	11.0	100.1	***
Comically	13.1	0.4	8	13	9.6	80.9	***
Stricken	13.1	0.4	8	13	9.6	80.9	***
Polio	13.1	0.4	8	13	9.6	80.9	***
Serious	11.5	0.2	7	10	9.7	80.4	***
Idea	13.1	0.6	8	15	8.9	68.3	***
Accuse	13.1	0.7	8	17	8.2	58.7	***

Gephi, as they both retrieve edges and nodes as main information, we controlled the total number of nodes to be 1,300. For Webometric Analyst: Following the advice of its developer, “To search for and download tweets, please use Mozdeh rather than Webometric Analyst” (“Twitter—Webometric Analyst,” n.d.), we did not collect data with this software tool; in the next data analysis section, data used in Webometric Analyst were retrieved from Mozdeh. SocioViz: A total number of 5,000 tweets were retrieved from October 16, 2019, at 07 hr 26 min 28 s, to October 17, 2019, 12 hr 59 min 53 s; however, there is no possibility to filter or select a practical sample with our data, so in the case of this software tool, we decided to analyze the whole retrieved data set.

Mozdeh

Mozdeh contains the following analysis functions: searching for specific texts within the collected data set by key word queries; sentiment strength and gender of the author; measuring the average sentiment for extracted or refined tweet contents or finding terms associated with positive or negative sentiment; drawing time series graphs of tweet activity or sentiment-based time series graphs of the whole data set or refined contents; creating a time line of Twitter activity for an individual user; mining word associations; finding gender differences in the texts; generating statistics including average retweets, citations, or number of likes; creating networks of users; and detecting spikes in key words (Thelwall, 2018b). However, before carrying out data analysis in Mozdeh, using the spam filtering function to mark and remove spam can make the research results considerably more efficient.

The minimum time interval unit in Mozdeh is an hour, thus Mozdeh’s time series graph function requires a large data set, and a pilot test data set may not be sufficient to build a graph.

The “mine word associations” function can be used to study the connections between words and to gain insights into important issues within the collected data set. There are three types of word association analyses that can be applied to the retrieved data: a key word query and/or filter, a single key word query and/or filter against another key word query and/or filter, and a comparison of the whole project to a reference set of text (Mozdeh Big Data Text Analysis, n.d.). Table 3 shows the top 10 words most closely associated with the query “harassment” in our collected data set, with words listed in descending order of statistical significance. Mozdeh uses the Benjamini–Hochberg significance (Benjamini & Hochberg, 1995) where three stars *** represent 0.1% significance, two stars ** represent 1% significance, one star * represents 5% significance, and words with no stars are ignored for this test. NoMatch gives the percentage of texts that do not match the search but do contain the word (Thelwall, 2018b).

Table 4 and Figure 1 present the sentiment analysis of the whole retrieved data set, where scores of 1 and –1 represent no positive and no negative (Thelwall, 2018b; Thelwall et al., 2010). It is

Table 4. Mozdeh Sentiment Analysis Results.

Sentiment Analysis Results		
Score	Positive (%)	Negative (%)
1	62.75	37.40
2	26.84	10.80
3	10.09	20.11
4	0.31	31.46
5	0.00	0.23
Average and 95% confidence intervals		
Positive	1.4797 [1.4421, 1.5173]	
Negative	2.4632 [2.3930, 2.5334]	
Average positive – average negative	-0.9836	

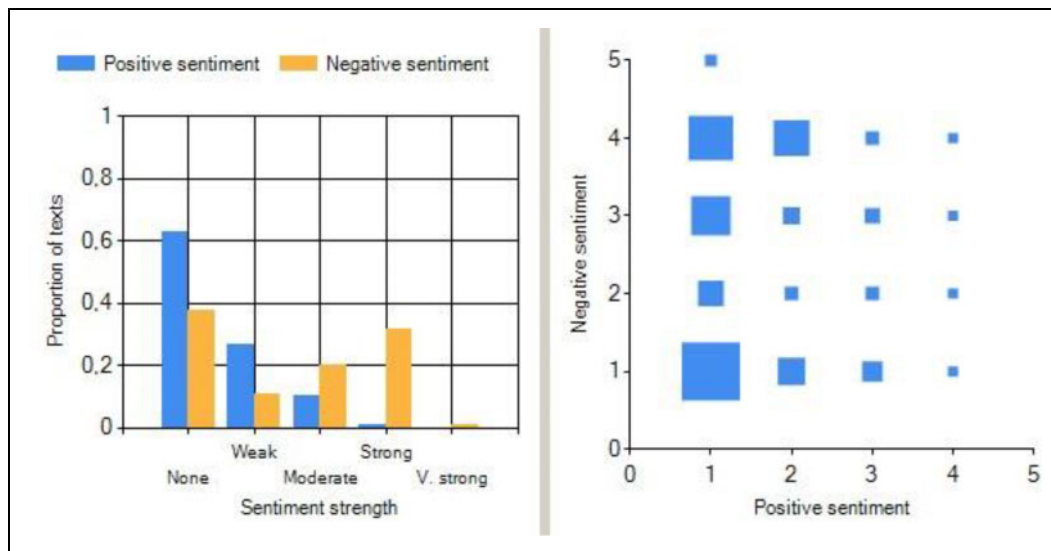


Figure 1. Mozdeh sentiment analysis bar and bubble charts (all data sets).

believed that Twitter users who posted about #MeToo showed a stronger negative sentiment than positive sentiment. As a plus, since the data set can be refined in Mozdeh by gender, key word, and so on, the results of other related analyses could be further explored, for example, a sentiment analysis based on gender differences.

Mozdeh developers recommend that when creating user networks, it is better to use Webometric Analyst rather than Mozdeh. This function will be discussed in the section on Webometric Analyst. However, the detection of a key word spike requires a large-scale data set, and a pilot test will not be able to provide useful results.

Webometric Analyst

Webometric Analyst can use data retrieved from Mozdeh and allows the user to analyze Twitter users' time lines and create networks from Twitter. The users' time line is created along with the users' tweet retrieval and can be saved automatically as a plain text file that can be viewed or processed using a spreadsheet or other related software tools (e.g., Mozdeh). In Webometric Analyst, the generation of networks from Twitter includes the creation of a Twitter conversation network and a following/follower network.

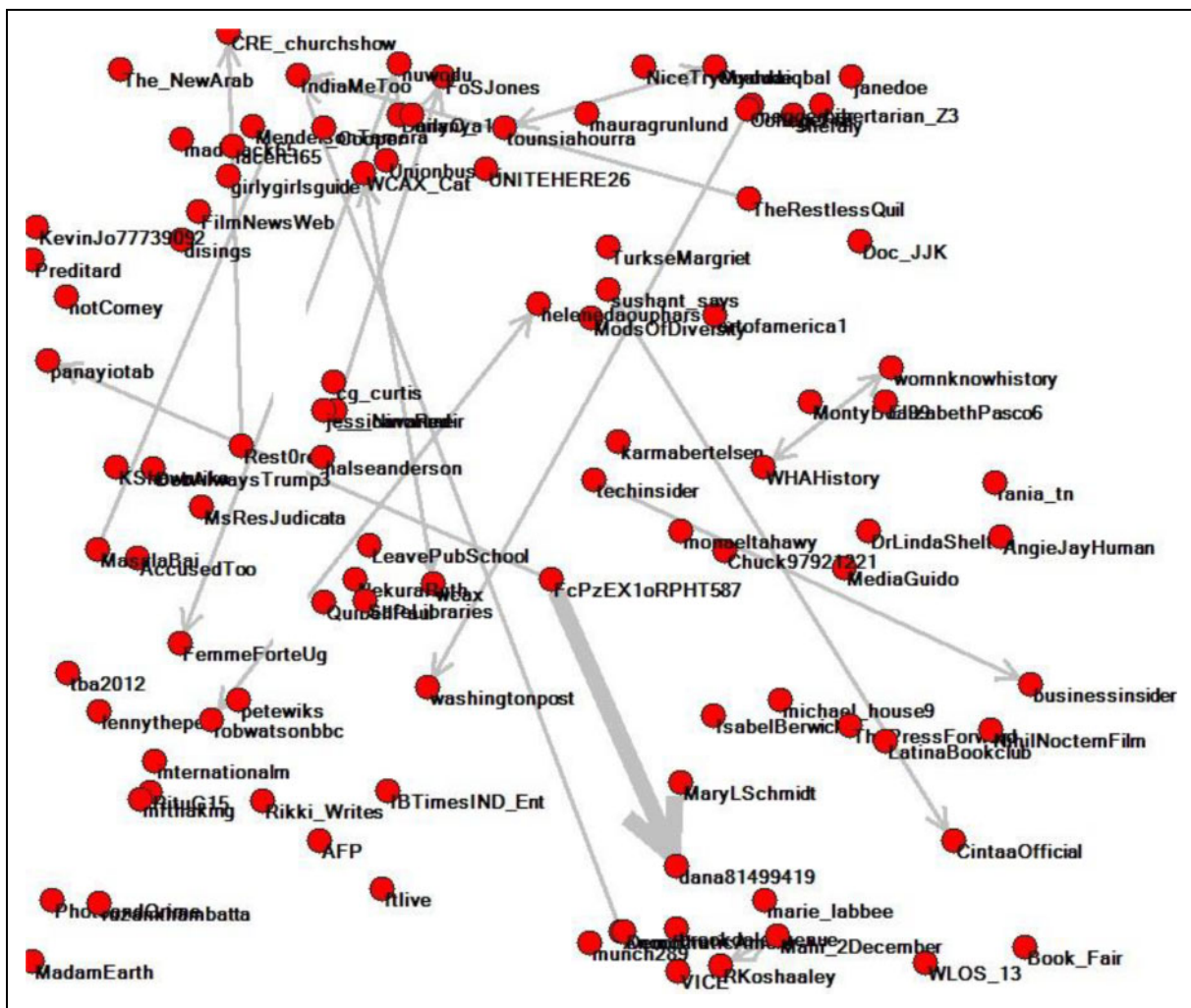


Figure 2. Webometric analyst Twitter conversation network.

Twitter conversation networks are of two types: co-mention networks for retrieved Twitter users and direct tweet networks of retrieved Twitter users (i.e., who tweets whom). Figure 2 shows the direct tweet network for the data previously collected via Mozdeh (to make the figure clearer and more understandable, here we only take a random 100 of the 1,278 original tweets as a test sample). Based on this network graph, we can easily identify the main line of conversation and the main Twitter users in this retrieved sample.

In Webometric Analyst, there are three types of following/follower network analysis: following networks, follower networks, and following/follower networks. Using Figure 2, we selected three conversationalists from the data set and created a user list as follows: @FcPzEX1oRPHT587, @dana81499419, and @panayiotab. Figure 3 shows the following/follower network for these three Twitter users (sample size: randomly 100 nodes). This function provides users with a more comprehensive understanding of the sociodemographic connections between Twitter users and the possibility of discovering the interpersonal connections behind a specific research topic.

NodeXL

The data depuration process implies the elimination of the duplicate edges; it is necessary to do so in some types of analysis, but in the case of Twitter data, researchers must be careful because the

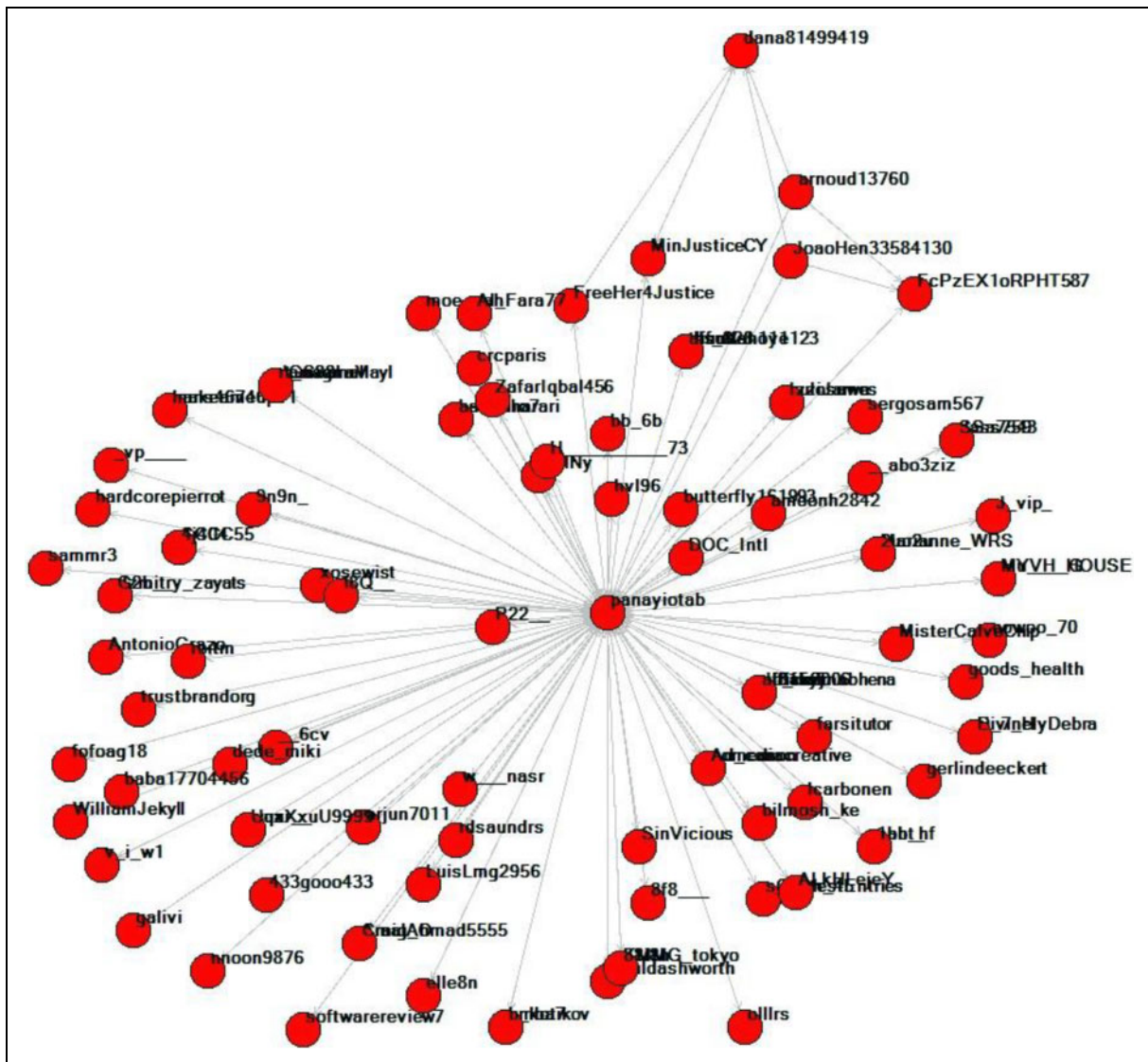


Figure 3. Webometric analyst following and follower network.

default configuration of NodeXL interprets duplicate cases where there is a match between a pair of nodes, which implies that if two users have more than one relationship (e.g., if User #1 mentions two different tweets of User #2, NodeXL would interpret it as a duplicate).

Given the features of the collected data, NodeXL allows only analysis of the network relationship between users; in other words, in the relationships tab, the nodes are only and exclusively users who can be related from mentions, replies to, or retweets (the relationship “replies to” is established only with the first person mentioned in a tweet).

The analysis can be started by performing metrics calculations at the node level. NodeXL offers a wide variety of metrics, including the most common centrality (degree, betweenness, and closeness) along with others such as PageRank or clustering coefficient. Network metrics such as graph density and graph reciprocity can also be calculated. Node-level calculations are included in their corresponding columns on the vertex sheet, while those related to the entire network will appear on the overall metrics sheet. In order to detect clusters or communities (on the groups tab), three algorithms can be chosen: Clauset–Newman–Moore, Girvan–Newman, or Wakita–Tsurumi.

The results can be viewed in the groups and group vertex sheets. The first one offers information about each one of the groups (metrics and visualization options), while the second

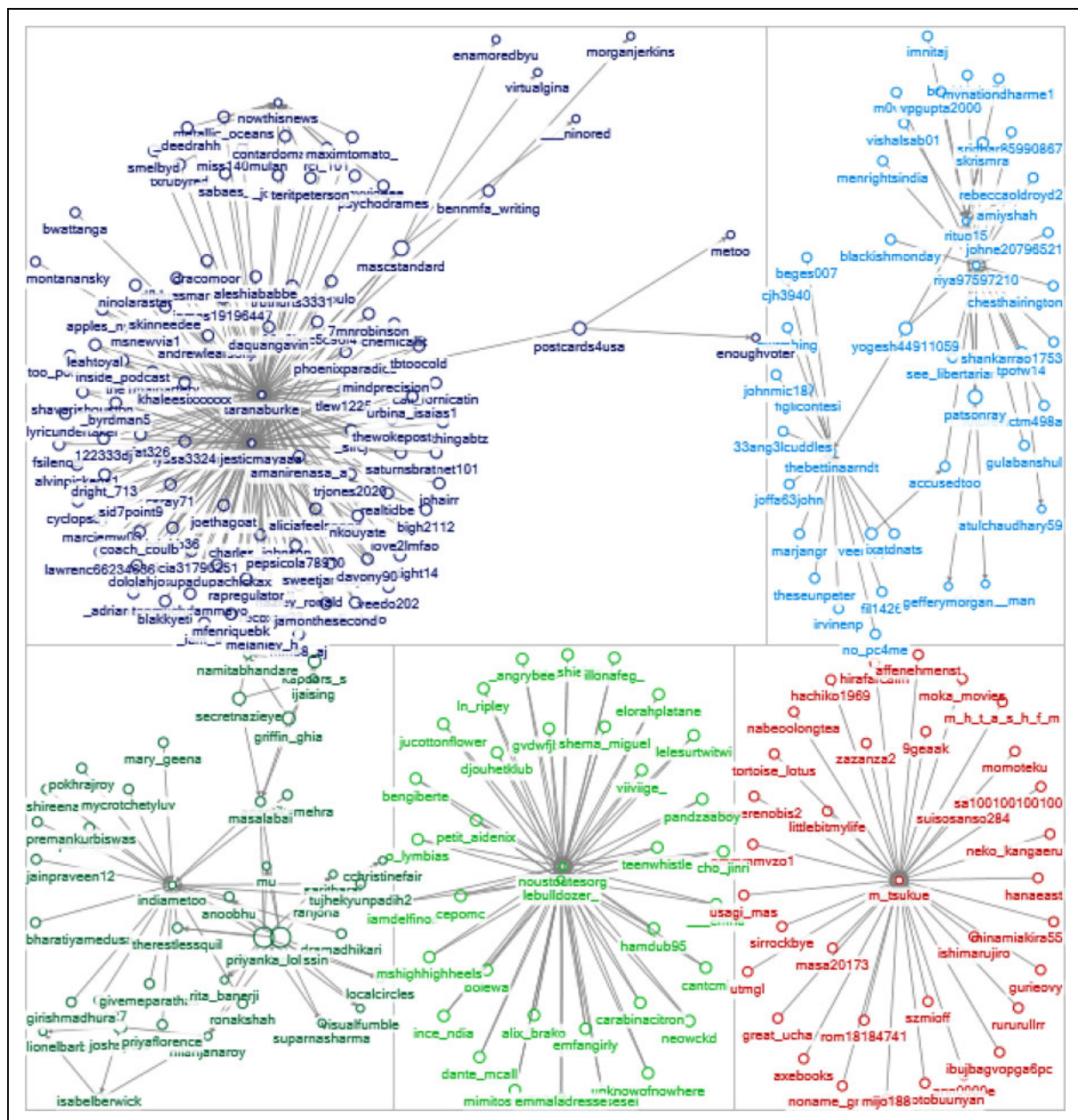


Figure 4. NodeXL network analysis treemap.

one shows the information about which (identified) group each node belongs to. A new “group edges” sheet is also generated showing the number of present relationships in each of (and between) the groups.

Once the calculations are done, we can visualize and adjust the aesthetics of the network graph. As in other programs, we can use the calculated metrics to, for example, modify the size, color, or shape of the nodes (for the properties of the nodes, we can use the metrics of the vertices or groups sheet), lay out the network (e.g., Fruchterman Reingold, Harel–Koren Fast Multiscale). Finally, by applying filters, the number of nodes and edges can be controlled.

Figure 4 shows an example in which the visualization option “Lay out each of the graph’s groups in its own box” (Treemap) has been used, in which each group is distributed in an individual virtual grid. A filter has been applied to show only the five groups with the highest number of members, and at the level of nodes, only those that have an in-degree value greater than 1 are displayed.

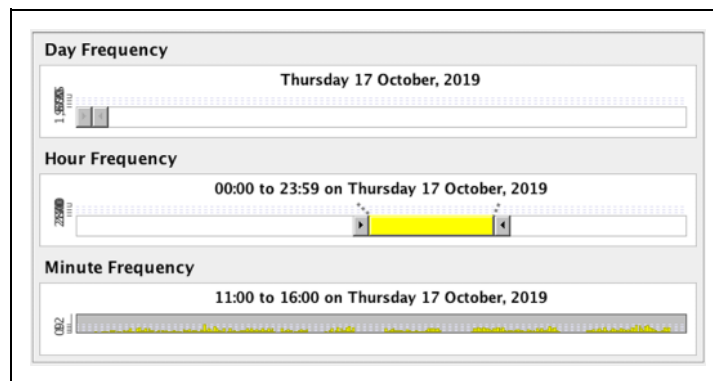


Figure 5. COSMOS frequency analysis.

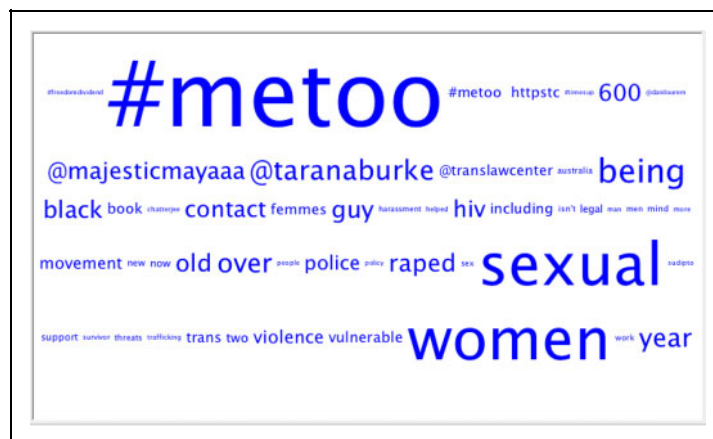


Figure 6. COSMOS word cloud.

One of the features that differentiates NodeXL from other software tools focused on network analysis is that it also allows content analysis of the tweets. With the option “network top items” (in metrics), the most frequent hashtags or URLs in the data set (and also previously calculated groups) can be easily obtained. Finally, NodeXL offers the possibility to perform a (limited) sentiment analysis, adding two new sheets (“words” and “word pairs”) in which it offers a list of words (or word pairs) with their positive or negative sentiment characterization.

COSMOS

After we retrieved our data, a tweet frequency analysis was first applied to monitor the Twitter activity on a specific topic. Figure 5 shows the tweet frequency of our sample. Unlike the time line functions of the other software tools, the COSMOS frequency analysis panel contains three categories: day, hour, and minute.

In addition to the tweet frequency time line, COSMOS can build a word cloud (Figure 6), with a view to identifying the most closely related and frequent key words in the sample by sizing textual and numeric data according to their frequency.

In the same way as the other software tools, COSMOS is able to build retweet networks and mentions networks. This provides researchers with the opportunity to discover the interactions between the Twitter users within the sample. In the geolocation analysis, COSMOS provides both an OpenStreetMap of the world, which is zoomable to street-level detail, and Environment System

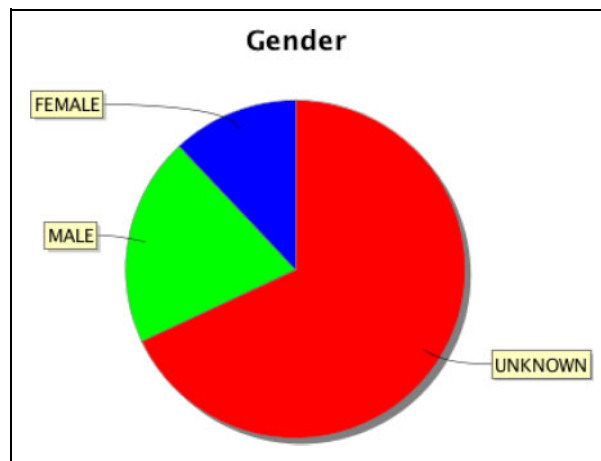


Figure 7. COSMOS Twitter gender analysis pie chart.

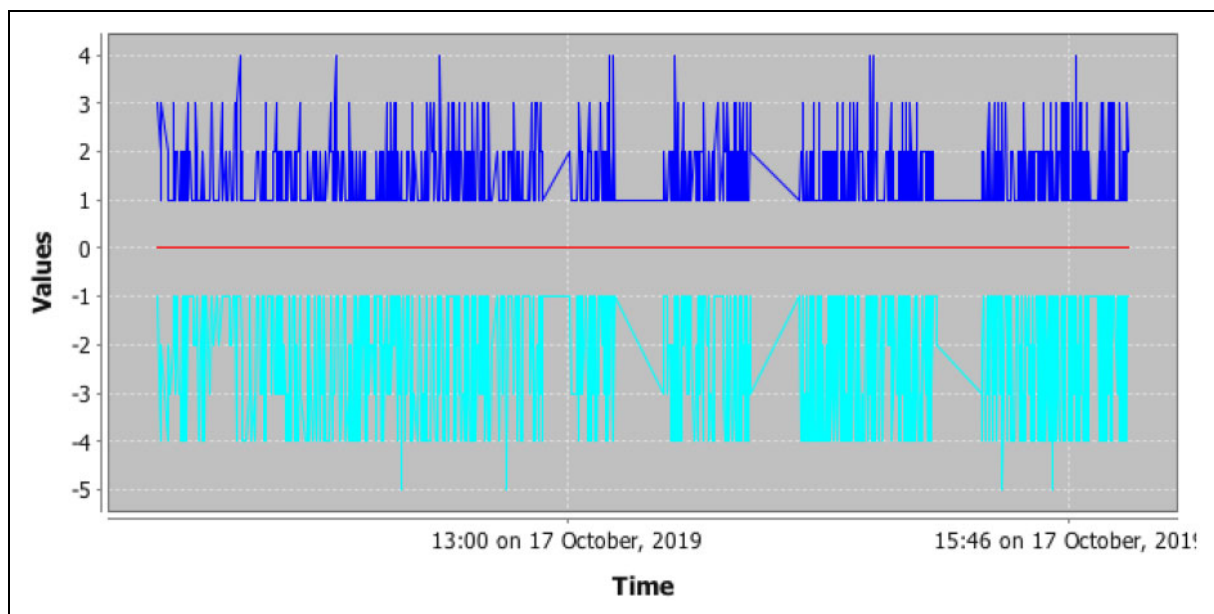


Figure 8. COSMOS sentiment analysis line chart.

Research Institute ShapeFiles for the UK. COSMOS users can easily find the core location of their research topic. However, a previous study has shown that the majority of Twitter users do not have location services enabled, and only 3.1% have been geotagged (Sloan & Morgan, 2015). When researchers are carrying out geolocation Twitter analysis, the sample size should be carefully considered. COSMOS can display a gender analysis in pie chart format (Figure 7), which may allow researchers to do Twitter research from a gender perspective.

COSMOS can also carry out Twitter sentiment analysis using sentiment scores by SentiStrength (Burnap et al., 2015; Thelwall et al., 2010). Sentiment criteria are scored from -1 (*no negative*) to -5 (*extremely negative*) and from 1 (*no positive*) to 5 (*extremely positive*). The COSMOS sentiment analysis is visualized using a line chart in which the x axis represents the tweet time and the y axis represents sentiment values. Figure 8 shows the sentiment analysis of the retrieved data set. It can be observed that within the period of our data sample, a negative sentiment dominates the retrieved #MeToo tweets.

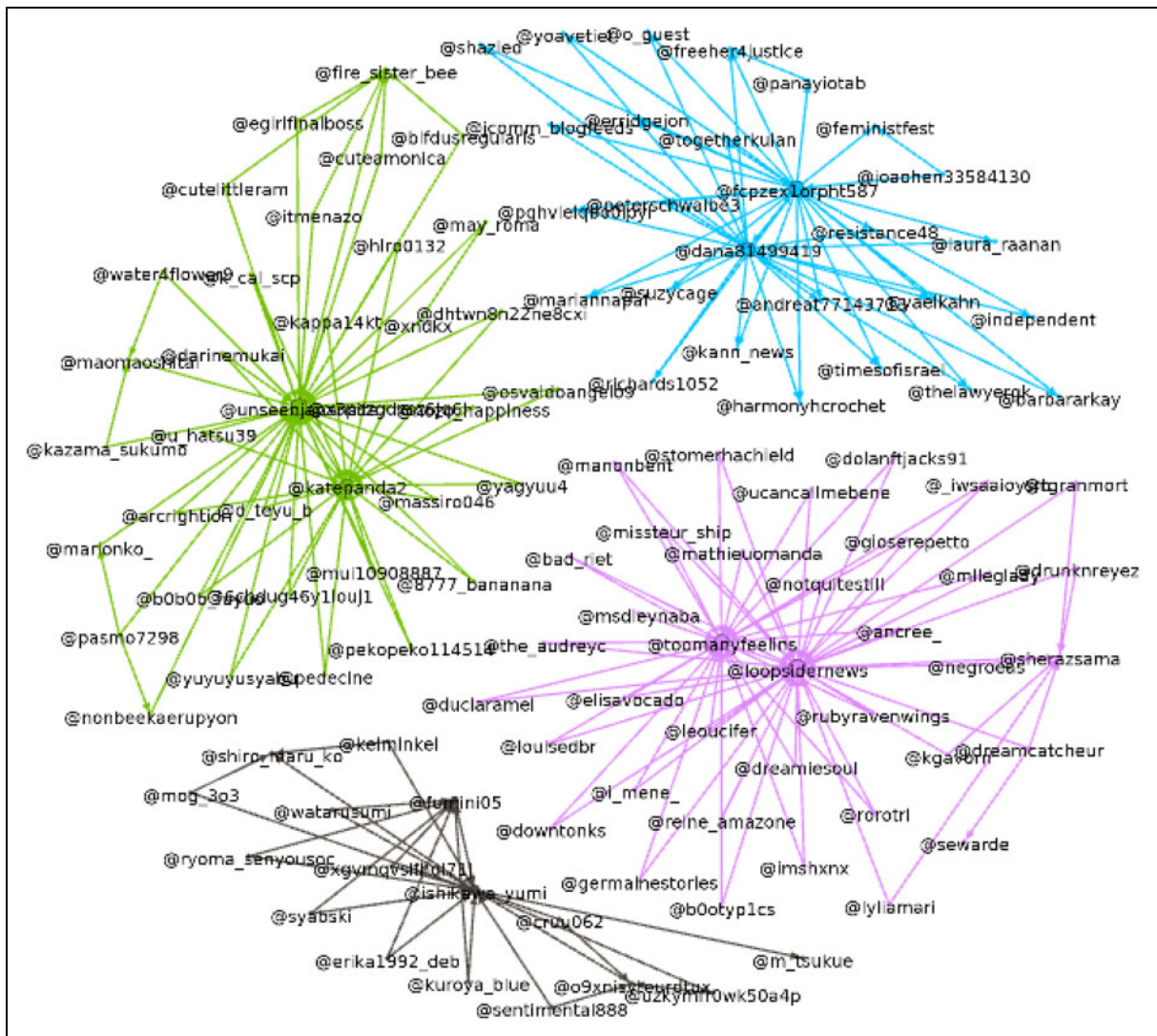


Figure 9. Gephi network analysis visualization.

Gephi

Gephi has three tabs: overview, where we can work interactively with the graph; data laboratory, where we find two tables with node and relationship data; and preview, where the user can see an improved version of the graph. Although it is mainly featured in network analysis, like the other abovementioned software tools, Gephi can also create Twitter activity time lines.

Regarding the retrieved network, to display the graph, several kinds of layouts (e.g., ForceAtlas, Fruchterman Reingold) are supported in Gephi. Network metrics (e.g., degree, centrality) can be easily calculated and adjusted through the “statistics” window. Community detection can be realized by using different algorithms such as Louvain, Girvan–Newman, or Leiden (the last two need plugin installation). All calculations performed and corresponding information can be viewed and manually edited in data laboratory.

Network graphs can be further processed in the appearance window (to adjust the appearance of the nodes and edges) and filter window (to filter the original network by different criteria/variables). Figure 9 shows a user network of our collected data (main parameters: layout = Fruchterman Reingold, community detection algorithm = Louvain, and min degree = 2).

Since Gephi was not specifically created for Twitter analysis, fewer Twitter data analysis functions are available in this software; however, all collected data can be exported for analysis with other software.

DMI-TCAT

DMI-TCAT is able to do many different types of data analysis. For the time line of tweets, DMI-TCAT provides three kinds of resolutions: by day, hour, or minute. Figure 10 shows the time line of the tweets by minute.

DMI-TCAT can offer a great variety of Twitter statistics and activity metrics including tweet statistics, user statistics (overall or individual), hashtag frequency, mention frequency, and so on. All reports are generated as .csv files, and the statistics can be grouped by minute, hour, day, week, and so on. Researchers can also define custom categories for statistics and activity metrics.

DMI-TCAT can generate different kinds of network analysis, described as “social graph by mentions,” “social graph by in_reply_to_status_id,” “co-hashtag graph,” “bipartite hashtag-mention graph,” and so on. All networks are generated as .gexf or .gdf files, which can be opened in Gephi or similar software.

There are several experimental data analysis functions: Cascade can be used to explore temporal structures and retweet patterns; “The Sankey Maker” produces an alluvial diagram that can be used for plotting the relation between various fields such as hashtags, sources, languages, and so on; and the associational profile is used to explore shifts in hashtag associations. Figure 11 shows the relation between the sources and hashtag of the retrieved data (cutoff = 10). In this figure, several interesting findings can be highlighted: Hashtags related to #MeToo have been presented; the Twitter Mobile App made up almost half of the tweet sources; paper.li and ifttt were the most popular nonofficial software in #MeToo tweets.

TAGS

The tweet analysis in TAGS mainly consists of three functions: the TAGS Summary Sheet, the TAGS Dashboard Sheet, and the TAGS Explorer. The first two of these are built based on Google Spreadsheets, and the TAGS Explorer uses external sources to visualize replies to tweets, mentions, and retweet networks.

The TAGS Summary Sheet displays statistical results for the collected sample. It contains a set of general statistics within the sample data, such as the number of links, number of retweets, number of tweets, tweet rate, and so on. In addition, it can provide each tweeters’ mention and retweet activities as well as the tweeters’ link. The TAGS Dashboard Sheet contains four main modules: top tweeters, Twitter activity, tweet volume over time (max 60 days), and a sheet of tweets with the most retweets from the last 1 or 2 days.

A tweet replies network was built in the TAGS Explorer (Figure 12), and from this network, we can identify the connections between the tweeters and the key player in this sample. Researchers can also visualize mention and retweet networks by changing the custom variable of the networks. However, in mapping network, TAGS Explorer does not provide a sample selection function nor can the network layout be switched.

The “top tweeters,” “top hashtags,” and “top conversationalists” rankings can be generated on the TAGS Explorer page. These are shown in bar graph format, which can provide researchers with a comprehensive view of the key information. The “search archive” function can be found alongside the “top conversationalists.” Two archive searching methods are available in this function: The first delimits the time interval, allowing the user to find a specific tweet archive(s) by its post time, and

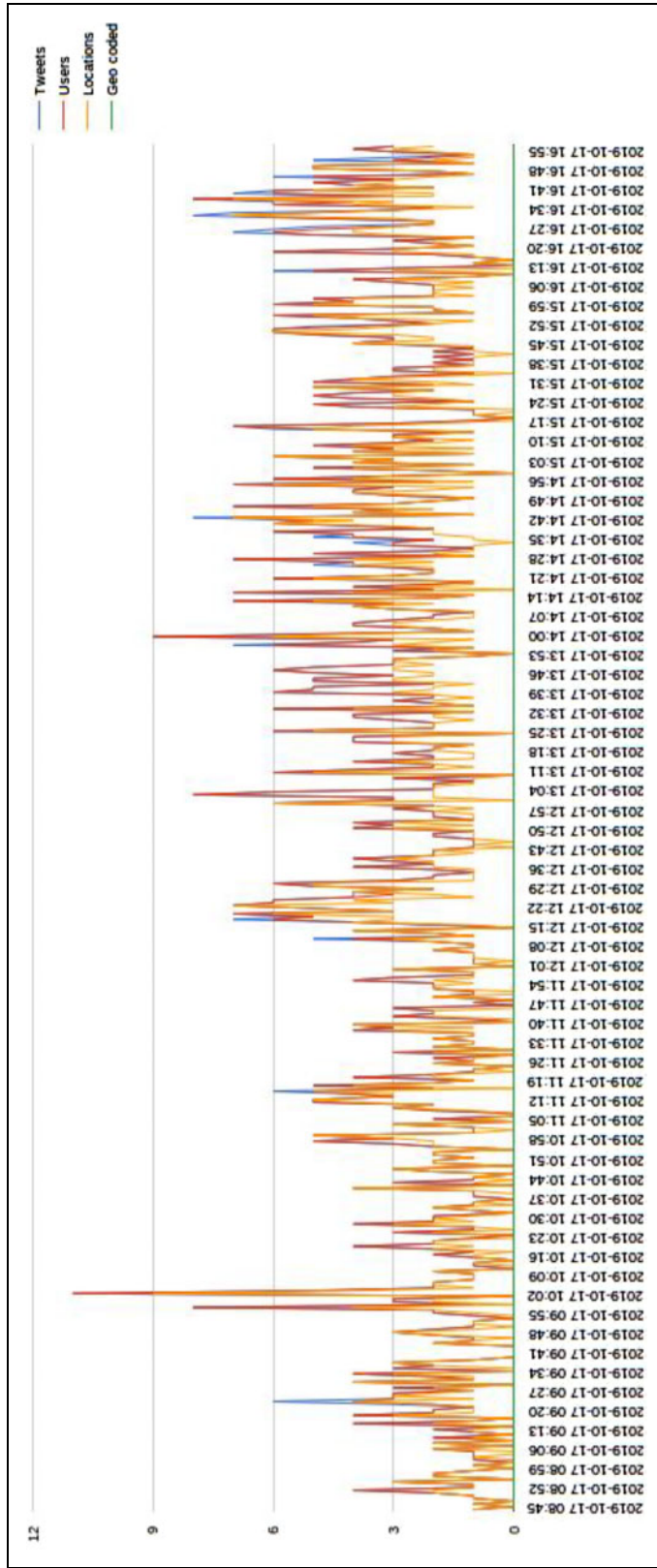


Figure 10. DMI-TCAT tweets time line by minutes.

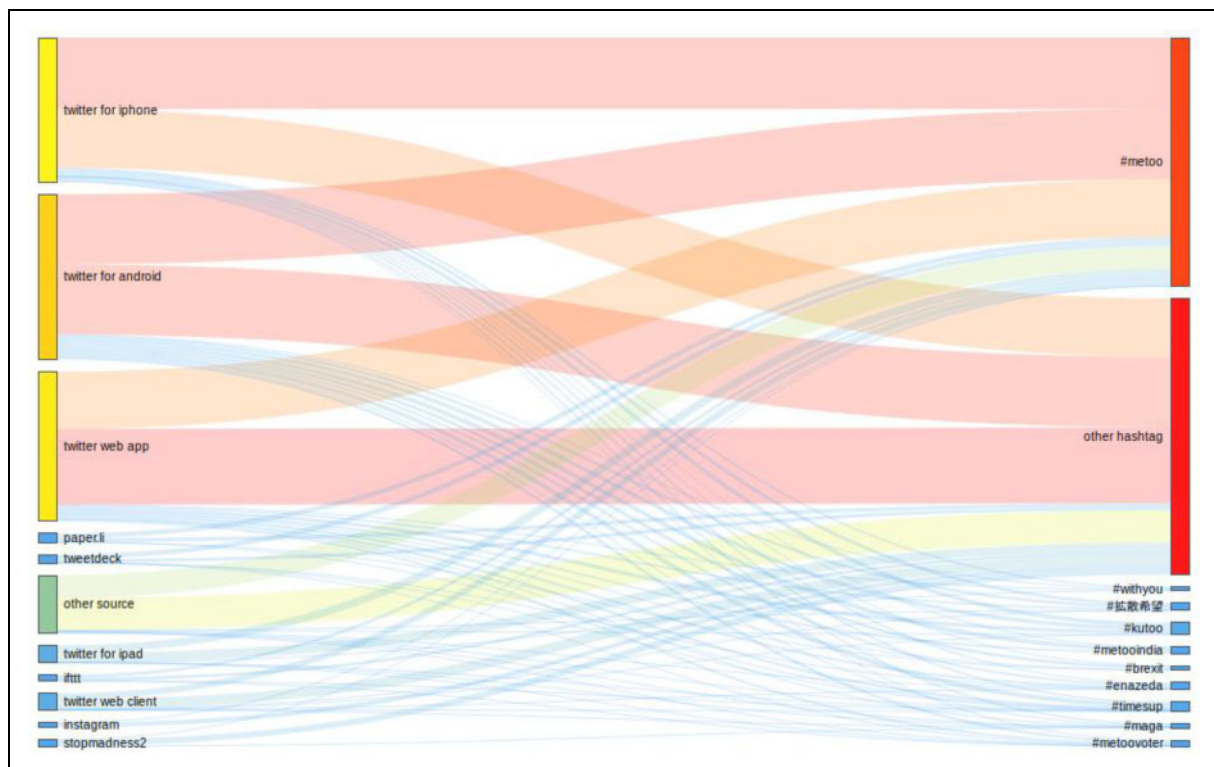


Figure 11. Relation between tweet sources and hashtags of the captured hashtag.

the second allows archive(s) to be searched by filter (tweet key word, screen name). Figure 13 shows the “search archive” panel.

Netlytic

There are mainly two kinds of analysis in Netlytic: text analysis and network analysis. Two functions are contained in text analysis: “key word extractor” and “manual categories.” Key word extractor can be used to identify popular topics of the data set by word frequency. “Word cloud” and “stacked graph” are the two options to visualize the results of key word extractor. Stacked graph is able to show the word frequency (y axis) over time (x axis); a maximum of 100 top topics (represented as key words) can be displayed. “Manual categories” aims to classify broader concepts of the text data (e.g., emotion detection) and can be visualized as an interactive Treemap; however, this function is not fully automated. Figure 14 shows the stacked graph (top topics = 100) of key word extractor; key word evolution alongside time series has been clearly presented. This analysis result can also be exported as a comma-separated values file.

There are two functions in network analysis: “name network” (who mentions whom) and “chain network” (who replies to whom). Both of them contain three kinds of layouts: “Fruchterman Reingold,” “DrL layout,” and “LGL layout.” Clusters (FastGreedy Algorithm) are automatically generated and represented in different colors. Mapped networks are zoomable and can be saved. Figure 15 shows the name network of our data set; @taranaburke received the most mentions in its cluster and can be identified as a key player in the whole data set. The mapped networks can also be exported to csv, gexf, and GraphML formats.

Finally, Netlytic contains a report panel, which includes several kinds of statistics about the collected/imported data set, such as geographic posts (shown on a world map, map data:

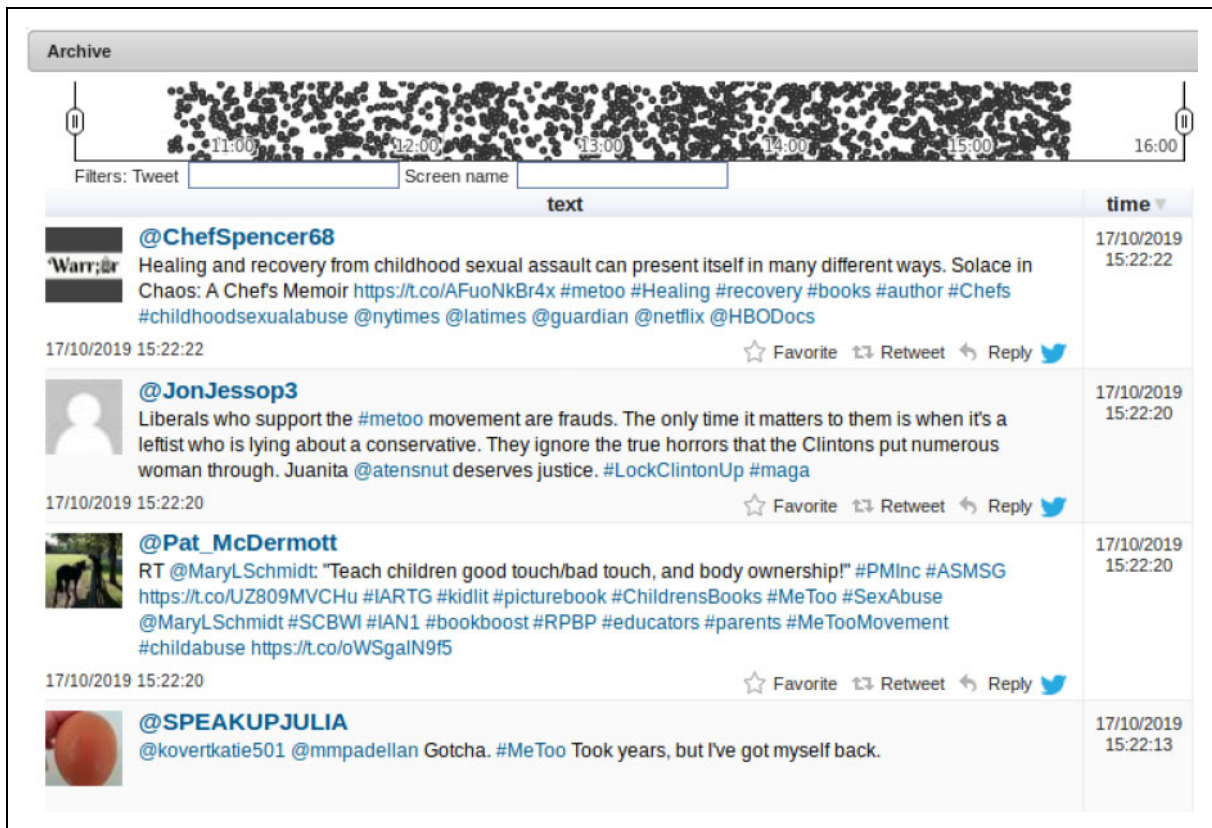


Figure 13. TAGS search archive function panel.

results can be exported as png, gml, and gexf formats, which may give users the possibility to further explore the network with other software. Figure 17 provides the hashtag network of our data set, in which several clusters are clearly displayed.

Table 5 presents a detailed comparison of the main data analysis functions of the nine software tools.

Discussion and Conclusion

The data retrieval features of each software tool are different; COSMOS and Gephi can only retrieve real-time data. While the other abovementioned software tools can retrieve both historic and real-time data, it should be mentioned that DMI-TCAT, Gephi, and SocioViz do not offer automatic data deuration or de-duplication functions, meaning that after retrieving Twitter data, researchers need to clean their data manually (or using other software tools). The majority of these software tools allow the use of Boolean operators when searching for Twitter data, which offer the possibility of retrieving more specific data. This provides an easy way of selecting the right data retrieval tool.

Mozdeh and Webometric Analyst can be seen as twin software tools. Both of these rely on each other; however, for data collection, it is recommended to use Mozdeh rather than Webometric Analyst, while for building networks, Webometric Analyst has a more important role. As stated on its official webpage, Mozdeh is a big data text analysis tool, and the most powerful feature of this software is its word association analysis (Thelwall, 2018b). However, several limitations should be mentioned: Tweet time line analysis with Mozdeh requires large-scale data, and small data sets are not sufficient to build a time line, meaning that Mozdeh is the best fit for large Twitter data sets.

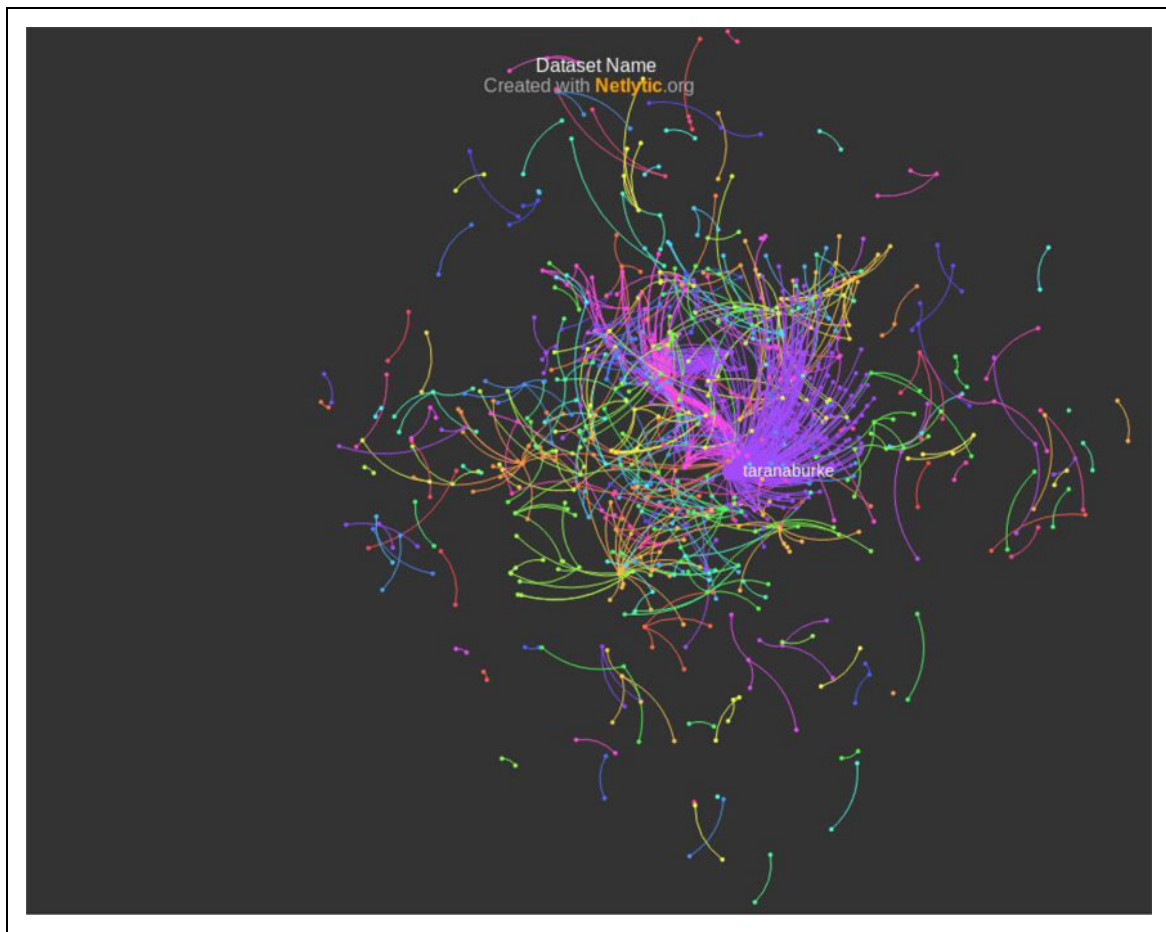


Figure 15. Netlytic name network.

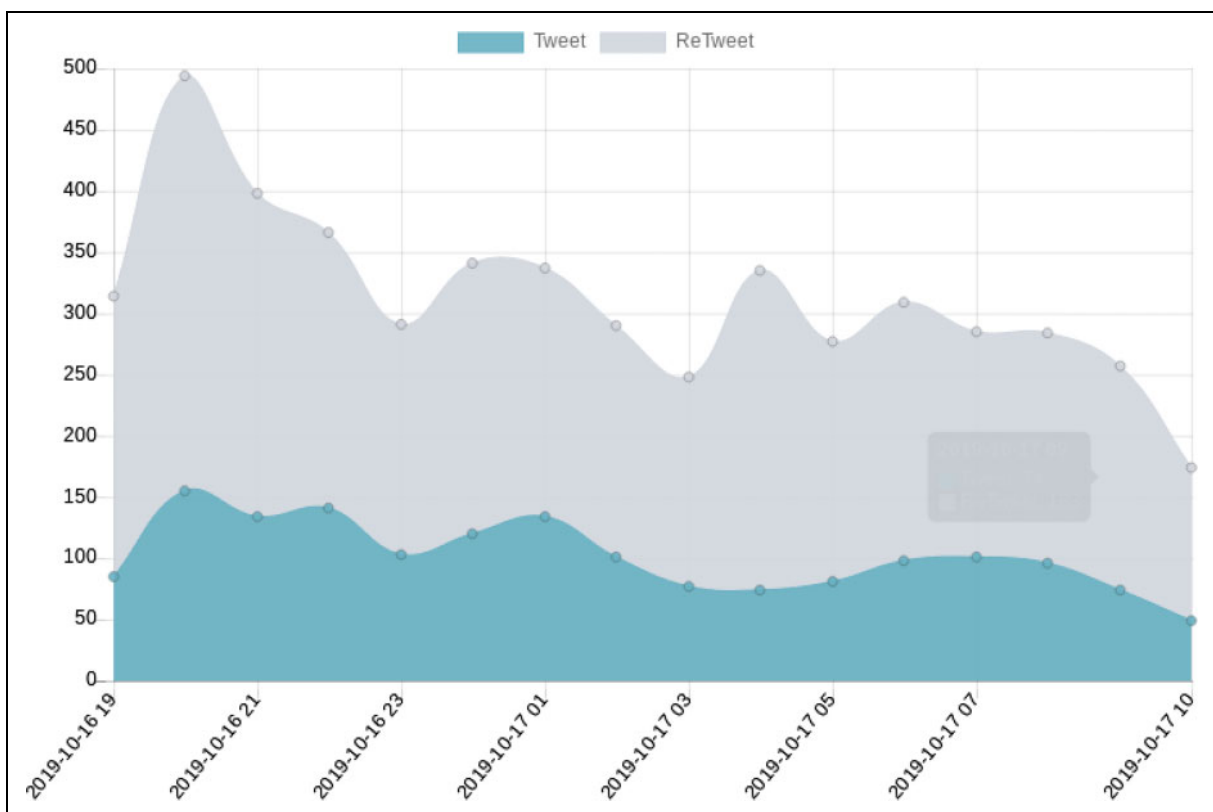


Figure 16. SocioViz tweet and retweet time line.

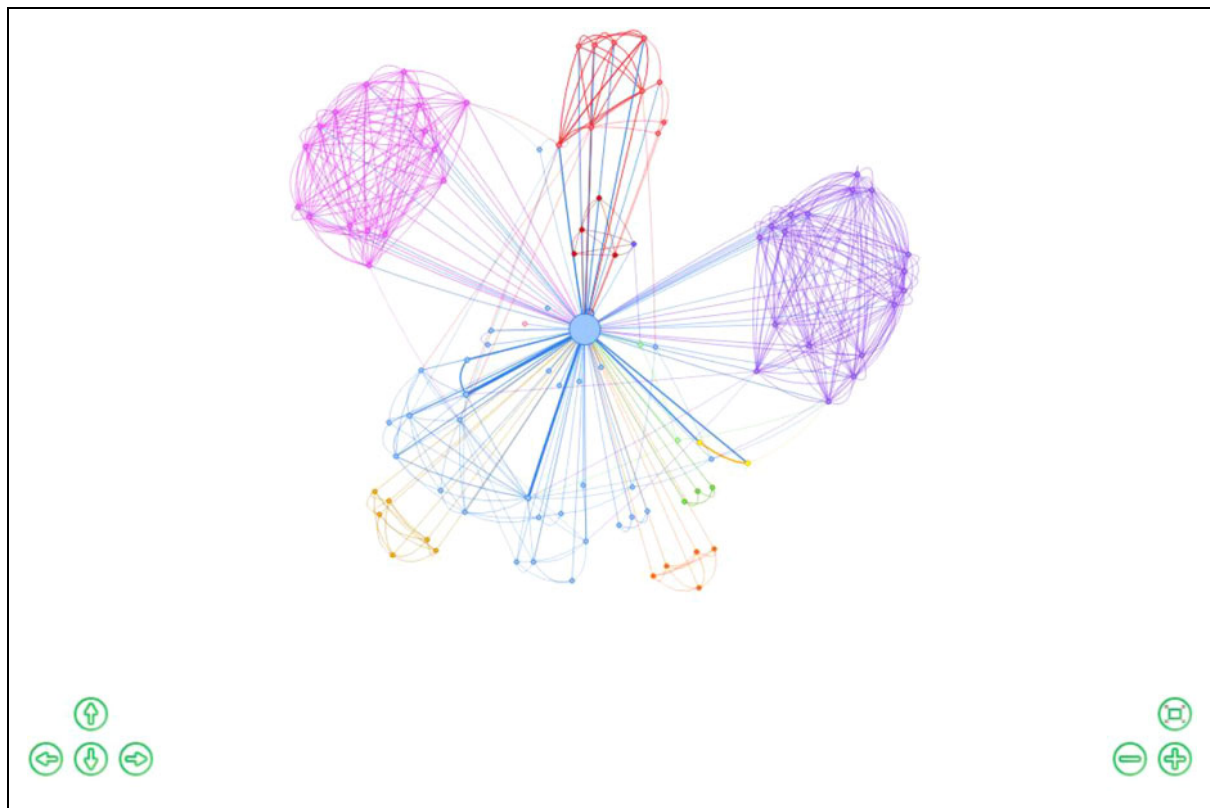


Figure 17. SocioViz hashtag network.

Table 5. Software Tools’ Data Analysis Features.

Software	Tweet Time Line Analysis	Gender Analysis	Content Sentiment Analysis	Geolocation Analysis	Tweet Source Analysis	Cluster Detection	User Network	Hashtag Network	Analysis Result Export
Mozdeh	Yes	Yes	Yes	No	No	No	Yes	No	Yes
Webometric Analyst	Yes	No	No	No	No	No	Yes	No	Yes
NodeXL	Yes	No	Yes	No	No	Yes	Yes	No	No
COSMOS	Yes	Yes	Yes	Yes	No	No	Yes	No	Yes
Gephi	Yes	No	No	No	No	Yes	Yes	Yes	Yes
DMI-TCAT	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes
TAGS	Yes	No	No	No	No	No	Yes	No	Yes
Netlytic	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes
SocioViz	Yes	No	No	No	No	Yes	Yes	Yes	Yes

During our research, we found that the data collection efficiency of Mozdeh was lower than that of other software, although this is still a powerful tool for Twitter research.

During our test and experiment period, NodeXL showed a stable performance. It offers a great variety of options to optimize Twitter network analysis. For less experienced scholars, it also provides a friendly way to explore Twitter networks with the “automate” function. As NodeXL runs on Microsoft Excel, it may be able to employ more types of statistics than other network analysis software tools. Being one of only two programs that mainly focused on network analysis in this review (the other one being Gephi), NodeXL can also do some content analysis, while Gephi cannot. However, several cautions or limitations should be recognized while using this program.

First, as we have discussed, the data deuration function of NodeXL may contain the risk of inaccuracy while removing duplicated edges. Second, NodeXL is only able to realize user network analysis, and compared to other tools, it lacks the ability to analyze the relationship of other elements (e.g., hashtags, emojis).

COSMOS gives good performance in both the data retrieval and data analysis phases. It also includes various types of analysis functions, one of the most notable of which is its geolocation analysis. This makes COSMOS one of only two applications (the other one being Netlytic) of the nine applications studied here that can carry out this type of analysis. However, its limitations are also important, since by default, COSMOS only provides a UK map in its geolocation analysis, and further geographical information needs to be added externally. Moreover, it also lacks several important data analysis functions such as cluster detection and hashtag networks.

Gephi was not created exclusively for social media analysis, but its powerful visualization function offers various kinds of layouts and algorithms for building networks and detecting clusters, which can greatly improve the quality of data visualization. Compared with NodeXL, Gephi is simpler, but it is more flexible in the sense of network types. Due to its nature, it can rarely be used as a primary tool for Twitter research, but as the network analysis results of other software can generally be exported as gexf or gdf or gml formats, it is recommended as a secondary tool in related studies.

DMI-TCAT showed good performance in both the data retrieval and data analysis phases. Compared to other software tools, the data visualization function of DMI-TCAT is one of the best in terms of aesthetics. It is the only tool that can carry out tweet source analysis. However, although it offers various kinds of data analysis functions, DMI-TCAT does not have certain important functions such as content sentiment analysis, cluster detection, and so on.

The three web-based software tools may be much easier to master than other software tools, but they all have limitations. TAGS (and SocioViz) provides very limited data set storage, meaning that large academic projects are difficult to realize. For TAGS, the network analysis does not provide data selection function; scholars who want to control visualization sample size must reduce data from the original data set. Furthermore, unlike Netlytic and SocioViz, TAGS users cannot configure network parameters nor can the network layout be switched. Its very limited types of analysis functions mean that it is difficult to provide in-depth insights. However, for small-scale data analysis (e.g., a pilot test), TAGS can be considered an efficient tool for quickly scanning a specific research theme, and especially for beginning Twitter researchers.

Netlytic is able to provide a much larger data set; it is also the only web-based software tool that can carry out sentiment and geolocation analysis. Although it cannot provide hashtag networks, it is still the most competitive web-based tool regarding rich kinds of data analysis functions. The only problem remains on the manual categories of text analysis, which may require researchers to invest more time on it.

SocioViz is the only tool that contains neither automatic data deuration/de-duplication function nor data import function; this limitation makes SocioViz data the least manipulable. Given that the text analysis of this tool provides only simple statistics, no in-depth insights can be reached. In network analysis, one inconvenient point is that users are not able to change the node label size, which makes it difficult in zoomed networks to identify key values (users, hashtags, etc.). Regarding the important limitations, SocioViz is perhaps the least flexible software tool for Twitter studies in this review.

Authors' Note

This work was carried out within the framework of the doctoral program entitled Person and Society in the Contemporary World at the Autonomous University of Barcelona/Este trabajo ha sido realizado en el marco del programa de doctorado en Personas y Sociedad en el Mundo Contemporáneo de la Universitat Autònoma de Barcelona.

Data Availability

All our research raw data are available at: <https://osf.io/t3krg/>

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Software Information

Mozdeh: Version: 2.0. Download url: <http://mozdeh.wlv.ac.uk/installation.html>

Webometric Analyst: Version: 4.1. Download url: <http://lexiurl.wlv.ac.uk/searcher/installingWebometricAnalyst.htm>

NodeXL: Version: 1.0.1.413. Download url: <https://www.nodexlgraphgallery.org/Pages/Registration.aspx>

COSMOS: Version: 1.5. Download url: <http://socialdatalab.net/COSMOS>

Gephi: Version: 0.9.2. Download url: <https://gephi.org/>

DMI-TCAT: Version: —. Download url: <https://github.com/digitalmethodsinitiative/dmi-tcat>

TAGS: Version: 6.1. Download url: <https://tags.hawksey.info/>

Netlytic: Version: —. Download url: <https://socioviz.net/>

SocioViz: Version: —. Download url: <https://netlytic.org/>

References

- Ahmed, W. (2019). *Using Twitter as a data source: An overview of social media research tools (2019) | Impact of Social Sciences*. <https://blogs.lse.ac.uk/impactofsocialsciences/2019/06/18/using-twitter-as-a-data-source-an-overview-of-social-media-research-tools-2019/>
- Barnard, S. R. (2017). Tweeting #Ferguson: Mediatized fields and the new activist journalist. *New Media & Society*. <https://doi.org/10.1177/1461444817712723>
- Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*. <https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 57, 289–300. https://www.jstor.org/stable/2346101?seq=1#metadata_info_tab_contents
- Blaszka, M., Burch, L. M., Frederick, E. L., Clavio, G., & Walsh, P. (2012). #WorldSeries: An empirical examination of a Twitter hashtag during a major sporting event. *International Journal of Sport Communication*, 5, 435–453. <https://doi.org/10.1123/ijsc.5.4.435>
- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2, 1–8. <https://doi.org/10.1016/J.JOCS.2010.12.007>
- Borra, E., & Rieder, B. (2014). Programmed method: Developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management*, 66, 262–278. <https://doi.org/10.1108/AJIM-09-2013-0094>
- Brooker, P., Barnett, J., & Cribbin, T. (2016). Doing social media analytics. *Big Data & Society*. <https://doi.org/10.1177/2053951716658060>
- Burnap, P., Rana, O., Williams, M., Housley, W., Edwards, A., Morgan, J., Sloan, L., & Conejero, J. (2015). COSMOS: Towards an integrated and scalable service for analysing social media on demand. *International Journal of Parallel, Emergent and Distributed Systems*, 30, 80–100. <https://doi.org/10.1080/17445760.2014.902057>
- COSMOS. (n.d.). <http://socialdatalab.net/COSMOS>

- Golder, S., & Macy, M. (2012, January). *Social Science with Social Media*. http://www.asanet.org/sites/default/files/savvy/footnotes/jan12/socialmedia_0112.html
- Gruzd, A. (2016). *Netlytic: Software for automated text and social network analysis*. <https://netlytic.org/>
- Himelboim, I., Smith, M. A., Rainie, L., Shneiderman, B., & Espina, C. (2017). Classifying Twitter topic-networks using social network analysis. *Social Media + Society*, 3, 1–13. <https://doi.org/10.1177/2056305117691545>
- McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., & Spiro, E. S. (2017). Using Twitter for demographic and social science research: Tools for data collection and processing. *Sociological Methods & Research*. <https://doi.org/10.1177/0049124115605339>
- Mozdeh Big Data Text Analysis. (n.d.). <http://mozdeh.wlv.ac.uk/>
- Roenneberg, T. (2017). Twitter as a means to study temporal behaviour. *Current Biology*, 27, R830–R832. <https://doi.org/10.1016/j.cub.2017.08.005>
- Sloan, L., & Morgan, J. (2015). Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PLoS ONE*, 10, e0142209. <https://doi.org/10.1371/journal.pone.0142209>
- Smith, M., Ceni, A., Milic-Frayling, N., Shneiderman, B., Mendes Rodrigues, E., Leskovec, J., & Dunne, C. (2010). *NodeXL: A free and open network overview, discovery and exploration add-in for Excel 2007/2010/2013/2016*. <https://www.smrfoundation.org>
- Social Media Data Stewardship. (2018). *Social Media Research Toolkit*. <http://socialmediadata.org/social-media-research-toolkit/>
- SocioViz. (n.d.). <https://socioviz.net/>
- TAGS—Twitter Archiving Google Sheet. (n.d.). <https://tags.hawksey.info/>
- Thelwall, M. (2009). Introduction to webometrics: Quantitative web research for the social sciences. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1, 1–116. <https://doi.org/10.2200/S00176ED1V01Y200903ICR004>
- Thelwall, M. (2018a). *Big Data and Social Web Research Methods*. <http://www.scit.wlv.ac.uk/~cm1993/papers/IntroductionToWebometricsAndSocialWebAnalysis.pdf>
- Thelwall, M. (2018b). *Social Web Text Analytics with Mozdeh*. <http://mozdeh.wlv.ac.uk>
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61, 2544–2558. <https://doi.org/10.1002/asi.21416>
- Thelwall, M., & Cugelman, B. (2017). Monitoring Twitter strategies to discover resonating topics: The case of the UNDP. *El Profesional de La Información*, 26, 649. <https://doi.org/10.3145/epi.2017.jul.09>
- Twitter. (n.d.). *Overview-Twitter developers*. <https://developer.twitter.com/en/docs/tweets/search/overview>
- Twitter—Webometric Analyst. (n.d.). <http://lexiurl.wlv.ac.uk/searcher/twitter.htm>
- Webometric Analyst. (n.d.). <http://lexiurl.wlv.ac.uk/searcher/twitter.htm>
- What is free software? (2018). <https://www.gnu.org/philosophy/free-sw.en.html>
- Wilson, C., & Dunn, A. (2011). Digital media in the egyptian revolution: Descriptive analysis from the Tahrir data sets. *International Journal of Communication*, 5, 1248–1272. <http://ijoc.org>

Author Biographies

Jingyuan Yu is a PhD candidate at the Universitat Autònoma de Barcelona. He received his master's degree in social communication from the Universitat Pompeu Fabra. His research focuses on social network analysis, data mining and visualization, political communication, and cyber-psychology.

Juan Muñoz-Justicia is an associate professor at the Universitat Autònoma de Barcelona. His main research interests include social psychology, qualitative research methodology, and social network analysis.

Chapter 4

Third Publication: Analyzing Spanish News Frames on Twitter during COVID-19-A Network Study of El País and El Mundo



Article

Analyzing Spanish News Frames on Twitter during COVID-19—A Network Study of *El País* and *El Mundo*

Jingyuan Yu ^{1,*} , Yanqin Lu ² and Juan Muñoz-Justicia ¹

¹ Department of Social Psychology, Universitat Autònoma de Barcelona, 08193 Barcelona, Spain; Juan.Munoz@uab.cat

² School of Media and Communication, Bowling Green State University, Bowling Green, OH 43403, USA; ylu@bgsu.edu

* Correspondence: jingyuan.yu@e-campus.uab.cat

Received: 26 June 2020; Accepted: 24 July 2020; Published: 28 July 2020



Abstract: While COVID-19 is becoming one of the most severe public health crises in the twenty-first century, media coverage about this pandemic is getting more important than ever to make people informed. Drawing on data scraped from Twitter, this study aims to analyze and compare the news updates of two main Spanish newspapers *El País* and *El Mundo* during the pandemic. Throughout an automatic process of topic modeling and network analysis methods, this study identifies eight news frames for each newspaper's Twitter account. Furthermore, the whole pandemic development process is split into three periods—the pre-crisis period, the lockdown period and the recovery period. The networks of the computed frames are visualized by these three segments. This paper contributes to the understanding of how Spanish news media cover public health crises on social media platforms.

Keywords: Twitter; news frame; network analysis; topic modeling; Spain

1. Introduction

As COVID-19 is becoming a global health crisis, it has been announced as pandemic by World Health Organization (WHO, Geneva, Switzerland) on 11 March [1]. Three days after, being one of the most infected countries, Spanish prime minister Pedro Sanchez declared state of alarm. This is the second time that Spain declared a national lockdown, so the influence of the pandemic on Spain is substantial. As the situation of the pandemic became stable, the Spanish government announced a 4-step plan for the transition to a new normality on 3 May (*Plan para la transición hacia una nueva normalidad*), signaling that the pandemic is gradually becoming under control.

News media are important information sources for the public during epidemic crisis [2], serving as interactive community bulletin boards, as well as global or regional monitors [3]. With the prevalence of social media, news media organizations have been using these emerging tools to reach and engage broader audiences during crises [4]. Twitter, being one of the most popular social media, has attracted a great number of traditional newspapers to digitalize real-time core information within 280 characters. While newspaper articles tend to use conflict, responsibility, consequence and savior frames in the coverage of epidemics, their Twitter accounts often post real-time updates, scientific evidence and actions [5]. The tones adopted in the two kinds of news are also different, with newspaper articles using more alarming and reassuring tones and Twitter updates using more neutral tones [5].

Scholars have been using the network analysis techniques to study news content. For example, Guo [6] proposed a Network Agenda Setting Model (NAS) to analyze the salience of the network relationships among objects and/or attributes. Inspired by this method, this study conducts network analysis on the Twitter posts, analyzing and comparing the news frames of the two most important

general-interest and nationally-circulated Spanish newspapers (*El País* and *El Mundo*) during different stages of the COVID-19 crisis. The two selected newspapers are considered different regarding their political stance [7], with *El País* representing the political center-left media and *El Mundo* seen as a political center-right media outlet [8,9]. Discussion on the two media would allow us to better explore their particular news focus regarding their divergent political ideologies, thus illustrating a more comprehensive landscape of Spanish news coverage on the pandemic. Moreover, as this study focuses on the analysis of their Twitter content, compared with other newspapers, *El País* and *El Mundo* have the largest number of online followers, reflecting their substantial influence online.

Two research gaps are filled in this paper. From the empirical approach, despite the fact that the two Spanish newspapers have been widely studied in the past epidemic crisis [10–12], their news posts on Twitter deserves more investigation in communication research. From the methodological perspective, manual coding process is generally applied in most of the network news agenda and news frame studies [13,14]. To enhance efficiency and minimize the biases involved in manual coding, this study combines unsupervised machine learning technique and network visualization method to make a fully automatic network study, which is a major methodological contribution to the news frame literature.

2. Literature Review

2.1. Framing and Health Communication

Framing is an important research focus in communication studies because how an issue is reported in news can influence how it is understood by audiences [15]. Entman [16] defined framing as “to select some aspects of a perceived reality and make them more salient in a communication text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described” (p. 52). Frames in news media coverage can affect the topical focus and evaluative implications perceived by the audience, as well as their subsequent decision making about public policy [17].

News frames about health issues and diseases have been found to affect audiences’ understanding of health problems and their attitudes and behaviors [18,19]. Regarding the ongoing COVID-19 pandemic, the severity of the virus and preventive actions should be communicated to the public effectively. In this case, news media play an important role in enhancing public’s understanding of the highly contagious disease, as well as in influencing the attitudinal and behavioral response on the prevention, containment, treatment and recovery [18].

Empirical studies about news frames have been conducted during the past epidemic crisis. For example, Lee and Basnyat [18] focused on the news articles of Singaporean *Straits Times* during H1N1 pandemic and identified nine dominant frames via manual coding—*basic information, preventive information, treatment information, medical research, social context, economic context, political context, personal stories* and *other* (open-ended). Their study revealed that the news coverage focused more on H1N1 information updates and prevention than on other frames. In another one of their articles [20], four additional news themes were found—*imported disease, war/battle metaphors, social responsibility* and *lockdown policy*. Shih, Wijaya and Brossard [21] focused on news coverage about the mad cow disease, West Nile virus and avian flu from the *New York Times* by examining six frames—*consequence, uncertainty, action, reassurance, conflict, new evidence*. The results of their study revealed that the newspaper emphasized the consequence and action frames consistently across diseases but media concerns and journalists’ narrative considerations regarding epidemics did change across different phases of development and across diseases.

2.2. Framing in Spanish News Media

According to the Association for Media Research (*Asociación para la Investigación de Medios de Comunicación*, <http://reporting.aimc.es/index.html#/main/diarios>), *El País* and *El Mundo* are the two

most read general-interest newspapers in Spain in the first quarter of 2020. Comparative studies about these two newspapers have been conducted in various context. For example, Baumgartner and Chaqués-Bonafont [7] found that there are important news coverage differences between these two newspapers when they make explicit reference to individual political parties. Regarding negative news about corruption, *El País* tends to mention right-wing political party, while *El Mundo* mentions left-wing political party more often. The comparison between these newspapers in their news coverage about cannabis have also shown significant differences, *El País* focused more on the news about marijuana legalization, while *El Mundo* focused more on police and crime news on drug consumption [22].

During the Ebola outbreak, Ballester and Villafranca [12] studied the two newspapers together by comparing their news coverage of Ebola with other rare diseases. The word “terror” appears more frequently in Ebola related news, generating a higher level of anxiety toward Ebola than other diseases. Catalan-Matamoros et al. [10] studied the visual contents of the two newspapers, two main conclusions are made by the authors. First, the “conflict” frame dominates the portal of the two newspapers, which revealed alarming messages for the audience. Second, they found the total number of visual content increased rapidly in the first two days of the crisis and decreased from the fifth day. In sum, the authors described the first two days as “high risk phase” of the epidemic outbreak and from the fifth day onward the “less severe phase.”

Regarding the ongoing COVID-19 crisis, researchers have found that there is a significant increase of coronavirus news in Spanish State of Alarm phase than the pre-alarm period and the total number of relevant news reported by *El Mundo* is much more than *El País* [23]. Thanks to the ease of information exchanges on social media platforms, Masip et al. [2] indicates that Spanish citizens are more informed during the coronavirus crisis than before. In this case, an in-depth analysis of social media news is warranted.

2.3. Methodological Background

Latent Dirichlet Allocation (LDA) is frequently used to extract latent topics from large scale textual data and has also been widely applied for social media studies [24–26]. According to the developers of this technique, “LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document” [27] (p. 993).

Previous research has suggested LDA an appropriate method to study news media coverage [28]. For example, Heidenreich, Lind, Eberl and Boomgaarden [29] used this method to identify 16 frames from European refugee crisis news across five countries. For the COVID-19 related studies, Poirier et al. [30] applied LDA to identify six news frames (*Chinese outbreak, economic crisis, health crisis, helping Canadians, social impact, Western deterioration*) from 12 Canadian media sources.

In addition, network analysis methods have been widely adopted on communication studies. For example, regarding the mad cow disease, Lim, Berry and Lee [31] visualized the core word network of four groups (bureaucrats, citizens, scientists and interest groups) across four policy stages based on 6400 newspaper articles. They found the four groups focused on different policy issues and the news coverage did change over different stages. This study demonstrated that semantic network analysis is a powerful method for understanding issue framing in the policy process. Fu and Zhang [32] used word co-occurrence network to study NGOs’ HIV/AIDS discourse on social media and website. Their study revealed overlapping themes about HIV/AIDS across social media and website and NGOs use social media to engage with the government, as well as other health care resources. Kang et al. [33] examined the vaccine sentiment on Twitter by constructing and analyzing semantic networks of related information and found that semantic network of positive vaccine sentiment has a greater cohesiveness than the less-connected network of negative vaccine sentiment. This study sheds the light on discovering online information with a combination of natural language processing and network methods.

On the other side, Bail [34] conceptualized a method to combine natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media.

The author's idea is to convert the content of different documents into bag-of-words and then find the similarities (edges) between the documents by word co-occurrence. This method is further developed as a visualization tool to display a text network at group-word level [34,35]. In our case, each of the computed news frame (latent topic) is considered as a group of their relevant content, represented as nodes on the network and the edges between the frames are visualized according to the co-occurrence of the content and weighted by term frequency-inverse document frequency (tf-idf). To be clearer, the tf-idf is a numerical statistic to measure how relevant a word to a document in a corpus [36], it has been widely applied in text mining research, also in the abovementioned Bail's work [34].

3. Methods

Our data are hydrated from open access institutional and news media tweet dataset for COVID-19 social science research [37], which includes the Twitter posts from the two selected Spanish newspapers from the end of February. The first step is data cleaning, in which all the retweets are removed. Then we deleted all the attached external website addresses, hashtags (#hashtags), mentions (@mentions), emojis, Arabic numbers and stopwords (e.g., prepositions, pronouns etc.), because such information is considered less meaningful in computational text analysis [38]. In addition, all the capital letters were converted to lower case (to standardize all the words) and we normalized the text with lemmatization (which refers to group together the inflected forms of a word) before the data are ready for the LDA model analyses.

Using the LDA function of R package "topicmodels" [39], we computed eight topics for each newspaper's Twitter posts. The decision made on the number of topics is because too few topics make news frames less specific and too many topics make the network less interpretable [40]. In order to make the performance of the topic model more efficient, we used the Gibbs sampling method [39,41], one of the most widely used statistical sampling techniques for probabilistic models and short-text classification [42–45].

After having obtained the computed topics (news frames), we re-assigned each of the news tweets into their belonging frames, so we have a new dataset with the tweets of each newspaper categorized by the news frames. As the news focus regarding epidemics did change across different phases of the pandemic's development [21], following the work of Pan and Meng that adopted a three-stage model to analyze news frames during a previous pandemic [46], we split each dataset by three periods. The pre-crisis period includes tweets before March 14 when Spanish national lockdown was announced. This is the period that the pandemic information has been reported but not been officially alarmed by Spanish government. The lockdown period includes tweets between 14 March and 11 May, the period that the Spanish government adopted a strict national confinement. The recovery period includes the tweets from 11 May (the day when Spain stepped into the first stage of social recovery) to 3 June (the last day of data collection). Finally, a network of relationships between news frames has been generated from their word co-occurrence matrix for each newspaper during each time period. Therefore, a total of six networks are constructed.

4. Results and Discussion

4.1. *El País*

For the *El País* dataset, a total number of 22,223 tweets are collected from 25 February 2020 to 3 June 2020. After removing retweets, 14,800 original tweets are saved for our in-depth analysis. Eight news frames have been successfully computed, they are "Livelihood" (family life and children), "Public Health Professional" (news about the department of public health), "Pandemic Update" (contagion and death poll), "Madrid" (news about Madrid), "Politics" (general political news), "State of Alarm" (Spanish government and PM's announcement and policy update), "Economy" (the effect of the pandemic on Spanish economy) and "Covid Information" (general information about the pandemic). Table 1 presents the details of the eight news frames of *El País* with their top seven relevant words.

Table 1. El País Twitter news frames with most relevant words (translated into English).

Livelihood	Public Health Professional	Pandemic Update	Madrid
years	simon	spain	madrid
life	public	country	pass
live	mask	death	week
child	fernando	hour	confinement
leave	change	contagion	exit
son/daughter	should	die	common
family	form	data	phase
Politics	State of Alarm	Economy	Covid Information
police	government	crisis	person
think	sanchez	million	pandemic
inform	doctor	month	sanitary
question	minister	arrive	world
politics	alarm	spanish	virus
video	health	work	hospital
ask	president	economy	covid

Figure 1 presents the news frame network of the three segmented periods. Each of the nodes represents a news frame and the size of nodes indicates the strength of the node, also known as weighted node degree, it is the sum of the edge weights of the adjacent edges for each node [47], reflecting the importance of a node in a weighted network. The edges between the nodes represent the connection strength between two frames (normalized by tf-idf), it is the sum of the tf-idf value of the co-words. Table 2 presents the detailed information about the news frames in each of the three periods, with the node strength, number of tweets in each news frame and their proportion of the total number in each segment. Table 3 presents the table of the most weighted edges in the three time segments; it is able to provide us the news frames with the highest similarity ties. Overall, “Livelihood,” “Public Health Professional,” “Pandemic Update” and “Politics” are the most important news frames of *El País*. As the crisis is gradually under control, the “Pandemic Update” turned to be less prominent in the recovery period.

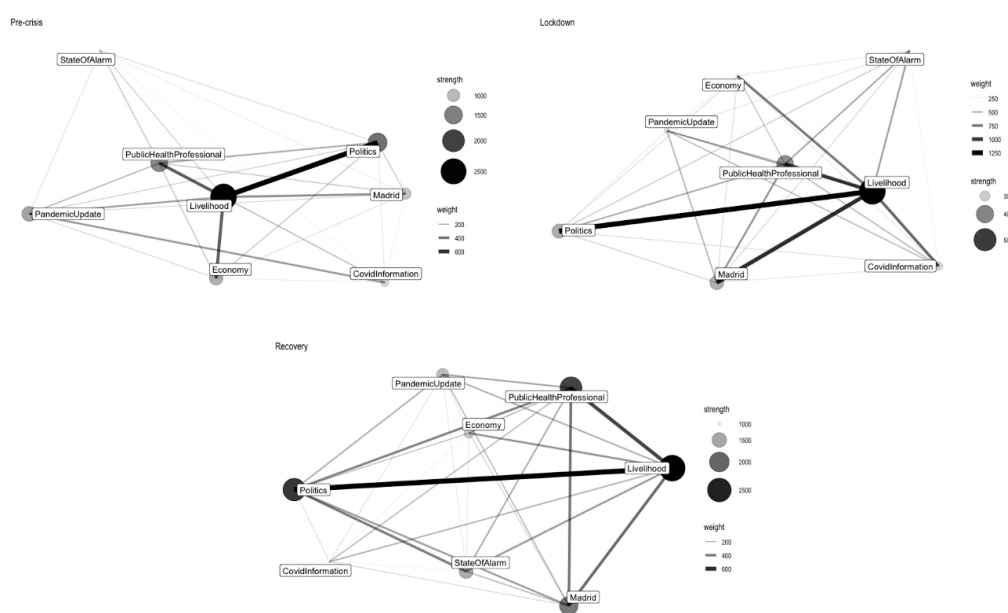


Figure 1. El País news frame network (from top to bottom and from left to right—pre-crisis period, lockdown period, recovery period).

Table 2. Detailed information of the (El País) news frame in each of the period (S: node strength. N: number of tweets. P: proportion.).

News Frames	Pre-Crisis Period			Lockdown Period			Recovery Period		
	S	N	P	S	N	P	S	N	P
Livelihood	2583.42	633	26.3%	5907.56	1625	18.1%	2808.26	618	18.4%
Public Health Professional	1414.49	356	14.8%	3857.47	1320	14.7%	2253.56	554	16.5%
Pandemic Update	1191.59	320	13.3%	2694.64	1226	13.6%	1317.59	368	11.0%
Madrid	915.14	206	8.6%	3436.31	1050	11.7%	1858.97	447	13.3%
Politics	1611.72	306	12.7%	3450.10	945	10.5%	2352.80	470	14.0%
State of Alarm	620.85	185	7.7%	2623.05	1107	12.3%	1452.55	389	11.6%
Economy	1087.55	229	9.5%	2631.96	791	8.8%	1199.97	265	7.9%
Covid Information	744.08	169	7.0%	2826.89	907	10.1%	996.43	249	7.4%

Table 3. Detailed information of the (El País) most weighted edges.

Pre-Crisis Period		Lockdown Period		Recovery Period	
Edge name	Edge weight	Edge name	Edge weight	Edge name	Edge weight
Livelihood–Politics	766.78	Livelihood–Politics	1273.50	Livelihood–Politics	738.83
Livelihood–Economy	468.20	Livelihood–Madrid	1052.63	Livelihood–Public Health Professional	536.00
Livelihood–Public Health Professional	463.99	Livelihood–Public Health Professional	956.40	Livelihood–Madrid	431.43
Livelihood–Madrid	321.80	Livelihood–Covid Information	800.83	Madrid–Public Health Professional	405.91
Covid Information–Pandemic Update	290.59	Livelihood–Economy	728.75	Politics–State of Alarm	386.56

“Livelihood” is the most prominent news frame of *El País* and it shows a strong connection with “Politics,” “Economy” and “Public Health Professional” in the pre-crisis stage, suggesting a close connection with government policy and economic situation. In the next two periods, it started to have a more significant relation with “Madrid.” This is understandable because the Spanish capital suffered the most during the COVID-19 pandemic. According to the actual policy, the Community of Madrid is one of the last regions that stepped into the recovery plan [48] and this can also explain why the proportion of “Madrid” increased across the three time segments.

In addition, we indeed observed a news framing change in different stages of the pandemic outbreak. For example, the “Politics” frame is less reported in the second period while the “State of Alarm” and “Covid Information” frames have been paid higher attention during this stage. It is worth noting that although both frames have connections with others, no connections are observed between these two during the three periods, suggesting they are independent from each other. “State of Alarm” is a policy oriented news frame while “Covid Information” focused more on general sanitary information.

As the crisis is gradually controlled, the pandemic related news frames (“Pandemic Update,” “State of Alarm,” “Public Health Professional” and “Covid Information”) are becoming less prominent in the recovery period. The media interests in general political news (“Politics”) decreased during the most difficult time but soon recovered with the crisis situation becoming stable. Regarding the network, the “Politics” frame has the strongest connection with “Livelihood” during all of the three periods. It also has significant relation with “Public Health Professional” (weight: 236.60) and “Economy” (weight: 154.96) during the pre-crisis period but the two connections have been developing in different trends. While “Politics” and “Public Health Professional” remained connected in the other two periods, the connection between “Politics” and “Economy” turned to be less significant. Instead, the “Politics” frame becomes more connected with “State of Alarm” and “Madrid.”

4.2. *El Mundo*

For the *El Mundo* dataset, a total number of 17,577 tweets are collected from 19 February 2020 to 3 June 2020. After removing retweets, 14,290 original tweets are saved for our in-depth analysis.

Eight news frames are computed, six of which are considered the same as *El País*. They are “Madrid,” “State of Alarm,” “Covid information,” “Economy,” “Pandemic Update,” “Politics.” The two unique *El Mundo* frames are “Lockdown” (news about the confinement) and “Hospital” (news related to hospital, doctor and patient). Table 4 presents the news frames with their most relevant keywords.

Table 4. El Mundo Twitter news frames with most relevant words (translated into English).

Madrid	State of Alarm	Lockdown	Covid Information
madrid	government	confinement	world
pass	sanchez	person	country
common	alarm	doctor	pandemic
phase	pedro	leave	inform
health	ask	social	virus
week	president	secure/insurance	china
de-escalation	announcement	quarantine	port
Economy	Pandemic Update	Hospital	Politics
sanitary	spain	years	minister
crisis	death	hospital	police
million	case	death	iglesias
spanish	covid	patient	pablo
mask	contagion	doctor	investigation
economy	die	resident	press
euro	italy	child	civil

Figure 2 presents the network of the three segmented periods, Table 5 provides the detailed information of the news frames across time and Table 6 presents the detailed information of the most weighted edges. Generally speaking, “Madrid,” “State of Alarm” and “Lockdown” are the three most prominent news frames during the pre-crisis period, along with the crisis becoming more severe, “Covid Information” is paid more attention by the newspaper. And finally these four frames are the most prominent news frames during the recovery period.

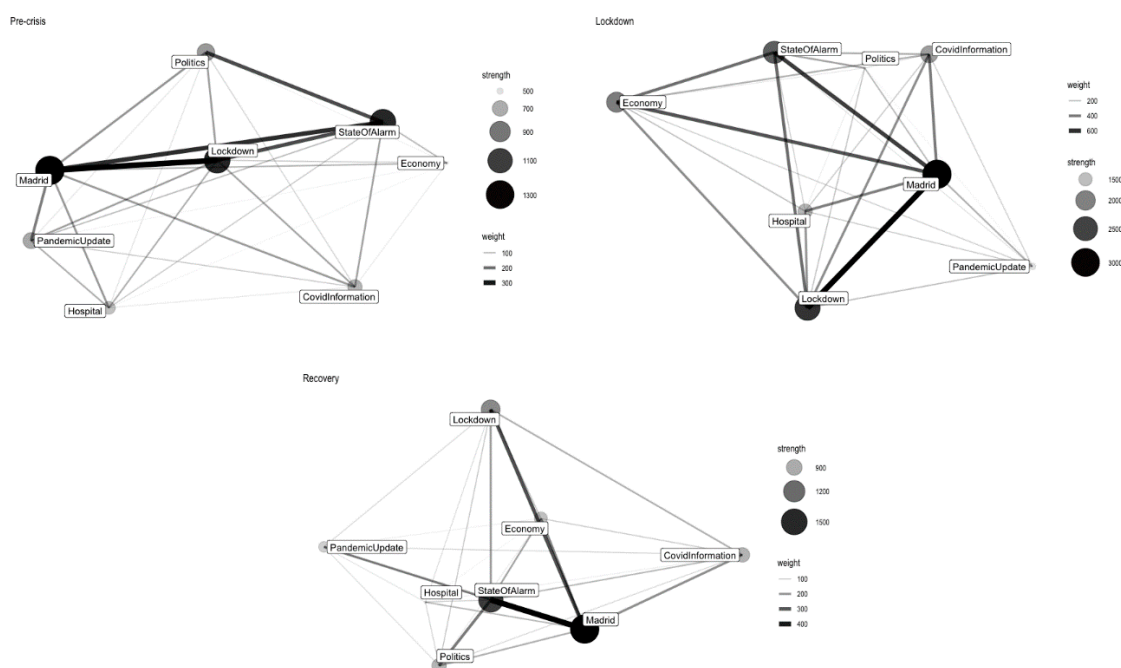


Figure 2. El Mundo news frame network (from top to bottom and from left to right: pre-crisis period, lockdown period, recovery period).

Table 5. Detailed information of the (El Mundo) news frames in each of the period. (S: node strength. N: number of tweets. P: proportion.).

News Frames	Pre-Crisis Period			Lockdown Period			Recovery Period		
	S	N	P	S	N	P	S	N	P
Madrid	1339.36	537	19.4%	3073.28	1389	17.6%	1713.88	694	21.9%
State of Alarm	1174.63	450	16.2%	2274.25	1126	14.2%	1407.71	542	17.1%
Lockdown	1177.45	426	15.4%	2616.23	1210	15.3%	1058.93	391	12.3%
Covid Information	688.13	319	11.5%	1760.66	958	12.1%	856.27	373	11.8%
Economy	482.75	200	7.2%	2039.09	927	11.7%	797.00	299	9.4%
Pandemic Update	714.74	296	10.7%	1212.13	806	10.2%	725.14	316	10.0%
Hospital	619.37	271	9.8%	1562.08	871	11.0%	615.52	245	7.7%
Politics	768.16	271	9.8%	1164.04	622	7.9%	846.89	308	9.7%

Table 6. Detailed information of the (El Mundo) most weighted edges.

Pre-Crisis Period		Lockdown Period		Recovery Period	
Edge name	Edge weight	Edge name	Edge weight	Edge name	Edge weight
Madrid–Lockdown	339.38	Madrid–Lockdown	744.75	Madrid–State of Alarm	455.67
Madrid–State of Alarm	276.40	Madrid–State of Alarm	560.53	Madrid–Lockdown	314.04
Lockdown–State of Alarm	240.89	Madrid–Economy	502.84	Politics–State of Alarm	244.29
Politics–State of Alarm	239.43	Lockdown–State of Alarm	436.85	Madrid–Economy	229.17
Madrid–Pandemic Update	178.53	Economy–State of Alarm	429.68	Madrid–Pandemic Update	214.04

“Madrid” is the most prominent news frame of *El Mundo* of all the time. The proportion of this topic is greatly changed from the second period to the third. As we have explained in the previous section, Madrid is the last region that stepped into recovery plan, so this change is understandable. The “Madrid” frame has the strongest connection strength with “Lockdown” and “State of Alarm” during the first two periods and the association between “Madrid” and “Covid Information” becomes more and more eye-catching during the last two periods. The second most important news frame is “State of Alarm,” it has been paid less attention during the lockdown period but still, shared a significant proportion of the total news coverage. The “State of Alarm” frame has the highest connection strength with “Madrid” and “Lockdown,” similar to “Madrid,” the relation between “State of Alarm” and “Covid Information” is becoming stronger during the second and third time segments (weight in the 2nd period: 259.72, in the 3rd period: 139.42).

As Spain started to get recovered from the strict national lockdown, the proportion of the relevant news frames “Lockdown” and “Hospital” decrease during the recovery period but the connection between these two topics have been strengthened in this stage. As the “Lockdown” frame is highly associated with “Madrid” and “State of Alarm,” we assume this frame is strongly policy orientated. On the other hand, the “Hospital” frame includes both health and social news, so it is naturally associated the most with “Madrid” and “Lockdown.” Regarding the “Economy” frame, the proportion of this topic arrived its peak at the second period. It is significantly different from the frame “Politics,” which has been less adopted during the same period. Both of them have strong ties with “Madrid” and “State of Alarm” but no significant connections have been exposed between these two frames.

Given that the frames “Covid Information” and “Pandemic Update” have almost no proportion changes during the three time periods, these two news frames are considered as stable news frames, tweets about “Pandemic Update” is slightly fewer than “Covid Information.” Regarding the network, like many other *El Mundo* news frames, both of the two have the strongest connection with “Madrid” and the tie between these two frames is getting more and more meaningful over time.

4.3. Comparative Discussion

Significant differences are observed between *El País* (EP) and *El Mundo* (EM) in the frames used in their Twitter news posts. First, the most prominent news frame of the two Spanish newspapers are different. While EP focused on “Livelihood,” EM tended to adopt the “Madrid” frame most frequently. Despite the fact that “Madrid” is also a frame in the EP dataset, it is considered as a peripheral news frame. Both of the two frames have the strongest connections with other topics in the networks, so these two frames can be seen as the motor themes of their newspapers on Twitter.

Second, both of the newspapers have two unique news frames. While the EP news coverage on Twitter focuses on “Livelihood” and “Public Health Professional,” we observed the “Lockdown” and “Hospital” frames in the EM Twitter posts. The “Livelihood” frame is somewhat similar to “Hospital,” because both of the two news frames contain social and living attributes. Nevertheless, their connection strength with the other common frames are different. While “Livelihood” associates the most with “Politics” and “Public Health Professional” in the EP networks, “Hospital” associates the most with “Madrid” and “Lockdown” in the EM networks. A possible interpretation of this difference is “Livelihood” is linked to government (including relevant government departments) policy but “Hospital” is more linked to the news about specific regions. Also, EP shows higher attention to the Ministry of Health and professional perspective by adopting the “Public Health Professional” frame while EM focuses more on the effect of confinement from social perspectives with the “Lockdown” frame.

Third, although there are six common news frames identified in the Twitter posts of both newspapers but the longitudinal changes in their proportion over time are different. For example, the “Economy” related news tweets are increasingly scarce over time in the EP dataset but for EM, such information is more posted during the second time period (the lockdown period). Another significant example can be seen from the “Politics” frame. The EP Twitter account posted more politics-related news during the recovery period than during the lockdown period. But for EM, the increasing trend during the same periods is not so salient as EP.

“State of Alarm” is the second most important news frame for EM on Twitter but this frame is not so prominent in EP Twitter posts. Although the most relevant keywords of this frame in the two datasets are almost the same but the connections are different in the networks. During the first two periods, “State of Alarm” is considered most associated with “Lockdown” and “Madrid” in the EM network, while it is mostly linked to “Livelihood” and “Public Health Professional” in the EP network. During the recovery period, the link between “State of Alarm” and “Politics” is strengthened in EP network, while the connection between “State of Alarm” and “Covid Information” is more eye-catching in the EM network. This finding implies that, with the pandemic crisis getting under control, Twitter posts about “State of Alarm” is more related to political news on EP but connected to health news more closely in the EM Twitter coverage.

5. Conclusions

This study analyzed and compared the frames of Twitter news posts in the two most important Spanish newspapers during Covid-19 pandemic crisis. With a combination of topic modeling and network analysis method, a general landscape of the news coverage of the two newspapers has been illustrated. We found that the center-left media focused the most on family life and living issues (“Livelihood”), while the center-right media focused the most on the Spanish capital news (“Madrid”). From the distribution and proportion of news frames, it can be concluded that *El País* focused the most on public health professionals and real-time alarming (“Pandemic Update”) information during the first two periods. The *El Mundo* coverage on Twitter focused on the state of alarm and confinement (“Lockdown”) related information. During the recovery period, the proportion of general political news (“Politics”) update is largely increased in *El País*, being the third most prominent news frame in this stage. Nevertheless, no such changes are observed in the results of *El Mundo*. Our results are consistent with the thesis proposed by Shih et al. [21] that media coverage about epidemics did

change across different phrases of the crises. Given our limited data collection timespan and the unique characteristics of Twitter data, a more comprehensive analysis is needed for future studies.

From the methodological approach, our method combination provides a dynamic overview of news frames' evolution over time. The weighted node degree and the most weighted edges in each of the stages have been reported. Each of the motor themes ("Livelihood" for *El País*, "Madrid" for *El Mundo*) is the leading topics of all of the three time segments. Given the strong connections of the two topics with other frames, we observed a more unbalanced network structure in *El Mundo* dataset. Specifically, a second-level community is identified, which consisted of "Madrid," "Lockdown" and "State of Alarm" in the pre-crisis period. The community is enlarged with "Covid Information" included in the last two periods. It implies that the content of the four news frames have a high degree of co-occurrence, they are relatively more independent from other frames. But the second-level community cannot be clearly observed in the *El País* network, thus, we believe that the news frames of *El Mundo* is more centralized than *El País*.

Finally, several limitations of our study should be mentioned. First, previous literature has indicated that Twitter based short-text news updates are different from their full length articles [5]. In this case, it is worth noting that our results are solely based on the Twitter posts, which may not be generalized to the comparison between the contents of the two newspapers' articles. Second, as the news coverage may less focus on the health issue in the pre-crisis period than in later stages and our adopted topic modeling method is highly depended on the vast dataset, the number of tweets in the pre-crisis period is much less than the two other periods, news frames on the first period may not be perfectly classified. Finally, although we have analyzed the two most important Spanish newspapers with different political stances, the number of research objects are still limited and we would like to include more newspapers and use a larger dataset as our improvement strategies for the future.

Author Contributions: Conceptualization, J.Y., Y.L.; methodology, J.Y.; software, J.Y.; validation, J.Y., Y.L. and J.M.-J.; formal analysis, J.Y., Y.L.; investigation, J.Y.; resources, J.Y.; data curation, J.Y.; writing—original draft preparation, J.Y.; writing—review and editing, Y.L., J.M.-J.; visualization, J.Y., J.M.-J.; supervision, J.M.-J.; funding acquisition, J.M.-J. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was partially funded by the Department of Social Psychology, Universitat Autònoma de Barcelona.

Acknowledgments: This work belongs to the framework of the doctoral programme in Person and Society in the Contemporary World of the Autonomous University of Barcelona.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. WHO. WHO Director-General's Opening Remarks at the Media Briefing on COVID-19-11 March 2020. Available online: <https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020> (accessed on 30 March 2020).
2. Masip, P.; Aran-Ramspott, S.; Ruiz-Caballero, C.; Suau, J.; Almenar, E.; Puertas-Graell, D. Consumo informativo y cobertura mediática durante el confinamiento por el Covid-19: Sobreinformación, sesgo ideológico y sensacionalismo. *El Prof. La Inf.* **2020**, *29*. [CrossRef]
3. Lee, A.Y.L. Between Global and Local: The Glocalization of Online News Coverage on the Trans-regional Crisis of SARS. *Asian J. Commun.* **2005**, *15*, 255–273. [CrossRef]
4. Nguyen, A.; Western, M. The complementary relationship between the Internet and traditional mass media: The case of online news and information. *Inf. Res.* **2006**, *11*, 151–183.
5. Zhang, X.; Bie, B.; Billings, A.C. Newspaper Ebola articles differ from Twitter updates. *Newsp. Res. J.* **2017**, *38*, 497–511. [CrossRef]
6. Guo, L. The Application of Social Network Analysis in Agenda Setting Research: A Methodological Exploration. *J. Broadcast. Electron. Media* **2012**, *56*, 616–631. [CrossRef]
7. Baumgartner, F.R.; Chaqués Bonafont, L. All news is bad news: Newspaper coverage of political parties in Spain. *Polit. Commun.* **2015**, *32*, 268–291. [CrossRef]

8. Widlak, E.; i Pont Sorribes, C.; i Guillaumet Lloveras, J. The Media portrayal of Queen Sofia of Greece in Spanish newspapers. Analysis of the press coverage of Queen Sofia in El País and El Mundo, May 2011 to August 2012. *Comun. Rev. Recer. i D'anàlisi* **2016**, *33*, 75–92.
9. Hallin, D.C.; Papathanassopoulos, S. Political clientelism and the media: Southern Europe and Latin America in comparative perspective. *Media Cult. Soc.* **2002**, *24*, 175–195. [[CrossRef](#)]
10. Catalan-Matamoros, D.; Do Nascimento, B.G.; Langbecker, A. Visual content published by the press during a health crisis: The case of ebola, spain, 2014. *Interface Commun. Health Educ.* **2020**, *24*, e190271. [[CrossRef](#)]
11. Barberá-González, R.; Cambra-Cuesta, U. El virus del ébola: Análisis de su comunicación de crisis en España. *Opción Rev. Ciencias Humanas y Soc.* **2015**, *31*, 67–86.
12. Ballester, M.C.C.; Villafranca, P.L. The impact of the ebola virus and rare diseases in the media and the perception of risk in Spain. *Catalan J. Commun. Cult. Stud.* **2016**, *8*, 245–263. [[CrossRef](#)]
13. Guo, L.; Chen, Y.-N.K.; Vu, H.; Wang, Q.; Aksamit, R.; Guzek, D.; Jachimowski, M.; McCombs, M. Coverage of the Iraq War in the United States, Mainland China, Taiwan and Poland. *J. Stud.* **2015**, *16*, 343–362. [[CrossRef](#)]
14. Guo, L.; Mays, K.; Wang, J. Whose Story Wins on Twitter? Visualizing the South China Sea Dispute. *J. Stud.* **2019**, *20*, 563–584. [[CrossRef](#)]
15. Scheufele, D.A.; Tewksbury, D. Framing, Agenda Setting, and Priming: The Evolution of Three Media Effects Models. *J. Commun.* **2007**, *57*, 9–20. [[CrossRef](#)]
16. Entman, R.M. Framing: Toward Clarification of a Fractured Paradigm. *J. Commun.* **1993**, *43*, 51–58. [[CrossRef](#)]
17. Price, V.; Tewksbury, D.; Powers, E. Switching Trains of Thought: The Impact of News Frames on Readers' Cognitive Responses. *Communic. Res.* **1997**, *24*, 481–506. [[CrossRef](#)]
18. Lee, S.T.; Basnyat, I. From Press Release to News: Mapping the Framing of the 2009 H1N1 A Influenza Pandemic. *Health Commun.* **2013**, *28*, 119–132. [[CrossRef](#)]
19. Berry, T.R.; Wharf-Higgins, J.; Naylor, P.J. SARS Wars: An Examination of the Quantity and Construction of Health Information in the News Media. *Health Commun.* **2007**, *21*, 35–44. [[CrossRef](#)]
20. Basnyat, I.; Lee, S.T. Framing of Influenza A (H1N1) pandemic in a Singaporean newspaper. *Health Promot. Int.* **2015**, *30*, 942–953. [[CrossRef](#)]
21. Shih, T.-J.; Wijaya, R.; Brossard, D. Media Coverage of Public Health Epidemics: Linking Framing and Issue Attention Cycle Toward an Integrated Theory of Print News Coverage of Epidemics. *Mass Commun. Soc.* **2008**, *11*, 141–160. [[CrossRef](#)]
22. Santos-Diez, M.T.; Camacho-Markina, I. Treatment of cannabis in the Spanish press. *Cuadernos. Info* **2017**, 153–171. [[CrossRef](#)]
23. Lázaro-Rodríguez, P.; Herrera-Viedma, E. Noticias sobre Covid-19 y 2019-nCoV en medios de comunicación de España: El papel de los medios digitales en tiempos de confinamiento. *El Prof. La Inf.* **2020**, *29*. [[CrossRef](#)]
24. Guo, L.; Vargo, C.J.; Pan, Z.; Ding, W.; Ishwar, P. Big Social Data Analytics in Journalism and Mass Communication. *J. Mass Commun. Q.* **2016**, *93*, 332–359. [[CrossRef](#)]
25. Hong, L.; Davison, B.D. Empirical study of topic modeling in Twitter. In Proceedings of the First Workshop on Social Media Analytics, Washington, DC, USA, 25 July 2010; ACM Press: New York, NY, USA, 2010; pp. 80–88.
26. Ramage, D.; Dumais, S.; Liebling, D. Characterizing Microblogs with Topic Models. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, Washington, DC, USA, 23–26 May 2010; AAAI Press: Menlo Park, CA, USA, 2010; pp. 130–137.
27. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
28. Daud, A.; Li, J.; Zhou, L.; Muhammad, F. Knowledge discovery through directed probabilistic topic models: A survey. *Front. Comput. Sci. China* **2010**, *4*, 280–301. [[CrossRef](#)]
29. Heidenreich, T.; Lind, F.; Eberl, J.-M.; Boomgaarden, H.G. Media Framing Dynamics of the 'European Refugee Crisis': A Comparative Topic Modelling Approach. *J. Refug. Stud.* **2019**, *32*, 172–182. [[CrossRef](#)]
30. Poirier, W.; Ouellet, C.; Rancourt, M.A.; Béchar, J.; Dufresne, Y. (Un)covering the COVID-19 pandemic: Framing analysis of the crisis in Canada. *Can. J. Polit. Sci.* **2020**, 1–7. [[CrossRef](#)]
31. Lim, S.; Berry, F.S.; Lee, K.-H. Stakeholders in the Same Bed with Different Dreams: Semantic Network Analysis of Issue Interpretation in Risk Policy Related to Mad Cow Disease. *J. Public Adm. Res. Theory* **2016**, *26*, 79–93. [[CrossRef](#)]
32. Fu, J.S.; Zhang, R. NGOs' HIV/AIDS Discourse on Social Media and Websites: Technology Affordances and Strategic Communication Across Media Platforms. *Int. J. Commun.* **2019**, *13*, 181–205.

33. Kang, G.J.; Ewing-Nelson, S.R.; Mackey, L.; Schlitt, J.T.; Marathe, A.; Abbas, K.M.; Swarup, S. Semantic network analysis of vaccine sentiment in online social media. *Vaccine* **2017**, *35*, 3621–3638. [[CrossRef](#)] [[PubMed](#)]
34. Bail, C.A. Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 11823–11828. [[CrossRef](#)] [[PubMed](#)]
35. Boy, J.; Foote, J.; pyup.io bot. jboynyc/textnets: Textnets version 0.4.9 (Version V0.4.9). *Zenodo* **2020**. [[CrossRef](#)]
36. Rajaraman, A.; Ullman, J.D. Data Mining. In *Mining of Massive Datasets*; Cambridge University Press: Cambridge, UK, 2011; pp. 1–17. ISBN 9781107737419.
37. Yu, J. Open Access Institutional and News Media Tweet Dataset for COVID-19 Social Science Research. *arXiv* **2020**. Available online: <https://arxiv.org/abs/2004.01791> (accessed on 3 May 2020).
38. Nothman, J.; Qin, H.; Yurchak, R. Stop Word Lists in Free Open-source Software Packages. In Proceedings of the Workshop for NLP Open Source Software (NLP-OSS), Melbourne, Vic, Australia, 20 July 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 7–12.
39. Grün, B.; Hornik, K. Topicmodels: An r package for fitting topic models. *J. Stat. Softw.* **2011**, *40*, 1–30. [[CrossRef](#)]
40. Steinskog, A.; Therkelsen, J.; Gambäck, B. Twitter Topic Modeling by Tweet Aggregation. In Proceedings of the the 21st Nordic Conference on Computational Linguistics, Gothenburg, Sweden, 22–24 May 2017; Association for Computational Linguistics: Gothenburg, Sweden, 2017; pp. 77–86.
41. Phan, X.-H.; Nguyen, C.-T. GibbsLDA++: A C/C++ Implementation of Latent Dirichlet Allocation (LDA) 2007. Available online: <http://gibbslda.sourceforge.net/> (accessed on 20 July 2020).
42. Poblete, B.; Guzman, J.; Maldonado, J.; Tobar, F. Robust Detection of Extreme Events Using Twitter: Worldwide Earthquake Monitoring. *IEEE Trans. Multimed.* **2018**, *20*, 2551–2561. [[CrossRef](#)]
43. Xiang, G.; Fan, B.; Wang, L.; Hong, J.; Rose, C. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In Proceedings of the ACM International Conference Proceeding Series, Maui, HA, USA, 29 October–2 November 2012; ACM Press: New York, NY, USA, 2012; pp. 1980–1984.
44. Ahmed, S.; Jaidka, K.; Cho, J. The 2014 Indian elections on Twitter: A comparison of campaign strategies of political parties. *Telemat. Inform.* **2016**, *33*, 1071–1087. [[CrossRef](#)]
45. van Ravenzwaaij, D.; Cassey, P.; Brown, S.D. A simple introduction to Markov Chain Monte-Carlo sampling. *Psychon. Bull. Rev.* **2018**, *25*, 143–154. [[CrossRef](#)]
46. Pan, P.-L.; Meng, J. Media Frames across Stages of Health Crisis: A Crisis Management Approach to News Coverage of Flu Pandemic. *J. Contingencies Cris. Manag.* **2016**, *24*, 95–106. [[CrossRef](#)]
47. Barrat, A.; Barthélemy, M.; Pastor-Satorras, R.; Vespignani, A. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 3747–3752. [[CrossRef](#)]
48. Viejo, M. Desescalada: Así llega Madrid a la fase 1: Paro disparado, sin turistas y con miles de ciudadanos en las colas del hambre. *El País*. 25 May 2020. Available online: <https://elpais.com/espana/madrid/2020-05-23/asi-llega-madrid-a-la-fase-1-paro-disparado-sin-turistas-y-con-miles-de-ciudadanos-en-las-colas-del-hambre.html>. (accessed on 14 June 2020).



Chapter 5

Conclusion

In this PhD thesis, we have proposed four specific research objectives, and three corresponded academic publications are included (the first two specific research objectives are carried out in the bibliometric study). It has been 14 years since the foundation of this microblogging service, a short time in the history of social science, but large source to amplify the way to do innovative researches. Throughout the first publication, we have witnessed an explosion of Twitter related scientific publications in the last decade, which have included many different research fields, for example, communication science, computer science, behavior science etc. The easy-access and timely short-text information have indeed provided researchers opportunities to study social phenomenon and to create new research methods drawn on big data. It is a fact that the scientific production of 2019 has turned down, as argued, Twitter study may have surpassed its advanced period, and stepped into a new stage, or, by a pessimistic meaning, maybe the "golden age" of Twitter research is gone.

The Twitter study has a high degree of international collaboration, two large clusters have been visualized in the publication, Asian-Pacific-North-America cluster, and the European cluster, it means Asian-Pacific countries collaborated the most with North American countries, and European countries prefer to collaborate internally. From another perspective, European countries and English speaking countries have a relatively high international collaboration degree. Since there are two giant collaboration groups in Twitter related-studies, the research difference between the European countries and Asian-Pacific-North-American countries remained uncertain, a more detailed research is required.

Twitter studies are highly event based, the Twitter research timeline is closely related to real world timeline. And the most important research topics are business, communication, disaster management, scientometrics and computer science. The popularity of these research topics are not unchanged. business-related research lines took an important place in the initial period (2006-2012), disaster management is one of the main research topics in the developing period (2013-2016), and scientometrics showed its representativeness in the advanced period (2017-2020). For communication and computer science, they are the predominant research fields during all the three periods of Twitter studies.

Compared to other similar scientific literatures, two remarkable improvements have been realized by this paper. First, we have largely expanded the research data, as having been reviewed before, the largest data sample adopted in Twitter study bibliometric analysis were 4709, and in our case, a total number of 19205 academic publications were analyzed. Second, a temporal evolution of the main research themes of Twitter related studies have been clearly visualized and analyzed, which, in case of other publications, was never done before.

In the second publication, we have analyzed 9 different free and low-cost Twitter research software tools, and categorized them into windows tools (Mozdeh, Webometric Analyst, NodeXL), multi-platform tools (COSMOS, Gephi, DMI-TCAT) and web-based tools (TAGS, Netlytic, SocioViz). We did so because we believe that researchers shall have the freedom to pick the most adequate software according to their operation system. A general Twitter research guideline is provided based on the comparison of their functionalities (regarding data collection and data analysis). We insist that there is no best Twitter research software, but the most suitable ones for different research purposes. For big data collection, we recommended COSMOS and Gephi Twitter Streaming Importer plugin as the data collection tool, because the data retrieval quantity of these two tools is unlimited. We have summarized nine different data analysis techniques according to the different characteristics of the nine tools, an easy-to-check table is provided inside the publication. For example, for social network analysis we recommended NodeXL and Gephi, because these two software are exclusively designed for social networks. Mozdeh, COSMOS and DMI-TCAT are the three tools that can conduct Twitter analysis from gender perspective, and DMI-TCAT is the only that can do source analysis. For Geolocation analysis, COSMOS and Netlytic are the only available tools.

However, it is worthwhile to mention that the software included in this study are "integrated frameworks", which means they allow both data access, data exploration, filtering and analysis (Antonakaki et al., 2020), they are excellent tools for junior social science researchers, because they don't require programming skills, which might be relatively easier to get started with. But they are not omnipotent, because they lack the flexibility to be adjusted into research purposes other than their functionalities. And not all the analyzed tools are free and/or open source software, which may take the research procedure into "black box", more uncertainty may be revealed regarding data analysis (Chan et al., 2020; Dienlin et al., 2020). Hence, researchers are strongly recommended to master open source programming languages (e.g. R, Python, Julia etc), not only because of their incomparable flexibility, but also for the purpose of open science.

Our third research article explored the evolution of Spanish news frames during the ongoing COVID-19 pandemic, by focusing on the Twitter news update of the two major Spanish newspapers (El País and El Mundo), our study divided their twitter information into three periods (pre-crisis period, lockdown period and recovery period) according to the pandemic timeline in Spain. The news frames were identified by topic modeling method (LDA), and the relations between the news frames are visualized by a tweet-term co-occurrence network. The research results are twofold: first, regarding the difference between the two newspapers' Twitter posts, El País (center-left media) focused more on family life and living issues (the "Livelihood" frame), and El Mundo (center-right media) tended to adopt the "Madrid" frame (Spanish capital news) most frequently. These two frames are considered as the prominent frames for these two media, and they are the most associated frames on the co-occurrence network. Second, the longitudinal changes of the frame proportion over time are different for both of the media Twitter posts. For example, the "Economy" frame related news are increasingly scarce over time in the El País tweet post, but for El Mundo, such frame is most adopted in the lockdown period. El País posted more politics-related news (the "Politics" frame) during the third period than the second period, but for the case of El Mundo, the increasing trend during the same periods is not so significant.

At the level of the methodology, we combined the topic modeling method with network analysis techniques, which have provided a dynamic overview of news

frames' evolution over time. We observed a more unbalanced network structure in El Mundo news frame network, in which a second-level community is identified, but such phenomenon cannot be clearly observed on El País network, it means the news frames adopted by El Mundo is more centralized than El País. By the time of writing this thesis, Spain is suffering from the second-wave of the pandemic outbreak, and our study solely focus on the "first-wave". In order to have a more comprehensive understanding of the news frames' evolution, it might be a good idea to adopt a larger dataset for an updating research.

Overall, the three included publications in this thesis are all closely linked to computational methods, and I have explained the current environment of how people study and use Twitter. For the general objective of "how people study", we did a bibliometric analysis and software review, to synthesize the current Twitter research environment from the level of the global scientific production and the level of general research methods/tools. For the general objective of "how people use", we followed the real-world timeline, and put our focus on framing study regarding COVID-19, but our research was not exactly for the explanation of "how people use" but "how news media use", it may be considered as an important limitation of this PhD thesis. But I believe that the proposed specific research objectives have all been successfully fulfilled by these scientific works.

Bibliography

- Ahmed, W., Bath, P. A., & Demartini, G. (2017). Using twitter as a data source: An overview of ethical, legal, and methodological challenges. In *The ethics of online research*. Emerald Publishing Limited.
- Ahmed, W., Vidal-Alaball, J., Downing, J., & Seguí, F. L. (2020). Covid-19 and the 5g conspiracy theory: Social network analysis of twitter data. *Journal of Medical Internet Research*, 22(5), e19458.
- Alves, H., Fernandes, C., & Raposo, M. (2016). Social media marketing: A literature review and implications. *Psychology & Marketing*, 33(12), 1029–1038.
- Antonakaki, D., Fragopoulou, P., & Ioannidis, S. (2020). A survey of twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert Systems with Applications*, 114006.
- Bajaj, S. G. (2017). The use of twitter during the 2014 indian general elections: Framing, agenda-setting, and the personalization of politics. *Asian Survey*, 57(2), 249–270.
- Barnard, S. R. (2018). Tweeting# ferguson: Mediatized fields and the new activist journalist. *New Media & Society*, 20(7), 2252–2271.
- Basch, C. H., Kecojevic, A., & Wagner, V. H. (2020). Coverage of the covid-19 pandemic in the online versions of highly circulated us daily newspapers. *Journal of community health*, 1–9.
- Belli, S., Mugnaini, R., Baltà, J., & Abadal, E. (2020). Coronavirus mapping in scientific publications: When science advances rapidly and collectively, is access to this knowledge open to society? *Scientometrics*, 124(3), 2661–2685. doi:10.1007/s11192-020-03590-7
- Blank, G. (2017). The digital divide among twitter users and its implications for social research. *Social Science Computer Review*, 35(6), 679–697.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993–1022.
- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.
- Boulianne, S. (2015). Social media use and participation: A meta-analysis of current research. *Information, communication & society*, 18(5), 524–538.
- Buettner, R., & Buettner, K. (2016). A systematic literature review of twitter research from a socio-political revolution perspective. In *2016 49th hawaii international conference on system sciences (hicc)* (pp. 2206–2215). IEEE.
- Burnap, P., Gibson, R., Sloan, L., Southern, R., & Williams, M. (2016). 140 characters to victory?: Using twitter to predict the uk 2015 general election. *Electoral Studies*, 41, 230–233.
- Burscher, B., Vliegthart, R., & Vreese, C. H. d. (2016). Frames beyond words: Applying cluster and sentiment analysis to news coverage of the nuclear power issue. *Social Science Computer Review*, 34(5), 530–545.
- Carter Olson, C. (2016). #bringbackourgirls: Digital communities supporting real-world change and influencing mainstream media agendas. *Feminist Media Studies*, 16(5), 772–787.

- Castillo-Esparcia, A., Fernández-Souto, A.-B., & Puentes-Rivera, I. (2020). Political communication and covid-19: Strategies of the government of Spain. *Profesional de la información*, 29(4).
- Chahrour, M., Assi, S., Bejjani, M., Nasrallah, A. A., Salhab, H., Fares, M., & Khachfe, H. H. (2020). A bibliometric analysis of covid-19 research activity: A call for increased output. *Cureus*, 12(3).
- Chan, C.-H., Zeng, J., Wessler, H., Jungblut, M., Welbers, K., Bajjalieh, J. W., ... Althaus, S. L. (2020). Reproducible extraction of cross-lingual topics (rectr). *Communication Methods and Measures*, 1–21.
- Charles-Smith, L. E., Reynolds, T. L., Cameron, M. A., Conway, M., Lau, E. H., Olsen, J. M., ... Suda, K. J., et al. (2015). Using social media for actionable disease surveillance and outbreak management: A systematic literature review. *PloS one*, 10(10), e0139701.
- Chew, C., & Eysenbach, G. (2010). Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. *PloS one*, 5(11), e14118.
- Christensen, C. (2011). Twitter revolutions? addressing social media and dissent. *The Communication Review*, 14(3), 155–157.
- Das, S., & Dutta, A. (2020). Characterizing public emotions and sentiments in covid-19 environment: A case study of India. *Journal of Human Behavior in the Social Environment*, 1–14.
- Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., ... Johnson, B. K., et al. (2020). An agenda for open science in communication. *Journal of Communication*.
- Domahidi, E., Yang, J., Niemann-Lenz, J., & Reinecke, L. (2019). Outlining the way ahead in computational communication science: An introduction to the ijoc special section on "computational methods for communication science: Toward a strategic roadmap". *International Journal of Communication* (19328036), 13.
- Egger, M., Smith, G. D., & Phillips, A. N. (1997). Meta-analysis: Principles and procedures. *Bmj*, 315(7121), 1533–1537.
- Enli, G. (2017). Twitter as arena for the authentic outsider: Exploring the social media campaigns of trump and Clinton in the 2016 US presidential election. *European journal of communication*, 32(1), 50–61.
- Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4), 51–58.
- Filo, K., Lock, D., & Karg, A. (2015). Sport and social media research: A review. *Sport management review*, 18(2), 166–181.
- Francia, P. L. (2018). Free media and twitter in the 2016 presidential election: The unconventional campaign of Donald Trump. *Social Science Computer Review*, 36(4), 440–455.
- Fung, I. C.-H., Tse, Z. T. H., Cheung, C.-N., Miu, A. S., & Fu, K.-W. (2014). Ebola and the social media. *The Lancet*, 384(9961), 2207. doi:10.1016/S0140-6736(14)62418-1
- Galvagno, M. (2017). Bibliometric literature review: An opportunity for marketing scholars. *Mercati & Competitività*.
- Goff, D. A., Kullar, R., Laxminarayan, R., Mendelson, M., Nathwani, D., & Osterholm, M. (2019). Twitter to engage, educate, and advocate for global antibiotic stewardship and antimicrobial resistance. *The Lancet Infectious Diseases*, 19(3), 229–231.
- Guidry, J. P., Jin, Y., Orr, C. A., Messner, M., & Meganck, S. (2017). Ebola on Instagram and Twitter: How health organizations address the health crisis in their social media engagement. *Public relations review*, 43(3), 477–486.

- Gupta, B., Kumar, A., Gupta, R., & Dhawan, S. (2016). A bibliometric assessment of global literature on "twitter" during 2008-15. *International Journal of Information Dissemination & Technology*, 6(3).
- Gutiérrez-Salcedo, M., Martínez, M. Á., Moral-Muñoz, J. A., Herrera-Viedma, E., & Cobo, M. J. (2018). Some bibliometric procedures for analyzing and evaluating research fields. *Applied intelligence*, 48(5), 1275–1287.
- Hand, L. C., & Ching, B. D. (2011). "you have one friend request" an exploration of power and citizen engagement in local governments' use of social media. *Administrative Theory & Praxis*, 33(3), 362–382.
- INE. (2019). Porcentaje de usuarios de internet en los últimos tres meses por tipo de actividad realizada y sexo. 2019. Retrieved from https://www.ine.es/jaxi/Datos.htm?path=/t00/mujeres_hombres/tablas_1/10/&file=c04003.px#!tabla
- Isa, D., & Himelboim, I. (2018). A social networks approach to online social movement: Social mediators and mediated content in# freeajstaff twitter network. *Social Media+ Society*, 4(1), 2056305118760807.
- Jin, S.-A. A., & Phua, J. (2014). Following celebrities' tweets about brands: The impact of twitter-based electronic word-of-mouth on consumers' source credibility perception, buying intention, and social identification with celebrities. *Journal of advertising*, 43(2), 181–195.
- Jungherr, A. (2016). Twitter use in election campaigns: A systematic literature review. *Journal of information technology & politics*, 13(1), 72–91.
- Kang, B., & Lee, J. Y. (2014). A bibliometric analysis on twitter research. *Journal of the Korean Society for information Management*, 31(3), 293–311.
- Kharroub, T., & Bas, O. (2016). Social media and protests: An examination of twitter images of the 2011 egyptian revolution. *New Media & Society*, 18(9), 1973–1992.
- Kilgo, D. K., Yoo, J., & Johnson, T. J. (2018). Spreading ebola panic: Newspaper and social media coverage of the 2014 ebola health crisis. *Health communication*.
- Kruikemeier, S. (2014). How political candidates use twitter and the impact on votes. *Computers in human behavior*, 34, 131–139.
- Kruspe, A., Häberle, M., Kuhn, I., & Zhu, X. X. (2020). Cross-language sentiment analysis of european twitter messages duringthe covid-19 pandemic. *arXiv preprint arXiv:2008.12172*.
- Lancet, T. (2014). The medium and the message of ebola. *The Lancet*, 384(9955), 1641. doi:10.1016/S0140-6736(14)62016-X
- Lázaro-Rodríguez, P., & Herrera-Viedma, E. (2020). Noticias sobre covid-19 y 2019-ncov en medios de comunicación de españa: El papel de los medios digitales en tiempos de confinamiento. *El profesional de la información (EPI)*, 29(3).
- Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., ... Van Alstyne, M. (2009). Computational social science. *Science*, 323(5915), 721–723. doi:10.1126/science.1167742. eprint: <https://science.sciencemag.org/content/323/5915/721.full.pdf>
- Li, P., Cho, H., Qin, Y., & Chen, A. (2020). #metoo as a connective movement: Examining the frames adopted in the anti-sexual harassment movement in china. *Social Science Computer Review*, 0894439320956790.
- Liu, B. F., & Kim, S. (2011). How organizations framed the 2009 h1n1 pandemic via social and traditional media: Implications for us health communicators. *Public Relations Review*, 37(3), 233–244.
- López-Meri, A., Marcos-García, S., & Casero-Ripollés, A. (2017). What do politicians do on twitter? functions and communication strategies in the spanish electoral campaign of 2016. *El profesional de la información*, 26(5), 795–804.

- Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., et al. (2011). The arab spring | the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International journal of communication*, 5, 31.
- Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., ... Häussler, T., et al. (2018). Applying lda topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2-3), 93–118.
- Mansfield, S. J., Bonning, M. A., Morrison, S. G., Stephens, H. O., Wang, S.-H., Perry, A. W., & Olver, R. C. (2011). Social networking and health. *The Lancet*, 377(9783), 2083.
- Masip, P., Aran-Ramspott, S., Ruiz-Caballero, C., Suau, J., Almenar, E., & Puertas-Graell, D. (2020). Consumo informativo y cobertura mediática durante el confinamiento por el covid-19: Sobreinformación, sesgo ideológico y sensacionalismo. *El profesional de la información (EPI)*, 29(3).
- Michael, Z., & John, P. N. (2014). A topology of twitter research: Disciplines, methods, and ethics. *Aslib Journal of Information Management*, 66(3), 250–261. doi:10.1108/AJIM-09-2013-0083
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Group, P., et al. (2009). Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *PLoS med*, 6(7), e1000097.
- Moreno, Á., Fuentes-Lara, C., & Navarro, C. (2020). Covid-19 communication management in spain: Exploring the effect of information-seeking behavior and message reception in public's evaluation. *El profesional de la información (EPI)*, 29(4).
- Motta, G., & Baden, C. (2013). Evolutionary factor analysis of the dynamics of frames: Introducing a method for analyzing high-dimensional semantic data with time-changing structure. *Communication Methods and Measures*, 7(1), 48–82.
- Murthy, D. (2012). Towards a sociological understanding of social media: Theorizing twitter. *Sociology*, 46(6), 1059–1073.
- Nicholls, T., & Culpepper, P. D. (2020). Computational identification of media frames: Strengths, weaknesses, and opportunities. *Political Communication*, 1–23.
- Park, H. W., Park, S., & Chong, M. (2020). Conversations and medical news frames on twitter: Infodemiological study on covid-19 in south korea. *Journal of Medical Internet Research*, 22(5), e18897.
- Park, S. “, Kim, H. J., & Ok, C. “ (2018). Linking emotion and place on twitter at disneyland. *Journal of Travel & Tourism Marketing*, 35(5), 664–677.
- Perrin, A., & Anderson, M. (2019). Share of us adults using social media, including facebook, is mostly unchanged since 2018. *Pew Research Center*, 10.
- Phua, J., Jin, S. V., & Kim, J. J. (2017). Uses and gratifications of social networking sites for bridging and bonding social capital: A comparison of facebook, twitter, instagram, and snapchat. *Computers in human behavior*, 72, 115–122.
- Poell, T., & Rajagopalan, S. (2015). Connecting activists and journalists: Twitter communication in the aftermath of the 2012 delhi rape. *Journalism Studies*, 16(5), 719–733.
- Poirier, W., Ouellet, C., Rancourt, M.-A., Béchar, J., & Dufresne, Y. (2020). (un) covering the covid-19 pandemic: Framing analysis of the crisis in canada. *Canadian Journal of Political Science/Revue canadienne de science politique*, 1–7.
- Putnam, R. D. et al. (2000). *Bowling alone: The collapse and revival of american community*. Simon and schuster.
- Qualman, E. (2012). *Socialnomics: How social media transforms the way we live and do business*. John Wiley & Sons.

- Rainie, L., & Smith, A. (2012). Social networking sites and politics. *Washington, DC: Pew Internet & American Life Project*. Retrieved June, 12, 2012.
- Rehm, J. (2018). Ten years after the economic crash, r&d funding is better than ever. *Nature*, September, 13.
- Ridley, D. (2012). *The literature review: A step-by-step guide for students*. Sage.
- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., . . . Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4), 1064–1082.
- Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of communication*, 49(1), 103–122.
- Schmidt, F. (2008). Meta-analysis: A constantly evolving research integration tool. *Organizational Research Methods*, 11(1), 96–113.
- Shuja, J., Alanazi, E., Alasmay, W., & Alashaikh, A. (2020). Covid-19 open source data sets: A comprehensive survey. *medRxiv*. doi:10.1101/2020.05.19.20107532. eprint: <https://www.medrxiv.org/content/early/2020/07/13/2020.05.19.20107532.full.pdf>
- Skoric, M. M., Zhu, Q., Goh, D., & Pang, N. (2016). Social media and citizen engagement: A meta-analytic review. *New Media & Society*, 18(9), 1817–1839.
- Smith, A., & Anderson, M. (2018). Social media use 2018: Demographics and statistics. Pew Research Center. Retrieved from <https://www.pewresearch.org/internet/2018/03/01/social-media-use-in-2018/>
- Tang, M., Liao, H., Wan, Z., Herrera-Viedma, E., & Rosen, M. A. (2018). Ten years of sustainability (2009 to 2018): A bibliometric overview. *Sustainability*, 10(5), 1655.
- Tejedor, S., Cervi, L., Tusa, F., Portales, M., & Zabolina, M. (2020). Information on the covid-19 pandemic in daily newspapers' front pages: Case study of spain and italy. *International Journal of Environmental Research and Public Health*, 17(17), 6330.
- Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election forecasts with twitter: How 140 characters reflect the political landscape. *Social science computer review*, 29(4), 402–418.
- Van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, 12(2-3), 81–92.
- Van Atteveldt, W., Strycharz, J., Trilling, D., & Welbers, K. (2019). Computational communication science | toward open computational communication science: A practical road map for reusable data and code. *International Journal of Communication*, 13, 20.
- Van Raan, A. (2003). The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments. *TATuP-Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis*, 12(1), 20–29.
- Veil, S. R., Buehner, T., & Palenchar, M. J. (2011). A work-in-process literature review: Incorporating social media in risk and crisis communication. *Journal of contingencies and crisis management*, 19(2), 110–122.
- Vicari, S. (2020). Is it all about storytelling? living and learning hereditary cancer on twitter. *New Media & Society*, 1461444820926632.
- Wallace, N. (2020). Eu leaders slash science spending in 1.8 trillion deal. *Science*. doi:10.1126/science.abd8830
- Walter, D., & Ophir, Y. (2019). News frame analysis: An inductive mixed-method computational approach. *Communication Methods and Measures*, 13(4), 248–266.

- Wang, W., & Guo, L. (2018). Framing genetically modified mosquitoes in the online news and twitter: Intermedia frame setting in the issue-attention cycle. *Public Understanding of Science*, 27(8), 937–951.
- Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019). Systematic literature review on the spread of health-related misinformation on social media. *Social Science & Medicine*, 240, 112552.
- Wehner, M. R., Chren, M.-M., Shive, M. L., Resneck, J. S., Pagoto, S., Seidenberg, A. B., & Linos, E. (2014). Twitter: An opportunity for public health campaigns. *The Lancet*, 384(9938), 131–132.
- Weller, K. (2014). What do we get from twitter—and what not? a close look at twitter research in the social sciences. *KO KNOWLEDGE ORGANIZATION*, 41(3), 238–248.
- Williams, D. (2006). On and off the net: Scales for social capital in an online era. *Journal of computer-mediated communication*, 11(2), 593–628.
- Williams, S. A. [Shirley A], Terras, M. M., & Warwick, C. (2013). What do people study when they study twitter? classifying twitter related academic papers. *Journal of Documentation*.
- Williams, S. A. [Shirley Ann], Terras, M., & Warwick, C. (2013). How twitter is studied in the medical professions: A classification of twitter papers indexed in pubmed. *Medicine 2.0*, 2(2).
- Xu, Z., Ellis, L., & Laffidy, M. (2020). News frames and news exposure predicting flu vaccination uptake: Evidence from us newspapers, 2011–2018 using computational methods. *Health Communication*, 1–9.
- Yu, J. (2020). Open access institutional and news media tweet dataset for covid-19 social science research. *arXiv preprint arXiv:2004.01791*.
- Yu, J., Lu, Y., & Muñoz-Justicia, J. (2020). Analyzing spanish news frames on twitter during covid-19—a network study of el país and el mundo. *International Journal of Environmental Research and Public Health*, 17(15), 5414.
- Yu, J., & Muñoz-Justicia, J. (2020). A bibliometric overview of twitter-related studies indexed in web of science. *Future Internet*, 12(5), 91.
- Zhang, X., Bie, B., & Billings, A. C. (2017). Newspaper ebola articles differ from twitter updates. *Newspaper Research Journal*, 38(4), 497–511.
- Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In *European conference on information retrieval* (pp. 338–349). Springer.
- Zhu, Q. (2017). Citizen-driven international networks and globalization of social movements on twitter. *Social Science Computer Review*, 35(1), 68–83.