Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN

CULTURE LETTERARIE E FILOLOGICHE

Ciclo 34

Settore Concorsuale: 01/B1 - INFORMATICA

Settore Scientifico Disciplinare: INF/01 - INFORMATICA

SCIENCE OF RETRACTED SCIENCE: A CITATION ANALYSIS OF THE ARTS AND HUMANITIES DOMAIN

Presentata da: Ivan Heibi

Coordinatore Dottorato	**Supervisore**
Marco Antonio Bazzocchi	Silvio Peroni
	Co-supervisore
	Nicola Grandi

Esame finale anno 2022

Abstract

In the scholarly publishing domain, a retraction is raised when a specific publication is considered erroneous by the venue in which it appeared after it was published. The aim of this work is uncovering new insights and learn new important information to help us understand the retraction phenomenon in the arts and humanities domain. Our investigation is based on a methodology defined using quantitative and qualitative measures derived from previous studies in the transdisciplinary research field of "science of science" (SciSci). The designed methodology takes into account a general case of retraction and applies a citation analysis based on five phases. Citations to retracted publications (before and after their retraction) are gathered and characterized with a set of attributes, including general metadata and information extracted from citing entities' full text. The annotated characteristics are further considered for a statistical and a textual analysis (i.e., a topic modeling analysis). The contribution of this thesis is grounded by addressing the following research questions: (RQ1) How did scholarly research cite retracted humanities publications before and after their retraction? (RQ2) Did all the humanities areas behave similarly concerning the retraction phenomenon? (RQ3) What are the main differences and similarities in the retraction dynamics between the humanities domain and the STEM disciplines? RQ1 and RQ2 are addressed by tuning and applying the methodology on the analysis of the retracted publications in the humanities domain. RQ3 is addressed on two levels, i.e., considering and comparing: (L1) the outcomes of the past studies on the retraction in STEM, and (L2) the results obtained from an analysis of a retraction case in STEM using the defined methodology.

Keywords: Retraction, Science of Science, Citation Analysis, Topic Modeling, Arts and Humanities

Introduction

Retraction is the action of withdrawing a statement which is now admitted being erroneous or unjustified. When a retraction is applied in the scholarly publishing domain, it indicates that a specific publication was withdrawn from the venue in which it appeared after it was published. In other words, retraction is a way by which the scientific record is corrected, and it is meant to ensure the integrity of the literature and alert readers and the scientific community on such records, rather than punish the authors of the retracted materials (COPE Council, 2019). Reusing the materials of erroneous publications may either waste future research effort or, in the worst case, have a negative repercussion on science and on the public, e.g., the continuing negative impact originated from the erroneous link between autism and childhood vaccines suggested by the publication of Wakefield et al. (1998) on the Lancet journal, later retracted in 2010.

COPE's Retraction guidelines state that reasons for retraction includes redundant publication, plagiarism, peer review manipulation, reuse of material or data without authorization, copyright infringement or other legal issues (e.g., privacy, illegality), unethical research, and a failure to disclose a major competing interest that would have unduly influenced interpretations or recommendations (COPE Council, 2019). It is noteworthy to mention that a retraction is not always a sign of a research misconduct; indeed, it could be the result of honest errors.

The editor(s) of the venue (e.g., a journal) in which the publication considered erroneous was originally published has the final decision on whether to raise a retraction. Editors have also the responsibility of alerting readers of such retraction and the type of mistakes noticed. This should be done by (1) amending with a “RETRACTED” watermark the original publication, and (2) publishing a retraction notice. The retraction notice should be linked to the retracted publication, be freely available to all the readers, state clearly who is retracting and what are the reason(s) for retraction.

Generally, retracted publications should not be used (i.e., cited) unless to highlight and study specific aspects of the retracted paper, such as an analysis of its flaws. Although a retracted publication represents a bad science product, hiding it from the community is not the right solution, whereas making it visible and accessible to be analyzed can benefit the future of science.

We strongly believe that studying the science flaws can help science itself improve in its quality. Indeed, the macro-objective of this thesis is uncovering new insights and learning new important information to help us understand the retraction phenomenon. In particular, this thesis is mainly concerned about the analysis of retraction in the humanities domain which gained less attention as compared to other domains, i.e., Science, Technology, Engineering, Mathematics (STEM) disciplines, including the medical and life-science domain.

Several works and a multitude of approaches have been adopted in the study of retraction in the past. These approaches have mainly involved several steps which could be formalized as following: (1) Definition of a case study and research question(s); (2) Detecting, collecting, and annotating the retractions; (3) Definition and application of a methodological analysis; (4) Presenting the results and discussing the outcomes. This thesis presents a definition of a methodology with respect to these four steps. The methodology has been conceived as a general mechanism to be adopted in the analysis of any case of retraction, i.e., does not have any restriction on the domain and neither on the nature/reason of retraction. The methodology represents a crucial part of the work of this thesis, it is conceived to be compliant with the principles of open science and it has been defined and formalized in a dedicated paper published during the doctoral period (Heibi & Peroni, 2021d).

The designated approach used in this thesis to study retraction and on which the defined methodology is conceptualized is inspired by the methods adopted by the past studies of a new emerging discipline called "science of science" (SciSci) – a transdisciplinary research field wherein science is used to study the dynamics and evolution of science itself using large datasets. The definition of SciSci given by the milestone article of Fortunato et al. (2018) is *"the science of science places the practice of* *science itself under the microscope, leading to a quantitative understanding of the genesis of scientific discovery, creativity, and practice and developing tools and policies aimed at accelerating scientific progress".* With SciSci our aim is to develop systems and polices which can improve the quality of science and prospects, through a deeper understanding of the factors that impact the science behaviors. The aim is to study science through a combination of quantitative methods used by scientometricians (which aim is quantifying, and predicting the scientific research) with qualitative ones used by other disciplines (e.g., social sciences, philosophy, etc.) (Zavaraqi & Fadaie, 2012). The combination of quantitative and qualitative research techniques provides a broader consistency, it combines depth (qualitative) with breadth (quantitative) in the analysis (Zavaraqi & Fadaie, 2012). For instance, the evolved techniques adopted in fields such as data science, network science, and artificial intelligence offer powerful tools to use in SciSci for the analysis of big data available nowadays.

Our defined methodology is based on a quantitative and qualitative citation analysis of retracted research. Citations to retracted publications (before and after their retraction) are gathered and characterized with their main metadata and other characteristics based on the textual contents of their full-text. The in-text citations (i.e., the textual segments included in the paper body in which a bibliographic reference is denoted by means of an in-text reference pointer, e.g., "[3]") are part of the analysis. These in-text citations are characterized with other attributes such as the section where they appear, the context, the sentiment, and the intent.

All the citing entities and their annotated characteristics are collected in a dataset. Based on the annotated values contained in the dataset, our methodology applies a statistical analysis and a textual analysis on the textual values gathered. The textual analysis is done through a topic modeling analysis (Mohr and Bogdanov, 2013) of the abstracts and the in-text citations contexts. A topic modeling analysis uses statistical modeling to automatically discover the topics (represented as a set of words) that occur in a collection of documents.

To foster the reproducibility of the topic modeling analysis and to let this process (once shared with other scientists) be understandable and modifiable by others who lack programming skills, or the ability to understand programming languages to a certain extent, we decided to use a dedicated software: MITAO – Mashup Interface for Text Analysis Operations, developed during the doctoral period (Ferri et al., 2020; Heibi et al., 2021).

Finally, another noteworthy aspect to mention regarding the designed methodology concerns the services used and the datasets queried. The retracted publications have been gathered using Retraction Watch (http://retractionwatch.com/) which keeps track and collects retractions of scholarly articles (Oransky & Marcus, 2012). To gather the entities which have cited the considered retracted publications, our methodology relies on three main services: Microsoft Academic Graph (MAG) (Wang et al., 2020), OpenCitations (Peroni & Shotton, 2020) and Crossref (Hendricks et al., 2020). The Semantic Web technologies used by OpenCitations permit the publication of bibliographic and citation data as Linked Open Data (LOD) (Bizer et al., 2018). The citations gathering process relies on a citations index developed during the doctoral period provided by OpenCitations: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations (Heibi et al., 2019). The dataset contains details of all the citations that are specified by the open references to DOI-identified works present in Crossref, and it is freely available to be queried through REST API calls.

The macro goal of this thesis is analyzing and understanding retraction dynamics in the humanities domain. This goal has been formalized into a serious of three research questions:

- RQ1: How did scholarly research cite retracted humanities publications before and after their retraction?
- RQ2: Did all the humanities areas behave similarly concerning the retraction phenomenon?
- RQ3: What are the main differences and similarities in the retraction dynamics between the humanities domain and the STEM disciplines?

To address RQ1 and RQ2, we tuned and applied our methodology on the analysis of the retracted publications in the humanities domain. This elaboration and the derived outcomes are based on the material and outcomes available in one article written during the doctoral period (Heibi & Peroni, 2021e). RQ3 was addressed on two levels. The first level (i.e., L1) answers RQ3 through a comparison of the findings emerged after the analysis of the retraction in humanities with the outcomes of the past studies of retraction in STEM. While the answers of the second level (i.e., L2) are based on a comparison of the findings in the analysis of the retraction in the humanities case with the results obtained after an analysis of a retraction case in STEM using our methodology. On the second level we applied our methodology to a popular and highly cited retracted paper from the health science domain: *"Ileal-lymphoid-nodular hyperplasia, non-specific colitis and pervasive developmental disorder in children"* by Wakefield et al. (1998). The analysis has been done using the material used in a dedicated article published during the doctoral period (Heibi & Peroni, 2021).

All the datasets and visualizations generated while treating RQ1 – RQ3 are published online. In addition, we have designed dedicated webpages to visualize the results and modify/filter the dynamic charts produced. The outcomes produced after the analysis of RQ1 and RQ2 (i.e., the humanities case) are available at https://ivanhb.github.io/thesis_results/hum/, and published on Zenodo (Heibi & Peroni, 2021c). Likewise the results of the L2 analysis in RQ3 (i.e., the Wakefield et al. retraction case), which are available at https://ivanhb.github.io/thesis_results/wakefield/, and have been published on Zenodo (Heibi & Peroni, 2020b).

The thesis is structured in three parts.

The first part "Retraction in science" provides a general introduction to retraction, numbers, and statistics regarding retraction, the nature of information and data encompassed in it, the multiplicity of analyses/approaches that could be done, the outcomes of some previous studies which have adopted such analysis, and the humanities case, thus, to motivate our designated approach which takes its cue from past studies in the science of science research field (SciSci).

The second part "Materials and methods" is dedicated to the definition of the methodology. The first chapter of this part (Chapter 3) gives a general overview of the methodology through the illustration of its five main phases, the principles it is compliant with (i.e., open science principles), the datasets and services used, and the terminology adopted. Chapter 4 discusses phase-by-phase the first four phases of the methodology, leaving the last phase to Chapter 5. The last phase concerns the application of the topic modeling analysis through MITAO. It needs a dedicated chapter (Chapter 5) to be introduced and described clearly.

The last part "Results, discussion, and conclusions" applies the methodology presented in the second part to two different case studies: (Chapter 6) retractions in the humanities domain to address RQ1 and RQ2, and (Chapter 7) the Wakefield et al. retraction case to answer the last research question RQ3 on L2. Yet, Chapter 7 will first compare the outcomes of retraction analysis in humanities with corresponding findings of other past studies on retraction in STEM, to answer RQ3 on L1. Finally, Chapter 8 summarizes the work of this thesis with a discussion of limitations, the adaptions that may improve the overall analysis, and the possible future work which can extend the discoveries and knowledge of this thesis.

Contents

Part I

RETRACTION IN SCIENCE

Chapter 1

What is retraction

This chapter provides a general introduction to retraction, numbers and statistics regarding the retraction, the nature of information and data encompassed in it, the multiplicity of analysis/approaches that could be done, and the outcomes of some previous studies which have adopted such analyses. Based on the observations of previous studies, the third section of this chapter defines a strategical standard process for the application of an analysis of retraction. Such a process is based on four steps, each step is introduced and then discussed through the presentation of the choices adopted in the past by other works.

The materials presented in this chapter serve as a background on retraction and the approaches that could be/have been adopted in its study, such as to motivate the chosen analysis of this work, to be further introduced in Chapter 2. The last section focuses on the retraction in the arts and humanities domain. It presents the different aspects of retraction in this research field compared to the Science, Technology, Engineering, and Mathematics (STEM) disciplines.

1.1 An introduction to retraction

A definition of retraction given by Oxford English Dictionary is "The action of withdrawing a statement, accusation, etc., which is now admitted to be erroneous or unjustified" (https://www.oed.com/view/Entry/164384). When a retraction is applied in the scholarly publishing domain, it indicates that a specific published resource (e.g., journal paper, conference paper, book, etc.) was withdrawn from the venue in which it appeared after it was published. To indicate that an article has been retracted, a retracted article may have "RETRACTED" appear before its title and its full text labeled "Retracted" as well. In addition, a document/page may also be linked to such article to explain the retraction motivations.

In other words, retraction is a way by which the scientific record is corrected, and it is meant to ensure the literature integrity and alert readers on such records, rather than punish the authors of the retracted materials. Ensuring the integrity of scientific publications is crucial since these published resources represent the basis of new studies. Indeed, new studies might use the methods/results presented by these past publications. Reusing past materials of erroneous publications may waste future research effort or have a direct negative effect on the public. This is particularly true for health science; publishing incorrect results may cause physical harm to the general population (e.g., the erroneous link between autism and childhood vaccines suggested by the study of Wakefield *et al.* (1998)). The harm caused by retractions in health science is also evident in a recent report regarding the high number of retracted papers about COVID-19 (Boschiero et al., 2021) (tracked at https://retractionwatch.com/retracted-coronavirus-covid-19-papers/), for instance, a study suggesting using a malaria drug to treat people with COVID-19 has been highly questioned and subsequently retracted (Mehra et al., 2020). Yet the findings of this study were publicly endorsed, with politicians such as the former US president Donald Trump endorsing the medication, as cheap and easy to use as a treatment (Ledford, 2020).

Articles reporting flawed data or content errors in small parts, are rectified by a correction. Formally this is done by raising a “partial retraction” or “retraction in part”. The correction of the erroneous article portions keeps the general information and the article’s stated conclusions uncompromised. The erratum/corrigendum publications are used to report such partial errors.

Reasons for retraction includes redundant publication, plagiarism, peer review manipulation, reuse of material or data without authorization, copyright infringement or other legal issues (e.g., privacy, illegality), unethical research, and a failure to disclose a major competing interest that would have unduly influenced interpretations or recommendations (COPE Council, 2019).

The retraction of an article is not always a sign of a research misconduct. It could be the result of honest errors. Retractions linked to honest error account for less than 20% of the total, and there is an open discussion on how to distinguish between ‘good’ and ‘bad’ retractions (Fanelli, 2016). In case of honest errors it is expected that the authors themselves inform the journal’s editor of such errors they have noticed (or have been informed of) in their published article. Table 1.1 lists the possible reasons of misconduct and honest errors (Cerejo, 2013). The list is not exhaustive, yet it underlines the most frequent types.

Table 1.1. The main reasons of retractions divided in: (first column) honest error/unintentional and (second column) intentional misconduct (Cerejo, 2013)

Honest error	Misconduct
Errors in sample or data	Undisclosed conflicts of interest
Skewed statistical analysis	Plagiarism or self-plagiarism
Inaccuracies or unverifiable information	Salami slicing (using the same data set to publish multiple studies)
Irreproducibility	Data fabrication or manipulation
Redundant publication (discovery that some aspects have already been published)	Lack of adherence to ethical protocols
Disputes over authorship attribution	Duplicate submissions (to different journals at the same time)

As suggested by Bar-Ilan & Halevi (2018), misconduct reasons of retraction could be further divided into scientific distortion and ethical misconduct. Ethical misconduct includes reasons such as duplicate publication, plagiarism, missing credit, etc., while scientific distortion includes reasons such as data manipulation, fraudulent data, unsupported conclusions, etc.

Retractions should also alert the readers of such erroneous article(s) and the type of mistakes noticed; this is done by (1) amending with a "RETRACTED" watermark the original publication, and (2) publishing a retraction notice. The retraction notice is a document which describes the reasons that justify such a retraction. According to the COPE retraction guidelines (COPE Council, 2019), this notice should: (1) be linked to the retracted article and clearly identify it, (2) be clearly identified as a retraction (i.e., distinct from other types of corrections), (3) be freely available to all the readers, (4) state who is retracting and the reason(s) for retraction, and (5) use objective, factual, and non-inflammatory language.

The editor(s) of the venue in which the bibliographic resource in consideration (e.g., an article in a journal) was originally published has the final decision on whether to raise a retraction, which needs to be formally accompanied by a retraction notice. The editor may decide to retract a publication even if some of the authors of the publication do not agree. In fact, they should not delay the retraction in case the authors are not cooperative. Publications should be retracted as soon as possible if the editor is convinced that the publication is involved in any of the reasons of retraction (COPE Council, 2019).

The COPE retraction guidelines states that "the original article should not be completely removed or 'replaced' but should be retained and linked to". A retracted paper is completely removed in rare exceptional cases, where leaving it online might cause significant harm or constitute an illegal act. Citing a retracted paper is not prohibited, yet it is important to acknowledge its retraction. Ideally, this could be done by citing both the paper and the retraction notice (that are two different documents). It is important that researchers are aware of the articles they are citing. Services such as Zotero (https://www.zotero.org/), which helps researchers collect, organize, and cite their bibliographies, can help detect and alert authors when citing a retracted publication (Mueen & Dhubaib, 2011).

Generally retracted scientific papers should not be used, or cited, unless to highlight an aspect about the retracted paper, such an analysis of its flaws. Although a retracted paper represents a bad science product, hiding it from the community is not the right solution, whether make it visible and accessible to be analyzed can benefit the researchers. In summary, we strongly believe that studying the science flaws is an analysis which can help science itself improve in its quality.

The next section gives a quantitative overview picture of the retraction phenomenon and discusses the numbers and main characteristics of it.

1.2 Numbers on retraction

The oldest retracted article goes back to 1756, and the next recorded retraction occurred in 1927 (Whelden, 1926). After this period, retractions were typically recorded as taking place every several years, and began to become more common in 1999, with constantly increasing numbers (Vuong et al., 2020).

Generally, retractions are rare events. According to one count done by (Hilgard & Jamieson, 2017) there were 300 retractions among 1.4 million papers published annually, i.e., one retraction in 4,666 published papers (i.e., 0.021%). However, the frequency of retractions in the last 20 years had a continuous increasing trend. This trend has been monitored between 1753 to 2019, with a collection of 18,603 retractions (Vuong et al., 2020), and it is shown for the last 20 years in Figure 1.1. This trend was also observed by (Brainard, 2018), indeed the rate roughly doubled from 2003 to 2009, yet it remained in a steady level since 2012. In part, that trend reflects a rising denominator: more scientific papers are published annually (doubled from 2003 to 2016) and, thus, the number of retractions raised as well. The exceptional numbers reported in 2010 are probably due to an investigation and a massive retraction of several articles, in particular the articles authored by Joachim Boldt (Brainard, 2018). Although, the rate takes a dramatic decline in 2019, this fact is almost certainly not true, and it is caused by delays in publishing retractions, as it is clarified by

Vuong et al. (2020). Considering these numbers and period (excluding 2019), the average number of retractions per year is 634.

The time of retraction is defined as the difference between the year of retraction and the year of publication. The year of retraction is the year of publication of the retraction notice. The work of Steen et al. (2013) monitored the time to retraction based on a collection of retractions from the biomedical literature and life science domains up to 2012. The mean time to retraction for the articles retracted from 2003 to 2012 was 33.81 months, on the other hand, for the period between 1973 and 2002 the average value was 29.6. Another analysis of all the retractions up to 2020 reported an average period of almost 45 months for a publication to be retracted (Bhatt, 2020). On a collection of data based on 35 years of publications in the journal *Science* (from 1983 to 2017), the work of Wray & Anderson (2018) got a mean time to retraction of 34 months. Thus, is very much in line with the 33.81 value observed by Steen at al (2013). An interesting fact that emerged from the work of Wray & Anderson (2018), is the fact that almost 76% of the retracted papers were retracted within four years of publication, and 30% of the retracted papers were retracted within a year of publication, not far from the 25% reported by (Bhatt, 2020). Bar-Ilan & Halevi (2018) reported that 50% of the collected articles were retracted within one year. This work made also a differentiation depending on the type of retraction, such that it pointed out the fact that misconduct retractions have a longer time of retraction while administrative errors are corrected within 5 years or less.

In general, relatively few authors are responsible for a disproportionate number of retractions. Just 500 of more than 30,000 authors (and co-authors), account for about one-quarter of the 10,500 retractions, taken from the dataset analyzed by Brainard (2018). One hundred of those authors have 13 or more retractions each.

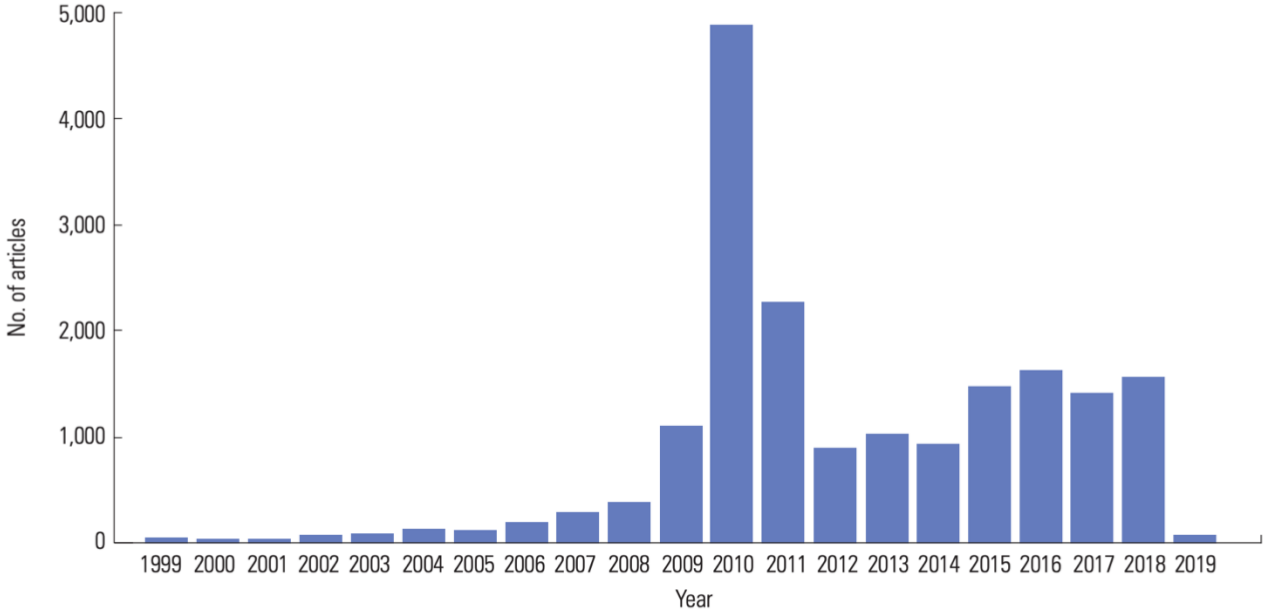


Figure 1.1. The number of retracted articles per year, starting from 1999 to 2019. The rate appears to decline in 2019, but this fact is almost certainly not true, and it is caused by delays in publishing retractions. This chart is taken from (Vuong et al., 2020)

United States and China are the two countries with the highest number of retractions (Vuong et al., 2020; Ribeiro & Vasconcelos, 2018, Bhatt, 2020). In an examined period from 2013 – 2015, the total retractions account for about 41% of total retractions, with 376 (United States) and 283 (China) retracted papers. In both cases, retractions due to misconduct was higher than the ones resulted from honest errors (Ribeiro & Vasconcelos, 2018). These numbers could be a consequence of the high number of publications that these countries have, yet this hypothesis needs a confirmation by further analysis on the subject.

When looking at the fields of study, most retractions are associated with business and technology, physical sciences, basic life science, and the health sciences (Vuong et al., 2020). Domains such as the social sciences and humanities relatively accounts for a small portion of all the retractions (Figure 1.2). Another important fact that emerged from the study of (Vuong et al., 2020), is the fact that most of the retractions in fields having few retractions (e.g., social sciences) are also associated with other fields with a high number of retractions (e.g., business and technology). This fact is possible since one retracted publication could be related to more than just one field of study, i.e., it treats arguments belonging to different fields of study.

A total of 753 publishers with retracted papers were detected by (Vuong et al., 2020). In this distribution IEEE, Elsevier, and Springer accounted for 56.81% (10,569 of 18,603). IEEE has the highest number of retractions, with a total of 6,763 (36.35%) retracted papers. Elsevier had the highest number of journals that have had papers retracted (877). The same work also studied the distribution of the number of retracted articles by journal impact factor (JIF). The obtained results indicates that more than three fourths of the retracted publications were published in journals with a JIF of 5 or lower. Yet we think that this result needs a further analysis through the consideration of the specific domain, since an apparently low JIF score for a journal in specific domain (e.g., computer science) have a completely different perception.

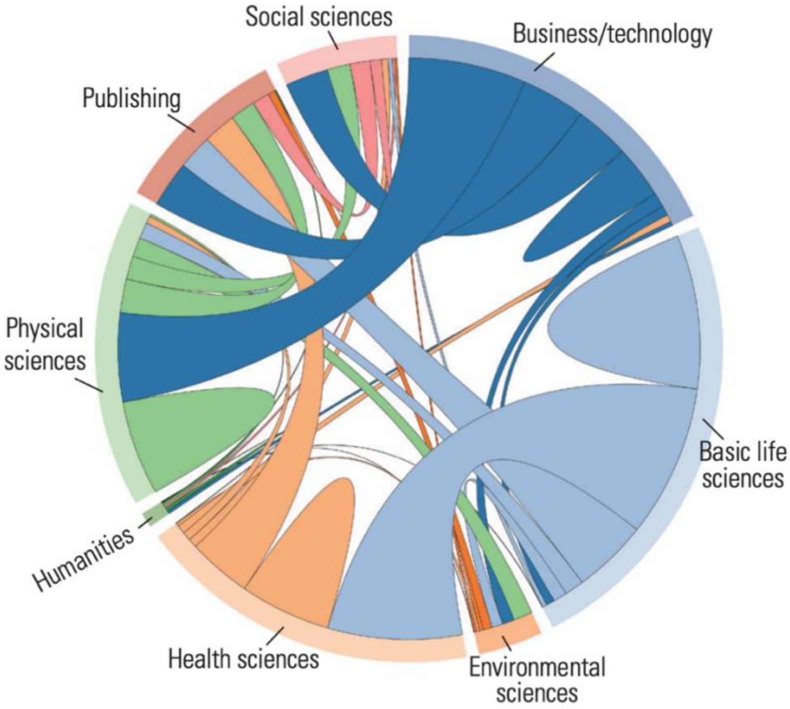


Figure 1.2. A Chord diagram for the distribution of the retracted papers in the different fields of study. The graphics highlights the relationships among retractions in different fields. The chart is taken from (Vuong et al., 2020) and is based on the collected retractions up to 2019.

Usually, a retracted paper can have more than one reason for retraction. According to (Hilgard & Jamieson, 2017), from 2012 to 2019, "fake peer review" has become a major reason of retraction, with a total of 676 retracted articles. In another collection of 1,082 retracted publications from 2013 to 2016 the main cause of retraction was misconduct (65.3%), and the leading reasons were plagiarism, data management and compromise of the review process (Campos-Varela & Ruano-Raviña, 2019). The analysis of Fang et al. (2012) on the period from 1977 to 2011 demonstrated that retractions due to fraud and errors was first evident starting from 1990s and had a dramatic rise until the last year considered (i.e., 2012), while a more modest increase in retractions because of error has been reported. In addition, reasons such as "plagiarism" and "article duplication" were noticed only starting from 2005.

1.3 Approaches and outcomes

Several works in the past studied and uncovered important aspects regarding the phenomenon of retraction. A multitude of approaches have been adopted in the study of retraction. These previous studies, which this work is part of, involve several steps, and could be formalized in the following ordered steps:

1. Definition of a case study(s) and research question(s);
2. Detecting, collecting, and annotating the retractions;
3. Definition and application of a methodological analysis;
4. Presenting the results and discussing the outcomes.

In this section we discuss each step separately, starting from a definition of the step and its expected outcomes, and through a review of some examples from the past studies. The reported examples help us highlight the most common choices that have been taken regarding each processing step.

1.3.1 Definition of a case study(s) and research question(s)

The case study/domain selection could be either based on a large-scale collection or on either one or limited number of retraction cases. In case of a limited number of retraction cases, usually the selection is driven by the popularity of the case(s) or by a common characteristic, e.g., a collection of retracted papers from the same author/journal/publisher.

Retractions cases from Science, Technology, Engineering, and Mathematics (STEM) disciplines took a lot of attention and have been highly preferred. Examples include the engineering domain (Rubbo et al., 2019), life sciences (Corbyn, 2012; Bhatt, 2020), computer science (Al-Hidabi & Teh, 2019), and economics and management (Karabag & Berggren, 2012).

Health science gained the higher attention compared to the other STEM disciplines. This is due to the importance and overall popularity of its thematic and the possible harm that might be caused by erroneous retracted publications related to this field of study. Several studies have considered this domain in general (Campos-Varela & Ruano-Raviña, 2019; Gasparyan et al., 2014; Gaudino et al, 2021), in specific fields such as genetics (Dal-Ré & Ayuso, 2021) and dentistry (Theis-Mahon & Bakker, 2020), or through the analysis of singular/limited popular retraction cases, such as Scott S. Reuben's retracted articles (Bornemann-Cimenti et al., 2016), Schön's series of papers retracted (Luwel et al., 2018), Wakefield's retracted paper (that fraudulently reported an association between vaccination and autism) (Suelzer et al., 2019; Heibi & Peroni, 2021), and the retracted paper of Narayan et al. (van der Vet & Nijveen, 2016). Recently, others have focused on all the retracted papers studying COVID-19 (Yeo-The & Tang, 2021; Boschiero et al., 2021; Ledford, 2020).

Other criteria have been also adopted in the selection of the retractions such as the country, e.g., retractions in Indian science (Elango et al., 2019), from the Arab regions (Al-Hidabi & Teh, 2019), or a selection based on the citation count, e.g., Teixeira da Silva & Dobránszki (2017) and Jan et al. (2018) respectively analyzing the top ten and seven highly cited retracted articles.

The selection of a case study/domain depends on the research question(s) related to a study. The main aim of the studies conducted on retraction is understanding such phenomenon and exploring its impact on the science behavior or on its main actors (e.g., authors, journals, publishers, etc.), e.g., an investigate toward the impact of retractions on the h-index of their authors (Mongeon & Larivière, 2016). We can summarize the possible research questions into three macro categories, presented as three major questions:

a) Why is a retraction raised? With minor related questions such as: do authors make mistakes with good/bad faith, or did they intended to cheat with deliberate misconduct?
b) How to raise a higher concern toward the retraction? With minor related questions such as: what operations could help the scientific community safeguard the integrity of scientific publications?
c) How a retraction affects its citations? With minor related questions such as: why other articles keep continuously citing a retracted article?

The first two categories (i.e., *a* and *b*) overlap on several occasions. The past studies of Casadevall et al., (2014) and Fang & Casadevall (2011) are good examples belonging to the first category. Based on the understanding of the errors that lead to retraction, they tried to guide and suggest some good practices to improve the integrity of the scientific literature, such as the need of an action by the scientific community in the adoption of protocols to rectify the scientific literature and ensure its integrity. Same has been done by the earlier work of Cokol et al (2007). Others monitored the journal impact factor of the retracted papers to understand the retraction behavior with respect to their venues, which helped suggest and discuss new solutions/approaches (Campos-Varela et al., 2020).

The third category is the one which we found out to be the most prominent. Understanding the impact of retraction in the scientific process, involves an understanding of citations and how retracted works are reused by other publications. Generally, this means an understanding of the quantitative impact of retraction into the citations (Lu et al., 2013; Mott et al., 2019; Schneider et al., 2020; Suelzer et al., 2019). This help having an insight on how “false science” is propagated and impacting (1) the scientific change in general (Azoulay et al., 2015; Schneider et al., 2020; van der Vet & Nijveen, 2016), (2) the works which used/cited retracted works before their actual retraction (Bolland et al., 2021), or (3) the other works of the authors which have been involved in a retraction (Azoulay et al., 2017). In addition, some had also investigated why other articles keep citing retracted papers (Teixeira da Silva & Bornemann-Cimenti, 2017) and why this trend is different depending on the domain (Karabag & Berggren, 2012). An interesting work by Bordignon (2020), investigated how negative citations and negative comments posted on other platforms can contribute to the correction of science. Jan et al. (2018) researched on whether the context of the citations (a passage or statement within the text of the citing document containing the reference) to retracted articles should be considered when analyzing the citations of the retracted publications.

1.3.2 Detecting collecting, and annotating the retractions

One of the basic techniques to detect retractions is to query large bibliographic databases, such as Science Direct by Elsevier, filtering the articles metadata or by searching for terms such as ‘‘RETRACTED’’ in the article title or in its retraction notice, as it has been adopted by several studies in the past (Moylan & Kowalczuk, 2016; Bar-Ilan & Halevi, 2017; 2018). Some services have labeled the retracted publication in their dataset (i.e., with the type “Retracted Publication”). This fact makes the search process for the retracted papers even easier and less error prone. This method was used by several studies, with services such as *PubMed* (https://pubmed.ncbi.nlm.nih.gov/) and MEDLINE (Campos-Varela et al., 2020; Dinh et al, 2019; Candal-Pedreira et al., 2020), or using the *Web of Science Core Collection* (https://clarivate.com/webofsciencegroup/solutions/web-of-science-core-collection/) (Candal-Pedreira et al., 2020).

Alternatively, an important available dataset is the Retraction Watch database (2018) (http://retractiondatabase.org) provided by Retraction Watch (Oransky & Marcus, 2012) – a blog founded by Ivan Oransky and Adam Marcus, launched in August 2010, with the aim of collecting, and making accessible a list of all the scientific papers retracted. The Retraction Watch Database has been used in several studies about the retraction phenomenon (Al-Hidabi & Teh, 2019; Jan et al.,2018; Teixeira da Silva & Dobránszki, 2017; Vuong et al., 2020). To foster a higher coverage, others used both the Retraction Watch database along with the other services mentioned above (Corbyn, 2012; Mott et al., 2019; Dinh et al., 2019; Bhatt, 2020).

Once the retractions are identified and collected, these need to be annotated and characterized with the related information/data. A collected retraction is characterized by:

1. The identification of the publication that has been retracted, for instance using a DOI, PubMedID, etc.
2. A characterization of the retracted article, e.g., the type of the publication (e.g., research article, conference paper, book chapter, etc.).
3. The date when the retraction was raised
4. The venue of the retracted publication (e.g., name of the journal)
5. The reason(s) of retraction.

The data used to characterize a retracted article could be extended depending on the information needed to answer the research question(s). Several attributes could be considered in the characterization of the retracted publication, such as the author(s), the field of study, the number of citations, the publisher, etc. For instance, if a particular study wants to investigate the usage of a specific retracted article *R* in its pre- and post-retraction periods, then we might need to collect the citations of *R* starting from the date of its publication, up to the last citation received (Suelzer et al., 2019).

Additional attributes might involve a review of the full text of the considered retracted publication(s). For instance, if the aim is to investigate the main topics treated in a collection of retractions, then annotation of the abstracts of all the retractions needs to be further analyzed (e.g., as has been done in the work of Bhatt (2020)).

The data needed could also involve in-depth analysis into the attributes that characterize the retracted publications. One might decide to get further information about the entities which have cited the retracted publication(s). For instance, labeling the venues of the citing entities could give further insights regarding the journals which have used the most retractions.

The Retraction Watch database provides many of these attributes. Each retraction in the dataset of Retraction Watch (i.e., a record) gives information about the retracted publication, with the attributes: title, corresponding subject(s), journal, publisher, affiliation(s), author(s), DOI/PubMedID, date of publication, type (e.g., research article), country, are paywalled policies. In addition, each record includes the reason(s) of retraction (taken from the list of reasons provided by Retraction Watch at https://retractionwatch.com/retraction-watch-database-user-guide/retraction-watch-database-user-guide- appendix-b-reasons/), the DOI/PubMedID of the retraction notice, and the date of publication of the retraction notice (i.e., the date of retraction).

To get the citations of a retracted article, external citation index services are used. Services such as Google Scholar, Elsevier's Scopus, and Web of Science are major examples which are highly used for this purpose.

1.3.3 Definition and application of a methodological analysis

The analysis that could be adopted in the study of the retraction, fall into two major categories – quantitative, and qualitative. In a qualitative analysis the data is non-statistical, therefore it is not necessarily measured using hard numbers, and it is mostly represented into a semi-structured form (i.e., organized in a naïve format, e.g., it does not follow a tabular structure). Typically, this analysis is used to answer "why" questions[1], and to investigate a specific aspect which is often open-ended and needs further research. The qualitative data can be subjective, therefore an analysis toward such

[1] *"Quantitative vs qualitative research questions"* https://unstick.me/qualitative-vs-quantitative-research-questions/

data is a complex process with numerous possibilities. For instance, qualitative information could be the information: the article *A* is authored by a group of PhD and Postdoc students.

On the other hand, quantitative data analysis is statistical and is typically structured (i.e., organized following a specific data model), therefore its definition is stricter and follows a specific data model. Numbers and strings are the common data types used in this case. The quantitative information are less prone to interpretation (contrary to the qualitative one). Typically, answers questions such as "how much/many". For instance, quantitative information could be "an article *A* is authored by two PhD students and one Postdoc".

From a methodological point of view, in quantitative research researchers utilize scientific methods, conduct experiments, and control phenomenon under investigation by altering the variables to achieve objective results. However, the quantitative researchers do not influence the research outcomes. On the other hand, the common aspects are that both quantitative and qualitative research use observation to collect data. Both research work on creating meaning, quantitative research follows a detached observer stance, whereas qualitative research co-creates the meaning by interpreting the outcomes (Arghode, 2012).

Some studies analyzed retractions from a qualitative point of view. In the work of Bar-Ilan & Halevi (2017) this is done through a subjective sentiment analysis of the context of the citations to retracted articles, in addition this work applied an ad-hoc qualitative analysis to a selection of popular retraction cases. A classification and analysis toward negative citations has been also done by Bordignon (2020). Boschiero et al., (2021) studied the typology (e.g., observational, or experimental) and the area of study (epidemiology, treatment, experimental, etc.) of the COVID-19 retractions. Casadevall et al., (2014) manually reviewed a collection of retraction notices, to have more details on the causes of error, and classified the reasons of retraction into 8 categories (irreproducibility, laboratory error, analytical error, etc.). A manual review of the contexts that used retracted publications was also done by van der Vet & Nijveen (2016), in this case the authors were interested to learn whether the retraction had been acknowledged.

The quantitative aspects and analysis of retractions took higher consideration. This prepossession to the quantitative research is likely to be related with the objective, fast, focused, and acceptable nature of this analysis[2]. Quantifying the reasons of retraction, time to retraction, and the fields of study of the retractions, are only some of the possible examples. In fact, the statistics presented in the Chapter 1.2 are based on quantitative analysis of data/information regarding the collected retractions. Quantitative analysis is used to predict outcomes using values and statistics, in case of an analysis toward retraction, and with the help of large datasets, it could uncover patterns and behaviors to answer complex research questions. For instance, Candal-Pedreira et al. (2020) performed a statistical/quantitative analysis using the characteristics of the retracted articles by reference to some variables of interest, such as the pre and post citation counts, and the retraction reason.

Citation analysis is an approach that many have used when the aim was to analyze the effects and impact of retractions on the scientific community and on the other works that used the retracted publications contents. In this regard the analysis was mainly focused on the quantitative aspects. Several works (Lu et al., 2013; Azoulay et al., 2017; Mongeon & Larivière, 2016; Shuai et al., 2017) focused on quantifying the pre and post citations and demonstrating the citations trends. Bolland et al., (2021) focused on the analysis of the citations made before the retraction. The citation analysis in other cases has been done not only on the direct citations to the retracted papers, yet also on the citing entities too (second level) (Schneider et al., 2020).

A citation analysis in many of the past works led also to a network analysis. This has been done considering as seeds several retraction cases (Chen et al., 2013), and it had a particular significant

[2] *"15 Reasons to Choose Quantitative over Qualitative Research"* https://www.formpl.us/blog/quantitative-qualitative-research

importance when the analysis of the citations included also the second level of citations, and not only the direct citations to the retracted publication(s) (Schneider et al., 2020).

From the above discussion we can notice that some of the works mentioned have combined both quantitative and qualitative analysis to address their research questions. Our work is part of these studies, and uses this strategy, details regarding the specific analysis that have been adopted are part of the next chapters.

1.3.4 Presenting the results and discussing the outcomes

To present complex quantitative insights of a specific analysis, usually studies use descriptive statistics. It helps describe the results obtained and understand the features of a specific data set by giving short summaries about the sample and measures of the data. Measures such as mean, median, and mode, are used at almost all levels of mathematical and statistical analysis.

Tables and charts are the way to present the descriptive statistics of an analysis. Bar/Pie charts, histograms, scatter plots are all possible graphical representations to use for representing quantitative results. On the other hand, for qualitative results the data is not relational and structured, therefore the visual representation of such data are ways to display the data such that they facilitate ease of reading and comprehension. For instance, word clouds (i.e., images composed of words) are a typical solution to visualize the results of a text analysis which is meant to emphasis the importance of the frequent words (i.e., frequent words are shown in larger and bolder font, while less-frequent words are shown in a smaller font)[3], e.g., the most frequent words used by a group of users to answer an interview.

A graphical representation for the data and obtained results helps infer more easily insights and answers to the research questions raised. It also improves the trustworthiness of the discussion with

[3] "*How to Visualize Qualitative Data"* https://depictdatastudio.com/how-to-visualize-qualitative-data/

visible facts. The work presented by Bhatt (2020) is a good example for the usage of several visualizations on the representation of its qualitative and quantitative outcomes, in respect to the multiplicity of the research questions raised. The importance of a good data representation as support for the discussion speculations, is demonstrated for instance by Feng et al. (2020), who worked specifically on building a multidimensional bubble diagram to illustrate the multidimensional results of his analysis. Van der Vet & Nijveen (2016) used a large graph visualization with blue and red colors to demonstrate the source of propagation difference between the directly citing articles and the indirect citations to a retracted article.

A discussion toward the obtained results to answer the research questions, can also lead to some suggestions for future strategies and policies to adopt regarding the retraction phenomenon. For instance, this is the case of Campos-Varela et al (2020) which suggests that: *"Measures before and after publication should be taken to limit misconduct",* same for Cokol et al. (2007) which suggests: *"However, the positive relationship between visibility of research and post-publication scrutiny suggests that the technical and sociological progress in information dissemination—the internet, omnipresent electronic publishing and the open access initiative— inadvertently improves the self-correction of science by making scientific publications more visible and accessible."*.

1.4 Retraction in arts and humanities

Retraction is also present in the arts and humanities domain, yet we have some significant differences with compared to STEM. From a methodological point of view, humanities research relies heavily on intuition and imagination (contrary to STEM which emphasizes ration and logical reasoning), although in some cases humanities research uses scientific methods derived from the STEM disciplines (e.g., studies in the digital humanities field) (Huang & Chang, 2008). In other words, arts and humanities research employs methods that are historical, interpretive, and analytical in nature, contrary to STEM where data and hard evidence are required to draw conclusions[4].

As a consequence of these methodological characteristics of the arts and humanities research, typically the humanists have a different approach in providing support for their arguments compared to their STEM colleagues. Humanists are less used to think on specific arguments as being right or wrong, instead their arguments can have different interpretable facets.

Considering these peculiarities, it makes less probable a situation in which the retraction of a publication in arts and humanities is warranted, whereas in STEM this is more frequent[5]. This disparity between STEM and arts and humanities is evident from the numbers presented previously in Figure 1.2.

The study of Halevi (2020) is a rare retraction analysis toward the arts and humanities domain. According to the outcomes of this work (based on the annotations provided by Retraction Watch), the most recurring reason for retraction is "significant overlap with previously published research" and "plagiarism" (representing 77% of the total). The detection and the forms of plagiarism are well defined and less prone to interpretation (Dhammi & Ul Haq, 2016). This fact and considering the methodological nature of the arts and humanities research previously discussed, might explain the outcomes of Halevi (2020). Yet, this final suggestion needs further analysis to be confirmed.

The work of Dougherty (2020) is the first book-length study fully dedicated to the analyzes of the forms of plagiarism in humanities disciplines. It provides a practical guide on the principal forms of disguised plagiarism using case studies and flow charts to assist researchers in protecting the integrity of their published research. Throughout the study of Dougherty (2020) we observed how plagiarism

[4] "*How is humanities research conducted?* " https://shc.stanford.edu/how-humanities-research-conducted

[5] "*How to Retract an Article in the Humanities* " https://case.edu/artsci/bakernord/events/past-events/2014-2015/how-retract-article-humanities

is more than just copy-and-paste, and its varieties involve additional concealment that creates further distance between the plagiarizing text and its source. The peculiarities regarding plagiarism, makes its analysis of its relation retraction very interesting.

Hopefully our work will in light and give new insights regarding the retraction phenomenon in the arts and humanities and answer some of the open questions derived from the lack of past studies that worked on a retraction analysis of this domain.

Chapter 2

Science of Science

This chapter introduces science of science (SciSci) – a transdisciplinary research field that uses large datasets to study the dynamics of the science production. The main principles and elaborations of our methodology are derived from previous studies in this research field.

The first section introduces science of science through a discussion toward the origins of this field of study, the methods, aims, expecting outcomes, and possible insights that could be inferred. We examine the subjects that have been mostly treated in SciSci literature, through a separation of the works into three main categories, based on the classification proposed by the key book on SciSci of Wang and Barabási (2021): (1) *The Science of Careers,* (2) *The Science of Collaboration*, and (3) *The Science of Impact.* The discussion of these three categories is based on a revisiting of the past studies through the illustration of important aspects, such as the domains, case studies, methods, and findings. The second section is dedicated to the citation analysis, a fundamental methodology adopted in the SciSci research field. The potentials of this analysis in the investigation of the retraction impact and the study of the error propagation, will become clearer in the last section, when reviewing SciSci studies which have worked on understanding retraction. In the final section, we also discuss how this thesis is intended to study retraction in the arts and humanities domain following the facts learned from past SciSci studies.

2.1 An introduction to Science of Science

The definition of science of science given by the key paper of Fortunato et al. (2018) is *"the science of science places the practice of science itself under the microscope, leading to a quantitative understanding of the genesis of scientific discovery, creativity, and practice and developing tools and policies aimed at accelerating scientific progress"*. The aim of the science of science is to develop systems and polices which can improve the scientific quality and prospects, through a deeper understanding of the factors that impact the science behaviors. The acronym SciSci, adopted by Fortunato et al. (2018), is the one we use for science of science throughout our work.

Larger data availability and the increasing collaboration between different disciplines such as natural, computational, and social sciences working on understanding science and the behaviors of the main actors of science (e.g., authors, publishers, etc.) are two major factors that led to the formulation of a new field of study such as SciSci. The continues evolving techniques and tools in data science, network science, and artificial intelligence offer powerful solutions to adopt in the analysis of the big data available. Scientists from different disciplines collaborate with the common goal of understanding science. With the help of such big data analysis techniques and methods, scientists study themselves and examine the successful and less successful projects, study the patterns that characterize the discovery process, and suggest new ways and policies to improve science as a whole.

Thomas Kuhn is considered one of the founding fathers of science of science. In 1962 he published a book titled *The Structure of Scientific Revolutions* (Kuhn, 1962), where he analyzed the science progress and concluded that such process is not gradual, yet it is much more like a punctuated equilibrium with important turning points. Kuhn defined the revolutionary processes in science with the term "paradigm shifts". Today this term is used in almost every creative activity and in the way other scientists think about the emergence and acceptance of new discoveries in science.

The idea behind the science of science is on the same line of the Kuhn’s aims – addressing questions that are of interest for the scientific community: What drive a good/bad scientific work? What is the life cycle of a new scientific discovery? When do authors reach their popularity peak in terms of citations? Which kinds of collaborations are more prolific? How can young researchers gain success in their careers?

In the book of Wang and Barabási (2021), the first exhaustive overview of SciSci, several of the above aspects and research questions about science have been discussed, such as the career path of individual scientists (e.g., what distinguishes one author from another), the collaborations (e.g., the advantages/disadvantages of teamwork and what are the benefits), the propagation of ideas/discoveries and their impact. In addition, Wang & Barabási (2021) highlighted the last innovative techniques used in SciSci, such as the role of the artificial intelligence.

Science in general is continuously evolving in its volume. This trend was demonstrated starting from the earlier important work of Price (1963), who showed the exponential growth in the volume of scientific literature. This volume growth of science, has a different impact on the emerging ideas and findings, which follows different dynamics. This fact makes the science of science research of high importance and demands a constant monitoring of the science evolution.

Scientific knowledge is defined by the contents and outcomes of the research publications, books, conference proceedings, etc. Yet, beside the inner contents and outcomes of these resources, they hold other meta information which represent an important source of knowledge and material in the science of science research. In addition, science is a dynamical system which involve cooperative and complex interactions between different actors (e.g., authors, institutions, fields of study, etc.). This social structure is a multiscale network continuously evolving which is defined by flows of information, methods, data, tools/software.

Understanding, quantifying, and predicting the scientific research and its resulting outcomes, is an object that has another existing discipline called scientometrics. Scientometrics works on measuring the scientific impact, understanding the citations, mapping the area and fields of study and developing indicators for decision makers. The definition given by Leydesdorff and Milojević (2015) is *"the study of science, technology, and innovation from a quantitative perspective"*. In the science of science research, the scope is broader than that, such that SciSci uses models to deeply probe the dynamics that drive science, from knowledge production to scientific impact, distinguishing predictable patterns from random ones (Zeng et al., 2017). In other words, science of science combines quantitative approaches, highly used by scientometricians, with qualitative approaches in the study of science. The combination of quantitative and qualitative research techniques provides a broader consistency and combines depth (qualitative) with breadth (quantitative). Qualitative research can also formulate problems and develop instruments for quantitative research, for instance (1) it may act as a source of hypotheses to be further tested from a quantitative point of view, (2) it can facilitate the definition of scales and new indices for quantitative research, or (3) data produced can assist the analysis of quantitative data.

Quantitative and qualitative research could be interpreted analyses that are relevant to macro and micro levels respectively. Bridging the gap between the two levels was suggested back in 1981 in the work of Duster (1981). The combination of macro and micro levels of science is one of the main reasons in the rising of the new science of science research field. The work of Zavaraqi and Fadaie (2012) presents an interesting theoretical review on the quantitative and qualitative approaches used in the study of science, thus, to define and discuss scientometric and science of science research methodologies. One of the oldest examples which combines qualitative and quantitative analysis is "Qualitative Scientometrics" (Callon et al., 1986).

The main aim of the SciSci research can be summarized in two words: "Understanding science". Authors, publishers, editors, and others want to study themselves through an examination of their works and projects, quantify the patterns that characterize their discoveries and outcomes, offer new insights, and follow the new lessons learned in order to improve themselves. From the applications point of view, SciSci adopts the quantitative techniques of the scientometric field of research combined with additional qualitative analysis. In this section we review some works which have used this mixed approach (i.e., quantitative, and qualitative), although in most of the cases, not explicitly defined as science of science studies. The aim is to give some concrete examples from the past and show the methods that have been combined, the outcomes, and the impact of such outcomes on the understanding of science and answering the research questions raised.

The overview of the past studies follows the thematic categorization proposed by Wang and Barabási (2021). Such that the works presented are divided in three different categories:

1. *The Science of Career*: works that put the spot on the careers of the scientists and try to study the successful or less successful paths through an analysis of the characteristics that distinguish these paths;
2. *The Science of Collaboration:* works that are part of this category study the collaboration networks between the main actors of science and the evolution of such networks as a function of other observed factors (e.g., the publication of a new discovery);
3. *The Science of Impact:* the subject of the works belonging to this category are the ideas and the discoveries, the aim is to exploare the dynamics underlying such scientific ideas and their general impact.

In the *Science of Career* the impact that significant events (e.g., winning a grant) can have on the career of scientists has been a matter of high interest. Herteliu et al. (2017) worked on understanding "*How much is the h-index of an editor of a well-ranked journal improved due to citations which occur after his/her appointment?"*. This study worked on quantifying the citations of a sample of authors on their prior and post join as editors to specific journals, and it showed how the activity of an author on the editorial board can increase positively on their *h*-index. Azoulay et al. (2011) analyzed how funding schemes that are tolerant of early failure and reward long-term success, are more likely to generate high-impact publications than grants that are subject to short review cycles. These findings are based on a description of a specific sample, the control set to use for comparison, the metrics that defines the scientific creativity, and finally some relevant descriptive statistics. The impact of the grant funding on the scientific productivity was subject of other studies as well (Jacob & Lefgren, 2011; Li & Agha, 2015).

Predictive analysis on the science career orientations have been also subject of several studies. For instance, through an analysis of distinct features across countries and cultures concerning which factors (e.g., personal motivation, family background, gender, etc.) can best predict students' future science careers orientation (Kjærnsli & Lie, 2011; Du & Wong, 2019). With respect to similar factors, DeWitt and Archer (2015) applied a quantitative and qualitative mixed approach to study the differences between the science career inspiration of primary and secondary school students.

Research on the impact of gender inequality has been highly addressed recently. Kalpazidou & Cacace (2017) developed an innovative assessment tool, which considers the complexity perspectives of this problem. The method is based on a quantitative analysis (e.g., toward roles/work positions) along with a large qualitative part, consisting in a semi-structured questionnaire. The results confirmed the multidimensional nature of the problem, and the processes complexity of gender equality policies, which highly depend on the interaction of a multiplicity of variables in dynamic contexts.

Finally, others studied the relation between new ideas/research creativity and the scientist's age. Packalen and Bhattacharya (2015) measured the emerge of new ideas using a textual analysis of the title and abstract of a collection of biomedical papers, the findings suggests that the papers published by younger researchers are more likely to build on new ideas, and that having a more experienced last author can positively increase this probability.

The category *The Science of Collaboration* studies the networks involving the main actors of science, the teamwork flow, the evolution of these networks and the impact of such collaborations. The work of Wu et al. (2017) is a perfect example belonging to this category. In this case a combination of different methods such as citation analysis and automatic text analysis techniques (i.e., Doc2vec) has been adopted to build the authors network and observe the differences between small and large teams' impact in the development and evolution of the ideas. A novel qualitative method such as autoethnography (i.e., *"a theoretical, methodological, and textual approach that seeks to experience, reflect on, and represent through evocation the relationship among self and culture, individual and collective experience, and identity politics and appeals for social justice"* (Holman Jones, 2007)) has been used in the analysis of sociology of science and to examine benefits, motivations, and challenges of teamwork (Dusdal & Powell, 2021). By applying text analysis on the abstracts of the papers, the work of Shi et al. (2015) has observed how science moves from questions answered in one year to issues to investigate the year after in a science network in the biomedical domain.

The last category we have listed is *The Science of Impact.* The analysis of some studies belonging to the previous two categories have a high probability to overlap with this one. In science the main factor used to measure and establish the scientific impact of a scholarly publication are citations. This fact is based on the simple concept that scientists cite papers that are relevant to their work (Wang & Barabási, 2021). Some studies focused on discussing and structuring ways to maximize the emphasizes and the impact of some specific typologies of research, such as the work done by Gregor and Hevner (2013) in Design Science Research (DSR: "*seeks to enhance technology and science knowledge bases via the creation of innovative artifacts that solve problems and improve the environment in which they are instantiated"* (Vom Brocke et al., 2020)). Foster et al. (2015) worked on the analysis of the factors that affect a scientist's choice regarding a research problem to study. Although innovative scientific research has a high impact, this study found out that this approach is less frequent, since pursuing innovation for many scientists is a gamble, without enough payoff, on average, to justify such risk. Another interesting factor to study, is the dynamic change of the collaborations and organization of the research in relation to the continuously increasing academic competition, demands and pressure (Degn et al., 2018).

From the above discussion we can clearly notice the prevalence of citation analysis in the methodology of several studies. Thus, we think that this type of analysis needs deeper understanding. The next section is especially dedicated to this topic, citation analysis methods, materials, and common strategies, will all be part of the discussion.

2.2 Citation analysis

A fundamental characteristic of science is its cumulative nature. New ideas and discoveries do not emerge in isolation, these are rather based and built upon previous discoveries of other past works. The standard mechanism used by scientists throughout the ages to address, use, and acknowledge the work of their colleagues is through citations. A network of citations characterizes patterns in the incidence of citations and references between papers according to factors such as the date of publication and subjects/areas of study (Price, 1965). This network represents an exchange of information about documents, subject relations, and scientific development. An analysis toward this network can trace back the history, scale, and trend of a scientific development. Citations are also the dominant measurable unit for establishing credit in science. Citation counts measure the impact and influence of an academic work. It is worth mentioning that some have also criticized the use of quantitative citation metrics in the evaluation of science. The Matthew effect*,* the non-consideration of the citation quality*,* and the differences between the local and global impacts, are some of the most common criticisms (MacRoberts & MacRoberts, 2018; Anauati et al., 2016). Other measures such as altmetrics have been considered as an alternative/enhancement to citations (Costas et al., 2015).

Fields of study such as bibliometrics and scientometrics work on the study of patterns of academic impact through citation analysis (Sengupta, 1992; Donthu et al., 2021). Considering the high impact that citations have to science, a correct analysis of it can enormously benefit the scientific community in many of the research questions that could be elaborated through a SciSci research.

Citation analysis combines mathematical, statistical, comparison, induction, abstraction, generalization, and logical methods, which are mainly applied to determine quantitative aspects (Qui et al., 2017). The statistical and quantitative analysis of citations has been adopted in several works and in the study of several behaviors. The first and main application was regarding the evaluation of the research (Moed, 2006), such as determining the impact factors of journals (Glänzel & Moed, 2002) or for measuring the impact of scholarly researchers and authors (Gasparyan et al., 2018). Other aspects have been studied throughout a citation analysis as well, for instance, an attempt for establishing an objective measure for evaluating the originality of a paper (Shibayama & Wang, 2020; Rossetto et al., 2018), mapping publications that have similar thematical subjects (Hjørland, 2013) or identify emerging research fields (González-Alcaide et al., 2016) through an analysis of the co-citations, the quantitative citations advantages of open access articles (Eysenbach, 2006), the characteristics of the in-text citations (the reference to the cited entity included in the textual body of the citing publication), e.g., position and distribution in the full-text, the counts per field of study, or the temporal trends (Boyack et al., 2018), etc.

The introduction and combination of the textual analysis and the natural language processing techniques with citation analysis has favored the analysis of the qualitative factors of citations (Jha et al., 2017). The analysis of the citation context is one of the most attractive subjects regarding the application of such techniques. This new analysis toward the concepts of the citations was also introduced and defined by Bornmann et al. (2020) (i.e., CCA, Citation Concept Analysis). Yet, is worth mentioning that the combination of an analysis toward the citation context and the traditional bibliometric methods was suggested and experimented with as early as the 1960s when Lipetz identified 29 different citation reasons (Lipetz, 1965). Also, the paper of Small (1978) was fundamental for the definition of the citation context and CCA: the work proposed an examination of the text around the in-text citations taken from a sample of very highly cited documents in chemistry to define an act of symbol usage. Other later works of Small focused on CCA, such as the identification of the discoveries in the biomedical sciences by the analysis of the citation context information (Small et al., 2017), and for the identification of the most relevant words used in the citations context for the classification of highly cited biomedical papers (Small, 2018). The recent works toward the citation context analysis has improved the understanding of the reasons of the citation and the topics discussed in the context that are linked to the cited publication (Anderson & Lemken, 2020; Crothers et al., 2020; Ding et al., 2014; Di Iorio et al., 2013). In addition, several works have also examined the sentiment of the citing entities toward the cited publications (Bar-Ilan & Halevi, 2017; Yousif et al., 2019). A sentiment analyses of citations contexts is a complex procedure to accomplish especially if it is meant to be done automatically. The work of Bornmann et al. (2020) on the analysis of the citation concepts provide a good example. The authors pointed out the importance of analyzing the sentiment of the citations, yet they mentioned how they tried to work on that without getting promising results, thus, they decided not to include such results in their study.

The citation reason or citation function is one of the most important features to consider for a qualitative evaluation of a citation. The citation function is defined as the author's reason for citing a given paper, e.g., to obtain support from the cited outcomes, to acknowledge and use the cited method, to use data/results of the cited publication (Teufel et al., 2006). The identification and classification of the citation reason has been a matter of debate by many in the last years (Di Iorio et al., 2013; Hassan et al., 2017; Tuarob et al., 2020) mainly due the large datasets availability and the new potential AI-based methodologies.

The graphical representation of the citation analysis outcomes especially regarding the construction of bibliometric maps has always been a crucial point. A good representation can be much enhanced by means of several important metrics such as zoom functionality, special labeling algorithms, and density metaphors. *VOSviewer* is one of the first programs which has focused on these graphical representations (Van Eck & Waltman, 2010). It displays a citation map which could be modifiable to emphasize different aspect, and permits operations such as zooming, scrolling, or searching to facilitate a detailed examination of the map. Same authors of *VOSviewer* worked on another tool called *CitNetExplorer*, this program was especially designated to analyze and visualize the development of a research field and track/review the literature regarding such topic (Van Eck & Waltman, 2014). *bibliometrix* is another interesting tool which supports a recommended workflow to perform in the application of a bibliometric/citation analysis (Aria & Cuccurullo, 2017). It is mainly used for performing comprehensive science mapping analysis.

2.3 Science of retracted science

Before presenting our approach in the study of retraction in humanities, inspired from the methods used by the past studies belonging to the SciSci field, we review other past studies in SciSci which have worked on the analysis of retraction in science (in general). This review goes through a discussion toward the aims, the datasets, the quantitative and qualitative methods used, and the outcomes of these studies, based on the research structure previously presented in Section 1.3. Although, studies that have adopted only quantitative analysis could be also part of the SciSci research family, in this review we focus only on those that have analyzed the retraction in a mixed approach (i.e., quantitative, and qualitative).

2.3.1 Review of past SciSci studies

The research questions and the objectives of the SciSci works which have studied retraction have been in most of the cases part of the category we called *The Science of Impact* (Section 2.1). The study of the impact of retraction on science and the implications that a retraction has on the community that uses the retracted publications (i.e., methods/outcomes) was highlighted by several works.

Bar-Ilan and Halevi (2017) investigated the sentiment of the studies that used (i.e., cited) the retracted papers, and following such discoveries they have summarized a list of general recommendations to authors, editors, publishers, etc. Also, the work of Bordignon (2020) was interested in the identification of the negative re-use of the retracted articles, yet in this case other sources such as the post-publication peer review platforms have been considered too. The findings suggests that post-publication peer review platforms (although not part of the scientific publication process) have a bigger contribution to the correction of science when compared to negative citations. The anonymous nature of the peer review platforms might represent a big contributor to these results. A deepen understanding of the use of retracted papers was also the goal of Hsiao and Schneider (2021). This case, this study worked on answering the macro question: "how the retracted papers have been cited over time?" This involves an analysis toward the quantitative aspects along with the qualitative ones, e.g., the citation purposes. Generally, this work demonstrated how retracted research can continue to spread and what might be the main factors contributing to such trend, for instance the responsibility of the information environment. Along the same line of Hsiao and Schneider (2021) the work of Schneider et al. (2020) worked on answering/uncovering closely related aspects. Schneider et al. (2020) used a network analysis combined with a retraction status visibility analysis to illustrate the potential harm of extended misinformation propagation. Candal-Pedreira et al. (2020) focused on the analysis of the citations to a specific typology of retractions – misconduct. One of the main findings is that such retractions do not have any association with the number of citations received in the long term, in other words they continue to be cited, thus circumventing their retraction.

Exceptional focus has been given in other cases also to specific domains such as dentistry (Theis-Mahon & Bakker, 2020). The investigation of Theis-Mahon and Bakker (2020) has been concentrated on the characteristics of retracted publications (e.g., reasons for retraction) and the sentiment of the post-retraction citations. The study observed positive citations in the post-retraction regardless of the designs or reasons for retraction, thus when compared to other fields, retractions in dentistry cannot be isolated to certain research methods or misconduct, instead it is a more widespread issue. Van der Vet and Nijveen (2016) examined the error propagation in the entire citation network of a specific retracted publication from the journal Nature which has been widely cited. Authors concluded that although the propagation of retracted results via direct citations was evident, regarding their specific case of study, they have not reported any propagation of the retracted error beyond the entities that have directly cited the retracted article (i.e., the error propagation didn't involve the entities which have cited those that have directly cited the original retracted publication).

From a methodological point of view all the above studies mainly applied a citation analysis and especially a citation context analysis. The context of the in-text citations was observed and annotated with characteristics such as the citation reason (Hsiao & Schneider, 2021), the sentiment of the citing entity (Bar-Ilan & Halevi, 2017; Bordignon, 2020; Schneider et al., 2020; Theis-Mahon & Bakker, 2020), the location, i.e., the section/chapter (Hsiao & Schneider, 2021), and whether the retraction was acknowledged in the context of the citation (Hsiao & Schneider, 2021; Schneider et al., 2020). In the majority of the cases, these characteristics needed a manual revision and annotation by a human annotator. A citation network analysis was also part of the methodology for some of the above studies (Van der Vet and Nijveen, 2016; Schneider et al., 2020). This turned out to be a crucial method for a visual observation of error propagation.

Other SciSci studies dedicated to the understanding of the retraction impact on science put the lens on the retracted entities themselves rather than their citations. Campos-Varela et al. (2020) analyzed the impact of retraction through a study of the journals characteristics that are associated with the retraction of papers. The findings showed the presence of retraction in both high and low impact factor journals (biomedical domain), but retractions due to misconduct (mainly plagiarism) is more frequent among the papers retracted from lower impact journals. Authors suggest such journals to implement and use more often plagiarism detection systems. The commentary of Yeo-The and Tang (2021) worked on describing and discussing the high rate of retractions for scientific publications on the Coronavirus Disease (COVID-19). The research addressed the reasons of this phenomenon with questions such as *"Why do COVID-19-associated publications have such a high retraction record appearance?"*. The approach was mainly qualitative combined with the analysis of the characteristics and general statistics of the retracted papers. The article concluded that one the possible reasons could be due to a rush by researchers in the submission of their papers in a global emergency, where human resilience and recovery are highly dependent on solutions to be provided by science.

Table 2.1 is a summary of the above review. We list all the past SciSci studies on retraction that we have mentioned above. For each study we define the case study/domain and the adopted methodology. The methodology is divided in two separated analyses: (1) citation analysis, and (2) analysis of the retracted publication(s) and/or the retraction notice(s). On one hand, the citation analysis is articulated in three types: (1.1) general statistical analysis (e.g., number of citations per year, post/pre citations, etc.), (1.2) context analysis, and (1.3) network analysis. On the other hand, the analysis toward the retracted publication(s) and/or the retraction notice(s) have two different types: (2.1) general statistics (e.g., time to retract, reasons of retraction, etc.), and (2.2) textual analysis of the full text of the retracted publication/retraction notice.

Table 2.1. A list of past SciSci studies on retraction. For each study we list: the case study/domain, and the adopted methodology. The methodology is further divided in two analyses: (1) citation analysis, and (2) analysis of the retracted publication(s) and/or the retraction notice(s). The citation analysis is articulated in three types: (1.1) general statistics analysis (e.g., number of citations per year, post/pre citations, etc.), (1.2) context analysis, and (1.3) network analysis. The retracted publication(s) and/or the retraction notice(s) have two different types of analysis as well: (2.1) general statistics (e.g., time to retract, reasons of retraction, etc.), and (2.2) textual analysis of the retracted publication/retraction notice full tex

Study	**Case study(s)/domain**	**Citation analysis**			**Analysis of the retracted publication(s) and/or the retraction notice(s)**	
		General statistics (e.g., citations per year)	**Context**	**Network**	**General characteristics (e.g., time for retraction)**	**Text analysis**
(Bar-Ilan & Halevi, 2017)	Articles retracted in 2014 with more than 10 citations between January 2015 and March 2016	Yes	Yes (sentiment)	No	Yes	No
(Bordignon, 2020)	Engineering field + 3 journals (*Science, Tumor Biology, Cancer Research*) + 3 authors (Sarkar, Schön, Voinnet)	Yes	Yes (sentiment)	No	No	No
(Candal-Pedreira et al., 2020)	304 retracted articles – indexed in MEDLINE from Jan 2013 to Dec 2015, and retracted between Jan 2014 and Dec 2016	Yes	No	No	Yes	No
(Campos-Varela et al., 2020)	Articles indexed in PubMed and retracted between Jan 2013 to Dec 2016	No	No	No	Yes	Yes (identifying reasons of retraction from the retraction notices)
(Hsiao and Schneider, 2021)	7,813 retracted papers retrieved from PubMed	Yes	Yes (retraction acknowledgment, section, purpose)	No	Yes	No
(Schneider et al., 2020)	One retracted clinical trial report (Matsuyama et al., 2005)	Yes	Yes (retraction acknowledgment, section, sentiment, and purpose)	Yes	Yes	No
(Theis-Mahon & Bakker, 2020)	Retracted publications in the dentistry domain up to Sep 2018	Yes	Yes (sentiment)	No	Yes	No
(Van der Vet & Nijveen, 2016)	One retracted paper published on *Nature* (Narayan et al., 2012)	Yes	Yes (retraction acknowledgment)	Yes	No	No
(Yeo-The & Tang, 2021)	Retractions on COVID-19	No	No	No	Yes	Yes

2.3.2 Science of retracted science in humanities

Now that we have pictured the SciSci research principles, discussed a standard approach to design a SciSci study, and addressed the case studies, methods/approaches adopted in the analysis of the problem, how obtained results are interpreted, and we have enriched our understanding through a review of the past studies in this field, we have enough knowledge to plan a methodological strategy to use in the analysis of our problem and research questions – retraction in humanities.

Our research is in line with the majority of the other studies presented in the previous section – it is part of the category that we have called *The Science of Impact.* The macro goal of our work is understanding the impact that retraction has on science. We do not want to limit our observations only to the analysis of such impact in the humanities domain, instead, based on the numbers and facts presented in the first two chapters, we have the feeling that the effects will go beyond the direct effects on the humanities community. In addition, we don't have any previous work in literature suggesting that such impact have a domain-localized nature, which strengthens our decision. Then, to have a complete vision of the retraction impact in the humanities domain our work cannot be limited to a one or to a limited number of retracted items, therefore the examined collection should include all the retractions with no restrictions or criteria toward the definition of such collection, for instance the selection of retracted publications through the consideration of features such as the reason of retraction, the venue, the type of publication, the year of retraction, etc.

As it has been demonstrated throughout the review of the previous section, the main mechanism/analysis that might help us understand and uncover the effects of retraction, the error propagation, and the reuse dynamics by others, is using a citation analysis approach. The methodology of our work is highly based on the citation analysis. Such analysis needs to embed a general statistical overview of the citations in the pre and post retraction and consider the features of the citing sources, for instance through the examination of the specific area of study of the citing entities. The context of the citations is another factor that should be considered. All the characterizations that have been considered in the observation of the citation context in literature are very likely to represent a significant role in the understanding of how retractions have been used. The citation context should be inspected in order to gather information about the section, the sentiment (i.e., positive, neutral, negative), the purpose (reason of citation), and whether the retraction was acknowledgment by the citing publication.

A general statistical overview of the retracted publications in the humanities is absolutely needed. To the best of our knowledge, only the work of Halevi (2020) highlighted the retraction in the humanities and analyzed some of the characteristics of the retracted articles in humanities. Yet, we think that a deeper characterization of the retracted publications through the inclusion of a wider set of aspects/attributes can be highly beneficial. For instance, through a further classification of the retracted publications into specific areas of study in the humanities domain (e.g., history, arts, literature, etc.)

Compared to other fields of study, humanities have several aspects that need a special consideration and might bring significant adjustments to the general retraction analysis methodology previously discussed for STEM disciplines. First, compared to STEM, bibliographic metadata in the humanities have a limited coverage in well-known citation databases (Hammarfelt, 2016). The main limitation of this fact is highly perceived during the citation analysis application (Archambault & Larivière, 2010). Considering this limit, is very plausible that the citing entities collected in the study might be fewer than those that had in fact cited the retracted articles. We also expect to have "plagiarism" as main reason of retraction (as it has been reported in Halevi (2020)), this fact and considering the different methodological/research nature of humanities (Section 1.4) are important facts to keep in mind in the analysis. This might lead to a higher possibility of introducing biases in the process. In addition, it will be hard to evaluate the reliability of the results with the outcomes of others considering the few works which have focused on the retraction in humanities.

Part II

MATERIALS AND METHODS

Chapter 3

Overview of the methodology

This chapter presents the methodology used in the thesis. The methodology takes one or more retracted publications as input and performs a quantitative and qualitative citation analysis. The elaborations of the methodology are inspired and defined following the approaches used by the SciSci studies (presented in Section 2.3).

The first section presents the methodology through the illustration of its five main phases: (1) defining a case study/domain, (2) identifying, retrieving, and extracting basic metadata from the entities which have cited retracted articles, (3) extracting and labeling additional data from the textual content of the citing entities (e.g., abstracts), (4) building a descriptive statistical summary, and finally (5) running a topic modeling analysis.

The methodology is compliant with the open science principles, presented and discussed in the second section. The third section describes the datasets and the services to be used with respect to the open science principles. Finally, the last section of the chapter defines the terminology and the abbreviations used to name some of the methodology components, which also will be used in the rest of the thesis.

The aim of this chapter is to give a theoretical background on the methodology and to outline its main structure to pave the way for the next two chapters (Chapter 4 and Chapter 5), which deepen into a pragmatical discussion of the five phases of the methodology.

3.1 Main phases

The methodology presented is a standard workflow we have defined for a quantitative and qualitative citation analysis of retractions. It is not restricted to a specific use case or domain (it can analyze an arbitrary number of retractions) nor to particular language, although this latter choice should be consistent throughout all the workflow, such that the generated dataset should contain values written in only one specific language (e.g., English).

The methodology is conceptualized into five different phases; each phase consists of one or more steps. The methodology takes one or more retracted publications as input and performs a SciSci analysis. It is a mixed strategy that combines quantitative and qualitative analysis on the citations of the considered set of retracted articles. The methodology workflow and its five phases are summarized graphically in Figure 3.1.

The first phase (i.e., "Defining a case study/domain") is dedicated to the definition of a case study or the domain we want to study using the methodology. This phase performs a general analysis over that retraction case taken into consideration and builds an overview picture to help in the definition of a collection of retracted publications to be further analyzed in the forthcoming phases. This overview picture relies on datasets holding information about retractions (e.g., the Retraction Watch Database).

The aim of the next two phases (i.e., Phase 2 and Phase 3) is to produce a dataset (i.e., "citing entities dataset" of Figure 3.1) to store all the citing entities accompanied by their attributes which have cited the retracted publications taken into consideration – each record (i.e., row) of the dataset will represent a citing entity characterized with several features (i.e., columns). More precisely, the second phase collects all the entities that have cited in either the pre or post retraction period any of the retracted publications contained in the set of retractions outputted from the first phase, then each entity (i.e., a record in the dataset) will be annotated with some initial main features (e.g., title, DOI, venue, etc.). The third phase characterizes each entity further with additional features based on the analysis of their textual content (i.e., full text) – mainly the abstracts and the in-text citation contexts which need to be extracted first and then analyzed for the annotation of other features such as: sentiment, citation reason, in-text citation reference etc.

The last two phases of the methodology take the citing entities dataset (produced and annotated in the previous two phases) as input for their elaborations. The fourth phase analyzes the information (characteristics of the citing entities) stored in the dataset to build a descriptive statistical summary around the quantitative aspects. The outcomes of this phase are represented as ulterior datasets along with a collection of static charts.

The first four phases of the methodology (i.e., Phase 1 – Phase 4) are discussed in detail in the next dedicated chapter (i.e., Chapter 4).

Finally, the last phase (i.e., Phase 5) consists in a topic modeling analysis – a statistical modeling for automatically discovering the *topics* (represented as a set of words) that occur in a collection of documents (Vayansky & Kumar, 2020). It takes the textual information (extracted from the full text of the citing entities and stored in the dataset during Phase 3), trains two topic models (i.e., abstracts and the in-text citation contexts) using this information as input, and builds a set of dynamic visualizations to have an overview on the generated topics and investigate the relation of such topics with the features of the citing entities. The topic modeling analysis and the details regarding the workflow of this phase are presented and discussed in a separated chapter (i.e., Chapter 5).

Table 3.1 summarizes the phases from 2 to 5 of the methodology. These are all phases which have an inner division of the process formalized in several steps to be followed. Each phase is represented in a different column and the steps of each phase are represented as cells. Each step is accompanied with a brief description, the inputs needed, and the expected outputs. The details mentioned in the table will become clearer in the next chapters (i.e., Chapter 4 and Chapter 5) when discussing each phase separately.

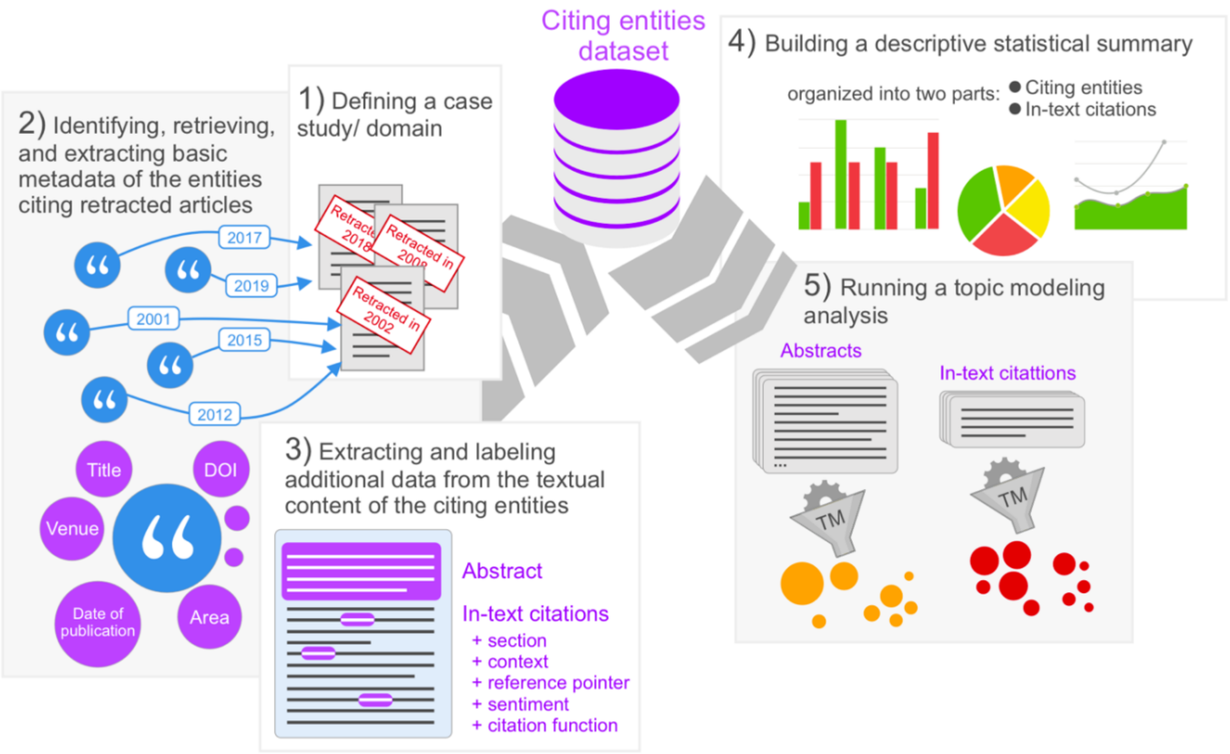

Citing entities dataset
1) Defining a case study/ domain
Retracted in 2018
Retracted in 2008
Retracted in 2002
2) Identifying, retrieving, and extracting basic metadata of the entities citing retracted articles
2017
2019
2001
2015
2012
Title
DOI
Venue
Date of publication
Area
3) Extracting and labeling additional data from the textual content of the citing entities
Abstract
In-text citations
+ section
+ context
+ reference pointer
+ sentiment
+ citation function
4) Building a descriptive statistical summary
organized into two parts:
● Citing entities
● In-text citations
5) Running a topic modeling analysis
Abstracts
In-text citattions
TM
TM

Figure 3.1. A graphical schema representing the methodology in its five phases (form left to right): (1) defining a case study/domain, (2) identifying, retrieving, and characterizing the citing entities, (3) defining additional features based on the citing entities contents, (4) building a descriptive statistical summary, and (5) applying a topic modeling (TM) analysis

Table 3.1. A description of phases 2 to 5 of the methodology. These are phases based on an inner division of their process formalized in several steps to be followed. For each phase (column), we list its corresponding steps (cells). Each step is accompanied with a brief description, the inputs needed, and the expected output

Phase 2: identifying, retrieving, and extracting basic metadata of the entities which have cited a retracted publication	**Phase 3: extracting additional data and labeling the textual content of the citing entities**	**Phase 4: building a descriptive statistical summary**	**Phase 5: running a topic modeling analysis**
Step 2.a: gathering the citing entities **Description:** Identifying the list of entities citing a retracted publication and storing their main metadata **Input:** the retracted publications **Output:** For each citing entity: (2.a.1) *DOI,* (2.a.2) *year of publication,* (2.a.3) *title,* (2.a.4) *venue id (ISSN/ISBN),* (2.a.5) *venue title*	**Step 3.a: extracting textual values from the citing entities** **Description:** Extracting the citing entities' abstracts and, for each in-text citation (i.e. the location where the citing article cites another work within its content) referencing to a retracted publication, the in-text reference pointer (i.e., the textual device denoting a bibliographic reference such as "[1]"), the textual context of the citation, and the title of the section where it appears. **Input:** The citing entities main metadata **Output:** For each citing entity: (3.a.1) *abstract,* (3.a.2) *in-text citation sections,* (3.a.3) *in-text citation contexts,* (3.a.4) *in-text reference pointers*	**Step 4.a: analyzing the citing entities** **Description:** Inferring some descriptive statistics from the data of the citing entities gathered in the previous phases **Input:** The citing entities dataset (produced after phases 2 and 3) **Output:** (4.a.1) a collection of charts to summarize the descriptive statistics of the citing entities	**Step 5.a: analyzing the abstracts of the citing entities** **Description:** Automatically extracting the topics through an analysis of the abstracts of the citing entities **Input:** The abstracts of the citing entities **Output:** (5.a.1) a dataset containing the N most important keywords of each topic, (5.a.2) a dataset containing a list of all the documents of the corpus and their representativeness against each topic, (5.a.3) a collection of dynamic visualizations to observe and investigate the topic modeling results
Step 2.b: marking the retracted entities **Description:** Annotating whether any of the citing entities has been or has not been retracted **Input:** The main metadata of the citing entities **Output:** For each citing entity: (2.b.1) *is / is not retracted*	**Step 3.b: annotating the in-text citations characteristics** **Description:** Annotating the intent (why an article is cited, e.g. because it *uses a method* defined in the cited article) and sentiment (positive, neutral, negative) of each in-text citation, and specifying whether the text of the citation context mentions the retraction of the cited article **Input:** In-text citations and their contexts **Output:** For each in-text citation: (3.b.1) *citation intent,* (3.b.2) *citation sentiment,* (3.b.3) *retraction is / is not mentioned*	**Step 4.b: analyzing the in-text citations** **Description:** Inferring some descriptive statistics from the data of the in-text citations gathered in the previous phases **Input:** The citing entities dataset (produced after phases 2 and 3) **Output:** (4.b.1) a collection of charts to summarize the descriptive statistics of the in-text citations	**Step 5.b: analyzing the in-text citation contexts** **Description:** Automatically extracting the topics through an analysis of the in-text citation contexts **Input:** The in-text citation contexts **Output:** (5.b.1) a dataset containing the N most important keywords of each topic, (5.b.2) a dataset containing a list of all the documents of the corpus and their representativeness against each topic, (5.b.3) a collection of dynamic visualizations to observe and investigate the topic modeling results
Step 2.c: classifying the citing entities into subject areas and subject categories **Description:** Classifying the citing entities into areas of study and specific subjects **Input:** The venues of the citing entities **Output:** For each citing entity: (2.c.1) *subject area,* (2.c.2) *subject category*			

3.2 An open science methodology

Open science is a movement to make science and scientific research, including its products (e.g., publications, data, software, etc.) and its dissemination accessible to all levels of society (i.e., amateur, or professional) with no restrictions. It is about bringing *"socio-cultural and technological change, based on openness and connectivity, on how research is designed, performed, captured, and assessed."* (Vicente-Saez & Martinez-Fuentes, 2018).

Our methodology is conceived to be compliant with the principles of open science. The adopted tools, the generated data and the produced outcomes should all be open and accessible. This is an essential requirement to foster the reproducibility of the process and the reproduction of the obtained results.More precisely, our approach is meant to follow the FAIR (Findable, Accessible, Interoperable, Reusable) principles (Wilkinson et al., 2016). These principles emphasize machine-actionability (i.e., computational systems can find, access, interoperate, and reuse data without the need of a human intervention). The need for FAIR data is continually increasing in conjunction with the increasing rely on computational support to deal with large data volume, complexity, and creation speed of data.

Citation analysis is crucial to our work; thus, it is important that the principles of open science are correctly applied in this analysis too. The citations gathered and analyzed in the methodology should be "open". Following the definition provided in (Peroni & Shotton, 2018b). A bibliographic citation is an open citation when the data needed to define the citation are:

- Structured: expressed in one or more machine-readable format such as JSON (a popular open standard file format to store and transmit data objects based on attribute–value pairs).
- Separate: available without the need to access the source article in which the citation is defined.

- Open: freely accessible and reusable without restrictions.
- Identifiable: the entities linked by an open citation must be clearly identified by using a specific persistent identifier scheme, such as a DOI, or a URL.
- Available: it must be possible to obtain the basic metadata of the entities involved in the citation by resolving their identifiers.

Open citations are provided by services such as Crossref: a nonprofit association with the stated scope of *"promote the development and cooperative use of new and innovative technologies to speed and facilitate scientific and other scholarly research"* (Hendricks et al., 2020), or by OpenCitations (Peroni & Shotton, 2020): an independent infrastructure organization for open scholarship, which is dedicated to the publication of open citation data using Semantic Web technologies (Berners-Lee et al., 2001), which facilitate the use of precise semantics for the encoding and creation of machine-processable data on the Web. The next section will have a deeper analysis of these services and others.

Text analysis and in particular the topic modeling analysis is yet another important operation of this methodology. We want to apply the open science principles to this analysis too, therefore use mechanisms that allow sharing the adopted topic modeling analysis workflow with other researchers to foster the reproducibility of the process. In addition to that, since the workflow must be reviewed and understandable by any researcher (with no relation to its field of study and background) we also want this process to be understandable and modifiable by others that lack of programming skills, or the ability to understand programming languages to a certain extent. Considering these aspects, we decided to build our own system to be compliant with such needs. We developed MITAO, a User friendly and modular software for textual analysis (Heibi et al., 2021; Ferri et al., 2020). Chapter 5 is entirely dedicated to the fifth phase of the methodology and MITAO.

3.3 Datasets and services

This section walks through and defines the main external (i.e., not implemented expressly for the work of this thesis) datasets and services. The various elaborations of the methodology may be run by combining automated computational approaches with manual elaboration. Our plan is to clearly devise and describe which approach (i.e., computational vs manual) to use when running the various steps of the methodology. This strategy fosters the possibility of a forthcoming and separated document which describes all the steps of the methodology that could be published in Protocols.io (https://www.protocols.io/) – an open-access platform for detailing, sharing, and discussing molecular and computational protocols that can be useful before, during, and after publication of research results (Teytelman et al., 2016). In particular, the Protocols.io document includes the automatable operations as scripts, while the manual operations are formalized as a set of conditions which can guide, simplify, and limit the ambiguities of the user on its annotation.

All the data gathered/produced through the execution of the methodology and the plotted visualization will be published in Zenodo (https://zenodo.org/) – a general-purpose open-access repository developed under the European OpenAIRE program and operated by CERN (European Organization For Nuclear Research, & OpenAIRE, 2013).

The first phase of the methodology defines the collection of retracted publications to be used. Our approach is in line with the strategy of several past studies (discussed in Chapter 1 and Chapter 2), thus an appropriate service to use in this case is Retraction Watch (http://retractionwatch.com/) which keeps track and collects retractions of scholarly articles. Using the Retraction Watch database (http://retractiondatabase.org/), we can look for and retrieve the retracted articles we want to use in our analysis. The Retraction Watch database keeps track of several information for each retracted article. In the context of our methodology, we need to keep note of (a) the DOI of the original publication (if any), (b) the year of publication, (c) the authors, (d) the subjects ("History", "Philosophy", etc.) and (e) the year of the retraction notice. A deeper discussion toward the use of Retraction Watch is done in the next chapter while presenting Phase 1 (Section 4.1).

To gather the entities which have cited the retracted publications, the methodology relies on three main services: Microsoft Academic Graph (MAG) (Wang et al., 2020), OpenCitations (Peroni & Shotton, 2020) and Crossref (Hendricks et al., 2020). The Semantic Web technologies used by OpenCitations permit the publication of bibliographic and citation data as Linked Open Data (LOD) (Bizer et al., 2018). These bibliographic and citation data are compliant with the OpenCitations Data Model (Daquino et al., 2020), which is implemented by means of the SPAR Ontologies (http://www.sparontologies.net) (Peroni & Shotton, 2018a). In particular, citations are described using Citation Typing Ontology (CiTO, http://purl.org/spar/cito) (Peroni & Shotton, 2012), which allows one to create metadata describing citations (that are distinct from the metadata describing the cited works themselves) and permits the *intent* of an author when referring to another document to be captured. OpenCitations provides COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations (http://opencitati ons.net/index/coci) (Heibi et al., 2019), which contains details regarding all the citations that are specified by the open references to DOI-identified works present in Crossref.

The Initiative for Open Citations (I4OC, https://i4oc.org) has dedicated the past four years to persuading publishers to provide open citation data by means of Crossref, obtaining the release of the reference lists of more than 50 million articles (as of October 2021), and it is this change of behavior by the majority of academic publishers that has permitted COCI to be created.

COCI and CiTO are two of the main components of the methodology presented in this thesis. COCI is used to gather the citations of the retracted article and CiTO is adopted for the definition and characterization of the citation intent of the citation occurrences inside the body of the citing document. It is based on a manual analysis of the citation context and examination of a guiding schema which fosters the facility of the citation intent selection and reduces the ambiguities that could potentially arise during such process.

MAG is a knowledge graph that contains the scientific publication records, citations, authors, institutions, journals, conferences, and fields of study. It also provides a free REST API service to search, filter, and retrieve its data (it is worth mentioning that Microsoft announced that it will shut down MAG by the end of 2021).

Several operations in the methodology will have the need to classify the items according to their fields of study. Two main classification systems that offer this possibility – the Scimago Journal Rank (SJR, https://www.scimagojr.com/)[6] for journal articles and the Library of Congress Classification (LCC, https://www.loc.gov/catdir/cpso/lcco/) for books/book chapters.

The Scimago Journal Rank groups the journals into 27 main subject areas (medicine, social sciences, computer science, etc.) and 313 subject categories (for medicine, psychiatry and anatomy, for social sciences, law and political science, etc.). These values define two different levels: (1) a macro layer for the subject area, and (2) a lower layer for the specific subject category. Using the query service provided by Scimago at https://www.scimagojr.com/, we can search for the venue of a given journal article to assign the subject area(s) and subject category(s).

The Library of Congress Classification (LCC) is a system of library classification, it assigns a LoC number, the first part is alphabetically, and the rest are numbers (e.g., E312). The letters are assigned to different subjects, and those subjects are further divided into two letters, and then into numbers. E.g., H is social sciences, HA is statistics and HA 29-32 is theory and method of social science statistics. Services such as ISBNDB (https://isbndb.com/) could be used to look up for the Library of Congress Classification code (LCC, https://www.loc.gov/catdir/cpso/lcco/) for a particular ISBN identifier of a given book.

[6] SCImago, SJR — SCImago Journal & Country Rank [Portal]. http://www.scimagojr.co

3.4. Terminology

This section lists and defines the terms used in the methodology and in the rest of the thesis to support the overall organization of the information in the methodology.

For simplicity, throughout our discussion, we refer to the set of the retracted articles (initialized in the first phase of the methodology) with the abbreviation RET-SET. We define four distinct events that concern a retracted publication (happening in a specific year):

- *E-RetPub*, i.e., the publication of the retracted article
- *E-PR*, i.e., its (possible) first partial retraction
- *E-FR*, i.e., its full retraction
- *E-LastCit*, i.e., the last citation it received.

In addition, we define the following five periods by combining the events mentioned above:

- P0, either [*E-RetPub*, *E-PR*[or [*E-RetPub*, *E-FR*[– from the year of publication of the retracted article to the year before the one of the first retraction notice, i.e., either from *E-RetPub* (inclusive) to *E-PR* (exclusive) or from *E-RetPub* (inclusive) to *E-FR* (exclusive) in case the retracted item does not have a partial retraction or if the partial retraction date coincides with the full retraction date;
- P1, [*E-PR*, *E-PR*] – the year of the first partial retraction
- P2,]*E-PR*, *E-FR*[– from the year after the first partial retraction to the year before the full retraction, i.e., between *E-PR* and *E-FR*, exclusive
- P3, [*E-FR*, *E-FR*] – the year of the full retraction

- P4,]*E-FR*, *E-LastCit*] – from the year after the full retraction to the year of the last citation received by the retracted article, i.e. from *E-FR* (exclusive) to *E-LastCit* (inclusive).

We use the annotation PERIOD-SET when referring to the above set of periods. Any retracted item in the RET-SET belong to one of these two categories: (RET-A) those that have received one partial retraction at least one year before being fully retracted, and (RET-B) those which have received only a full retraction with no partial retractions. The retracted items in RET-B do not have *P1* and *P2* specified and *P0* is set as [*E-RetPub*, *E-FR*[– i.e., its PERIOD-SET is equal to {*P0*, *P3*, *P4*}.

Each citing entity has two corresponding values:

- P_{CIT} is the list of the years included in the period PX (among P0-P4) in which the citing entity has been published
- P_{CUT} is a number between -1 and +1 which identifies if the publication date of the citing entity (and, consequently, the date of the citation it makes to the retracted article) is closer to the left margin of P_{CIT} (i.e., when P_{CUT} is close to -1) or if it is closer to the right margin of P_{CIT} (i.e. when P_{CUT} is close to +1).

Considering the years [Y_1, Y_2, ..., Y_{i}, $Y_{\mathrm{i+1}}$, ..., Y_{f}] in P_{CIT}, the formula to compute P_{CUT} is defined as follows:

$$P_{CUT}(citing) = \frac{(citing_{date} - Y_1) - \left(Y_f - citing_{date}\right)}{|P_{CIT}| - 1}$$

In practice, the P_{CUT} of a citing entity is equal to the difference of the distance in years between the publication date of the citing entity and the first year in P_{CIT} and the distance in years between the publication date of the citing entity and the final year in P_{CIT}, over the number of years in P_{CIT} minus one. It is worth noticing that if the publication year of the citing entity is less than the publication year of the cited retracted article (while rare, this situation may happen when the citing article cites the retracted one when the latter has not been formally published yet – but its text was available as, for instance, a preprint), then $citing_{\mathrm{date}}$ should be rounded up to the publication year of the cited retracted article.

For instance, let us consider a citing entity A published in 2010 that cites a retracted article R published in 2002, which has received a partial retraction notice in 2008 and a full retraction notice in 2012. Since the publication date of A is included in P2, $P_{\mathrm{CIT}}(A)$ is [2009, 2010, 2011] and $P_{\mathrm{CUT}}(A)$ is 0 (i.e., 0 /2), meaning that the citation that A made to R happened in the middle of P2. Similarly, if another citing entity B published in 2009 cites R, then $P_{\mathrm{CIT}}(B)$ is equal to $P_{\mathrm{CIT}}(A)$ and $P_{\mathrm{CUT}}(B)$ is -1 (i.e., -2/2), meaning that the citation that B made to R happened in the initial part of P2. It is worth mentioning that the computation of P_{CUT} is possible only for a P_{CIT} that includes more than one year. Thus, according to the formula P_{CUT} is undefined in P1 and P3 and when the P_{CIT} calculated in P0, P2, and P4 contains only one year. In these cases, we assign the value 0 to P_{CUT}, in practice, this means that the corresponding citations are collocated in the middle of the periods.

Using this approach, we can calculate P_{CIT} and P_{CUT} for each citing entity we have collected. These two metrics let us treat all the citing entitles consistently independently from their specific publication dates and the other dates related to the retracted articles they cite. Thus, we outline a general picture of the entire PERIOD-SET by mapping all the citing entities using the P_{CIT} and P_{CUT} values.

Chapter 4

Data collection, features annotation and descriptive statistics

This chapter walks through the first four phases of the methodology. Each phase is treated in a separate section following their corresponding order. The first section (i.e., defining a case study/domain) discusses the first phase of the methodology by sketching out the procedure to follow in order to define case study and collect the retractions to be processed using the methodology (i.e., RET-SET). Once the case study is clearly pictured and the RET-SET is defined we move to discuss the next phases: (second section) identifying, retrieving, and extracting the basic metadata of the entities which have cited any retracted article in RET-SET, (third section) extracting and labeling additional features based on the textual content of the citing entities, and (fourth section) building a descriptive statistical summary to represent the values of the features annotated in the previous two phases.

4.1 Phase 1: defining a case study/domain

The aim of this phase is defining a collection of retracted publications (i.e., RET-SET) to be further analyzed and processed in the next phases of the methodology. The set of the retracted publications RET-SET to be defined can contain from one to an arbitrary number of retracted publications. On one hand, using only one retracted publication, could be due to the fact that we are interested in one specific popular retracted document. On the other hand, a collection of several retracted articles might have in common a specific field of study (e.g., computer science, health science, social sciences, etc.) or a macro area (e.g., STEM), a particular venue (e.g., journal or proceeding of a conference), a specific period (e.g., from 2000 to 2010), a country (e.g., China), an affiliation (e.g., a specific university), etc.

Each retracted publication included in the RET-SET should be characterized with: (1) a DOI (if any), (2) the author(s), (3) the year of publication (i.e., *E-RetPub*), the year of the first partial retraction (if any, i.e., *E-PR*), the year of full retraction (i.e., *E-FR),* the year of the last citation received (i.e., *E-LastCit*), the subject area, and the subject category.

Our methodology uses the Retraction Watch Database to identify and gather all the retracted publications that follow any of the criteria of selection we have established. The query service enables a filtering toward the value of several attributes including the ones used in the selection of the retraction set. From the results returned by Retraction Watch we can infer all the characteristics that we need out of the year of the last citation received (*E-LastCit*). We will have enough information to annotate *E-LastCit* after the elaboration of the next phase (i.e., Phase 2).

Retraction Watch assigns to each retracted publication a macro area and a category of study. One retracted publication can have multiple areas/categories of study yet is difficult to determine the most representative area/category for that specific retracted item. For instance, a retracted article R in the Retraction Watch Database may be labeled with the areas of study “Heath Sciences (HSC)” and “Social Sciences (SOC)” so it is difficult to determine whether R is closer to “Heath Sciences (HSC)” or “Social Sciences (SOC)”, in some cases either “Social Sciences (SOC)” or “Heath Sciences (HSC)” could be marginally represented in the retracted item.

If our case study is based on the analysis of a specific domain, then this last case might generate an unwanted bias for the rest of the analysis; thus, it is important to be aware of and handle these situations. Throughout our experience with Retraction Watch, we have noticed some concrete examples of this problematic case. For instance, the following retracted article *"The Good, the Bad, and the Ugly: Should We Completely Banish Human Albumin from Our Intensive Care Units?"* (Boldt, 2000). In Retraction Watch, the subjects associated with it are Medicine and Journalism. However, when we checked the full-text of the article, we noticed that the content related to Journalism is very limited and, as such, the article could represent a bias and a source of error if considered in the RET-SET.

To avoid this problematic situation, we devised a mechanism which helps evaluate the affinity of each retracted item in RET-SET to the domain *D* taken in consideration for the study. We assign to each retracted item in RET-SET an initial score of 1, named *domain_affinity* – this value ranges from 0 (i.e., very low) to 5 (i.e., very high). The final value of *domain_affinity* for each retracted item is calculated as follows:

1. We assign to each retracted item additional subject categories obtained by searching the venue where it was published in external databases – using the Scimago classification (https://www.scimagojr.com/) for journals and the Library of Congress Classification (LCC, https://www.loc.gov/catdir/cpso/lcco/) for books/book chapters.
2. If both the Retraction Watch subjects and those gathered in (1) included at least one subject identifying a discipline under the domain D, we add 1 to *domain_affinity* of that item.
3. If all the Retraction Watch subjects referred to disciplines under the domain D, we add another 1 to *domain_affinity* of that item.
4. If the title of the retracted item has a clear affinity to D (e.g., "The Origins of Probabilism in Late Scholastic Moral Thought" has a high affinity to humanities and history), we add another 1 to *domain_affinity* of that item.

5. Finally, we provide a subjective evaluation ranging from -1 to 1 based on a manual review of the abstract of the item.

Once the *domain_affinity* value is calculated for each retracted item in RET-SET, we can decide a threshold used to filter/exclude the unwanted items from RET-SET. For instance, one can decide to analyze the retractions that have a high affinity to computer science by establishing a threshold value of 4, i.e., *domain_affinity* >= 4.

4.2 Phase 2: identifying, retrieving, and characterizing the citing entities

This section is dedicated to the illustration and discussion of the second phase of the methodology. Starting from the RET-SET, this phase first identifies all the entities which have cited it and characterize them with their main metadata, such as the title and the year of publication. Then, we check whether any of the citing entities has been or has not been retracted as well, and finally, we classify the citing entities according to their areas of study and specific subject categories, following the Scimago classification.

The output of this phase is a dataset containing the citing entities (the records) which have cited the RET-SET, with their main metadata (the columns). The same dataset will be further populated with additional data/features on the next phase.

4.2.1 Step 2.a: gathering the citing entities

To get the list of the entities which have cited the RET-SET, we rely on and query only open repositories storing scholarly bibliographic data compliant with the open science principles. Notable examples in this category are: Microsoft Academic Graph (MAG, https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/) (Wang et al., 2020), Crossref (https://www.crossref.org/) (Hendricks et al., 2020), and COCI provided by OpenCitations (http://opencitations.net/) (Peroni & Shotton, 2020). OpenCitations provides a free APIs service to query and retrieve the COCI data at http://opencitations.net/index/coci/api/v1. Crossref makes available a REST API service (https://api.crossref.org) that exposes the metadata that Crossref members (i.e., publishers) deposit in it, such as basic descriptive metadata of publications, funding data, license information, full-text links, ORCIDs, reference lists, etc. Users can search and filter publication information contained in Crossref, that are returned by the API in JSON. MAG is a knowledge graph which contains the scientific publication records, citations, authors, institutions, journals, conferences, and fields of study. MAG provides a free REST API service (https://www.microsoft.com/en- us/research/project/academic-knowledge/) to search, filter and retrieve such data.

These are only some notable examples, additional services (which are open) could be integrated and used in this phase. The choice of the most suitable services relies on the nature of our analysis. For instance, if we are only interested in collecting the citing entities having a DOI value, we might decide to use only COCI which stores only DOI-to-DOI citations. We need to make sure that the adopted service can return the list of entities which have cited a given set of articles (i.e., the RET-SET) and that some basic metadata of such citing entities can be retrieved, in particular: (a) the DOI value, (b) the year of publication, (c) the title of the article, (d) the ID (ISSN/ISBN) of the publication venue, and (e) the title of the publication venue. For instance, the COCI API provides a specific operation, i.e., http://opencitations.net/index/coci/api/v1#/metadata/, which takes a DOI of an entity as input and returns its main metadata, including the citing and cited DOIs of such entity.

If any of the collected entities is a bibliography, retraction notification, a presentation, or data repository, then it should not be considered in the analysis. Indeed, these items have a limited (or a totally insignificant) impact for the quantitative and qualitative analysis of the methodology, and it might also produce noise in the data and results, e.g., a retraction notification document could cite the retracted article, although it must not be included in the analysis with the other citing entities. Apart from these publication types that are excluded from the analysis, the methodology takes into consideration several publication types – journal articles, books, book chapters, conference papers, pre-prints, theses, editorials, etc.

Finally, before moving to the next step, now that we gathered the citing entities of the RET-SET and annotated information regarding the year of publication of each one of them (i.e., the year of the citation) we can fill the value of the year of the last citation received (i.e., *E-LastCit*) of each retracted publication in RET-SET which have been omitted in the previous phase.

4.2.2 Step 2.b: marking the retracted entities

We look over all the citing entities of the dataset and manually check them in the Retraction Watch database using their DOIs, to mark them as retracted if included in such a database. In case a citing entity does not have a corresponding DOI, we search for any other metadata attribute, e.g., the title or the author. The corresponding entity will be marked with "yes" only if it has received a full retraction notice.

4.2.3 Step 2.c: classifying the citing entities according to subject areas and subject categories

At this step we want to annotate the subject areas and subject categories of each citing entity in the dataset. To do this we consider the titles and IDs (either ISSN or ISBN) of the venues and classify them into specific subject areas and subject categories using the Scimago Journal Classification (https://www.scimagojr.com/).

We map each venue of the dataset having an ISSN value into its corresponding area and category following the values specified in the Scimago journal classification. This process is done in a semi-automatic way. If we have an ISSN associated to a citing entity, this check can be automated by matching such value with the venues index of Scimago (https://www.scimagojr.com/journalrank.php). Otherwise, we need to manually look for the venue title (and any other additional information, e.g., publisher) using the Scimago Journal Rank service at https://www.scimagojr.com/. In case that journal is assigned with more than one subject area or subject category, we will take into consideration all these values. In case we did not find any corresponding value in Scimago for a specific venue, then we search for it using an external source (such as JSTOR at https://www.jstor.org), we annotate it with only one major subject area which is the most consistent with our findings, and we use the same value of the subject area followed by the postfix "(miscellaneous)" (following the Scimago schema) to annotate the subject category.

We also need to classify books and book chapters. We use the ISBNDB service (https://isbndb.com/) to look up for the related Library of Congress Classification (LCC, https://www.loc.gov/catdir/cpso/lcco/). Then, we map the LCC categories found into Scimago subject areas as follows:

1. We consider only the starting alphabetic segment of the LCC code and find the corresponding LCC discipline using a pre-built lookup index (Heibi & Peroni, 2021b). For instance, for "RC360" we consider "RC" (i.e., Medicine).

2. We check whether the value of the LCC subject matches the exact value of a Scimago area using a pre-built Scimago mapping index (Heibi & Peroni, 2021b). If the corresponding value is present, we annotate the subject area with such a value, while the subject category will have the same value with the addition of suffix "(miscellaneous)". In case no corresponding Scimago area is found, we continue to point 3, otherwise we stop.

3. We check whether the value of the LCC subject is a Scimago category using the same pre-built Scimago mapping index. If the corresponding value is present, we annotate the corresponding category with such value, while the area will have the same value used in the

Scimago classification to denote the macro area of such a category. In case no corresponding Scimago category is found, we continue to point 4, otherwise we stop.

4. The ISBNs that were not processed in the previous steps need to be manually annotated through the consultation of the complete LCC index (http://www.loc.gov/catdir/cpso/lcco/). In case we did not find any entry in the LCC index for the corresponding ISBN value, we continue to point 5, otherwise we stop.

5. We manually search for the entity using any of its available metadata (e.g., title) and we annotate its subject area with one major category (the one we found to be the most suitable). The subject category is annotated with the same value followed by the word "(miscellaneous)".

4.3 Phase 3: Extracting and labeling content-based features

This phase requires accessing the full text of the citing entities. In case the full text is not accessible at all, e.g., due to paywalls restrictions, the corresponding entity will still be part of the dataset, but it will not take part in some of the quantitative analysis of the fourth phase (such as those related to the in-text citations), and it will be completely excluded from the textual analysis of the fifth phase.

The process is divided in two main steps: (a) extracting the textual values from the full-text of the citing entities, and (b) characterizing the in-text citations. More precisely, once we have access to the full-text of a citing entity, we will first extract the article abstract, then the section title and its type, the in-text reference pointer, and context of its in-text citations to the RET-SET. In the second part of this step, we will characterize each in-text citation with three main features: the intent, the sentiment, and whether the context mentions or does not mention the retraction of the cited retracted article.

4.3.1 Step 3.a: extracting textual values from the citing entities

From the full text of the citing entities, we first need to extract the abstracts. Some of the entities might lack an abstract – for example, if we consider the editorials or some book chapters. In this case the abstract slot will be left as empty. Then, we focus on the in-text citations: for each in-text citation to RET-SET, we extract and annotate its in-text reference pointer, context, and the title of the section where it appears in.

We define the in-text citation context as the sentence that contains the in-text reference pointer to the bibliographic reference of any article in the RET-SET (i.e., the anchor sentence), plus the preceding and following sentences. There are some exceptions which need to be handled differently, according to where the in-text reference pointer has been specified:

- In a title of a section – the citation context is the entire title
- In a cell of a table – the citation context is the entire cell
- In the first sentence of a section – the citation context is the anchor sentence plus the following sentence
- In the last sentence of a section – the citation context is the anchor sentence plus the previous sentence.

The sections containing the in-text citation are specified according to their type – using the categories “introduction”, “method”, “abstract”, “results”, “conclusions”, “background”, and “discussion” listed in Suppe (1998). These categories are used when the intent of the section is clear from its title, otherwise we use other three residual categories, i.e., “first section”, “final section” and “middle section” combined with the original title of the section. If the examined full text of the citing entity is not organized into sections (for instance, when we consider some editorials), then the value of its in-text citation section is set to “None”. The section title to associate with an in-text citation should always be the heading of the first-level section, even if the in-text citation appears in a lower subsection. For instance, if a citation occurs inside the Section 2.1, then the heading to consider is the one of Section 2.

4.3.2 Step 3.b: annotating the in-text citations characteristics

The characteristics we want to analyze are inferred from the observation of the citation context of each in-text citation gathered in the previous step. We examine each in-text citation retrieved and annotate:

- the author's sentiment regarding the cited entity of the RET-SET
- whether at least one citation context of a particular citing entity does explicitly mention the fact that the cited entity has been retracted, i.e., the citation context contains the word “retract” or one of its derivative words – “retractions”, “retracted”, etc.
- the citation intent (or citation function), defined as the authors’ reason for citing a specific article (e.g., the citing entity *uses a method* defined in the cited entity)

To specify the citation sentiment, we follow the classification proposed in the work of Bar-Ilan and Halevi (2017). Thus, we annotate each in-text citation with one of the following values:

- *positive*, when the retracted article was cited as sharing valid conclusions, and its findings could have been also used in the citing study
- *negative*, if the citing study cited the retracted article and addressed its findings as inappropriate and/or invalid
- *neutral*, when the author of the citing article referred to the retracted article without including any judgment or personal opinion regarding its validity

The importance of citation intent was acknowledged by several past studies, in general and regarding the analysis of the retraction phenomenon. For instance, the work of Chen et al., (2013) pointed out the importance of distinguishing whether citations made by authors are unknowingly building their work on false claims or are fully aware of the problems that led to the retraction of the article that they are citing.

To annotate the citation intent, we use the citation functions specified in the Citation Typing Ontology (CiTO, http://purl.org/spar/cito) (Peroni & Shotton, 2012), this ontology is used for the characterization of factual and rhetorical bibliographic citations. Although an in-text citation might refer to more than one citation function at the same time, we annotate each in-text citation with one citation function only. This decision has been taken in order to avoid possible ambiguities when analyzing these values.

To annotate the in-text citation intent we use a decision model which serves as a guiding scheme for the annotator, as it is represented in Figure 4.1. This decision model is based on a priority ranked strategy that works as follows:

1. we match each in-text citation against at least one of the three macro-categories, i.e., "Reviewing ...", "Affecting ..." and "Referring ..." (first row in Fig 2);

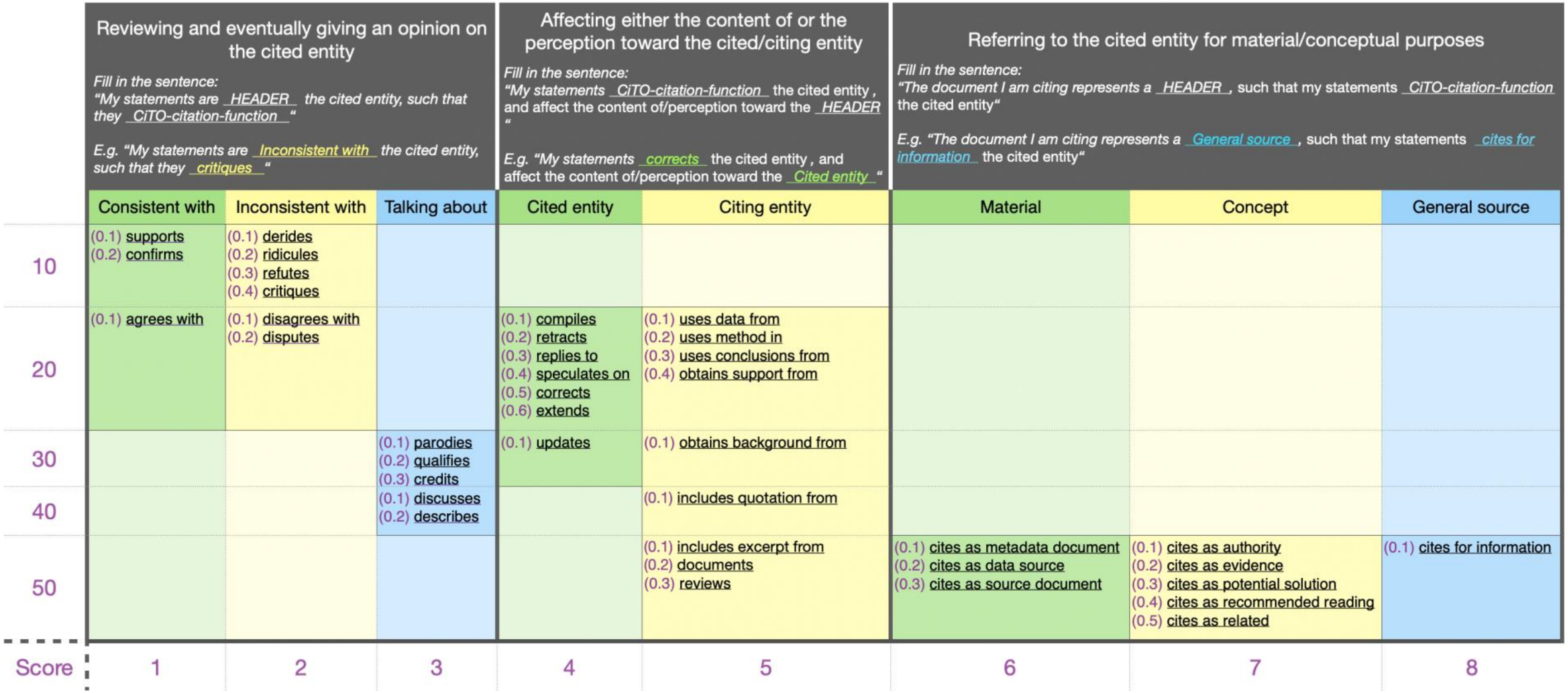
2. for each macro-category we have selected in (1), we choose one or more citation functions from those provided by CiTO;

3. in case we select only one citation function, then we annotate the in-text citation intent with such a value; otherwise

4. we calculate the priority of each citation function we have selected by summing its value in parenthesis (from 0.1 to 0.6) with the corresponding value in the y-axis (from 10 to 50) and in the x-axis (from 1 to 8) shown in Fig 2. The smaller the sum, the more priority the citation

function has. For instance, the priority of the citation function “confirms” is 11.2 that is higher than the one of the citation function “describes”, which is 43.2.

5. Finally, we select the citation function that has higher priority and annotate the in-text citation function with it.



Reviewing and eventually giving an opinion on the cited entity

Fill in the sentence:
"My statements are _HEADER_ the cited entity, such that they _CiTO-citation-function_ "

E.g. "My statements are _Inconsistent with_ the cited entity, such that they _critiques_ "

Affecting either the content of or the perception toward the cited/citing entity

Fill in the sentence:
"My statements _CiTO-citation-function_ the cited entity , and affect the content of/perception toward the _HEADER_ "

E.g. "My statements _corrects_ the cited entity , and affect the content of/perception toward the _Cited entity_ "

Referring to the cited entity for material/conceptual purposes

Fill in the sentence:
"The document I am citing represents a _HEADER_ , such that my statements _CiTO-citation-function_ the cited entity"

E.g. "The document I am citing represents a _General source_ , such that my statements _cites for information_ the cited entity"

	Reviewing and eventually giving an opinion on the cited entity			Affecting either the content of or the perception toward the cited/citing entity		Referring to the cited entity for material/conceptual purposes		
	Consistent with	Inconsistent with	Talking about	Cited entity	Citing entity	Material	Concept	General source
10	(0.1) supports (0.2) confirms	(0.1) derides (0.2) ridicules (0.3) refutes (0.4) critiques						
20	(0.1) agrees with	(0.1) disagrees with (0.2) disputes		(0.1) compiles (0.2) retracts (0.3) replies to (0.4) speculates on (0.5) corrects (0.6) extends	(0.1) uses data from (0.2) uses method in (0.3) uses conclusions from (0.4) obtains support from			
30			(0.1) parodies (0.2) qualifies (0.3) credits	(0.1) updates	(0.1) obtains background from			
40			(0.1) discusses (0.2) describes		(0.1) includes quotation from			
50					(0.1) includes excerpt from (0.2) documents (0.3) reviews	(0.1) cites as metadata document (0.2) cites as data source (0.3) cites as source document	(0.1) cites as authority (0.2) cites as evidence (0.3) cites as potential solution (0.4) cites as recommended reading (0.5) cites as related	(0.1) cites for information
Score	1	2	3	4	5	6	7	8

Figure 4.1. The decision model for the selection of a CiTO citation function to use for the annotation of the citation intent of a an examined in-text citation based on its context. The first large row contains the three macro-categories: (1) "Reviewing ...", (2) "Affecting ...", and (3) "Referring ...". Each macro-category has at least two subcategories, and each subcategory refers to a set of citation functions. The first row defines what are the citation functions suitable for it through the help of a guiding sentence which needs to be completed according to the chosen sub-category and citation function

4.4 Phase 4: Descriptive statistics

As a result of phases two and three of our methodology, we produce a dataset containing all the entities which have cited the RET-SET accompanied by their gathered information (basic metadata, textual content, citation intent, etc.). In Table 4.1, we show a summary of all the features contained in the dataset, accompanied by some examples. Based on the collected data, the phase presented in this section generates a set of charts which represent a descriptive overview of all the entities gathered.

Table 4.1. A summary of all the features included in the dataset, generated after the execution of phases two and three of our methodology. The steps of the two phases are mentioned in separated cells. Each step adds a list of features in the dataset (the first column of the inner tables in each cell). An example of the expected value is given next to each feature (the second column of the inner tables in each cell)

<table>
<tr><th>Phase 2: identifying, retrieving, and extracting basic metadata of the entities which have cited RET-SET</th><th>Phase 3: extracting additional data and labeling the textual content of the citing entities</th></tr>
<tr><td>
<b>Step 2.a</b>
<table>
<tr><td><b><i>DOI:</i></b> the DOI of the citing article</td><td><i>None</i></td></tr>
<tr><td><b><i>year of publication:</i></b> the year of publication of the citing article</td><td><i>2021</i></td></tr>
<tr><td><b><i>title:</i></b> the title of the citing article</td><td><i>A qualitative and quantitative citation analysis toward retracted articles: a case of study</i></td></tr>
<tr><td><b><i>venue id (ISSN/ISBN):</i></b> the ID (ISSN/ISBN) of the venue of publication of the citing article</td><td><i>1588-2861</i></td></tr>
<tr><td><b><i>venue title:</i></b> the title of the venue of publication of the citing article</td><td><i>Scientometrics</i></td></tr>
</table>
<b>Step 2.b</b>
<table>
<tr><td><b><i>is / is not retracted:</i></b> a yes/no value depending on whether the citing article has (“yes”) or has not (“no”) received at least one retraction notification.</td><td>no</td></tr>
</table>
</td><td>
<b>Step 3.a</b>
<table>
<tr><td><b><i>abstract</i></b><i>:</i> the abstract of the citing article</td><td><i>“In this article, we show the results of a quantitative and qualitative ... “</i></td></tr>
<tr><td><b><i>in-text citation section</i></b><i>:</i> the kind of section in the citing entity which includes the in-text citation, taken from the list in the work of Suppe (1998)</td><td><i>Introduction</i></td></tr>
<tr><td><b><i>in-text citation context</i></b><i>:</i> the textual context in the citing entity which includes the in-text citation</td><td><i>“... to those introduced in the latter set of studies. In particular, we want to focus on a highly cited retracted article, i.e. [1], that suggested ...”</i></td></tr>
<tr><td><b><i>in-text reference pointer</i></b><i>:</i> the string representing the in-text reference pointer (e.g., “James et al. 2020”) to the retracted article</td><td><i>[1]</i></td></tr>
</table>
</td></tr>
<tr><td>
<b>Step 2.c</b>
<table>
<tr><td><b><i>subject area:</i></b> the subject areas of the venue of publication of the citing article, based on the Scimago Journal Classification</td><td>+ <i>Computer Science</i>
+ <i>Social Science</i></td></tr>
<tr><td><b><i>subject category:</i></b> the subject categories of the venue of publication of the citing article, based on the Scimago Journal Classification</td><td>+ <i>Computer Science Applications</i>
+ <i>Library and Information Sciences</i>
+ <i>Social Sciences (miscellaneous)</i></td></tr>
</table>
</td><td>
<b>Step 3.b</b>
<table>
<tr><td><b><i>citation intent:</i></b> the citation intent of each in-text citation in the citing entity, according to the citation functions defined in CiTO</td><td><i>uses data from</i></td></tr>
<tr><td><b><i>citation sentiment</i></b><i>:</i> the sentiment of each in-text citation in the citing entity, classified as positive/negative/neutral, conveyed by the citation context of an in-text citation</td><td><i>neutral</i></td></tr>
<tr><td><b><i>retraction is / is not mentioned</i></b><i>:</i> a yes/no value that indicates if at least one of the citation contexts of the citing article explicitly mentions (“yes”) or does not mention (“no”) the fact that the cited entity is retracted</td><td>yes</td></tr>
</table>
</td></tr>
</table>

4.4.1 Step 4.a: citing entities

We want to analyze the distribution of the citing entities as a function of two features: (D1) the period P_x characterizing P_{CIT} and the related P_{CUT}, and (D2) their subject areas. These distributions are built as a function of the values in the PERIOD-SET. The results should also highlight, visually, the citing entities that have/have not mentioned the retraction.

Considering D1 and the periods in PERIOD-SET, we count the number of citing entities for each fifth in the [-1.0, +1.0] interval characterizing the P_{CUT} values (i.e., [-1.0, -0.61], [-0.60, -0.21], [-0.20, 0.20], [0.21, 0.60], [0.61, 1.0])[7], and classify them into those that have/have not mentioned the retraction. For instance, a citing entity having P_{CUT} equal to 6/7 and P_x equal to P4 is classified in D1 in the [0.61, 1.0] slice (since 6/7 is equal to 0.857) of the period P4.

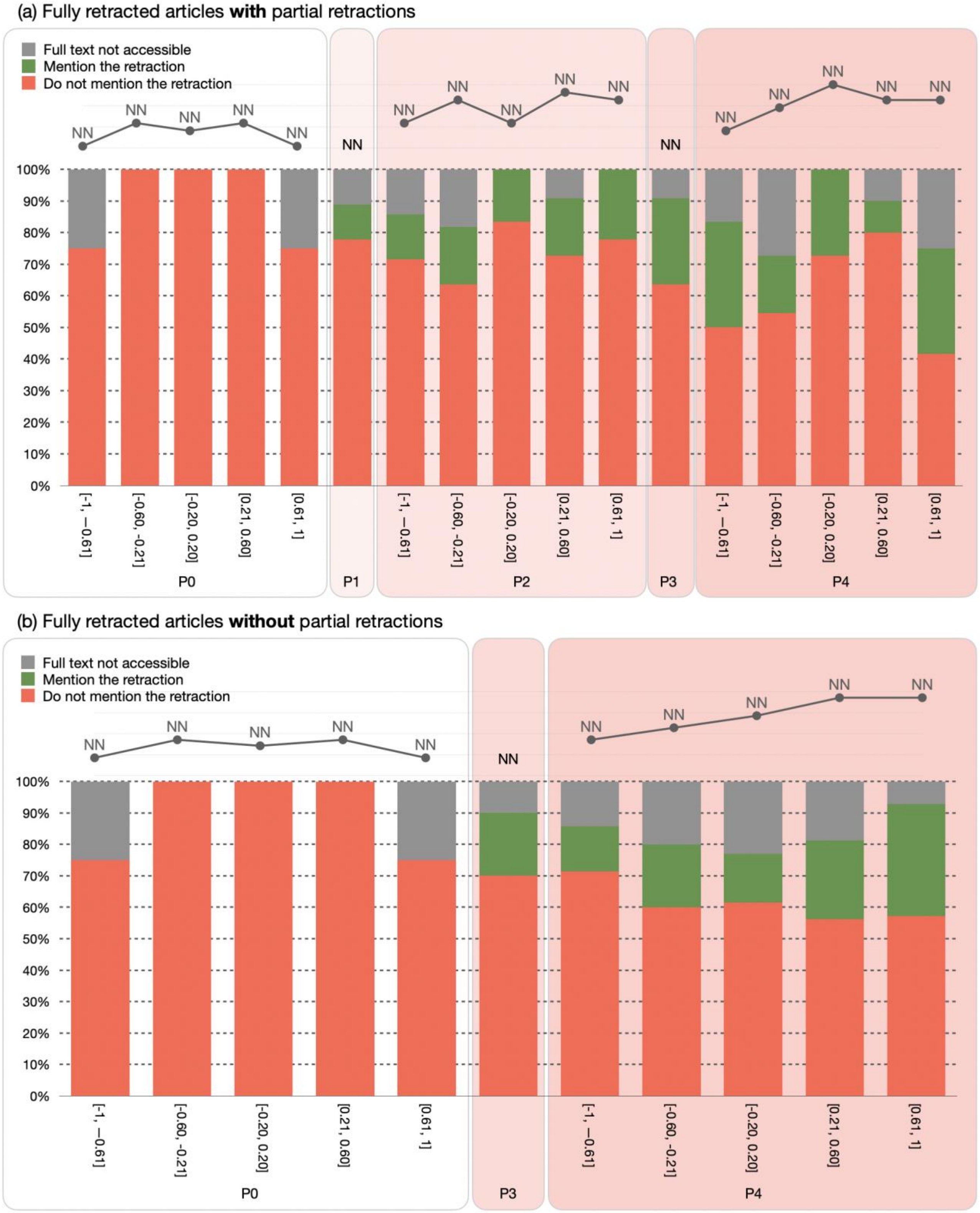
The visualization created using these values should have a similar format to the one in Figure 4.1. The results are sketched using a grouped bar chart which represents the number of entities as a function of their P_{CUT} and P_x values. The visualization has two versions according to the category of the retracted articles: RET-A (having at least one partial retraction) and RET-B (*with no partial retractions*). In case the RET-SET contains retracted articles from both types, then each type is treated separately and has its own visualization.

The visualization also highlights the citing entities that have/have not mentioned the retraction using a green bar and a red bar, respectively. Of course, the indication of this last feature could be applied only if we have successfully accessed the full text of such entities. To highlight the entities for which we do not have a corresponding full text for, we use a gray bar. The line chart on top of the bar chart of the periods shows the absolute number of citing entities in each slice. The values of the periods P1

[7] It is worth mentioning that the choice of splitting the [-1.0, 1.0] interval in balanced fifth is subjective. In principle, according to the analyzed data, one can even decide to split interval in less or more groups.

and P3 are not part of the line chart, since these values do not follow the [-1.0, 1.0] intervals defined for the other periods.

To represent the distribution of the citing entities among the subject areas we use a pie chart graphic, as in Figure 4.2. We show the 10 most representative subject areas (each one represented by a color). The rest of the subject areas are grouped under one category “Other subject areas”. In addition, we show the percentage of entities that have/have not mentioned the retraction (using the same colors of Figure 4.1), along with the absolute number of entities part of each different subject area. A similar visualization must be generated for each period in the PERIOD-SET. In case RET-SET contains retracted articles of the two types (i.e., RET-A and RET-B), then this visualization is repeated for each type.



(a) Fully retracted articles **with** partial retractions

(b) Fully retracted articles **without** partial retractions

Figure 4.1. A graphical representation for the distribution of the citing entities in the PERIOD-SET. The graphic has two different versions sketched according to the retracted articles categories, i.e., RET-A (in the top) and RET-B (in the bottom). It also highlights the citing entities that have/have not mentioned the retraction, along with the citing entities that do not have an accessible full text. The number of the citing entities of each period should be specified on top of the corresponding bar instead of "NN".

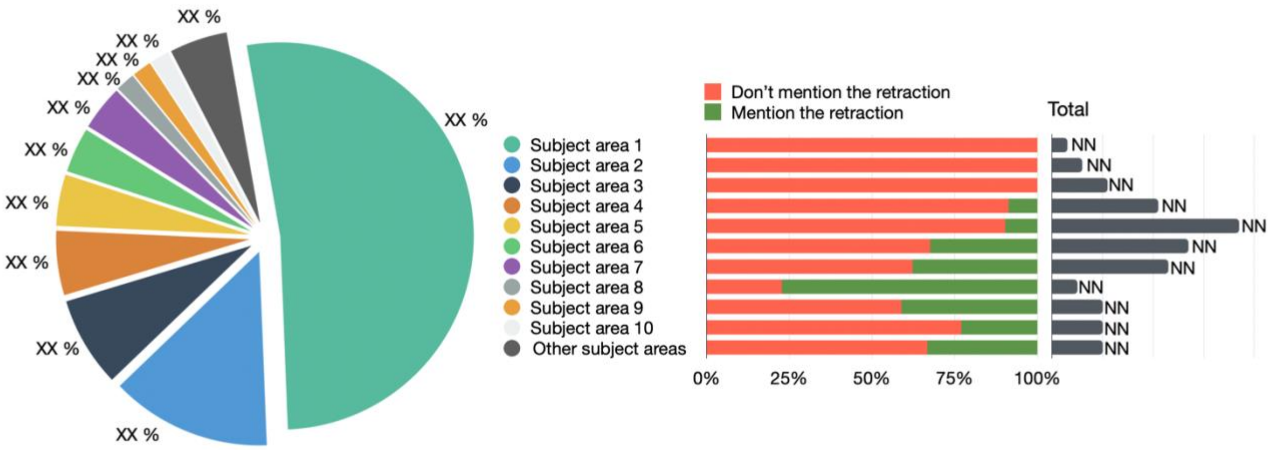


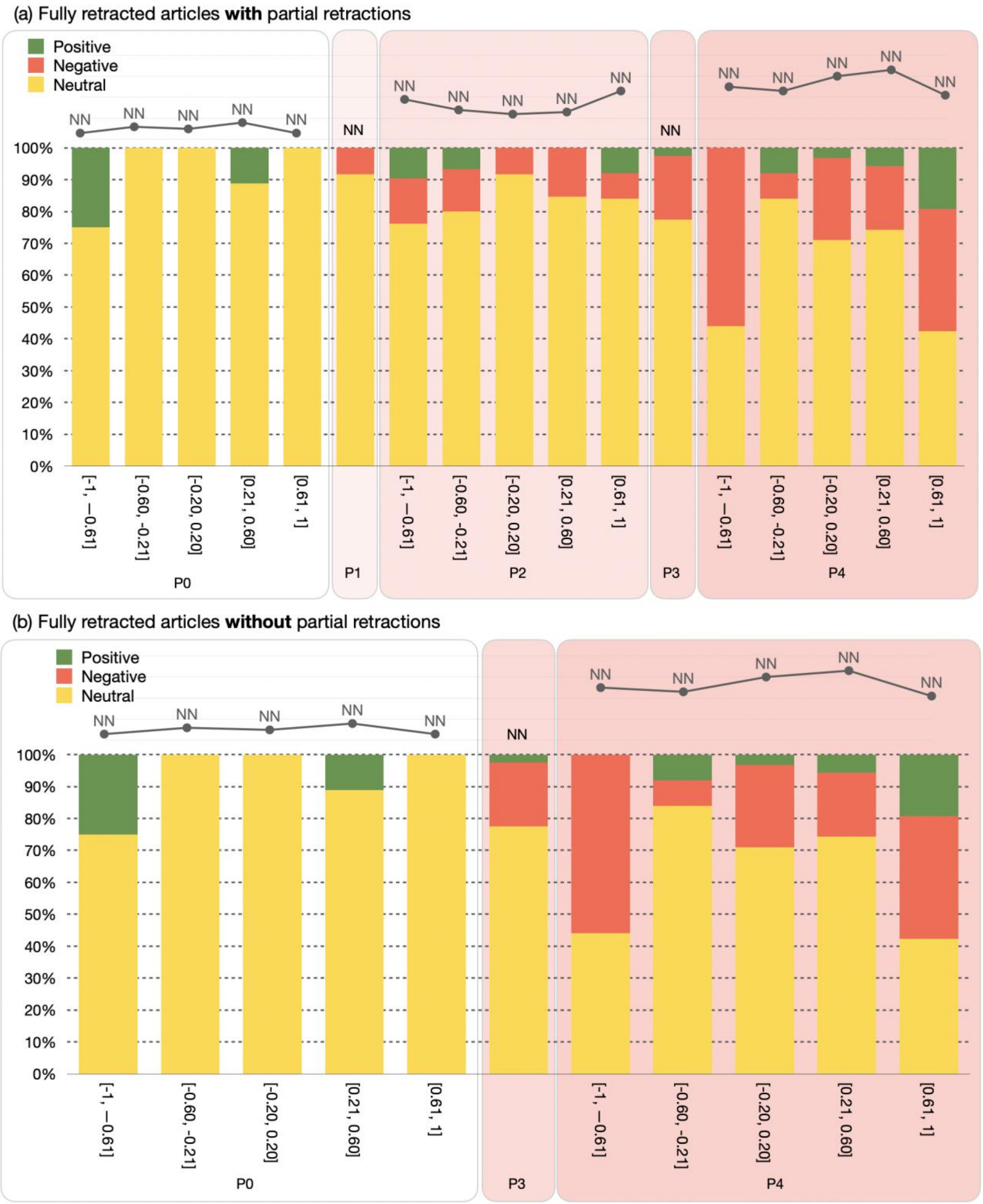
Figure 4.2. A pie chart used to represent the distribution of the citing entities across the subject areas. The chart shows the 10 most representative subject areas and groups the rest under the "Other subject areas" category. The graphic shows also the absolute number of entities for each category, along with the percentages of entities which have/have not mentioned the retraction.

4.4.2 Step 4.b: in-text citations

Considering the in-text citations and their related data, we want to analyze the distribution as a function of three features: (D1) the period P_X characterizing P_{CIT} and the related P_{CUT} of the citing entities containing the in-text citations, (D2) citation intent of each single citation, and (D3) the section where the in-text citation is contained. The distributions are divided according to the periods in the PERIOD-SET, as we have shown in the previous subsection. These distributions are accompanied with the value of the citation sentiment.

Figure 4.3 shows how to represent D1. The visualization is based on a grouped bar chart and the number of in-text citations as a function of the P_X and P_{CUT} of the related citing entities for which we can access their full text. As for the citing entities, the visualization has two versions based on the retracted articles category. In case the RET-SET contains retracted articles of type RET-A and RET-B, then each type is treated separately and has its version. The in-text citations are classified under the three citation sentiments: *negative*, *neutral*, and *positive*, indicated in red, yellow, and green, respectively. In addition, we also display the absolute number of in-text citations for each period in the same way we do for citing entities, as described in the previous subsection. In particular, the periods P0, P2, and P4 are characterized by means of fifth slices in the [-1.0, +1.0] interval, while the periods P1 and P3 are represented using one single bar chart and their values are not part of the line chart.

We use a horizontal bar chart to plot D2 and D3, as shown in Figure 4.4 and Figure 4.5, respectively. The bars in the chart are grouped considering the sentiment (i.e., negative/neutral/positive) and use the same colors adopted in D1. The length/percentage of the bars is proportional in relation to the total number of in-text citations of related to a particular slice. The absolute numbers are listed next to the chart. Both the visualizations in Figure 4.4 and Figure 4.5 need to be built for each period of the PERIOD-SET and refers only to the citing articles for which we can access their full text. For instance, in Figure 4.4, *citation function 4* represents almost 14% of the total in-text citations of that period, and it contains around 3% of the negative in-text citations of that same period. Figure 4.5 plots only the section values listed in the study of Suppe (1998), the other values are part of a separated group labeled “unclassified”. In case RET-SET contains retracted articles of RET-A and RET-B, then both the visualizations are repeated for each type.



(a) Fully retracted articles **with** partial retractions

(b) Fully retracted articles **without** partial retractions

Figure 4.3. A graphical representation for the distribution of the in-text citations in the *PERIOD-SET*. The periods P0, P2, and P4 are split in fifths, while P1 and P3 are represented using in one slice. The graphic has two different based on the categories of the retracted articles, i.e., *RET-A* (in the top) and *RET-B* (in the bottom). The graphic also highlights the neutral, negative, and positive in-text citations. The number of the in-text citations of each period should be specified on top of the corresponding bar instead of “NN”.

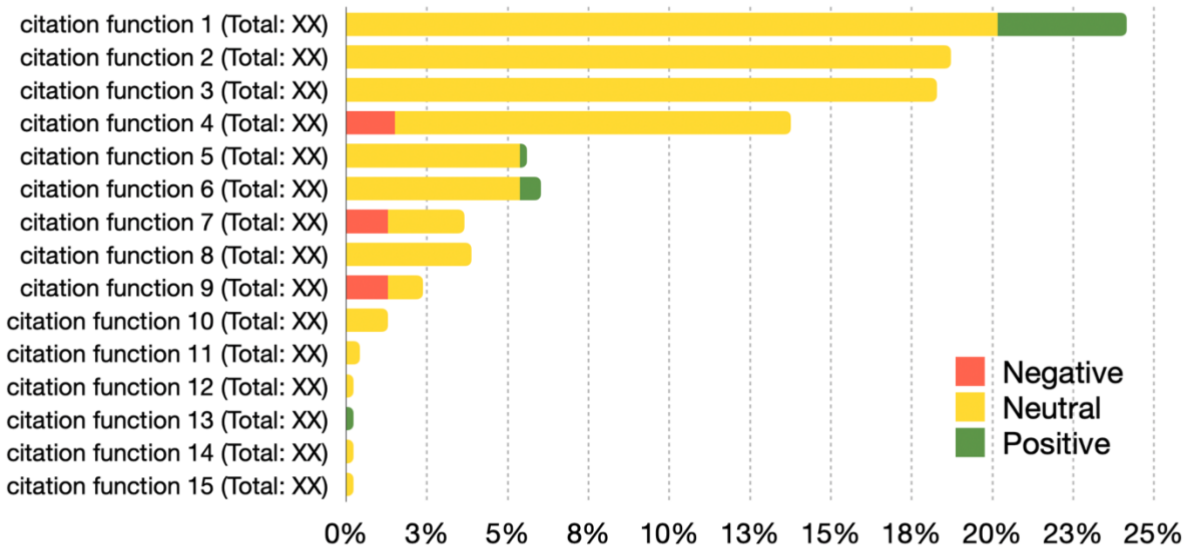


Figure 4.4. A horizontal bar chart representing the distribution of the in-text citations according to their citation functions. The graphic highlights the negative/neutral/positive percentages of in-text citations and mentions, between brackets, the total number on in-text citations annotated with such a citation function. The length (total percentage value) of the bars is in relation to the total number of in-text citation for the period in the PERIOD-SET shown by the graph.

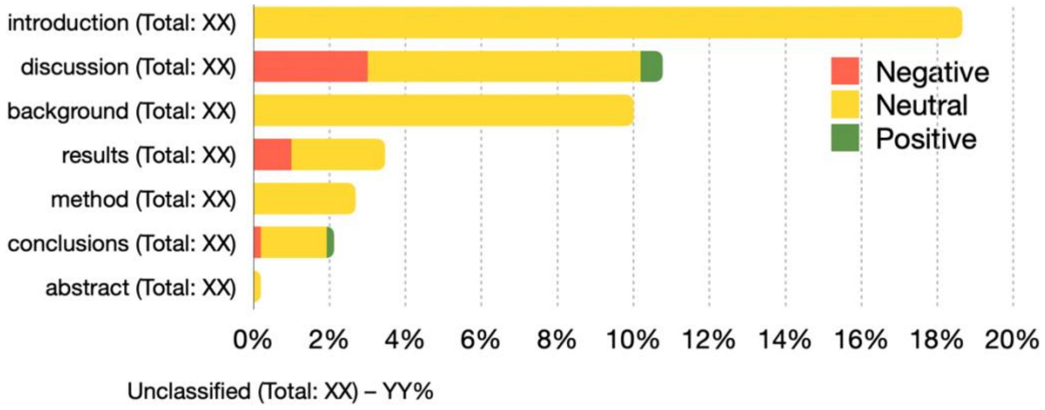


Figure 4.5. A horizontal bar chart representing the distribution of the in-text citations according to the section in which they appear. The graphic highlights the percentages of negative/neutral/positive in-text citations and mentions, between brackets, the total number on in-text citations annotated with such a section. The length (total percentage value) of the bars is in relation to the total number of in-text citation for the period in the PERIOD-SET shown by the graph and is used to sort the values of the sections.

Chapter 5

Topic modeling analysis

The topic modeling analysis is a statistical approach for automatically discovering the topics (represented as a set of words) that occur in a collection of documents. This analysis represents a crucial part of the methodology, and it is applied during its last phase (i.e., Phase 5). To accomplish such analysis a specific tool called MITAO, a user-friendly and modular software for topic modeling, was developed and used. This chapter starts with an introduction and definition of topic modeling analysis in its main steps (Section 5.1), then it presents how this process can be implemented using MITAO (Section 5.2). Finally, the last part of the chapter resumes the discussion of the methodology phases and presents its last phase. This phase applies a topic modeling analysis on two different textual corpora (built using the textual values included in the citing entities dataset produced in the first three phases): (1) the abstracts of the citing entities, and (2) the context of the in-text citations.

This chapter closes the second part of the thesis and concludes the definition of the entire methodology. The next part (i.e., Part III) configures and uses the methodology to analyze specific retraction cases to help us investigate the research questions presented in the introduction of the thesis.

5.1 Introduction to topic modeling

Texts are one of the most relevant data sources for several fields of study, for instance social scientists adopting qualitative research methods in their studies (Silverman, 2015; Flick, 2014). Traditionally, for interpreting texts and extracting meanings from textual sources, scholars rely on content analysis

(Krippendorff, 2018): here data interpretation stems from an in depth reading of texts, which can be aided by the production of a set of themes guided by a research question and by prior assumptions by the researcher (DiMaggio et al., 2013). While these methods produced wonderful and worthy research, they also have some limitations, such as dealing with epistemological issues (e.g., the relevance of intercoder reliability and the extent to which it is possible to have a real 'objective' stance toward data), but it is worth nothing that traditional content analysis is unable to deal with big corpora of texts, and that the imposition of a priori categories can help the researcher in dealing with data, but at the same time presumes that the researcher knows what to search for in advance. Automatic computing methods were therefore developed to help coping with these issues. The easiest technique is a word count of some keywords within texts (Stone et al., 1966), but while this technique is widely used, relying on word count alone infringes one of the most relevant principles of sociology, which is that the meaning of a word depends also on the surrounding words (DiMaggio et al., 2013). More sophisticated computer-based methods can better support research.

Topic Modelling (TM) is a technique based on Bayesian statistics, it analyses texts and creates 'topics', which are bags of words that often co-occur together in the original texts (Mohr and Bogdanov, 2013). The underlying idea is that the algorithm can elicit a latent structure of topics, that constitute the texts in the corpus (Blei et al., 2003). As a result, all the words in the textual sources are coded to a topic, and all the original texts are constituted by the topic in different percentages. Topics can be ontologically different, as they can be interpreted as themes, or discourse, or frames (Gamson, 1992). More in general, "the sets of terms that constitute topics index discursive environments, or frames, that define patterns of association between a focal issue and other constructs" (DiMaggio et al., 2013). DiMaggio et al. (2013) identify four relevant features of Topic Modelling: first, the results are explicit and the reproducibility of results is assured; second, the automated stage of TM permits dealing with big amounts of texts; third, at the same time, induction is just postponed to topics' interpretation, and this permits both discovering unexpected results, and using the same corpus of data for answering different research questions; finally, Topic Modelling is able to deal with polysemy, meaning that it is able to recognize that the meaning of a term depends on the surrounding words, and that the same word can have different meanings in different contexts.

Topic modeling has been used in several domains. For instance in arts and humanities, when in 2013, "Topic Models and the Cultural Sciences", a special issue of Poetics paved the way for a wider use of Topic Modelling: here the technique was used to analyze newspapers' coverage of arts funding (DiMaggio et al, 2013), the different uses of language in academia (McFarland et al., 2013), the meanings regarding the nature of violence during the Qing Dinasty in China (Miller, 2013), the media and public attention to a terrorist alert (Bonilla & Grimmer, 2013), the 'grammar of motives' in national security strategic texts (Mohr & Bogdanov, 2013), the comparison of cross-national disciplinary evolutions of themes in texts (Marshall, 2013), the application of TM to arts and humanities (Tangherlini & Leonard, 2013), and the analysis of the themes of about 3200 19th century novels (Jockers & Mimno, 2013). Topic Modelling was used in several other fields of study: in management for detecting novelties and emergence, developing inductive classification systems, understanding online audiences, analyzing frames and social movements, and understanding cultural dynamics (Hannigan et al., 2019), in political science, for example to reconstruct frames use by individual and organizations in comments on a public website dealing with a political debate on the US truck industry (Levy & Franklin, 2014), or for studies on institutional logics, where the impact of cultural meanings at field level is recognized (Thornton et al., 2012), for instance such as the work done by Croidieu and Kim (2018), which used Topic Modelling and archival data analysis to study the emergence of the U.S. wireless radiobroadcasting field.

To summarize, as the language used in a document represents its cognitive content (Whorf, 2012), Topic Modelling is increasingly used in all those contexts where researchers are interested in constructing meanings starting with words (Hannigan et al., 2019).

5.1.1 Running a topic modeling analysis

A standard workflow for building a topic model is based on three main steps: tokenization, vectorization, and topic model (TM) creation.

Tokenization is the process of converting the text into a list of words, by removing punctuations, unnecessary characters, and stop words. This operation lets the topic model analysis focus on the main concepts, by omitting the words that do not have a meaningful impact on the interpretation of the generated topics. In this step, we can also decide to lemmatize and stem the extracted tokens. The lemmatization converts words from third person to first person and verbs in past and future tenses to present. Stemming reduces the words to their root form. The aim of these two operations is to reduce inflectional forms and the derivative forms of a word to a common base form.

Then, we create vectors for each of the tokens retrieved. On this step, we can choose between two main models: (a) the term frequency-inverse document frequency (TF-IDF) model (Ramos, 2003), or (b) a Bag of Words (BoW) model (Brownlee, 2019). This choice depends on the type and size of our corpus. Works such as Bengfort et al. (2018) and Truica et al. (2016) have investigated this aspect. On the one hand, Bengfort et al. (2018) suggested that BoW is an appropriate choice when we want to represent the most frequent words but not always the most informative. On the other hand, TF-IDF is a better choice for general purpose cases and when the moderately frequent terms may not be representative of document topics. One of the main findings in the work of Truica et al. (2016) is the fact that LDA accuracy depends on the size of the vocabulary and TF-IDF works better with a larger one.

Finally, we build the topic model using the Latent Dirichlet Allocation (LDA) model (Jelodar et al., 2019). As a preliminary step, we need to determine in advance the number of topics to retrieve. Several approaches have been proposed to determine the correct number of topics (Zhao et al., 2015; Arun et al., 2010). A popular approach is based on the value of the topic coherence score, as suggested in the work of Schmiedel et al. (2019). The coherence score is used to measure the degree of the semantic similarity between high scoring words in the topic, and it helps us compare topics that are semantically interpretable from the topics that are artifacts of a mere statistical inference. Throughout this methodology, we decided to adopt this approach. Thus, we calculate the average coherence score for a range of topic models trained with a different number of topics. Then, we plot these values and observe the number of topics for which the average score plateaued, and we select the number of topics indicated in the plateau. For instance, Figure 5.1 shows the coherence score values of different LDA topic models built with a number of topics ranging from 1 to 40. The coherence score plateaued around 22-23 topics (the point of intersection between the average line and the line connecting the points of the chart). Thus, we might decide to train the LDA topic model with 22 topics.

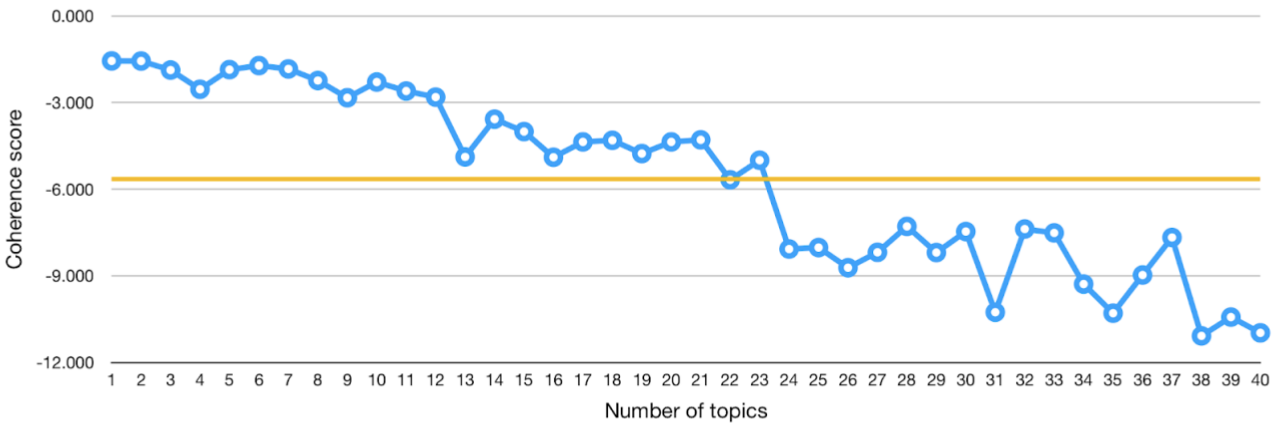


Figure 5.1. A chart example which plots the coherence score (y-axis) of different LDA topic models built using a variable number of topics, from 1 to 40 (x-axis). The orange line is the average value, and it plateaus around 22-23 topics.

5.2 MITAO: a user friendly and modular software for topic modeling

A researcher has plenty of options for performing Topic Modelling, with tools such as MALLET (Graham et al., 2012). Yet, unluckily, all the available free software resources that we have reviewed do not provide a mechanism for sharing the adopted workflow with the research community in order to foster the reproducibility of the process, which is a fundamental aspect of our study (following the open science principles) that we want to integrate also in the application of a topic modeling analysis. In addition, we also want this process (once shared) to be understandable and modifiable by others that lack of programming skills, or the ability to understand programming languages to a certain extent. These reasons prompted us developing MITAO – Mashup Interface for Text Analysis Operations.

MITAO is a tool which makes Topic Modelling techniques reusable by scholars with no or limited coding skills by means of a visual interface which enables them to create a workflow for processing textual content (Heibi et al., 2021; Ferri et al., 2020). MITAO uses the Latent Dirichlet Allocation (LDA) Topic Modelling, one of the earliest and well-known methods (Jelodar et al., 2019). Any workflow produced using MITAO can be stored and shared among scholars for reproducibility purposes. We give an overview of MITAO, its architecture, discuss its main features available, and introduce how to perform a Topic Modelling analysis and to get the tabular and graphical results using it.

MITAO aims at being open source, user-friendly, modular, and flexible software for performing several kinds of text analysis. The current release of MITAO (version 2.0) is able to convert data (from PDF to txt), clean data (stopword removal or the removal of parts of text through the use of regular expressions), perform Topic Modelling, provide a quantitative measure of the results (trough perplexity score and topic coherence), visualize and save data. MITAO is a Python based web application, which could be installed on any operating system (e.g., Windows, MacOS, or Linux) and run locally on any modern web browser. It provides a human-friendly interface for creating a customizable visual workflow that can be shared among scientists for reproducibility purposes. While users without coding expertise deal with a user-friendly visual interface, Python developers can further customize and extend it, adding new features for handling other elaborations. The source code and documentation of MITAO is available on GitHub at https://github.com/catarsi/mitao. MITAO is currently licensed under the ISC License (https://www.isc.org/licenses/).

Considering the state of the art and the available solutions for performing Topic Modelling, users with no or poor skills in coding are likely to encounter problems when they try to: (a) use such methods for their own works using a particular programming language, or (b) describe the technical workflow of all their processing phases and share it with other colleagues. The main objective of MITAO is to overcome these limits and: (a) enable scientists to use text analysis operations, especially topic modelling, without having strong knowledge in all the technicalities of programming and software development, (b) build a comprehensive workflow containing text analysis operations, which could be shared with other colleagues as to provide the reproducibility of the results obtained.

In this section, we will first talk about the architecture of MITAO and the conceptual model behind it. Next, we introduce a topic modelling analysis built using MITAO, as to clarify its potentials.

5.2.1 Architecture of MITAO

Two are the levels used to represent the general architecture of MITAO as it is represented in Figure 5.2: (a) frontend, and (b) backend. The backend level is the core of MITAO: its definition and further customization requires the expertise of a software developer. The frontend level is the part of MITAO which will be experienced by the user. Here the features and operations exposed are easily usable without the need of any software programming skill. The frontend level does not have any effect on the level underneath (i.e., the backend), it only needs to have a reading access to the procedures defined in the backend level and convert/interpret such procedures into a usable dynamic interface.

The backend level contains two layers: (a) server, and (b) configuration. The configuration layer enables the definition of three types of entities to embed in MITAO: data, tools, and parameters. These entities are characterized through a set of attributes and functions which are implemented in the underneath layer (the server layer); therefore, the two layers should be defined in parallel: every attribute/function must have its corresponding implementation on the server layer. The *Data* entity represents a typology of data a user can work with and use in their MITAO workflow, (e.g., text files, PDFs, images, etc.). These data typologies are also the ones a *Tool* can produce on its output or can have as input *Tools* represent the operations MITAO is able to perform on data (e.g., Topic Modelling, data conversion, and data cleaning). *Parameters* are needed by tools to perform their task. Obviously, each tool needs different parameters. The main idea of MITAO is to enable the final users to connect different *Data* and *Tool* entities following the compatibilities between inputs and outputs. Users should also set the necessary attributes for the entities (e.g., *Data* or *Tool*) they embed into the workflow. To let this happen, the configuration file should settle the *Data* and *Tool* entities with their corresponding *Parameters*. Users have no control on the *Data* entities which are produced as output of a *Tool*, therefore users cannot set the parameters of such *Data* entities. The scheme of Figure 5.3 summarizes the interconnections between these three entities.

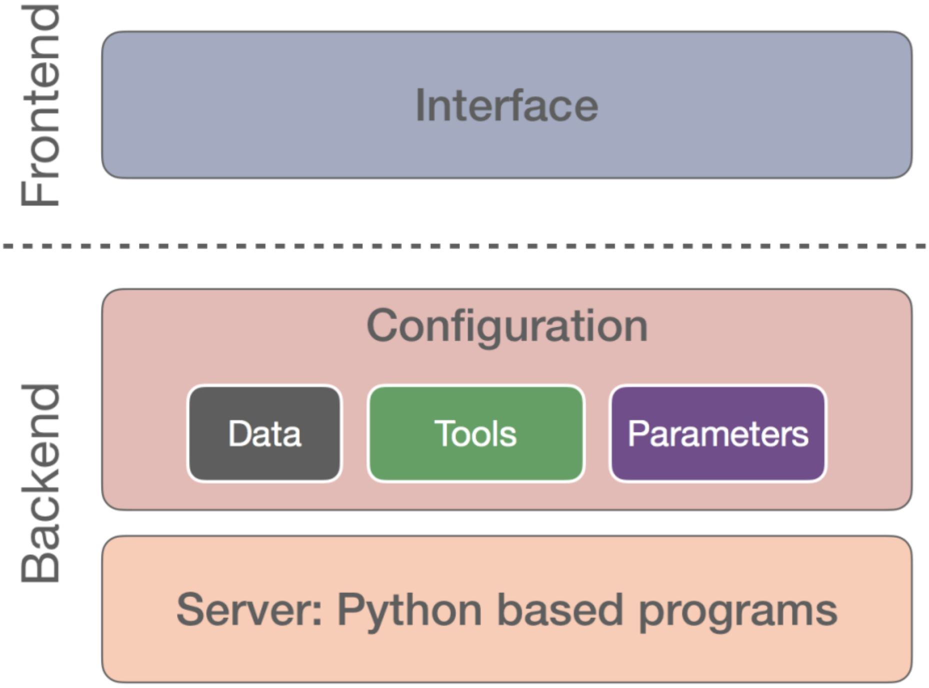


Figure 5.2. Architecture of MITAO – the backend level is the core of MITAO the definition and customization of this level requires coding expertise. The frontend level is the part of MITAO which will be experienced by the user

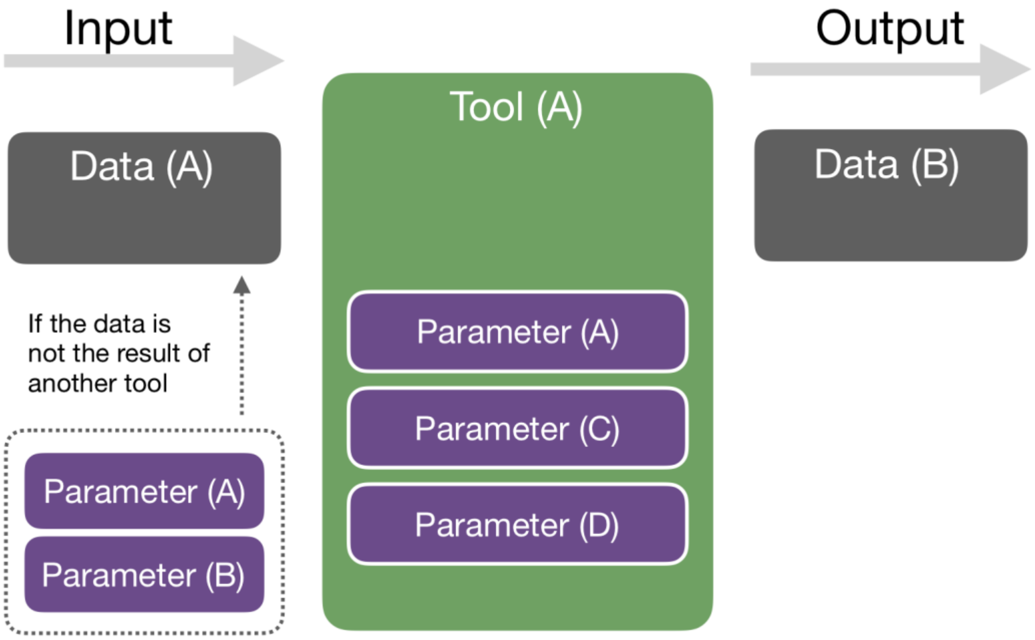


Figure 5.3. Logic connections between the main entities in MITAO: *Data, Tool, Parameter.* A *tool* takes as input and outputs a *Data* entity and could be customized using a set of *Parameter* entities (same parameters could be used for different components, see *Parameter A*). A *Data* entity could also be customized using *Parameter*(s) only if such *Data* is not the output of a *Tool.*

The underlying idea of MITAO is that each of the three entities – *Data*, *Tool* and *Parameter* – covers a wide number of possible usages. *Data* includes all the possible data sources, *Tool* all the possible operations, and *Parameter* all the necessary parameters. At the moment each new instance of these entities can be characterized as follows (Figure 5.4):

- *Tool*: any instance of this entity can be part of the typology:
    - Filter: these tools take a *Data* entity then perform some filtering operations on it and return the same *Data* instance but filtered.
    - Text analysis: all the tools which perform text analysis operations are part of this instance. They take a *Data* entity as input and return any other type of *Data*.
    - Terminal: these tools have no outputs: their goal is to visualize (e.g., charts/diagrams) or store the input *Data* entities

- *Data*: any instance of this entity can be part of the typology:
    - PDF: document/s in .PDF format
    - Textual: document/s in .TXT format
    - Image: image/s in .PNG or .JPEG format
    - Tabular: document/s with the inner information organized in a field separated list table records (e.g., .CSV format).
- *Parameter:* any instance of this entity can have one of the following types, which represent the input typologies a MITAO user can make: (a) Free text, (b) List of options, (c) File, or (d) Checklist.

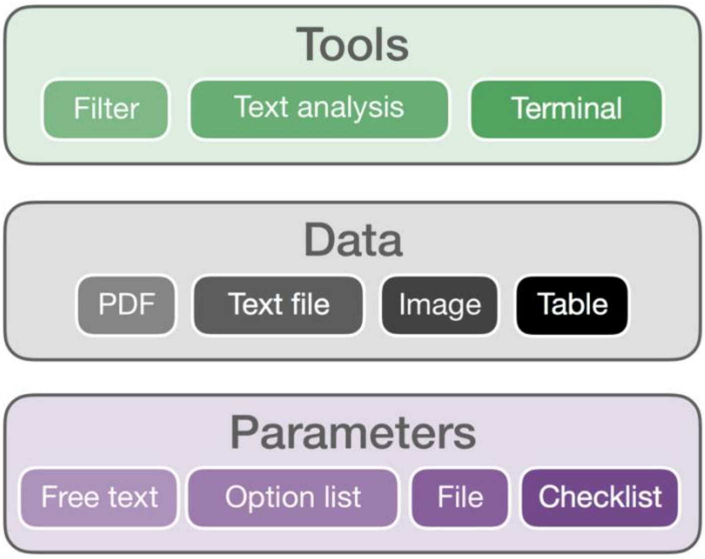


Figure 5.4. Available features for each instance of a *Tool, Data,* or a *Parameter*

The frontend level of MITAO is defined on one interface layer which is based on the idea of having a dynamic graphical user interface (GUI) which interprets the functionalities of backend level into a comprehensive system for building and customizing a complete text analysis workflow. Here we list the operations MITAO makes available to the users throughout its interface:

1. Adding and connecting *Tool* and *Data* entities in a directed graph network-based diagram which represents the complete workflow
2. Setting and editing the attributes of each workflow entity
3. Running, saving, and opening the workflows.

Considering the above operations, we designed a GUI as represented in Figure 5.5. The middle panel containing the workflow skeleton is based on a graph network schema, users can move and connect the inner nodes to create the final workflow to run. Along the above features, MITAO users should respect some rules/policies while using the interface:

- *Tool* or *Data* nodes can be connected to other *Tool* nodes of the workflow only if their output is compatible with the input of the target *Tool* node.
- MITAO will not allow users to create a directed circuit, cycle (the first and last vertices are repeated) in the network graph.

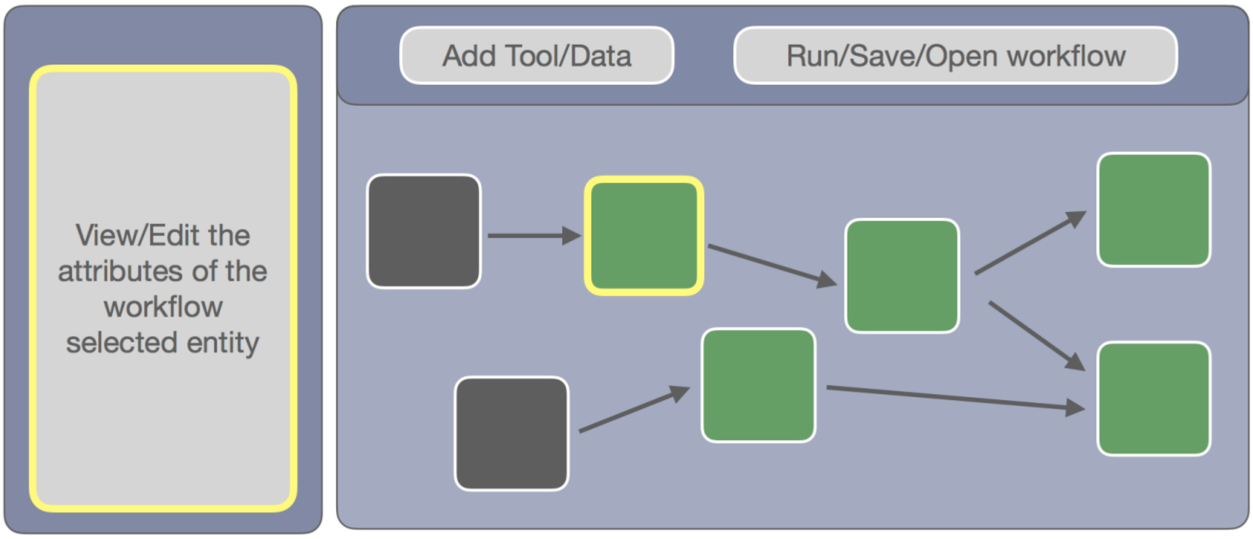


Figure 5.5. Scheme of the Graphic User Interface of MITAO. The main part is the middle panel which contains the workflow skeleton (i.e., a graph network schema), users can move and connect the inner nodes to create the final workflow to run

5.2.2 Topic Modelling with MITAO

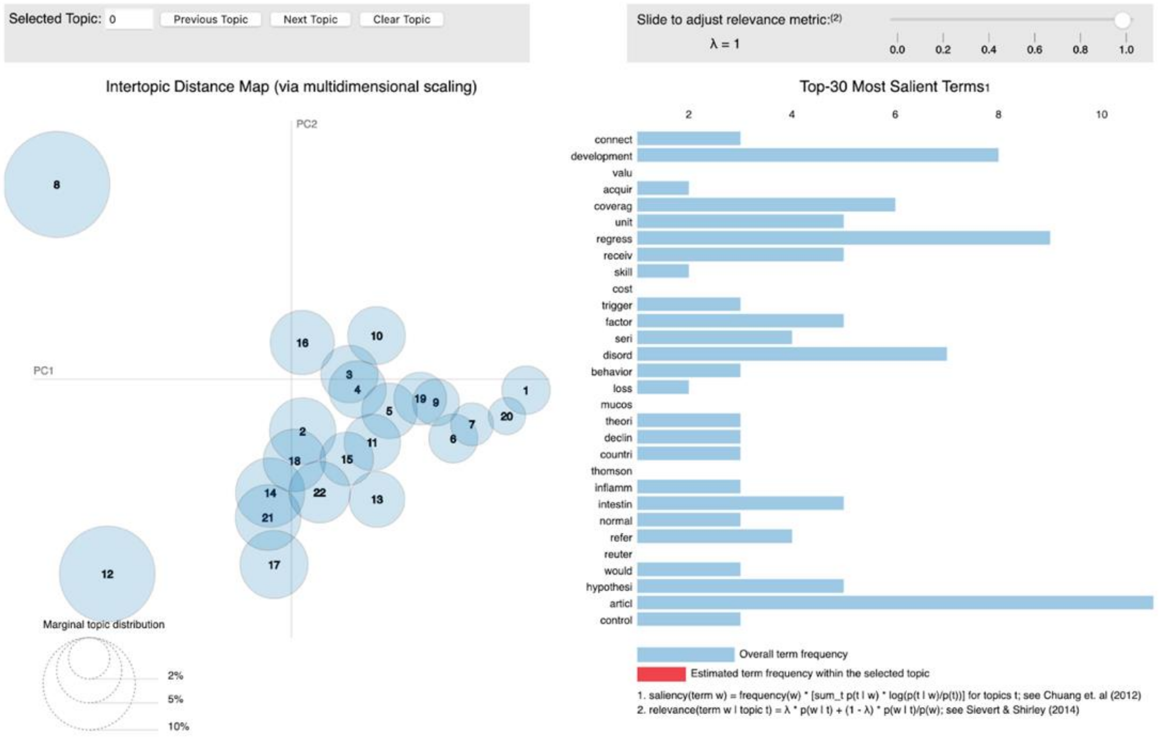
Using MITAO we can run a topic modeling analysis and manage its outputs. The results are represented either as tabular data or as visualizations. On one hand, in case of tabular data, the possibilities are two:

- the *N* most important keywords of each topic, which represent the *N* most useful and probable terms for interpreting a topic, ranked according to their probability value.
- document representativeness, i.e., the lists of all the documents of the corpus and their representativeness against each topic.

On the other hand, using MITAO we can also generate two types of dynamic visualizations to foster an analysis of the topic modeling results from different perspectives and help us infer new insights less evident from a pure observation of the tabular data. These two visualizations are named LDAvis (Latent Dirichlet Allocation Visualization) and MTMvis (Metadata-based Topic Modeling Visualization).

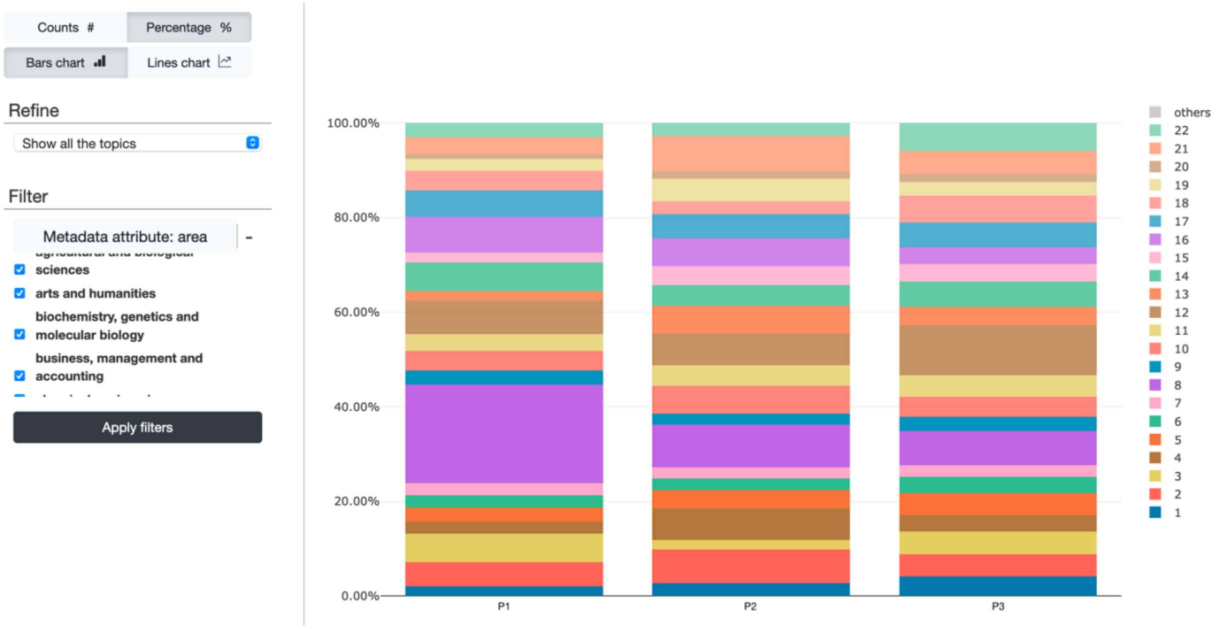
LDAvis, shown in Figure 5.6, provides a graphical overview of the topic modeling results (Sievert & Shirley, 2014). This visualization plots the topics as circles in a two-dimensional plane whose centers are determined by computing the distance between the topics and uses a multidimensional scaling to project the inter-topic distances onto two dimensions. The topic prevalence is represented by the dimension of the area of each circle. In addition, LDAvis shows a global list of 30 terms ranked using the *term saliency* measure. This saliency measure combines the overall probability of a term with its distinctiveness: how informative is a specific term for determining the generation of a topic versus any other randomly selected term (Chuang et al., 2012). In addition, by selecting a particular topic, LDAvis shows a list of 30 terms ranked using the *relevancy* measure which is used to rank terms within each topic to aid users in topic interpretation activities. This measure is controlled by a weight parameter λ, which allows one either to rank the terms in a decreasing order of their topic-specific probability if close to 1 or to rank terms solely by their lift if close to 0.

MTMvis, shown in Figure 5.7, is a dynamic and interactive visualization that shows the representativeness of the topics in the documents based on customizable metadata accompanying these documents. The visualization could be filtered using a predefined set of attributes. In MTMvis only the documents that have a corresponding value (any value but null) for all the attributes used for filtering are part of the MTMvis categorization process and will consequently appear in the visualization. For instance, if MTMvis uses only the attribute "Subject area" as filtering option, then all the documents displayed in the visualization have a corresponding value for such attribute in their descriptive metadata document.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 5.6. The LDAvis interface. The left side of the visualization plots the topics in a two-dimensional plane whose centers are determined by computing the distance between topics. On the right side LDAvis lists 30 terms ranked using the *term saliency* measure, this list is modified to show the 30 terms ranked using the *relevancy* measure of a specific topic when a topic (circle) is selected from the left graphic.

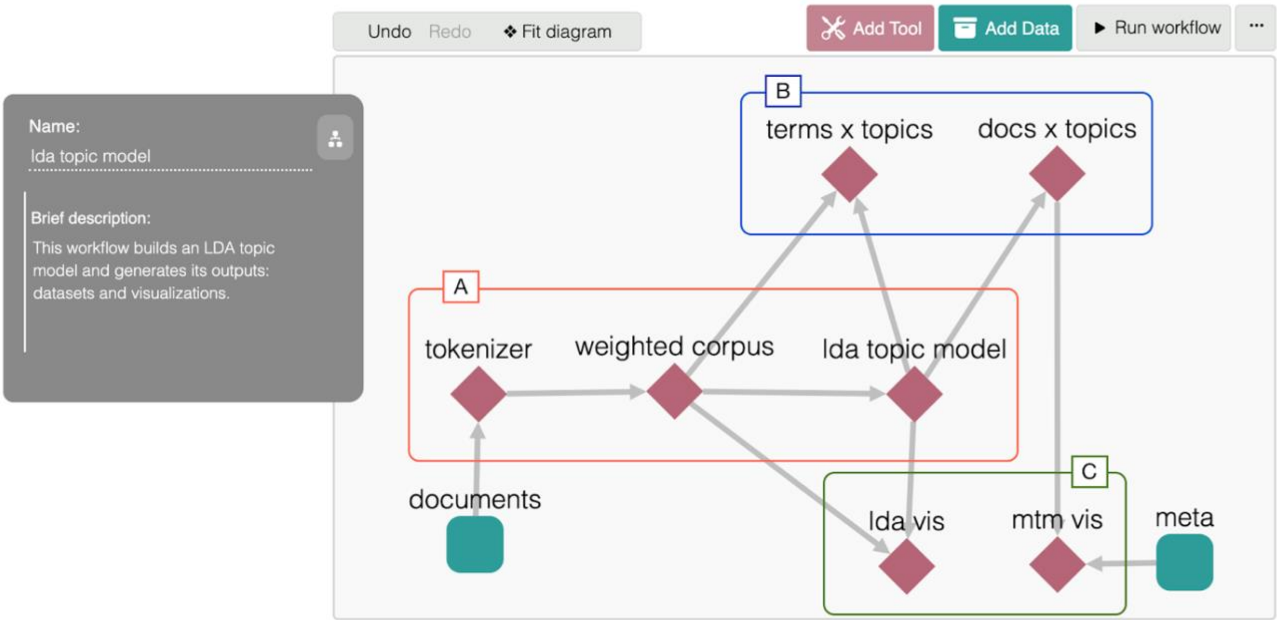


Metadata attribute: area

☑ agricultural and biological sciences
☑ arts and humanities
☑ biochemistry, genetics and molecular biology
☑ business, management and accounting

Figure 5.7. The MTMvis interface. On the left side, users can modify some visual and filtering parameters to dynamically change the main visualization. Each topic is colored differently. The chart plots the topics as a function of an established metadata attribute (X-axis values), e.g., the PERIOD-SET

Figure 5.8 shows how the topic model workflow is defined using MITAO. The workflow takes a collection of documents (i.e., the node "documents"), builds the LDA topic model (i.e., the "A" rectangle in red), generates the two datasets (i.e., the "B" rectangle in blue), and builds the two visualizations LDAvis and MTMvis (i.e., the "C" rectangle in green) using the metadata of the original documents (i.e., the node "meta").



Name:
lda topic model

Brief description:
This workflow builds an LDA topic model and generates its outputs: datasets and visualizations.

Figure 5.8. The MITAO workflow used for running the LDA topic model (i.e., rectangle A) and then consider the outcomes of the analysis for generating the datasets (rectangle B), and the visualizations (rectangle C). The workflow takes two inputs: the documents of the corpus and the metadata of the documents

5.3 Phase 5: running a topic modeling analysis on abstracts and citation contexts

We apply a topic modeling analysis using MITAO on the abstracts and the in-text citation contexts of the citing entities which have cited a retracted publication part of RET-SET. Each analysis follows the process we have discussed in the Section 5.2.2, with some adaptations and re-configurations which depend on the nature of the given corpus, i.e., either the abstracts or the in-text citation contexts.

5.3.1 Step 5.a: abstracts

In this case, the "documents" in Figure 5.8 represent the collection of the abstracts of all the citing entities which have cited *RET-SET*. Citing entities with no abstract, due to the scenarios that we have discussed in Section 4.3 "*Phase 3: Extracting and labeling content-based features*" (e.g., if the citing entity is an editorial), are not part of the analysis. Each document contains one abstract, and the total number of documents considered in the creation of this topic model is equal to the number of citing entities for which we were able to retrieve the abstract. In the tokenization process, we need to remove the stop words (such as “the”, “they”, and “you”) and other words which are very common in abstracts and cause noise in the obtained results, such as “background”, “summary”, “method”, “results”, etc. These words are likely to occur when dealing with structured abstracts. Finally, we build two MTMvis graphs on top of the two metadata attributes: (a) PERIOD-SET, and (b) subject area. In case RET-SET contains retracted articles of type RET-A (having at least one partial retraction) and RET-B (*with no partial retractions*), then these visualizations are done for each type.

5.3.2 Step 5.b: in-text citation contexts

In this case, the “documents” in Figure 5.8 represent the in-text citation contexts where a citation to RET-SET is contained. Each document contains the context of one in-text citation, and the total number of documents considered in the corpus given to the topic model is equal to the total number of in-text citation contexts we were able to retrieve. Citing entities with no in-text citations, because their full text was inaccessible, are not part of this analysis. In the tokenization process, we need to remove the stop words and other additional words which are part of the bibliographic reference of the cited retracted articles, e.g., name of the author(s) or the title of the publication cited. Finally, we build two MTMvis graphs using the same metadata attributes: (a) PERIOD-SET, and (b) subject area. In case *RET-SET* contains retracted articles of type RET-A (having at least one partial retraction) and RET-B (*with no partial retractions*), then these visualizations are done for each type.

Part III

RESULTS, DISCUSSION, AND CONCLUSIONS

Chapter 6

Retraction in the arts and humanities domain

This chapter investigates the retraction dynamics in the arts and humanities domain. This process answers the first two research questions: (RQ1) How did scholarly research cite retracted humanities publications before and after their retraction? and (RQ2) Did all the humanities areas behaving similarly concerning the retraction phenomenon?

Our approach is based on the application of the methodology introduced in the second part of the thesis. The methodology is tuned and applied (with respect to its phases) in the analysis of retraction in the arts and humanities domain. Phase by phase we describe how it has been configured and executed and illustrate the partial results obtained. The application of the methodology is divided in two parts, in the first part (Section 6.1) we discuss the three initial phases of the methodology which led to the generation of the dataset containing all the citing entities characterized by their related attributes. Then, in the second part (Section 6.2), we move to the last two phases which take the information stored in the citing entities dataset to build a descriptive statistical summary (i.e., fourth phase) and apply a topic modeling analysis on the abstracts of the citing entities and on the in-text citation contexts (i.e., fifth phase).

In the last part of this chapter (Section 6.3) we address research questions RQ1 and RQ2. Our answers are based on the findings and outcomes of our methodology.

6.1 Building the citing entities dataset

This section goes through the first three phases of the methodology which lead to the production of a dataset containing all the entities which have cited a retracted publication in the humanities domain. The entities in the dataset (i.e., records) are characterized with a set of attributes (i.e., columns). For each phase of the first three of our methodology we illustrate how such phase has been configured and executed. Each phase is discussed in a separated section. Phase by phase, throughout this chapter we discuss the extracted attributes used to characterize the citing entities, which are part of the final dataset produced after the execution of the first three phases.

The dataset named *cits.csv* has been published in Zenodo (Heibi & Peroni, 2021c) under a Creative Commons public domain dedication (CC0, https://creativecommons.org/publicdomain/zero/1.0/legalcode), features that are based on the textual content extracted from the full-text are part of a different dataset (i.e., *content.csv*) which keeps such data with their original license (the one provided by the publisher).

The values of the dataset (i.e., *cits.csv*) are subject of the analysis presented in the next section (i.e., Section 6.3) for the fourth and the fifth phases of the methodology.

6.1.1 Phase 1: defining a case study/domain

First, we wanted to have a descriptive statistical overview of the retractions in the humanities as a function of crucial features (e.g., reasons for retraction) to help us define the set of retractions to use as input in the next phases. Thus, on June 2021 we queried the Retraction Watch database searching for all the retracted articles labelled as humanities (marked with "HUM" in the database). Then we have classified the results as a function of three parameters: (a) the year of the retraction, (b) the subject area of the retracted articles (architecture, arts, etc.), and (c) the reason(s) for the retraction.

We collected an overall number of 474 articles, the earlier retraction occurred in 2002, while the last year of retraction we obtained was 2020.

As shown in Figure 6.1, we noticed an increasing trend throughout the years, with some exceptions, in particular we observed that the highest number of retractions per year was 119 in 2010, probably due to an investigation and a massive retraction of several articles belonging to one author, i.e., Joachim Boldt (Brainard, 2018). While looking at the subject areas, we noticed that most of the retractions are related to *arts* and *history*, while plagiarism motives were by far the most representative ones, confirming the observations of the work done by Halevi (2020).

After querying COCI and MAG[8], we found that 85 retracted items (out of 474) had at least one citation (a total of 2054 citations). We manually checked the dataset for possible mistakes introduced during the collection process. Indeed, some of the citing entities identified in MAG either did not include a bibliographic reference to any of the retracted articles, or the retracted article considered was not cited in the content of the citing article (although present in its reference list), or the citing entities' type did not refer to a scholarly article (e.g., bibliography, retraction notice, presentation, data repository). There was also one retracted article that received 1,050 citations, that we decided to exclude from the study to reduce the bias in the results. Following these considerations, the final number of retracted articles considered was 84, involving a total number of 935 unique citing entities. As shown in the bubble chart in Figure 6.2, most of the citing entities (i.e., 891) were included in MAG, 388 were included in COCI, and MAG and COCI shared 344 entities.

Although the retracted items identified so far were all in the humanities domain according to the categories specified in Retraction Watch, an item might have other non-humanities subjects associated with it. Sometimes, these non-humanities subjects might be more representative of the content of the retracted document and, thus, they might generate unwanted bias for the rest of the

[8] We used their REST APIs in June 2021 for retrieving citation information

analysis. For instance, consider the retracted article we gathered with the title *"The Good, the Bad, and the Ugly: Should We Completely Banish Human Albumin from Our Intensive Care Units?"* (Boldt, 2000). In Retraction Watch, this article is associated with two subjects: *Medicine* and *Journalism*. Yet, when we checked and read the full text of the article, we noticed that the contents related to *Journalism* are very limited and, as such, the article should not be considered as belonging to the humanities research (instead it is mostly related to *Medicine*).

To avoid considering these peculiar publications in our analysis, we used the *domain_affinity* score (introduced in Section 4.1) to evaluate the affinity of each retracted item we have collected. We calculated the *domain_affinity* score of each retracted item in the list (84).

The cake chart in Figure 6.2 shows how we classified the retracted articles and those citing them according to their *domain_affinity* score. To narrow our analysis and reduce the bias, we decided to consider only the retracted articles (and their corresponding citing entities) having a medium or high *domain_affinity* score (i.e., ≥ 2). At the end of this phase, the final number of retracted items we considered was 72, with a total of 678 citing entities. These 72 retracted items are all part of the RET-SET.

Only two out of the 72 retracted publications received a partial retraction before getting the full retraction notice: "*Men's music ability and attractiveness to women in a real-life courtship context*" (Guéguen et al., 2014), and *"Impact of Spiritual Leadership on Unit Performance"* (Fry et al., 2011). The full retraction notice was raised in the same year in the first case, and two years after the partial one in the second case. To have a homogenous analysis and reliable results, we decided to consider only the year of the full retraction of all the retractions gathered. In Table 6.1 we have a summary for the values of some attributes characterizing the 72 retracted publications of the RET-SET, such as the DOI, year of publication, year of full retraction, and the year of the last citation received.

Table 6.1. A descriptive summary of the values in RET-SET containing 72 retractions from the humanities domain. The listed attributes (first column) are: DOI, year of publication, year of full retraction, and the year of the last citation received.

Attribute	**Value**
DOI	94.4% (68) publications have a corresponding value
Year of publication (*E-RetPub*)	Values ranging from 1990 to 2020
Year of full retraction (*E-FR)*	Values ranging from 2006 to 2020
Year of the last citation received (*E-LastCit*)	2020

Since we considered only the year of the full retraction for all the retractions gathered, our analysis is applied to the following three periods of the methodology:

- P0, from the year of publication of the retracted article to the year before its retraction (the year of the retraction is not part of this period).
- P3, the exact year of the full retraction
- P4, from the year after the full retraction to the year of the last citation

To reduce the ambiguity emerging considering the names of the periods, i.e., P1 and P2 are not included which might cause a little of confusion as it is logically expected to have P1 after P0. Therefore, we decided to rename the periods for the rest of this analysis: P0 as P-Pre, P3 as P-Ret, and P4 as P-Post.

Each citing entity (further characterized in the next phase) is part of one of the above three periods and has a corresponding P_{CIT} and P_{CUT} value following the formula of Section 3.4. E.g., considering our case study, if a citing entity *C* published in 2003 cited a retracted article *R* (contained in RET-SET) published in 2000 and fully retracted in 2005, then C is part of P0, $P_{\mathrm{CIT}} = [2000,2001,2002,2003,2004]$ and $P_{\mathrm{CUT}} = 0.5$.

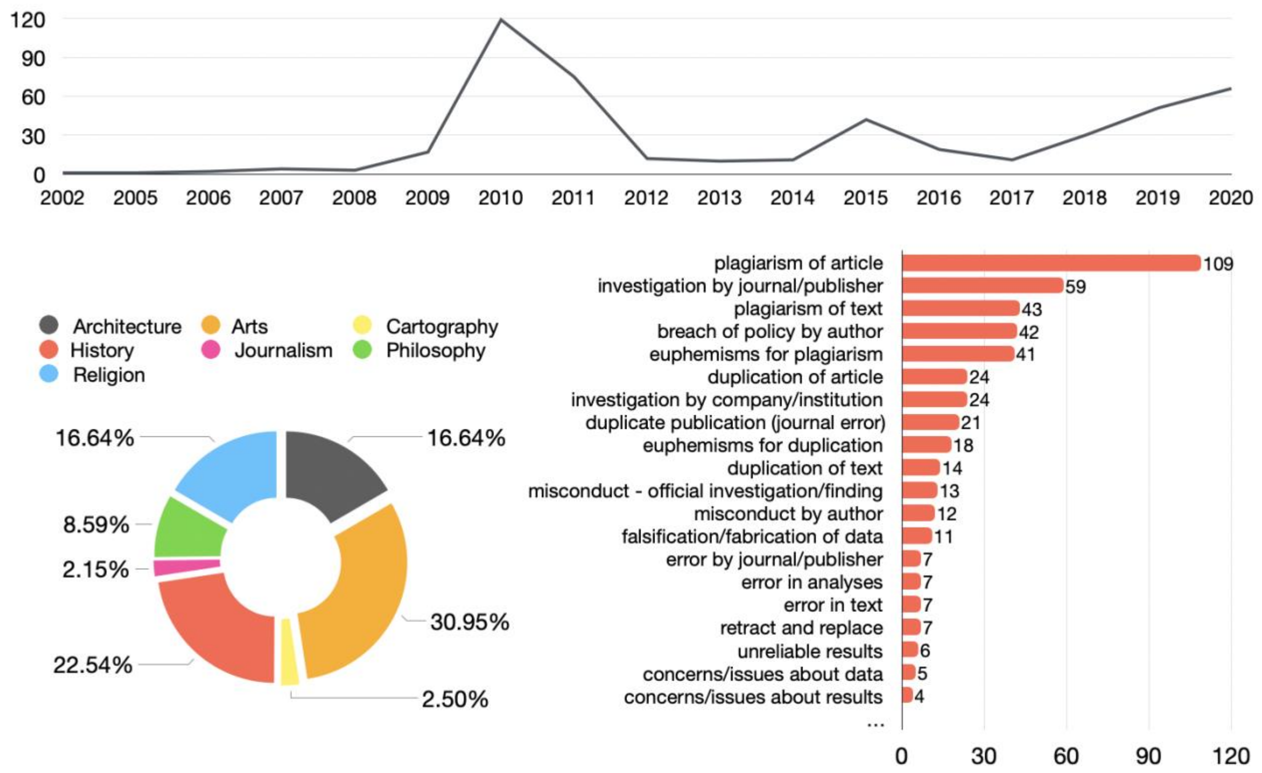


Figure 6.1. Retractions in the humanities domain as a function of three features: the year of retraction (line chart), the subject areas of the retracted articles (ring chart), and the reasons for retraction (horizontal bar chart). Based on the data retrieved from the Retraction Watch database in June 2021.

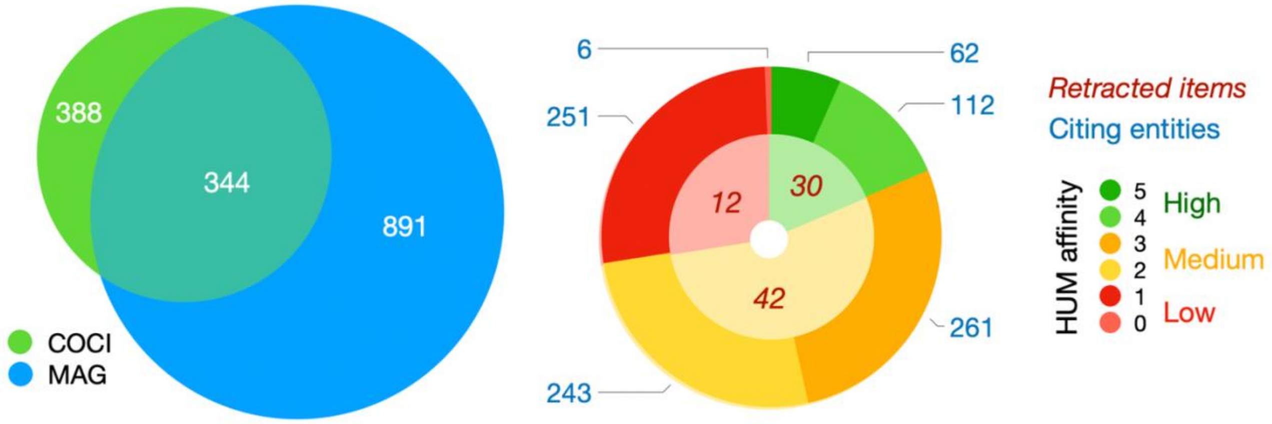


Figure 6.2. The citing entities of the retracted publications gathered from MAG and COCI (bubble chart) and their distribution according to the *domain_affinity* score of the retracted article they cite (cake chart).

6.1.2 Phase 2: identifying, retrieving, and characterizing the citing entities

Following the process of the previous phase we have already identified the citing entities of the RET-SET (i.e., 678), yet in this second phase we used again COCI/MAG REST APIs to get the basic metadata of each citing entity, i.e., DOI (if any), year of publication, title, venue id (ISSN/ISBN), and venue title. The metadata retrieved are those that were available in COCI and MAG in June 2021.

We queried the COCI REST API (http://opencitations.net/index/coci/api/v1) to get the metadata of the entities having a DOI using the "metadata" operation provided by the COCI REST API service (http://opencitations.net/index/coci/api/v1#/metadata/{dois}). On the other hand, we used the operation "evaluate" provided by the MAG REST API to get the metadata of the entities which are only part of MAG (https://api.labs.cognitive.microsoft.com/academic/v1.0/evaluate). The entities which are only part of MAG are mainly those that do not have a corresponding DOI (therefore not included in COCI).

Then, using the Retraction Watch database, we annotated whether the any citing entity was fully retracted as well. We also classified the citing entities into areas of study and specific subjects, following the Scimago Journal Classification (https://www.scimagojr.com/), using the 27 main subject areas (medicine, social sciences, etc.) and 313 subject categories (psychiatry, anatomy, etc.). We searched for the titles and IDs (ISSN/ISBN) of the venues of publication of all the citing entities and classified them into specific subject areas and subject categories. For books/book chapters, we used the ISBNDB service (https://isbndb.com/) to look up the related Library of Congress Classification (LCC, https://www.loc.gov/catdir/cpso/lcco/), and then we mapped the LCC categories into a corresponding Scimago subject area using the established set of rules described in Section 4.2. Table 6.2 lists the basic metadata extracted in this phase (first column), and for each attribute it summarizes the total number of entities having a corresponding value and the nature of such values.

Table 6.2. A summary representing the status of the citing entities dataset after the execution of the second phase. Each citing entity is characterized with the attributes (first column): DOI, year of publication, title, venue id, venue title, subject area, and subject category. For each attribute, on the second column, is summarized the total number of entities with a corresponding value and the nature of such values.

DOI	**Total:** 587 (86.57%) citing entities had a value specified
Year of publication	**Total:** All the citing entities had a value specified (678) **Values:** From 1990 to 2021
Title	**Total:** All the citing entities had a value specified (678)
Venue ID	**Total:** All the citing entities had a value specified (678) **Values:** ISSNs (547), ISBNs (83), Others (7 pre-prints + 41 dissertations)
Venue title	**Total:** All the citing entities had a value specified (678)
Is/is not retracted	**Total:** 7 citing entities
Subject area	**Total:** All the citing entities had a value specified (678) **Values:** 26 different values: "social sciences" (207), "arts and humanities" (170), "medicine" (99), "psychology" (81), "business, management and accounting" (69), "earth and planetary sciences" (32), "engineering" (31), "environmental science" (24), "economics, econometrics and finance" (23), "nursing" (18), "agricultural and biological sciences" (17), "computer science" (16), "multidisciplinary" (14), "neuroscience" (13), "mathematics" (9), "biochemistry, genetics and molecular biology" (7), "energy" (6), "materials science" (5), "decision sciences" (5), "physics and astronomy" (4), "chemical engineering" (3), "pharmacology, toxicology and pharmaceutics" (2), "geography, planning and development" (2), "chemistry" (2), "geology" (1), "dentistry" (1)
Subject category	**Total:** All the citing entities had a value specified (678) **Values:** 168 different values

6.1.3 Phase 3: extracting additional data and labeling the textual content of the citing entities

We used the attributes annotated in the previous phase and contained in the citing entities dataset by the end of the phase process (summarized in Table 6.2) to search for and retrieve the full text of all the citing entities collected (678). From the full-text, we extracted the abstracts of the citing entities, the in-text reference pointers denoting a bibliographic reference referencing to one of the retraced publications in RET-SET (i.e., a retraction in the humanities domain), the citation contexts of the in-text citations, and the sections where the citation contexts are contained. In case we could not access the full text of a citing entity, e.g., due to paywall restrictions, the corresponding entity was still considered in our dataset. However, we did not use it for the topic analysis of phase five.

A corresponding abstract was not found for 43 (6.34%) out of 678 citing entities, such as the editorial documents or book chapters with no abstracts. All the 678 citing entitles have been annotated with at least one in-text citation. We collected a total of 1020 in-text citations, defined by their in-text reference pointers (e.g., “[3]”, “Heibi et al. (2020)”, etc), the context (based on the sentence that contains the in-text reference, i.e., the anchor sentence, plus the preceding and following sentences), and the section (using the list of (Suppe, 1998) if such section rhetoric was clear by looking at its title, otherwise using other three residual categories, i.e., “first section”, “middle section” and “final section” depending on their position in the citing article). Some of the in-text citations did not have a corresponding section at all, e.g., an editorial document not structured in sections.

Considering such results, we had an average of 1.5 in-text citations per citing entity, each in-text citation has been manually characterized (following the rules introduced in Section 4.3) with its citation intent, citation sentiment, and whether the retraction has been/has not been addressed in the context of the citation.

In Table 6.3 we listed the new attributes added to the citing entities dataset after the execution of this third phase. For each new attribute added to the dataset (listed in the first column) we have summarized the total number of entities with a corresponding value and the nature of such values (the second column).

Table 6.3. A summary of the new attributes added to the dataset of the citing entities after the execution of the third phase. For each attribute (listed in the first column), on the second column, we summarize the total number of entities with a corresponding value and the nature of such values.

Abstract	**Total:** All the citing entities had a value specified (i.e., 678)
In-text citation contexts	**Total:** all the in-text citations had a value specified (i.e., 1020)
In-text citation sections	**Total:** 594 (58%) in-text citations had a value specified in the categories of (Suppe, 1998) **Values:** 10 different values: *abstract* (11), *introduction* (281), *background* (75), *method* (79), *results* (16), *discussion* (117), *conclusions* (15)
In-text reference pointers	**Total:** all the in-text citations had a value specified (i.e., 1020)
In-text citation intents	**Total:** all the in-text citations had a value specified (i.e., 1020) **Values:** 21 different values: *obtains background from* (286), *cites for information* (230), *discusses* (193), *describes* (86), *cites as evidence* (71), *obtains support from* (46), *qualifies* (39), *uses method in* (18), *disputes* (14), *cites as source document* (9), *includes quotation from* (6), *confirms* (4), *critiques* (3), *uses data from* (3), *credits* (2), *speculates on* (2), *extends* (2), *uses conclusions from* (2), *supports* (2), *parodies* (1), *agrees with* (1)
In-text citation sentiments	**Total:** All the in-text citations had a value specified (i.e., 1020) **Values:** *neutral* (989), *negative* (22), *positive* (9)
Retraction is/is not mentioned	**Total:** All the citing entities had a value specified (i.e., 678) **Values:** *no* (670), *yes* (8)

6.2 Descriptive statistics and topic modeling

In this section we discuss the analysis of the last two phases of the methodology (i.e., fourth and fifth phases) and show their outcomes. Both the phases consider the data contained in the citing entities dataset produced by the first three phases (see the previous section). All the data and visualizations are available in the Zenodo repository (Heibi & Peroni, 2021c).

We provide a dedicated webpage (https://ivanhb.github.io/thesis_results/hum/) to enable readers to use the dynamic visualizations that we present in this chapter. The dynamic visualizations (i.e., HTML documents) are part of the results obtained after the topic modeling analysis (i.e., fifth phase) which could be customized using the filters and parameters available in such visualizations. During the discussion we present only screenshots of such visualizations. The dedicated webpage for the generated outcomes also contains the static graphics of the descriptive statistical analysis (i.e., fourth phase), yet these can't be modified dynamically (i.e., PNG images).

6.2.1 Phase 4: descriptive statistics

First, we have classified the distribution of the citing entities in the three periods (i.e., P-Pre, P-Ret, and P-Post) as a function of the humanities disciplines used in Retraction Watch, as shown in Figure 6.3. *Religion* was the discipline that received the highest number of citations (375), while *History* had the highest number of retracted items (20).

In Figure 6.4 we have classified the entities citing a retracted article in each discipline according to their subject areas. *Arts and humanities* and *Social Sciences* (AH&SS) were highly represented in both the P-Pre and P-Post periods of almost all the retracted articles' disciplines. However, we noticed some exceptions to this rule in P-Pre in *Journalism* (10% of citing entities were AH&SS publications), P-Post in *Arts* (13% AH&SS publications), and P-Pre and P-Post of *Architecture* (no AH&SS publications in both periods).

Since we expected, as also highlighted in previous studies (Ngah & Goi,1997), that a good part of the citations to humanities articles should come from AH&SS publications, we decided to deepen into the obtained results before moving on the next stage. As shown in Figure 6.4, we noticed that *Journalism* has a completely different behavior compared to the other disciplines. Indeed, the citations of *Journalism* have cited three retracted articles: two with a *dom_affinity* of 3, and one with a *dom_affinity* of 2. The latter article was *"Personality, stress and disease: description and validation of a new inventory"* (Grossart-Maticek & Eysenck, 1990). This article has 130 citations (almost 95% of all the citations in *Journalism*). Retraction Watch has labeled this article with the additional two subject areas: *Public Health and Safety*, and *Sociology*, therefore *Journalism* represents the only humanities subject. A further investigation in the full text of the paper revealed the fact that this article is highly related to health sciences, and *Journalism* has a marginal (almost absent) relevance in it. Considering these discovered facts, we felt that this article could represent a significant bias to our analysis. Therefore, to limit its impact on the results we decided to exclude it from our analysis.

As a further check, we have investigated all the retracted articles of all the humanities disciplines in Figure 6.4 having fewer than 20% citations from *Arts and Humanities* publication in either P-Pre or P-Post. *Arts* and *Architecture* are the two disciplines falling in this category. After a manual check, we detected the article *"A systematic review on post-implementation evaluation models of enterprise architecture artefacts"* (Nikpay et al., 2020), classified under *Architecture,* yet while reading its full text we found little evidence supporting the proposed labelling, since it was a computer science study. Therefore, we decided to also exclude this article from our analysis.

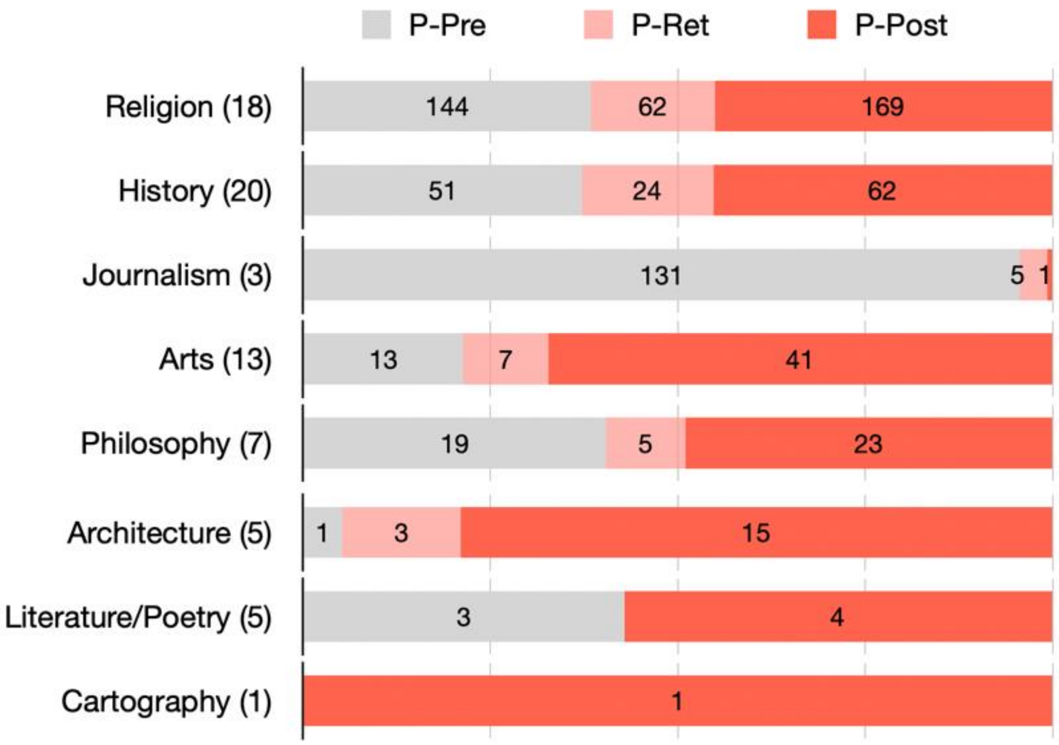


Figure 6.3. The number of citing entities in P-Pre, P-Ret, and P-Post for each different humanities discipline specified to the retracted article as gathered from Retraction Watch.

After this data refinement, our final data have been reduced to a total of 546 citing entities and 786 in-text citations to 70 retracted articles. We treated the citing entities and the in-text citations they contain as two different classes, and we present descriptive statistics of these two classes separately.

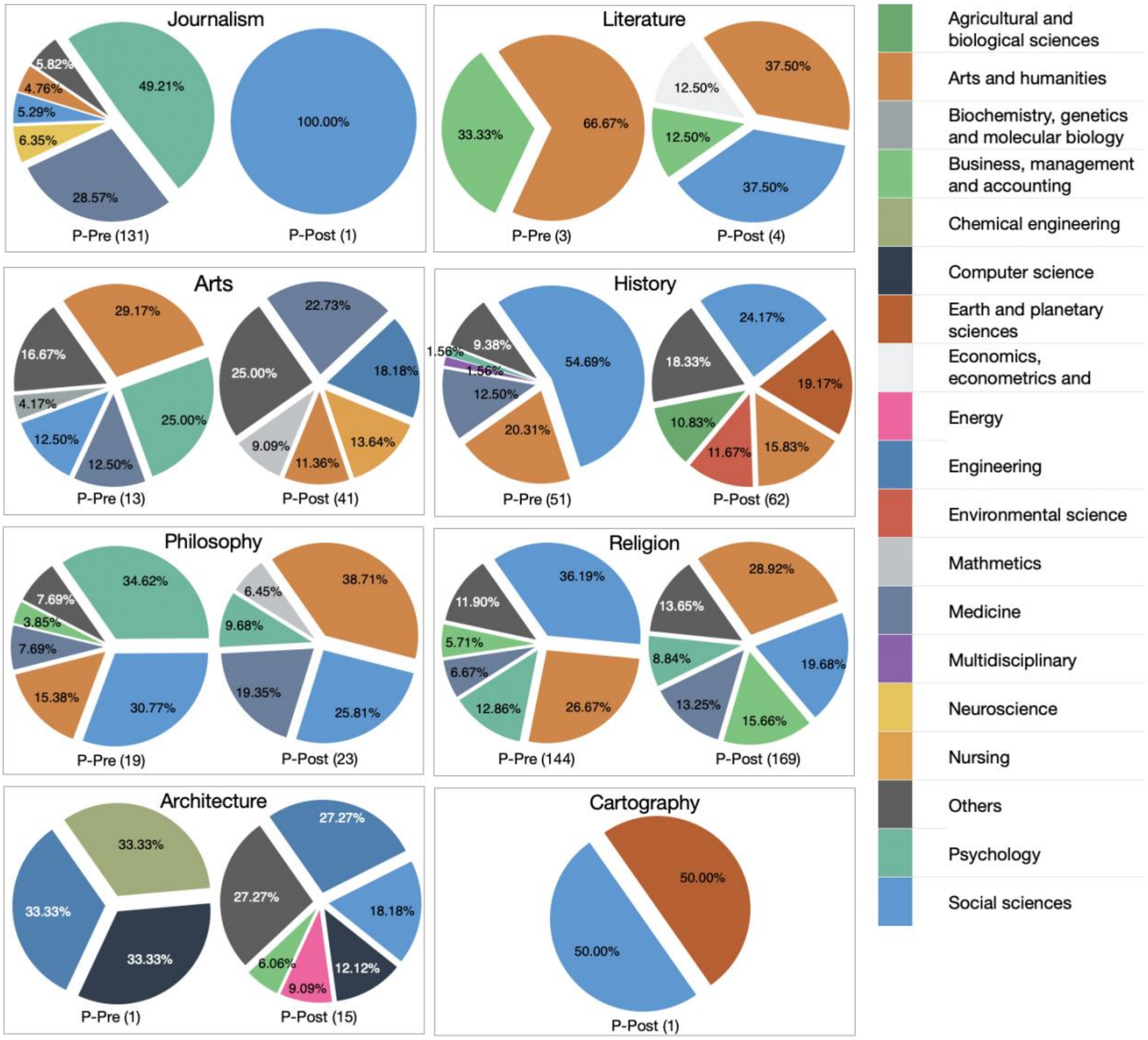


Figure 6.4. The subject areas distribution of the citing entities of the retracted articles in P-Pre and P-Post for each different humanities discipline as specified in Retraction Watch. The number of citing entities is mentioned between parentheses.

We examined the distribution of the citing entities to retracted articles as a function of two features: (1) the periods (i.e., P-Pre, P-Ret, and P-Post), further classified into those that mentioned the retraction or for which we could not access their full-text, and (2) their subject areas. The results are shown in Figure 6.5.

The number of the citing entities before the retraction (192, period P-Pre) was lower than the number of the citing entities after the retraction (260, period P-Post). Along P-Pre and P-Ret, we noticed a continuous increment in the overall number of citing entities, that suddenly started decreasing after the first fifth of P-Post, yet the numbers were in line with the ones observed in the third and fourth fifth of P-Pre. The last fifth of P-Post is an exception to the declining trend, with an unexpected high pick. This result was since 27 retracted items received only one citation in *P-Post* and, in these cases, that citation always represented the last citation received, which is the final border of P-Post.

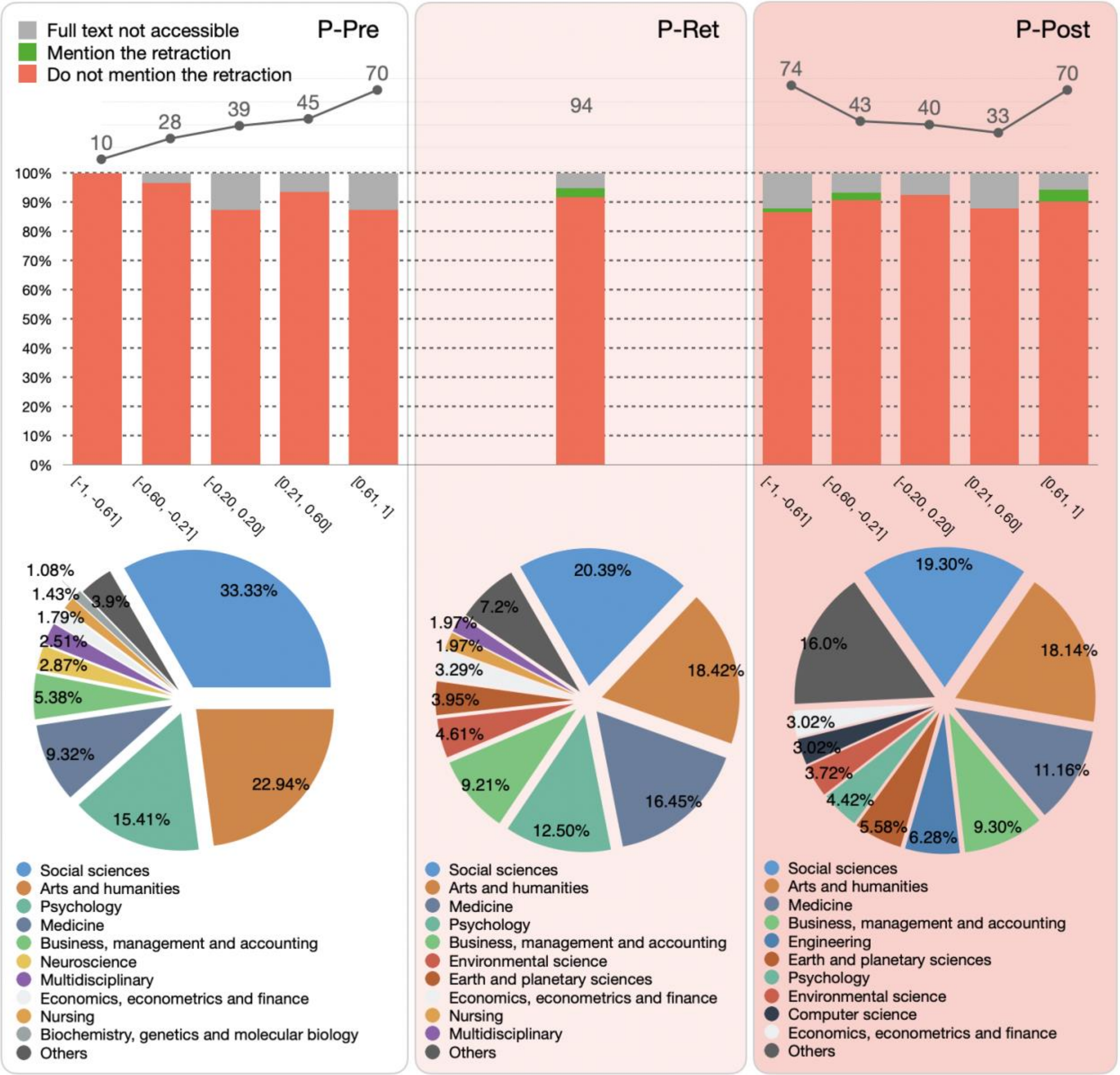


Figure 6.5. A descriptive statistical summary for the distribution of the citing entities to retracted articles in the three periods (i.e., P-Pre, P-Ret, and P-Post), also considering their subject areas. The bar charts on top highlights the citing entities that either did/did not mention the retraction and those for which we could not retrieve the full text.

The full text of 8.42% of the citing entities was not accessible. For those that we have successfully retrieved the full-text, our results showed that a relatively low percentage mentioned the retraction of the cited entity – 2.25% of the total number of citing entities in P- Ret and P-Post.

Looking at their subject areas, we noticed that the citing entities started to spread into a higher number of subject areas (i.e., additional 9) in P-Post compared to P-Pre, where the residual category *Others* contained 16% of the citing entities. The *Arts and Humanities* subject area had a similar percentage throughout all the three periods (22.94%, 18.42% and 18.14%), and it represents, together with *Social Sciences*, the 2 most representative subject areas in P-Ret and P-Post. We also noticed an important drop-down in *Psychology*, from 15.41% in *P-Pre* to a 4.42% in P-Post.

Regarding the in-text citations, we focused on their distribution as a function of three features: (1) the period (i.e., P-Pre, P-Ret, or P-Post), (2) the citation intent, and (3) the section containing the in-text citation. The results of the three distributions have been further classified according to the in-text citation sentiment (i.e., *negative/neutral/positive*), as shown in Figure 6.6.

The overall trend in the number of in-text citations along the three periods was close to the one we observed for the citing entities (shown in the previous section), although the differences between P-Pre and P-Post were even more marked. As introduced in the previous section, the pick in the last fifth of P-Post was due to the retracted items receiving only one citation in P-Post. Even though the overall percentage of negative citations was low, it had a higher presence in P-Pre (4.5%). Generally, most in-text citations were tagged as *neutral*, and very few were *positive* (0.75%).

The citation intents "*obtains background from*" and "*cites for information*" were the two most dominant ones in the three periods, and they respectively represented 31.29% and 22.64% of the total number of in-text citations, respectively. The citation intent "*cites for information*" increased its presence moving from 17.8% in P-Pre to 27.20% in P-Post.

Considering the citation sections, we can clearly see that the in-text citations were mostly located in the "*Introduction*" section in all the three periods. The in-text citations in the section "*Introduction*" decreased a lot after P-Ret moving from 30.15% in *P-Pre* to 22.13% in P-Post. Instead, the in-text citations contained in the section "*Discussion*" have an increasing trend, from 6.87% in P-Pre to 15.20% in P-Post.

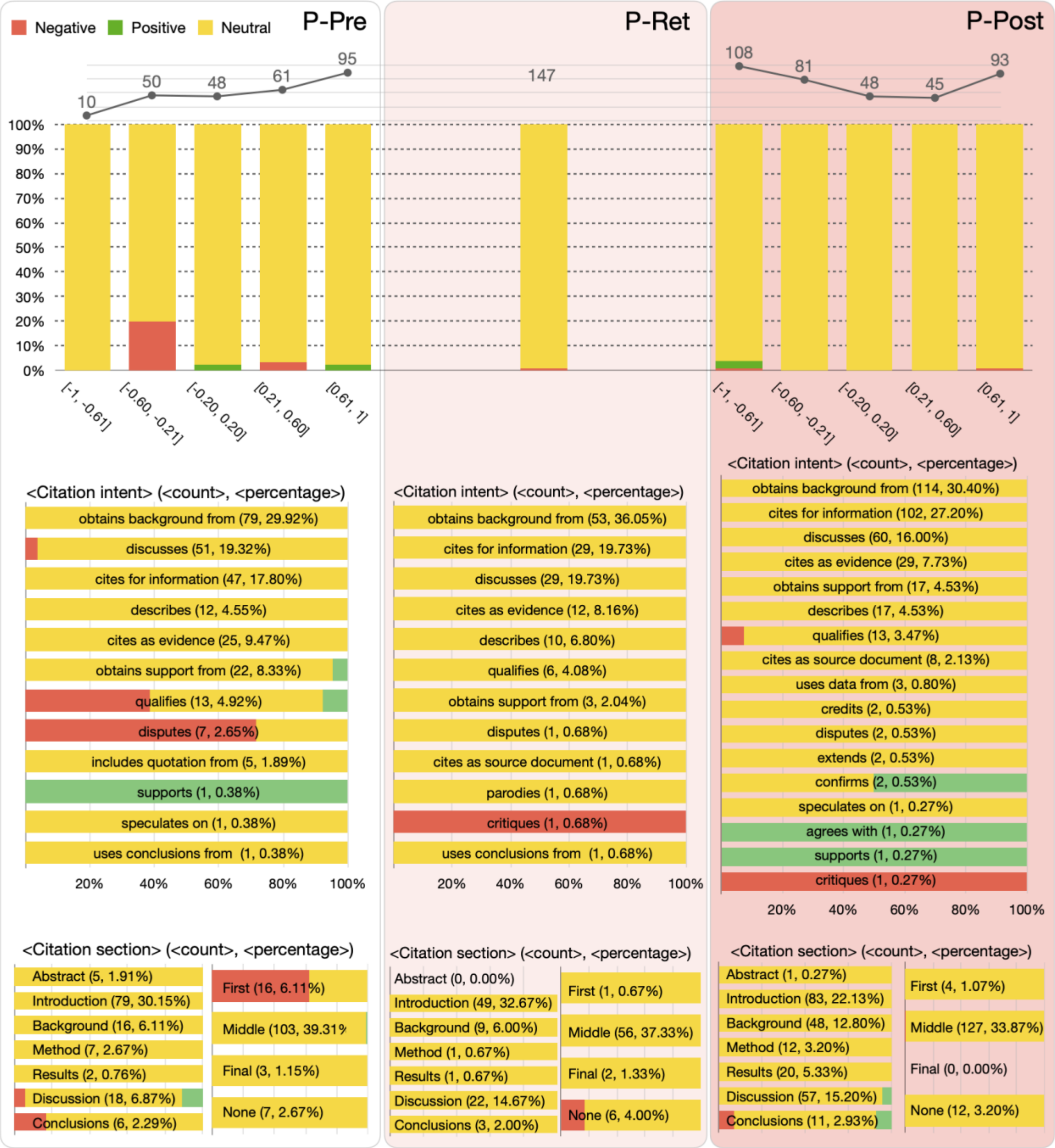


Figure 6.6. A descriptive statistical summary for the distribution of the in-text citations contained in the citing entities to the retracted articles in the three periods (i.e., P-Pre, P-Ret, and P-Post), according to their intent, and section. The sentiment of the in-text citations is also highlighted.

6.2.2 Phase 5: running a topic modeling analysis on abstracts and citation contexts

Before running a topic modeling analysis on the abstracts of the citing entities and the in-text citation contexts we needed to choose the best number of topics to use in the training process of the topic modeling by calculating the coherence score of different topic models (discussed in Section 5.1). Based on the results obtained (Figure 6.7 and Figure 6.8), we built and executed two LDA topic models, one using the abstracts of the entities citing the retracted articles (with 16 topics) – named *TM-Abs,* and another using the citation contexts where the in-text reference pointers to retracted articles were contained (with 20 topics) – named *TM-Cits*.

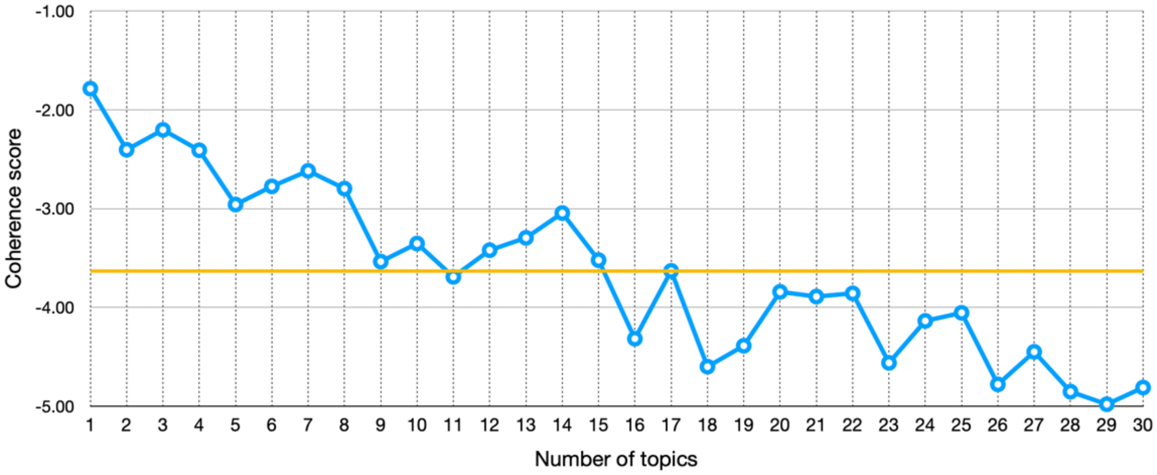


Figure 6.7. The coherence score (y-axis) of different LDA topic models built using a variable number of topics, from 1 to 30 (x-axis). The topic models are built using the corpus and dictionary of the citing entities abstracts. The orange line represents the average value, and it plateaus around 15–16 topics

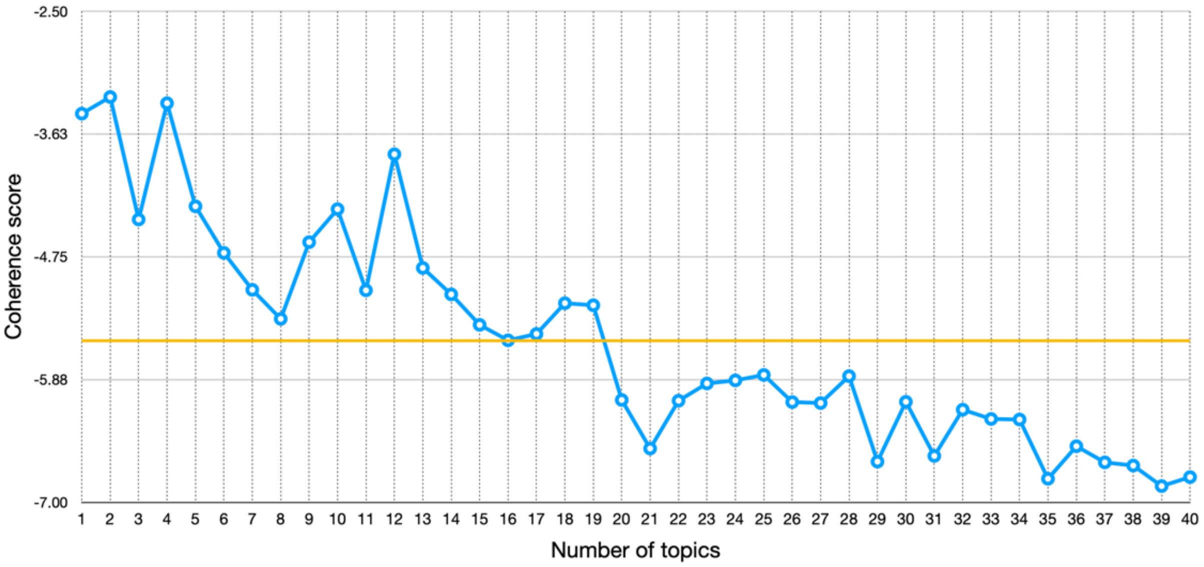


Figure 6.8. The coherence score (y-axis) of different LDA topic models built using a variable number of topics, from 1 to 40 (x-axis). The topic models are built using the corpus and dictionary of the in-text citation contexts. The orange line represents the average value, and it plateaus around 19–20 topics

The total number of abstracts extracted and annotated in our dataset was 509. We extended the list of MITAO's default English stop-words (e.g., "the", "is", etc.) with other ad-hoc stop-words such as "method", "results", "conclusions", etc., which represents the typical words that might be part of a structured abstract.

Figure 6.9 shows the topics represented in the two-dimensional space of LDAvis. Using LDAvis interface, we set the parameter λ to 0.3 to determine the weight given to the probability of a term under a specific topic relative to its lift (Sievert & Shirley, 2014), and retrieved the 30 most relevant terms of each topic. We gave an interpretation and a title to each topic by analyzing its related terms (Table 1 in Appendix) also available in (Heibi & Peroni, 2021c). Topic 6 ("Leadership organization, and management") was the dominant topic. The topics were distributed in four main clusters, as shown in Figure 6.9:

- one composed by topics 2 (“Socio-political issues related to leadership”) and 6, concerned issues related with leadership, work organization, and management form a socio-political point of view;
- a large one composed by topics 1 (“Socio-political issues possibly related to Vietnam”), 4 (“History of the Jewish culture”), 5 (“Music and psychological diseases”), 11 (“Family and religion”), etc.
- other two clusters composed by one topic each, i.e., topic 16 (“Geography and climatic issues”) and topic 3 (“Colonial history”).

Figure 6.10 shows the chart generated using MTMvis. We plotted the topics distribution as a function of the three periods. As a first analysis, we noticed how topics 6 and 16 increased their distribution along the three periods. On the other hand, topics 1 and 11 decreased their percentage throughout the three periods.

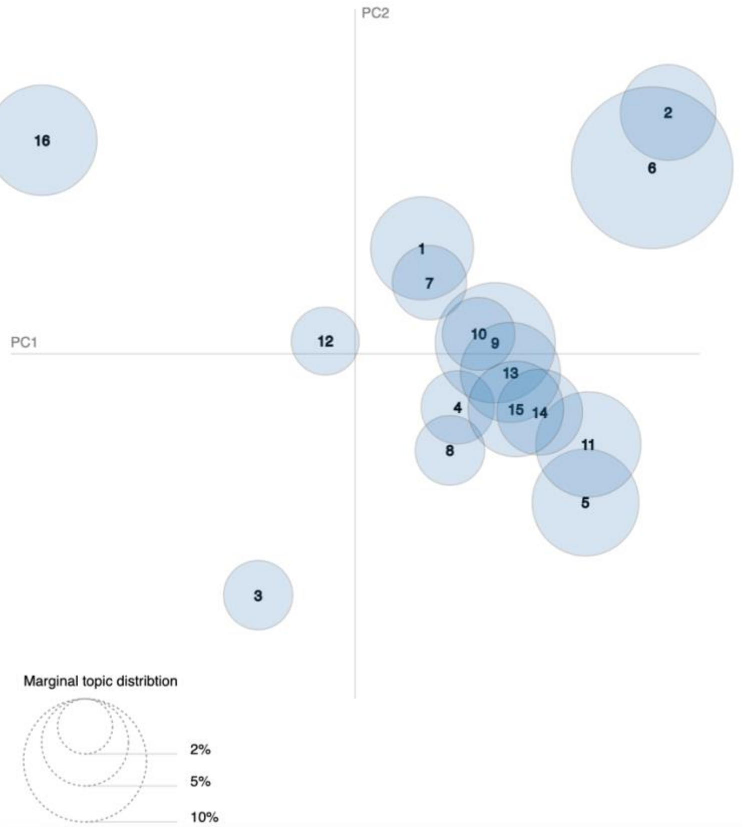


Figure 6.9. The 16 topics of *TM-Abs*. The visualization is taken from LDAvis, and it shows the topic distribution in a two-dimensional space.

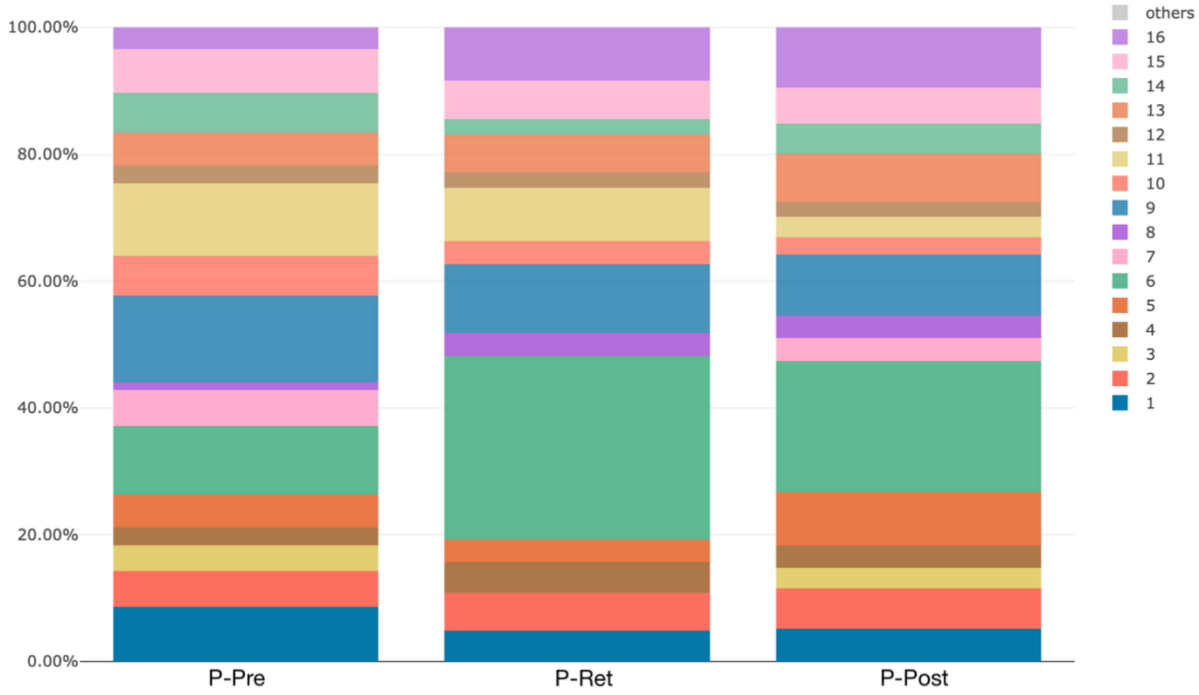


Figure 6.10. The MTMvis chart created over the 16 topics of *TM-Abs*. The topics are plotted as a function of the three periods (represented on the x-axis)

The total number of in-text citation contexts in our dataset, we used as input to produce the second topic model, was 786. As we did with the abstracts, we have defined and used a list of ad-hoc stop-words, which included all the given and family names of the authors of the cited articles.

Figure 6.11 shows the topics represented in the two-dimensional space of LDAvis. As we did for the abstracts' topic modeling, we set λ to 0.3 and interpreted each topic by analyzing its 30 most relevant terms (Table 2 in Appendix) also available in (Heibi & Peroni, 2021c). In this case, we noticed that the topics are less overlapping and more distributed along all the axis of the visualization. Topic 12 ("Leadership organization, and management") is the most representative (11.7%) and was very distant from the other topics. The bottom right part of the graphics – with topics 2 ("Countries in conflict"), 15 ("War and terrorism"), 17 ("War and history"), 18 ("History of Europe"), 20 ("War and army conflicts") – are mostly close to the history studies, especially discussion toward army conflicts. The part on the top of the graphics contains several single- topic clusters, such as topic 5 ("Gender social issues") and 9 ("Geography and climatic issues").

Figure 6.12 shows the chart generated using MTMvis, where we plotted the topic distribution as a function of the three periods. We noticed a continuous decrease in topics 7 ("Family and religion") and 18 along the three periods. Topic 3 ("Drugs/Alcohol and psychological diseases") had a high decrease right after P-Ret. On the other hand, we noticed an increment in topics 5, 9, and 11 ("Music and psychological diseases") – although the latter topic had a higher percentage in P-Ret than in P-Post.

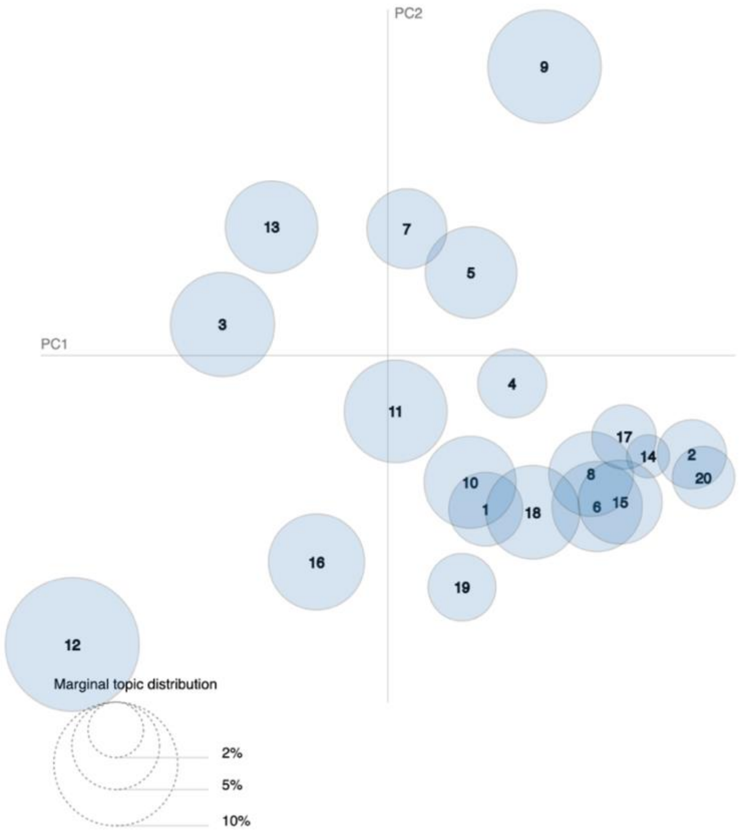


Figure 6.11. The 20 topics of *TM-Cits*. The visualization is taken from LDAvis, it shows the topic distribution in a two-dimensional space.

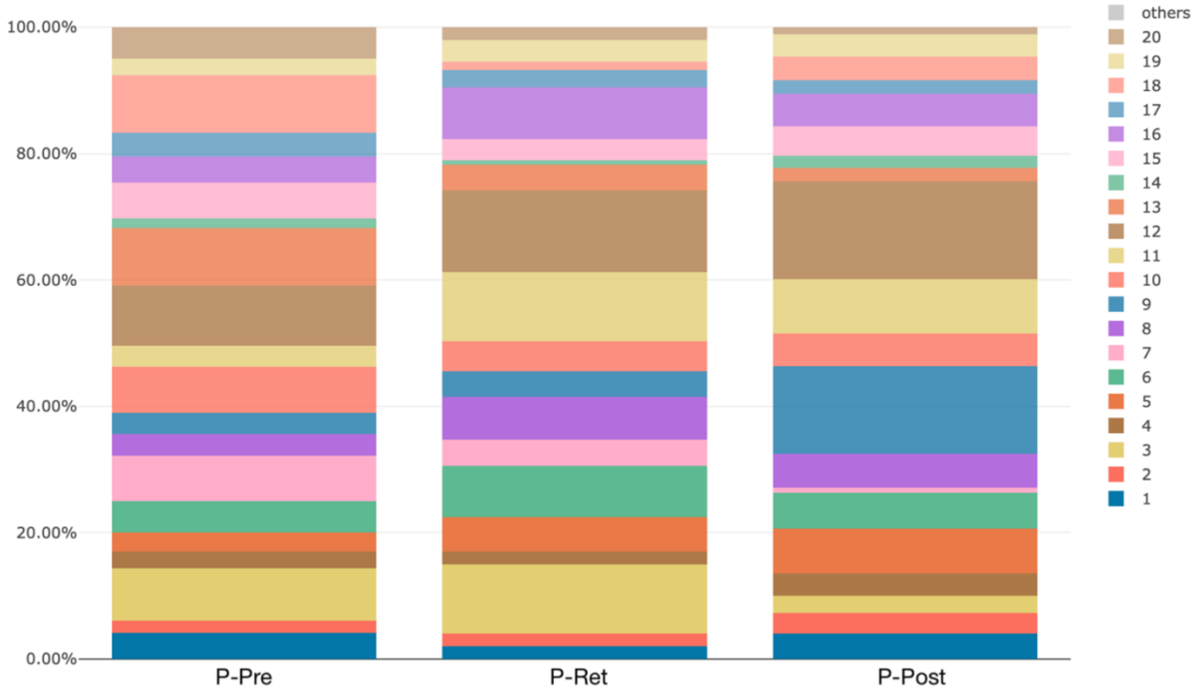


Figure 6.12. The MTMvis chart created over the 20 topics of *TM-Cits*. The topics are plotted as a function of the three periods (represented on the x-axis).

6.3 Addressing the research questions

In this section, we address separately the two research questions: (RQ1) How did scholarly research cite retracted humanities publications before and after their retraction? (RQ2) Did all the humanities areas behave similarly concerning the retraction phenomenon?

To answer both the research questions we consider the results (i.e., data, descriptive statistics, and the outcomes of the topic modeling analysis) of the methodology toward the retractions in the humanities domain, presented in the previous section (Section 6.2).

6.3.1 RQ1: citing humanities retractions

As shown in Figure 6.5, it seems that, on average, the articles in the humanities does not have a drop of citations after their retraction. Indeed, citing retracted articles is not prohibited and is reasonable when the retracted article is subject of a study, such as we have done in this thesis. However only 2.25% of the citing entities from P-Ret and P-Post – 5 *arts and humanities* publications and 3 related to health sciences subject areas (e.g., *medicine, psychology, nursing*, etc.) – mentioned the retraction in the citation context. In addition, we noticed that the negative perception of a retracted work, although limited in the data we have, happened before the actual retraction of the publication if the cited entity had a low affinity (i.e., *domain_affinity* = 2) to the humanities domain. Thus, by these facts, we can speculate that citing entities have been more inclined to explicitly give a negative sentiment toward the cited retractions only when the latter were less affiliated to the humanities domain.

Most of the in-text citations marked as *discusses* occurred in the *Discussion* section, which explains the higher distribution of this section in P-Ret and P-Post (as shown in Figure 6.14). By analyzing the distribution of the topics over the in-text citation sections, we noticed the emerging of topic 6 ("The retraction phenomenon") in *Discussion* sections only in P-Post – in other words, the retraction was not mentioned in the *Discussion* section before the retraction, and the retraction event might have been the trigger of a higher discussion from the citing entities. The results shown in Figure 6.15 suggested that the citing entities talking about retraction (mentioned the retraction in the contexts of their citations) have usually *discussed* the cited entity rather than obtaining background (i.e., *obtain background from*) material from it or generic informative claims (i.e., *cites for information*). In addition,

From the distribution of the subject areas of the citing entities over the three periods (Figure 6.5), we noticed that *social sciences* and *arts and humanities* had almost the same percentages in the *P-Ret* and *P-Post* periods, which is less than their percentages in *P-Pre*, suggesting that the retraction event had an impact on these subject areas. In addition, other subject areas such as *psychology* had a more significant decrease in *P-Ret* and *P-Post*, that may be an indicator of a higher concern by these subject areas toward the citation of retracted articles. In addition, if we take a look at the topic distribution of the abstracts topic model over the three periods for the citing articles assigned with the subject area *psychology* (shown in Figure 6.13), there was a clear decrement in the topics related to health sciences issues such as topics 10 and 11, while others such as topics 6 and 9 (close to socio-historical discussions with no relation to health sciences) increased their presence in P-Ret and P-Post. In other words, not only the overall number of the citing entities from the health sciences domain decreased after the retraction, but also the overall topics treated moved from the health sciences thematic to other subjects closer to *social sciences* and *arts and humanities*.

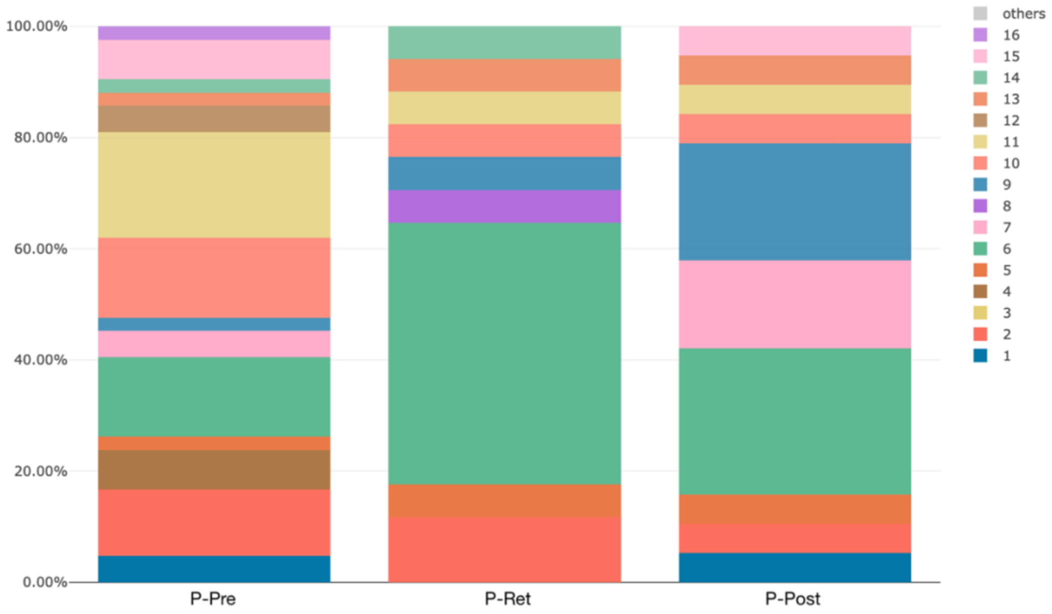


Figure 6.13. A filtered MTMvis to show the distribution of the topics of *TM-Abs* as a function of the three periods. The visualization is built considering only the documents (i.e., abstracts) that have *Psychology* as subject areas.

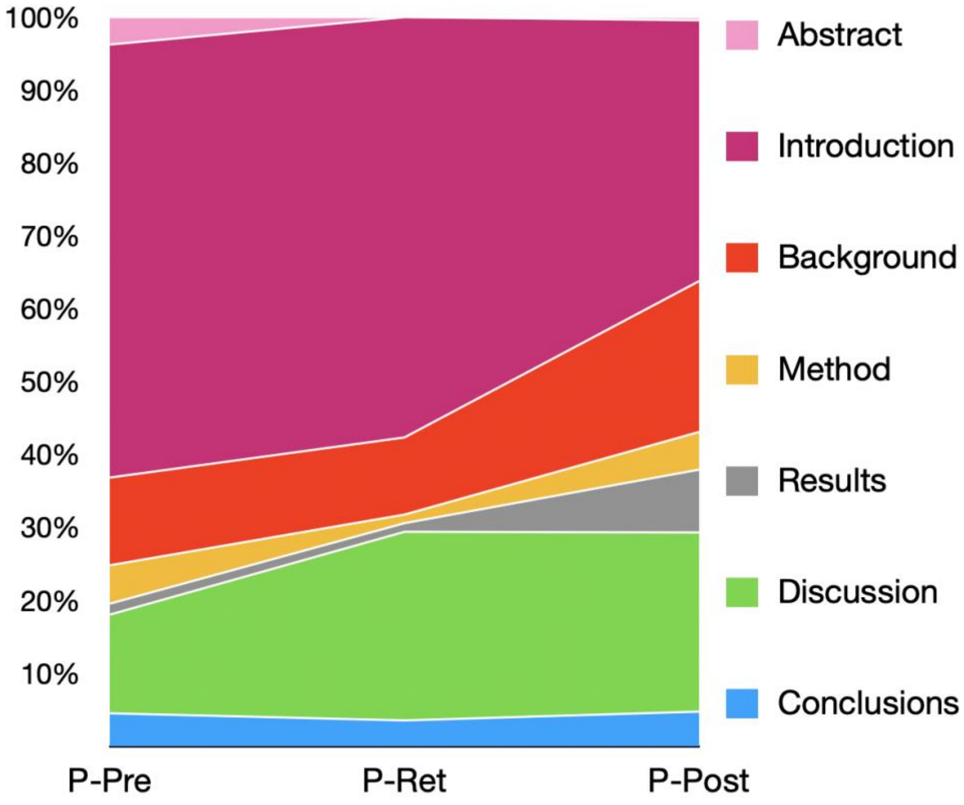


Figure 6.14. The distribution of the main (positional sections are not included, e.g., *first section*) in-text citation sections over the three periods.

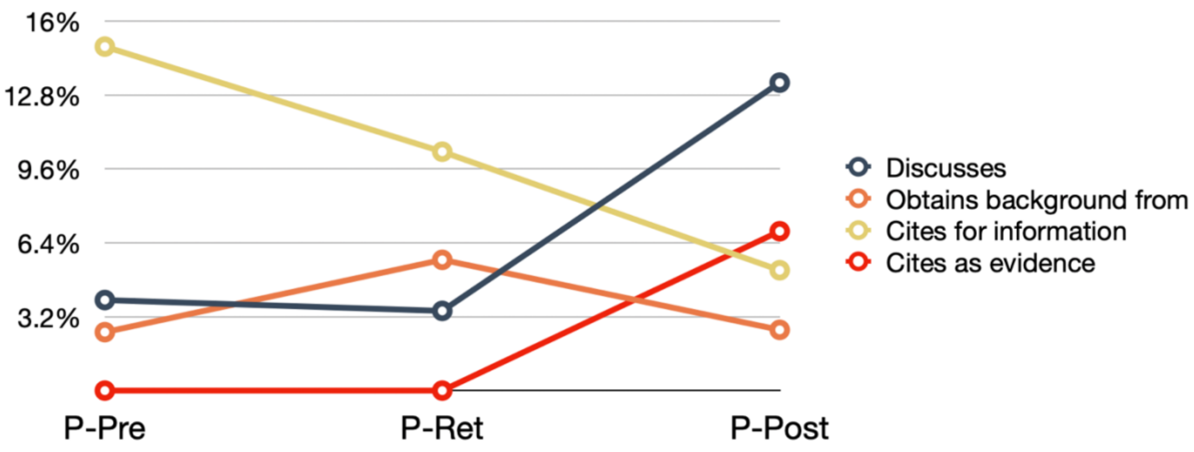


Figure 6.15. The distribution of topic 6 (“The retraction phenomenon”) of *TM-Cits* over the three periods for the 4 citation intents which have been used the most.

6.3.2 RQ2: behavior of the humanities areas

As shown in Figure 6.4, *Religion* and *History* had a very similar distribution pattern, in both, their citing entities labeled under the *social sciences* subject area had an important decrement in *P-Post*. After *P-Ret* the *TM-Cits* of these entities does not include topic 3 (“Drugs/Alcohol psychological diseases”) for *religion* and topic 7 (“Family and religion”) for *history.* We can speculate that *social sciences* studies significantly reduced its percentage due to a higher concern toward sensitive social subjects such as health care, family, and religion.

Arts had the highest number of citations in P-Post, although we reported an important drop in the citing entities associated to *Arts and humanities* in favor of subject areas such as *Medicine, Nursing* and *Engineering*, as shown in Figure 6.4. On the other hand, for *Philosophy,* we had a completely different situation: citing entities labeled as *Arts and humanities* incremented a lot in *P-Post* at the expense of citing entities from *Psychology*. In the case of *Arts*, the growing representation of topic 11 (“Music and psychological diseases”) of *TM-Cits* is the reason for the positive trend reported in P-Post. In other words, arts (and especially music) had been discussed with relation to psychological and medical diseases. Although in different ways, for both the situations (i.e., *Arts* and *Philosophy*), we can suggest that these facts are a demonstration of a high concern toward retraction. In both the situations this is true mainly thanks to the fact that the retracted publications from the humanities domain included other topics close to STEM disciplines in their studies which have been used by their citing entities.

In Figure 6.16, we show the distribution of topic 6 ("The retraction phenomenon") as a function of the three periods and considering the four humanities disciplines with the higher number of citing entities. Topic 6 increased a lot in P-Post in *Philosophy*, in *Religion* it showed a steady trend, while in *History* and *Arts* it had its pick in P-Ret and a lower, yet relatively high, percentage in P-Post. These results might suggest that the entities which cite retracted articles in *Philosophy*, *Arts*, and *History*, were those showing major concerns toward the retraction – in the case of *History* and *Arts*, starting from the year of the retraction. The sharp increase of *Philosophy* in P-Post is linked to our previous hypothesis, and it reflects the high sensitivity of the issues that this discipline used to talk about in its retracted publications.

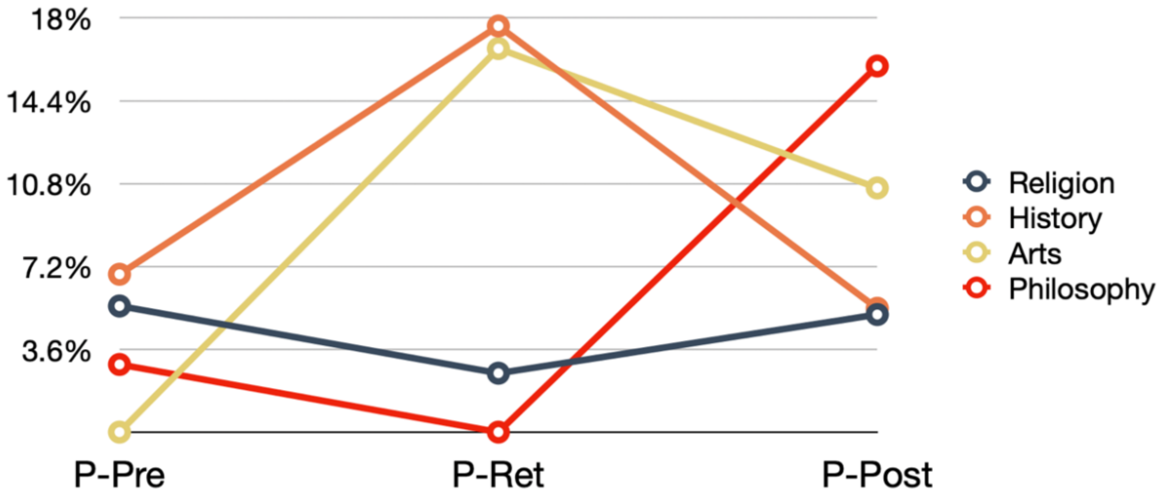


Figure 6.16. The distribution of topic 6 ("The retraction phenomenon") of *TM-Cits* over the three periods for the humanities disciplines: Religion, History, Arts and Philosophy.

Chapter 7

Humanities vs STEM: differences and similarities in retraction dynamics

In the previous chapter (Chapter 6) we worked exclusively on the analysis of retraction in the humanities domain by addressing the first two research questions of the thesis (i.e., RQ1 and RQ2). However, these outcomes have not been compared with other findings derived from the analysis of retraction in other domains (i.e., STEM). This chapter compares and brings out the differences and the similarities between the retraction in humanities and the retraction in STEM (that, for our analysis, includes disciplines in the life sciences, physical sciences, engineering, mathematics, computer science, and the health sciences). This analysis is formalized in the third research question of the thesis: (RQ3) What are the main differences and similarities in the retraction dynamics between the humanities domain and the STEM disciplines?

This comparison relies on the observations and findings learned from the analysis of retraction in the humanities domain (introduced in Chapter 6). This chapter splits the comparison and the answer of RQ3 into two levels: (L1) is based on a comparison with the outcomes of other past studies of retraction in STEM (Section 7.1), and (L2) works on a comparison with the results obtained on the analysis made using our methodology on a popular retraction case in STEM (Section 7.2).

In the analysis presented in the second level (i.e., L2), the methodology (defined in PART II) is tuned and applied to investigate a popular and highly cited retracted paper from the health science domain:

“Ileal-lymphoid-nodular hyperplasia, non-specific colitis and pervasive developmental disorder in children” by Wakefield et al. (1998). The workflow of L2 is divided in four main steps. In the first step we discuss and apply the three initial phases of the methodology which led to the generation of a dataset containing all the citing entities accompanied by their related attributes. In the second step, we discuss the fourth and the fifth phases which consider the information stored in the citing entities dataset, to respectively build a descriptive statistical summary and apply a topic modeling analysis on the abstracts of the citing entities and on the in-text citation contexts. Then on the third step, we present an open discussion and some general observations regarding the outcomes of the analysis. Finally, in the last step we address RQ3 by discussing and comparing the findings of the Wakefield et al. (1998) retraction analysis with the results and observations learned from the analysis of retraction in the humanities domain.

7.1 L1: comparison with past studies

The retraction of humanities publications showed a positive impact on the citation trend since the overall number has slightly increased after P-Ret (as shown in Figure 6.5). The work by Bar-Ilan and Halevi (2018) analyzed the citations of 995 retracted articles and found the same growing trend in the citations number in the post retraction period. However, they did not analyze the retraction through a differentiation of the retracted entities’ subject areas. As such, we might consider such results as a representation of a general trend for the retractions (with no-domain dependency), which is in line with the hypothesis and observations learned from our analysis of retraction in humanities.

However, the opposite trend was observed when the analysis focused on specific subject areas, for instance, when considering the prior studies conducted on the retractions of biomedicine (Dinh et al., 2019) and psychology (Yang et al., 2020). Based on these differences, it seems like the potential threats and damage of retracted materials has been perceived more seriously by others (i.e., citing entities) when the retracted publications have been linked to a sensitive area of study.

It is noteworthy to mention that other prior studies based on an analysis of general retractions in science (not belonging to a precise domain), e.g., the work of Chen et al. (2013), observed a decrease in the citations rates after the retraction (contrary to the results of Bar-Ilan and Halevi (2018)). Yet, in this case the results could be highly influenced by the fact that most of the retracted publications have been collected from PubMed, therefore might follow our hypothesis regarding the publications that deal with sensitive arguments. Generally, this case gives an idea on how several features can play a crucial rule in the interpretation of the results such as the resources used to search for retractions as well as those used to collect their citations.

The increasing trend in the number of citations has been observed in STEM when the analysis focused on one/limited number of retracted publications, for instance, in the results presented by the study of Schneider et al. (2020) on a retracted publication of falsified clinical trial data. This case might suggest that the discipline related to the retracted article is not the only central factor to consider for predicting the citation trend after the retraction, and that other factors might play a crucial role, such as the popularity and the media attention to the retraction case, as it has been discussed in the studies by Mott et al. (2019) and Bar-Ilan and Halevi (2017). The analysis toward the Schön case on a series of 30 retracted publications is also on the same line. However, this suggestion is very subjective to the specificity of the case, e.g., the work of Bornemann-Cimenti et al. (2016) on a series of 25 retracted publications authored by Scott S. Reuben is a counterexample to our speculation, since the rate of citation to his retracted articles distinctly dropped after the retraction.

Other important aspects concerning the characteristics of the in-text citations to retractions in STEM such as the sentiment, the intent, and whether or not they mentioned the retraction of the cited paper, have been investigated by some past studies, therefore is worthy to compare these findings with our results on the retraction in humanities. In the citation analysis presented by Bar-Ilan and Halevi (2017) on 15 cases of retraction (all in STEM), the average percentage of positive and negative citations reported is 87.7% and 3.2%, respectively. The low percentage of negative citations, which also implies the fact that the retraction was acknowledged in the text (according to the methodology of these studies), is almost in line with our findings on humanities (2.25% mentioned the retraction and 2% has been labeled as negative). Our percentage of citations mentioning the retraction is also close to the 4.5% reported by Schneider et al. (2020), as well as also to the results of Budd et al. (2011) reporting that 94% of the citations made no reference to the retraction of the entity they have cited, considering a collection of retracted articles in biomedicine.

On the other hand, the case is different when considering the results of studies on singular popular retraction cases. The analysis of Teixeira da Silva and Dobránszki (2017) done on 10 highly cited retracted papers (till 2017) demonstrated the positive high impact that retraction has on the number of citations even after just 1 month from the official retraction, e.g., the 819 citations gained in the 1 month of the post-retraction period of the retracted article of Fukuhara et al. (2005) in front of the 247 collected before the retraction. The work of Suelzer et al. (2019) on the Wakefield et al. retraction case, reported a relatively high number of citing entities mentioning the retraction in the citation contexts (38.3%), in addition 72.7% of the citing entities expressed a negative sentiment toward the retracted article. This latest case confirms our previous speculation regarding the importance of the popularity and the media attention to the retraction case and how this effect can vary subjectively depending on the retraction case.

7.2 L2: comparison with a new study

On the previous section we introduced the first level and we worked on presenting our first answers to RQ3 through a comparison of our findings in the analysis of retraction in humanities with the results presented by other previous studies on STEM retractions.

Our methodology and some of the previous approaches (presented in L1) investigated several common features such as the distribution of the citing entities over the time, the sentiment, and whether the retraction of the cited entity was addressed in the context of the citation. However, several other features that we have analyzed throughout our methodology have not been addressed in these past analysis, such as the citation intent, the subject area of the citing entities, etc.

To overcome this deficit and extend our answer/discussion toward RQ3, we worked on the application of the same methodology for a retraction case in STEM. This operation allows us to have the same methodological analysis and infer results with respect to the same features as it has been done with the analysis of the retraction in humanities. In addition, to let the analysis in STEM be as close as possible to the humanities analysis (presented in Chapter 6), we worked on the selection of a retraction case in STEM having an amount of data (i.e., citations) with an order of magnitude that matches the quantity of data collected for the humanities case.

This second level of analysis of RQ3, is based on applying and tunning the methodology (presented in PART II), with respect to its phases, for the investigation of a popular and highly cited retracted paper: *"Ileal-lymphoid-nodular hyperplasia, non-specific colitis and pervasive developmental disorder in children"* by Wakefield et al. (1998). This elaboration is split into four main parts. First, we apply the three initial phases of the methodology which led to the generation of the dataset containing the citing entities accompanied by their related attributes. In the second part, we discuss the fourth and the fifth phases of the methodology which consider the information stored in the citing entities dataset, to respectively build a descriptive statistical summary and apply a topic modeling analysis on the abstracts of the in-text citation contexts. Then in the third part, we discuss the obtained results and highlight the most significant outcomes. Finally, in the last part, we address RQ3 through a comparison of the obtained outcomes with the findings observed in the retraction analysis of the humanities domain.

7.2.1 Building the citing entities dataset

This section goes through the first three phases of the methodology which lead to the production of a dataset containing all the entities which have cited the Wakefield et al. (1998) retracted paper accompanied by their related attributes. Each phase is discussed in a separate section. For each phase, we illustrate how it has been configured and applied in the analysis of our retraction case. Throughout this chapter we list all the features used to characterize the citing entities. Such features are part of the final dataset outputted after the execution of the first three phases and they are being included in it gradually phase-by-phase. The dataset has been published also into a dedicated online repository in Zenodo (Heibi & Peroni, 2020b).

Phase 1: defining a case study/domain

We focused on a highly cited retracted article in STEM, i.e., Wakefield et al. (1998), this paper suggested a link between autism and childhood vaccines. The article was partially retracted in 2004 and subsequently fully retracted in 2010. In the rest of the discussion of this thesis, we refer to it with the abbreviation WF-PUB-1998.

The retraction of WF-PUB-1998 is an important case that deserves to be analyzed considering its popularity among several anti-vaccine movements and the implications it has had for society (Chen et al. 2013). WF-PUB-1998 is also one of the most cited retracted papers and is an important example of retraction in the STEM domain. Unlike the humanities case (to be analyzed in Chapter 7), in STEM we have a high number of retractions which makes the analysis of our methodology and especially the one that requires a manual elaboration (e.g., the intent of each in-text citation) highly expensive from a time perspective (since the feature annotation is done by only one reviewer). Therefore, we focused on one singular case with enough data to collect which needs a reasonable time for its analysis using our methodology. Its popularity helped us collect a large quantity of citations which is crucial for us to have a higher quality of data (less prone to bias) and consequently conduce important analysis for the generation of meaningful results that can help us answering the research questions we have raised in this thesis (i.e., especially RQ1 and RQ4).

We used the Retraction Watch Database to search for WF-PUB-1998 and got two different records. In Table 7.1 the two records are shown in two separate columns, i.e., the second and third which represent respectively the full and the partial retraction records. The records in the Retraction Watch Database are characterized with the attributes listed in the first column. The main differences between the two records are reported in the values of the attributes “Reason(s)” (second row), “Retraction or Other Notices; Date/PubMedID/DOI” (fifth row), and “Article Type(s); Nature of Notice” (sixth row).

The RET-SET defined in this phase consists in one retracted publication only (i.e., WF-PUB-1998), and we used the data of Retraction Watch to characterize it (i.e., Table 7.1). On Table 7.2 we show how the retracted publication of WF-PUB-1998 has been defined in RET-SET. The required attributes are listed in the first column, while the corresponding values are annotated in the second column. Although the attribute “Year of the last citation received” appears with a corresponding value, this value was inferred just after the end of the second phase of the methodology, since it relies on the quantity of citations gathered.

Table 7.1. The records stored in Retraction Watch Database regarding the retraction case of WF-PUB-1998. The second and third columns represent respectively the full and the partial retraction records. Each record in the Retraction Watch Database is characterized with the attributes listed in the first column.

Retraction or Other Notices **Title/Subject(s)/Journal ---Publisher/Affiliation(s)/Retraction Watch Post URL(s)**	Ileal-lymphoid-nodular Hyperplasia, Non-specific Colitis, and Pervasive Developmental Disorder in Children (BLS) Biology - Cancer; (BLS) Biology - Cellular; (BLS) Neuroscience; (HSC) Medicine - Immunology; (HSC) Medicine - Neurology; (HSC) Medicine - Pediatrics; Lancet ---Elsevier Inflammatory Bowel Disease Study Group, University Departments of Medicine and Histopathology, … http://retractionwatch.com/the-retraction-watch-leaderboard/top-10-most-highly-cited-retracted-papers/ http://retractionwatch.com/2015/02/03/frauds-long-tail-measles-outbreak-shows-important-look-downstream-retractions/	Ileal-lymphoid-nodular Hyperplasia, Non-specific Colitis, and Pervasive Developmental Disorder in Children (BLS) Biology - Cancer; (BLS) Biology - Cellular; (BLS) Neuroscience; (HSC) Medicine - Immunology; (HSC) Medicine - Neurology; (HSC) Medicine - Pediatrics; Lancet ---Elsevier Inflammatory Bowel Disease Study Group, University Departments of Medicine and Histopathology, … http://retractionwatch.com/the-retraction-watch-leaderboard/top-10-most-highly-cited-retracted-papers/
Reason(s)	+Falsification/Fabrication of Data; +Investigation by Company/Institution; +Investigation by Third Party; +Lack of Approval from Company/Institution; +Lack of IRB/IACUC Approval; +Manipulation of Results; +Upgrade/Update of Prior Notice	+Concerns/Issues About Results; +Error in Results and/or Conclusions
Author(s)	Andrew J Wakefield, Simon Harry Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, John Angus Walker-Smith	Andrew J Wakefield, Simon Harry Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, John Angus Walker-Smith
Original Paper; Date/PubMedID/DOI	02/28/1998; 9500320; 10.1016/S0140-6736(97)11096-0	02/28/1998; 9500320; 10.1016/S0140-6736(97)11096-0
Retraction or Other Notices; Date/PubMedID/DOI	02/06/2010; 20137807; 10.1016/S0140-6736(10)60175-4	03/06/2004; 15016483; 10.1016/S0140-6736(04)15715-2
Article Type(s); Nature of Notice	Clinical Study; Retraction	Clinical Study; Correction
Countries; Paywalled?; Notes	United Kingdom; No	United Kingdom; No

Table 7.2. The RET-SET represented by one item (i.e., retraction of WF-PUB-1998) characterized by the values (second column) defined for the required attributes.

DOI	10.1016/S0140-6736(97)11096-0
Author(s)	Andrew J Wakefield, Simon Harry Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, John Angus Walker-Smith
Year of publication (*E-RetPub*)	1998
Year of partial retraction (*E-PR*)	2004
Year of full retraction (*E-FR)*	2010
Year of the last citation received (*E-LastCit*)	2017
Subject area	Biology, Medicine
Subject category	Cancer, Cellular, Neuroscience; Medicine, Immunology, Neurology, Pediatrics

Considering the attributes of Table 7.2 and the fact that WF-PUB-1998 has a partial and a full retraction, we define five different periods (using the terminology of the methodology):

- P0, with $P_{\mathrm{CIT}} = [1998{,}1999{,}2000{,}2001{,}2002{,}2003]$
- P1, with $P_{\mathrm{CIT}} = [2004]$
- P2, with $P_{\mathrm{CIT}} = [2005, 2006, 2007, 2008, 2009]$
- P3, with $P_{\mathrm{CIT}} = [2010]$
- P2, with $P_{\mathrm{CIT}} = [2011{,}2012{,}2013{,}2014{,}2015{,}2016{,}2017]$

Phase 2: identifying, retrieving, and characterizing the citing entities

To gather the citation data for our analysis we used the COCI dataset of OpenCitations. We queried the COCI REST API (http://opencitations.net/index/coci/api/v1) when COCI was populated with citation data from its November 2018 release (OpenCitations 2018), that contained 445,826,118 citation links coming from 46,534,705 bibliographic resources. We retrieved all the required features for each citing entity of RET-SET (i.e., the retracted paper of WF-PUB-1998) starting from its DOI value. For doing that, we used the "citations" operation of the OpenCitations COCI API (http://opencitations.net/index/coci/api/v1#/citations/{doi}) to get the list of all citing entities, then we used the "metadata" operation (http://opencitations.net/index/coci/api/v1#/metadata/{dois}) to get the metadata of each citing entity.

At the end of this phase, we gathered and populated the citing entities dataset with a total of 616 citing entities. In Table 7.3 we present a summary representing the status of the dataset. Each citing entity (i.e., a record in the dataset) has been characterized with the main attributes required by this phase (listed in the first column): DOI, year of publication, title, venue id, venue title, subject area, and subject category. Each attribute (i.e., a column in the dataset) is summarized with the total number of entities having a corresponding value and a brief description of the nature of the collected values.

Table 7.3. A summary representing the status of the citing entities dataset after the execution of the second phase. Each citing entity in the dataset is characterized with the attributes (first column): DOI, year of publication, title, venue id, venue title, subject area, and subject category. For each attribute, on the second column, is summarized the total number of entities with a corresponding value and the nature of such values.

DOI	**Total:** All the citing entities had a value specified (616)
Year of publication	**Total:** All the citing entities had a value specified (616) **Values:** From 1998 (year of publication of WF-PUB-1998) to 2017
Title	**Total:** All the citing entities had a value specified (616)
Venue ID	**Total:** 599 (97%) citing entities had a value specified **Values:** ISSNs (548), ISBNs (51)
Venue title	**Total:** 603 (98%) citing entities had a value specified
Is/is not retracted	**Total:** 1 citing entity
Subject area	**Total:** 576 (93%) citing entities had at least a value specified **Values:** 24 different values: "medicine" (380), "social sciences" (90), "nursing" (81), "biochemistry, genetics and molecular biology" (59), "psychology" (58), "pharmacology, toxicology and pharmaceutics" (54), "immunology and microbiology" (52), "arts and humanities" (28), "neuroscience" (24),

	"environmental science" (17), "agricultural and biological sciences" (16), "health professions" (15), "computer science" (13), "mathematics" (10), "business, management and accounting" (8), "engineering" (7), "dentistry" (7), "multidisciplinary" (7), "decision sciences" (7), "economics, econometrics and finance" (5), "earth and planetary sciences" (1), "chemical engineering" (1), "materials science" (1), "physics and astronomy" (1)
Subject category	**Total:** 576 (93%) citing entities had a value specified **Values:** 170 different values

Phase 3: extracting additional data and labeling the textual content of the citing entities

Using the attributes (mainly the DOI value) of the citing entities in the dataset (discussed in the previous section and summarized in Table 7.3) we searched for and retrieved the full text of the citing entities. From the full-text, we extracted the abstracts of the citing entities, the in-text reference pointers denoting a bibliographic reference referencing the RET-SET (i.e., WF-PUB-1998, e.g., "Wakefield et al. 1998"), the citation contexts of the in-text citations, and the sections where the citation contexts are contained. We have collected a total of 870 in-text citations (1.4 in-text citations per citing entity on average) and we annotated 464 citing entities with a corresponding abstract. After the extraction of the textual values (step 3.a), following the rules we have established and discussed previously in Section 4.3, we have annotated the citation intent, citation sentiment, and whether the retraction has been/has not been addressed in the context of the citation. In Table 7.4 we have listed the new attributes added to the citing entities dataset after the execution of this third phase. For each new attribute added to the dataset (listed in the first column) we have summarized the total number of entities with a corresponding value and the nature of such values (the second column).

Table 7.4. A summary of the new attributes added to the dataset of the citing entities after the execution of the third phase. For each attribute (listed in the first column), on the second column, we summarize the total number of entities with a corresponding value and the nature of such values.

Abstract	**Total:** 464 citing entities had a value specified
In-text citation contexts	**Total:** all the in-text citations had a value specified (i.e., 870)

In-text citation sections	**Total:** 757 (87%) in-text citations had a value specified **Values:** 10 different values: *abstract* (5), *introduction* (167)*, background* (34)*, method* (15), *results* (28)*, discussion* (61), *conclusions* (17)
In-text reference pointers	**Total:** all the in-text citations had a value specified (i.e., 870)
In-text citation intents	**Total:** all the in-text citations had a value specified **Values:** 17 different values: *discusses* (226)*, disputes* (114)*, credits* (95)*, cites for informa- tion* (90)*, cites as evidence* (74)*, qualifies* (70)*, describes* (60)*, obtains background from* (56)*, critiques* (55)*, includes excerpt from* (8)*, obtains support from* (6)*, uses data from* (5)*, uses conclu- sions from* (4)*, ridicules* (4)*, extends* (1)*, updates* (1)*, refutes* (1)
In-text citation sentiments	**Total:** All the in-text citations had a value specified **Values:** *neutral* (549), *negative* (300), *positive* (21)
Retraction is/is not mentioned	**Total:** All the citing entities had a value specified (i.e., 616) **Values:** *no* (465), *yes* (151)

7.2.2 Descriptive statistics and topic modeling

In this section, we discuss the analysis and show the results of the last two phases of the methodology (i.e., fourth and fifth phases) which are based on the observations and data contained in the citing entities dataset produced by the first three phases (see the previous section).

Although the graphics we present regarding the results of the topic modeling analysis (i.e., fifth phase) are screenshots of the visualizations, these are also provided in dynamic HTML documents and each visualization could be customized using the filters and parameters it makes available. We provide a dedicated webpage (https://ivanhb.github.io/thesis_results/wakefield/) to enable readers to use such dynamic visualizations that we present in this chapter. The dedicated webpage includes also the static charts presented in the fourth phase (i.e., descriptive statistics). These visualizations are also part of the online repository of Zenodo (Heibi & Peroni, 2020b).

Phase 4: descriptive statistics

We organize the presentation of the results in two parts by first describing the citing entities and then their in-text citations. For both, we generate a descriptive statistics summary based on the data that characterize them which are part of the citing entities dataset we have produced.

Figure 7.1 and Figure 7.2 introduce some descriptive statistics of the citing entities based on the values contained in the citing entities dataset. Figure 7.1 is dedicated to periods P0, P2, and P4 (with an inner distribution in fifths), while Figure 7.2 is dedicated to the periods P1 and P3 (the exact same year of the partial and full retraction of WF-PUB-1998). The charts of both the figures are organized in distinct rows, one for each period considered. For both the figures, the second column contains the distribution per year of the citing entities according to the fact that they either mention the retraction (in green) or they do not (in red). On top of each bar in the chart, we also specify the number of citing entities the bar refers to. The third column contains the subject areas of the citing entities. The chart shows the ten most represented areas of study, while it groups all the other values (if any) in the last slice of the pie with the "Others" label.

From Figure 7.1 we noticed that the overall number of citing entities increments as moving from P0 (106), P2 (120) to P4 (337), with a high significant increment in P4. The number of the citing entities tend to be higher in the fifths close to P1, with 27 in [0.61, - 1] of P0 and 33 in [-1, -0.61] of P2, or which are close to P3, with 28 in [0.61, - 1] of P2 and 95 in [-1, -0.61] of P4. The incrementing trend is also reflected by the percentage of entities mentioning the retraction, the highest percentage (51%) is reported in [0.61, 1] of P4, on the other hand the lower percentage of entities which have mentioned the retraction after its partial retraction (i.e., P1) is reported in [0.61, 1] of P2 (7%).

Considering the distribution of the areas of study, we observed a slightly decreasing presence of the *medicine* area in favor of other areas of study which gained much more relevancy in P2 and P4 (e.g., *social sciences* being 1.35%, 8.38% and 12.59% in P0, P2 and P4 respectively). In addition, we noticed the emerging of many new areas in P2 and P4, such as *economics* and *environmental science, computer science, mathematics,* etc.

Data regarding the periods P1 and P3 and their descriptive statistics in Figure 7.2 are hard to interpret and could be highly biased, for instance a citing entity in P3 could has been written in the period P2 (in the last fifth) and published later in P3. Therefore, is hard to assume that such article had previous knowledge on regarding the full retraction of the WF-PUB-1998 while citing it. Despite these facts, it is worth mentioning that while looking at Figure 7.2 we noticed that the results are in line with the ones reported for the other periods (i.e., P0, P2, and P4).

Figure 7.3 and Figure 7.4 highlight some descriptive statistics of the in-text citations based on the features that characterize them contained in the citing entities dataset. As we have already done for the descriptive statistics of the citing entities, the data are separated in two figures and the charts of the figures are organized same as before (i.e., a distinct row for each period). Figure 7.3 is dedicated to periods P0, P2, and P4 (which have a distribution in fifths), and Figure 7.4 is dedicated to periods P1 and P3 (the exact same year of the partial and full retraction of WF-PUB-1998).

From the distribution of the in-text citations over the fifths (second column of Figure 7.3) we notice almost the same trend we have observed previously for the descriptive statistics of the citing entities (it is reasonable since the in-text citations are contained in the citing entities). The negative sentiment toward WF-PUB-1998 is reported even before its partial retraction (the highest percentage, i.e., 19.15%, belongs to the fifth [-0.20, 0.20] of P0). The percentage of negative in-text citations keeps incrementing along the periods P2 and P4 (the highest percentage, i.e., 55.93%, is reached in [-0.60, -021] of P4). In addition, it is worth mentioning the fact that we have reported a low percentage of positive in-text citations in P2, i.e., 6.7% in both the fifths [-0.61, -021] and [-0.20, 0.20].

When observing the citation intent distribution in Figure 7.3, we notice that *discusses* is by far the highest reason for citing WF-PUB-1998 with 22.89%, 23.63%, and 27.60% respectively in P0, P2 and P4. When moving from P2 to P4 *disputes* is the citation intent which have lost the most in its representatives (losing 4 positions in the overall rank) in favor of other intents such as *cites for information* and *qualifies*. The citation reason *critiques* emerges starting from P2, with 6.59% and 8.82% respectively in P2 and P4.

The in-text citations appeared mostly in the *introduction* section in all the periods. In-text citations appearing in the *discussion* section dropped significantly after P0, moving from 11.45% in P0 to 1.65% and 7.47%, in P2 and P4 respectively. We also reported a small percentage of in-text citations in P2 appearing in the *abstract* (i.e., 1.65%).

We omit the interpretation and discussion of the descriptive statistics of Figure 7.4 (related to the periods P1 and P3) for the same reasons we have mentioned previously when analyzing the descriptive statistics of the citing entities. Yet again despite these facts, from Figure 7.4 we noticed that generally the negative sentiment had a significant and higher impact in P3 compared to P1.

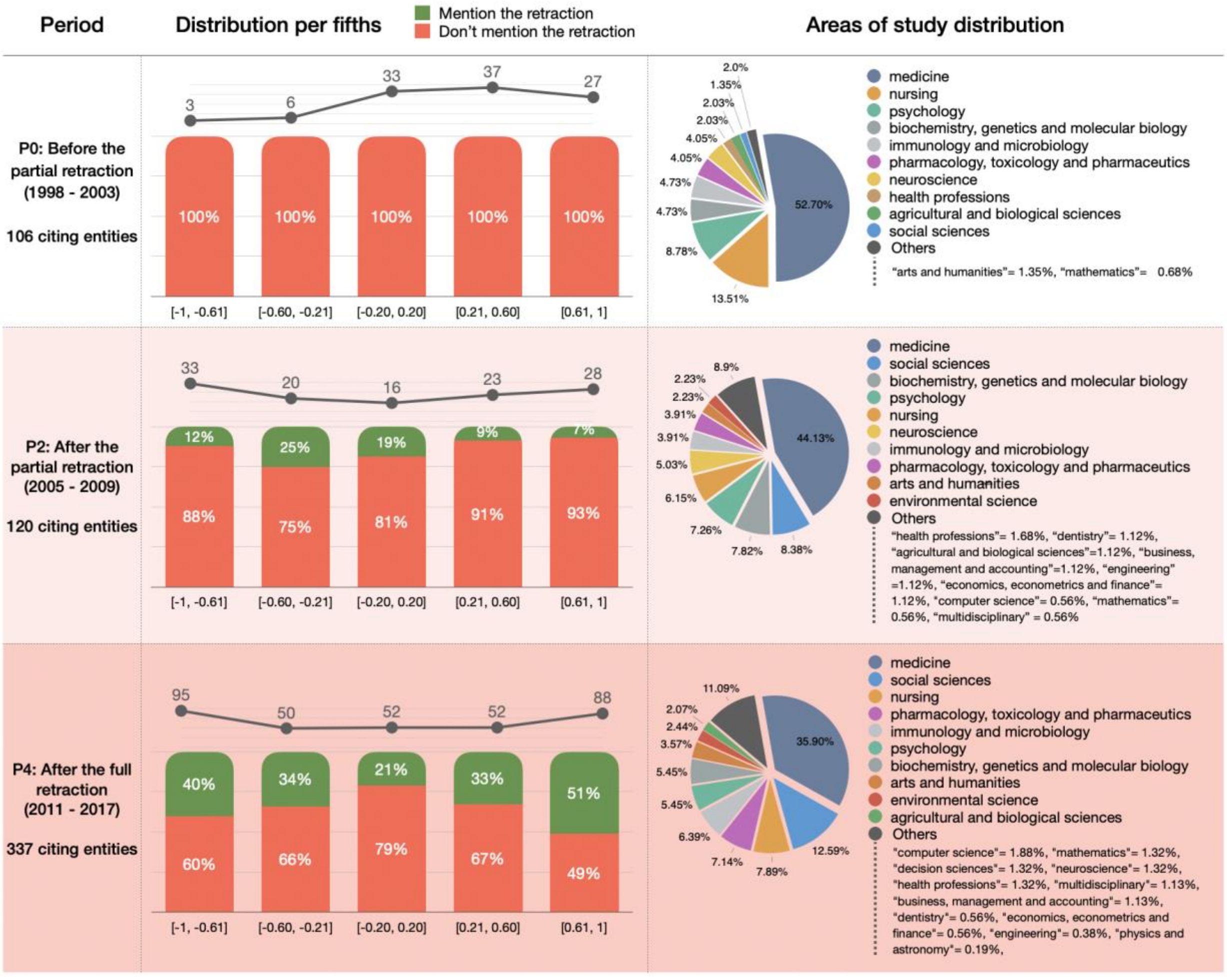


Figure 7.1. A summary of the citing entities. The first column contains the periods P0, P2 and P4, the second column shows the distribution per fifths of the citing entities that do mention (in green) or do not mention (in red) the retraction, while the third column shows the distribution of the subject areas of the citing entities

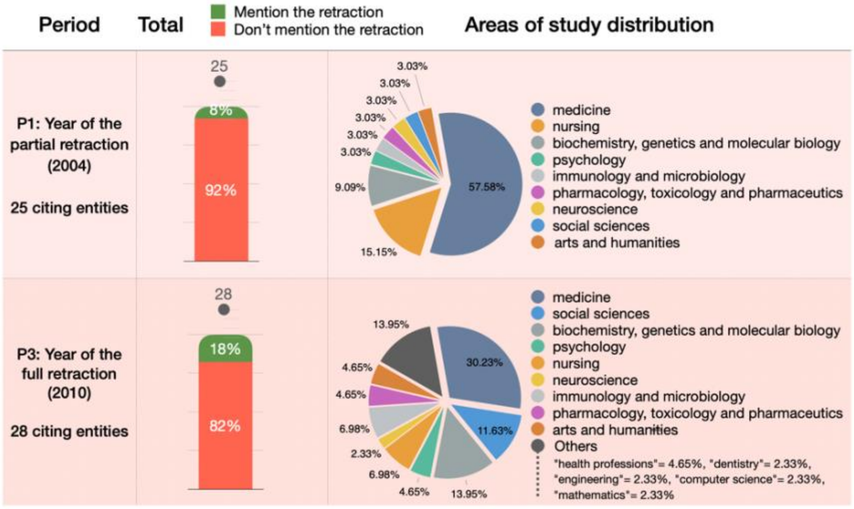


Figure 7.2. A summary of the citing entities. The first column contains the periods P1 and P3, the second column shows the number of citing entities that do mention (in green) or do not mention (in red) the retraction, while the third column shows the distribution of the subject areas of the citing entities

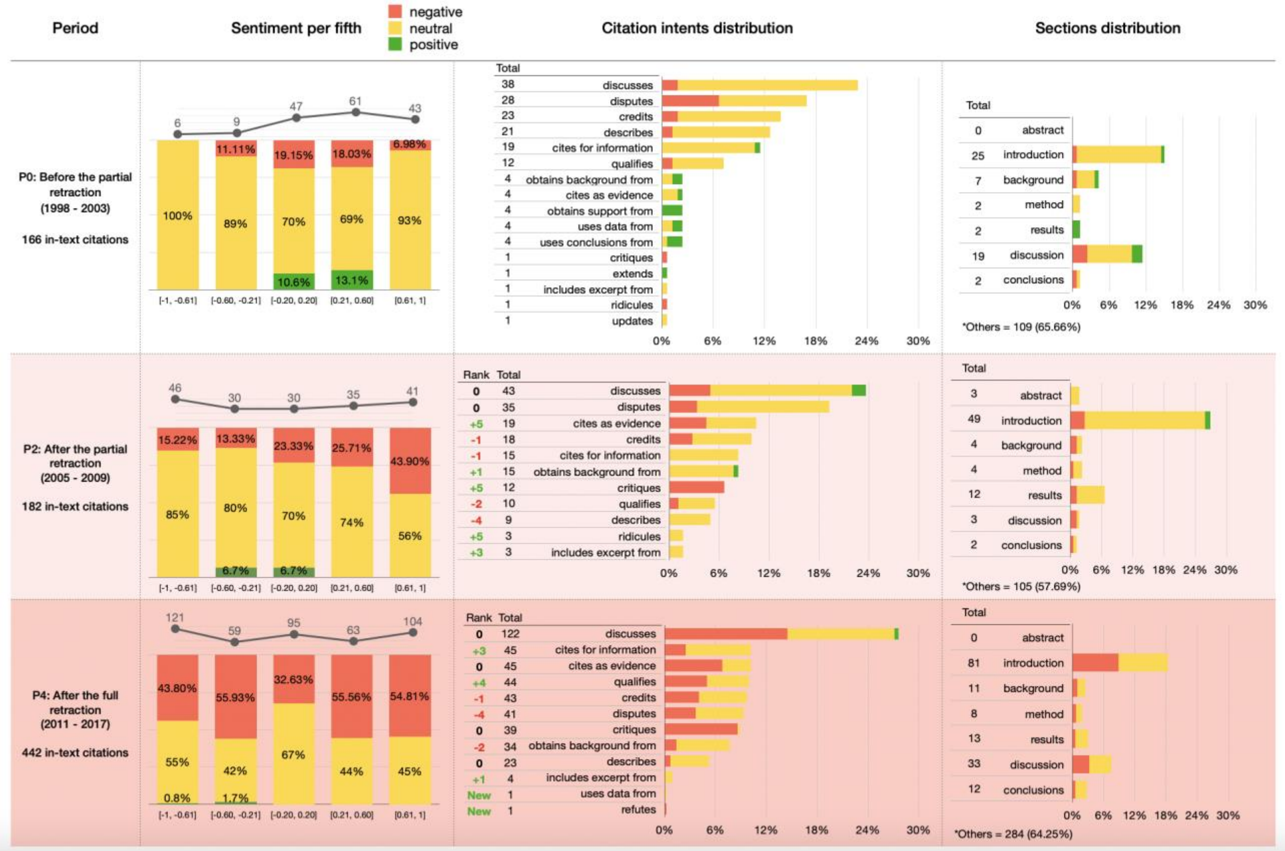


Figure 7.3. A summary of the in-text citations. All the data are classified under the three sentiments: negative (red), neutral (yellow) and positive (green). The first column contains the periods P0, P2 and P4, the second column shows the distribution per fifths of the in-text citations, the third column shows the citation intents distribution, and the last column shows the in-text citation sections distribution

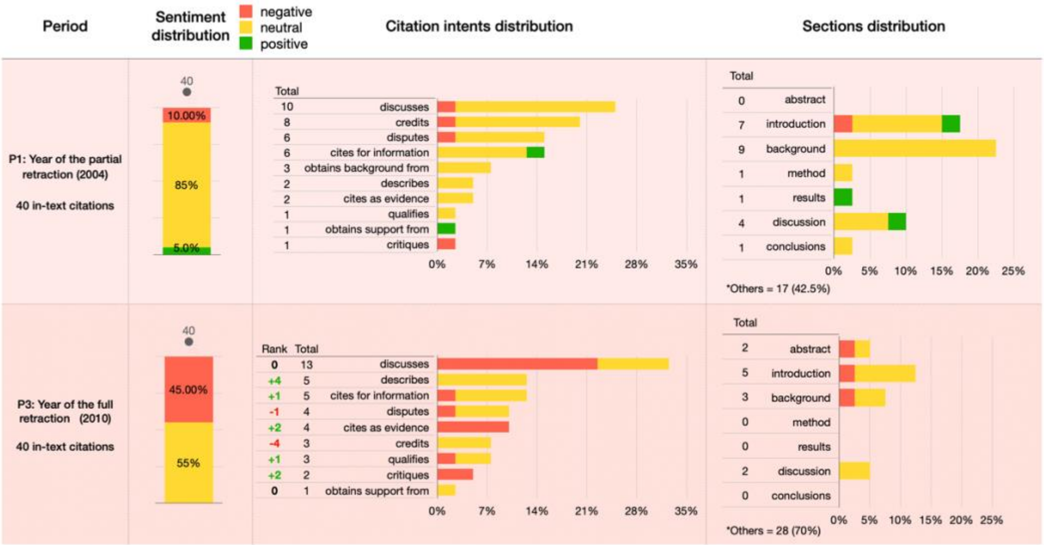


Figure 7.4. A summary of the in-text citations. All the data are classified under the three sentiments: negative (red), neutral (yellow) and positive (green). The first column contains the periods P1 and P3, the second column the total number of in-text citations, the third column shows the citation intents distribution, and the last column shows the in-text citation sections distribution

Phase 5: running a topic modeling analysis on abstracts and citation contexts

The topic modeling analysis has been done on the abstracts of the citing entities and the in-text citation contexts. Yet before running such analysis, we first needed to choose the best number of topics to use in the training process based and the calculation of the coherence score of different topic models (discussed in Section 5.1). Based on the results obtained (Figure 7.5 and Figure 7.6) we decided to run a topic modeling analysis on 22 and 13 topics, respectively for the collection of abstracts and in-text citation contexts.

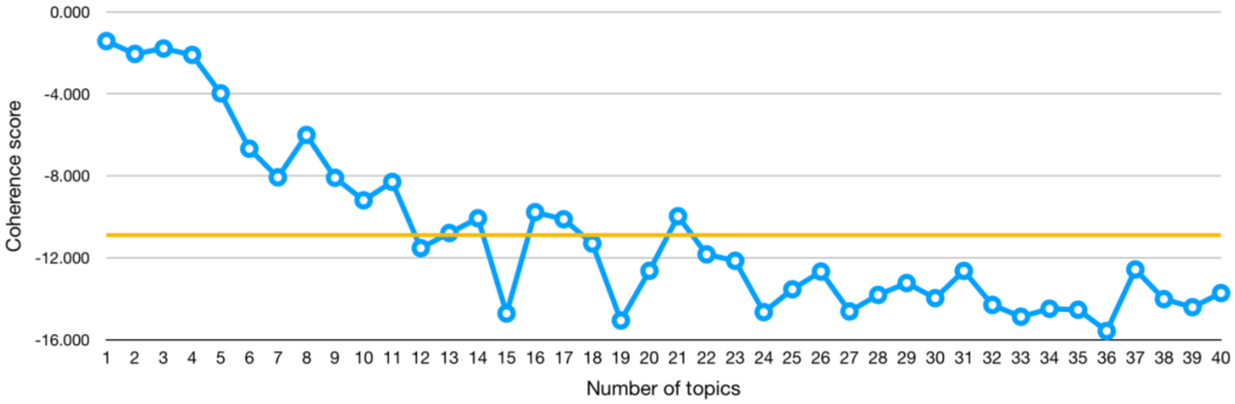


Figure 7.5. The coherence score of different LDA topic models built using a variable number of topics, from 1 to 40. The topic model is based on the corpus and dictionary of the abstracts of the citing entities. The orange line is the average value, and it plateaus around 12–13 topics

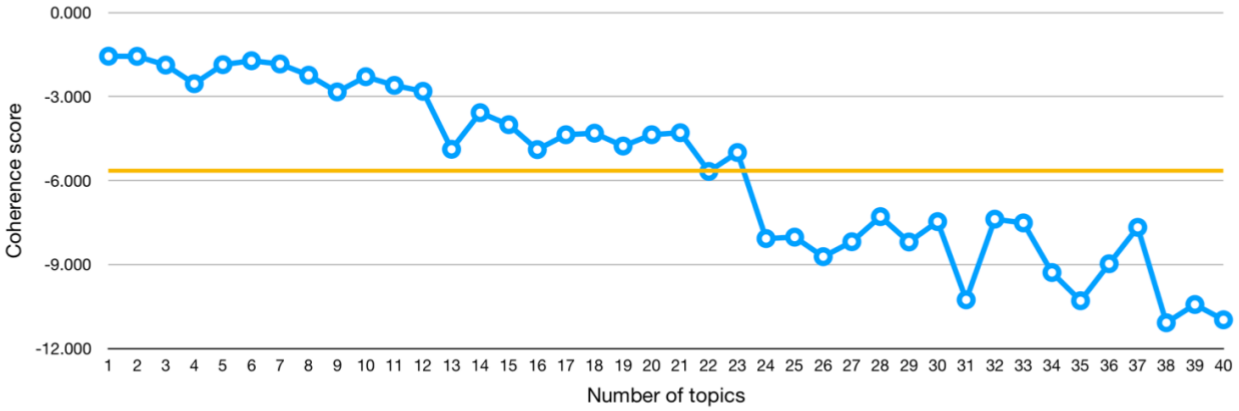


Figure 7.6. The coherence score of different LDA topic models built using a variable number of topics, from 1 to 40. The topic model is based on the corpus and dictionary of the in-text citation contexts. The orange line is the average value, and it plateaus around 22–23 topics

We obtained the topic model using the abstracts of all the publications having it, i.e., 464 citing entities in the dataset. Using the outcomes of the topic model, we were able to explore each individual topic and give a possible interpretation to it by analyzing its 30 most probable terms, as shown in Table 3 in Appendix.

Figure 7.7 shows the LDAvis visualization built over the topic model of the abstracts contained in the citing entities. The left part of it shows two different clusters and one of the clusters is composed of one big topic, i.e., topic 3 ("Discussion toward the contents of WF-PUB-1998"), which was by far the larger topic identified by the process. Looking at the 30 most salient terms, the term "retract" is in the 5th position, meaning that some of the citing entities talked about the retraction of WF-PUB-1998 or, more generally, the retraction phenomenon. The same list includes terms such as "social", "movement", "debat", "media" and "cultur" which seem not to be strictly related with medical jargon. This scenario may be an indicator that some of the citing entities are not medical publications. Finally, among these 30 most salient terms, we found terms with a strong negative connotation, such as "fraud".

The MTMvis visualization of the topic model built over the abstracts of the citing entities is shown in Figure 7.8. The distribution of the topics is represented as a function of the periods (i.e., P0-P4). If we consider only the periods P0, P2 and P4, then from Figure 7.8 we can observe a constantly increasing percentages over the time for the topics 1 ("Social science studies"), 2 ("Bibliometrics and retraction"), and 5 ("General discussion toward medical and non-medical subjects"), while, on the contrary, topics 4 ("Medicine and pharmacology") and 9 ("General description of the WF-PUB-1998") were decreasing. Topics 3 and 11 ("General medical background of WF-PUB-1998") showed a very similar pattern along the three periods. Later while discussing these results (i.e., when addressing the research question in Section 7.3), we will use the filters provided in MTMvis to modify the topics distribution as a function of the values of other meaningful attributes (e.g., "Subject area") and to highlight some specific aspects.

LDAvis and MTMvis has been built also on the topic model obtained from the in-text citation contexts contained in the citing entities as shown in Figure 7.9 and Figure 7.10 respectively. Figure 7.9 shows the distribution of the 22 topics in the two-dimensional plane. In contrast to the previous observations over the topic model of the abstracts, the 30 most salient terms did not include any term related to the retraction phenomena. The sparsity of the topics in this LDAvis is higher than the one observed with the abstracts and allowed us to spot three different clusters. In particular, we observed two topics with a high prevalence which are also very distant among them, i.e., topics 8 ("Description toward the medical topics of WF-PUB-1998") and 12 ("Negative impacts of the WF-PUB-1998 retraction"). Beside these two topics, from LDAvis of Figure 7.9 we noticed a big cluster which includes the rest of the topics. The closest topic contained in the cluster to topic 8, i.e.,16 ("Background and outcomes of WF-PUB-1998"), is indeed close to the interpretation we have given to topic 8. On the other hand, the topics of the cluster that are close to topic 12, e.g., topic 17 ("Impact of the WF-PUB-1998 outcomes") or topic 21 ("Bibliometric analysis of the WF-PUB-1998 retraction"), are close to the subjects of topic 12. Table 4 in Appendix lists all the topics and provides our own interpretation according to their 30 most probable terms.

The prevalence of topics 8 and 12 emerged also form the MTMvis visualization in Figure 7.10. Both are distributed along all the periods (i.e., P0-P4). Topic 12 which is highly correlated to subjects analyzing the retraction of WF-PUB-1998 has a higher distribution after the full retraction of WF-PUB-1998 (i.e., P3 and P4). On the other hand, topics labeled as promoting a general discussion of WF-PUB-1998 from a medical point of view have decreased along the five periods and especially after P1, e.g., topic 8 is the most evident example belonging to this category.

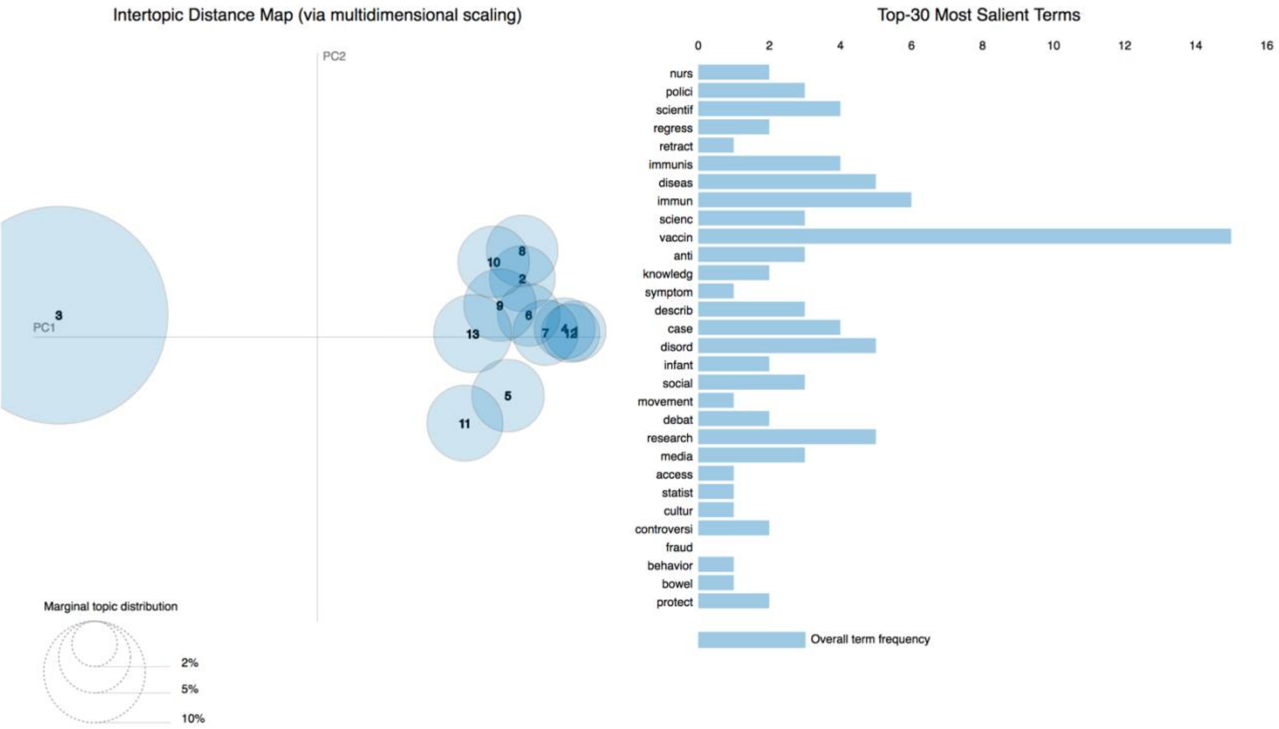


Figure 7.7. LDAvis visualization built over the topic model obtained from the abstracts of the citing entities

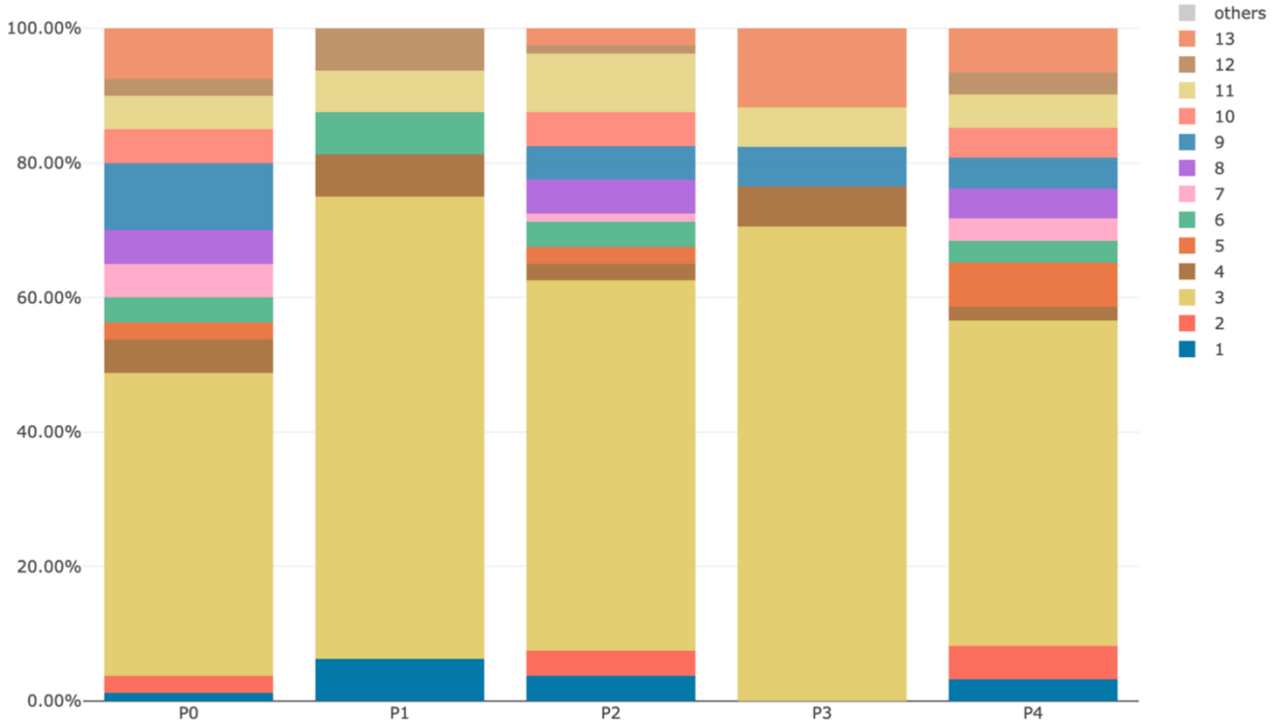


Figure 7.8. MTMvis built on the topic model obtained from the abstracts of the citing entities, shown against the all the periods P0-P4. For each period the visualization plots the topics distribution (e.g., topic 3 is the dominant topic in all the periods: P0-P4)

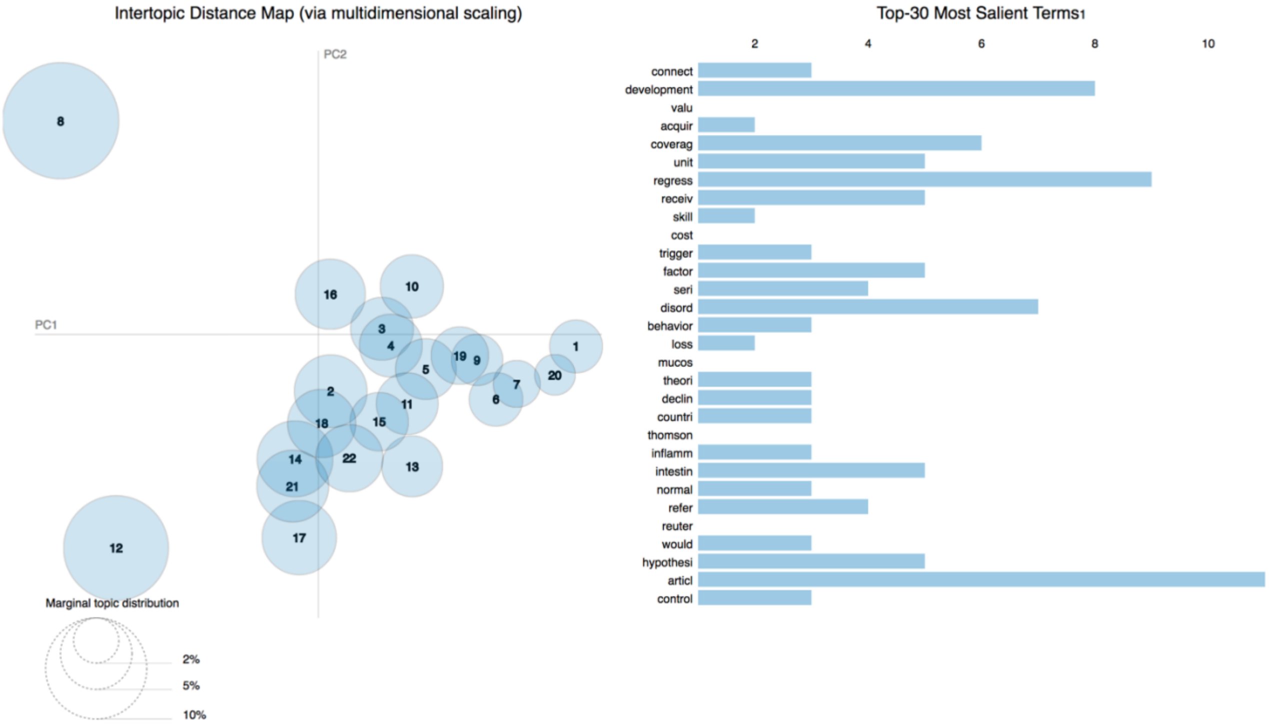


Figure 7.9. LDAvis visualization built over the topic model obtained from the in-text citation contexts contained in the citing entities

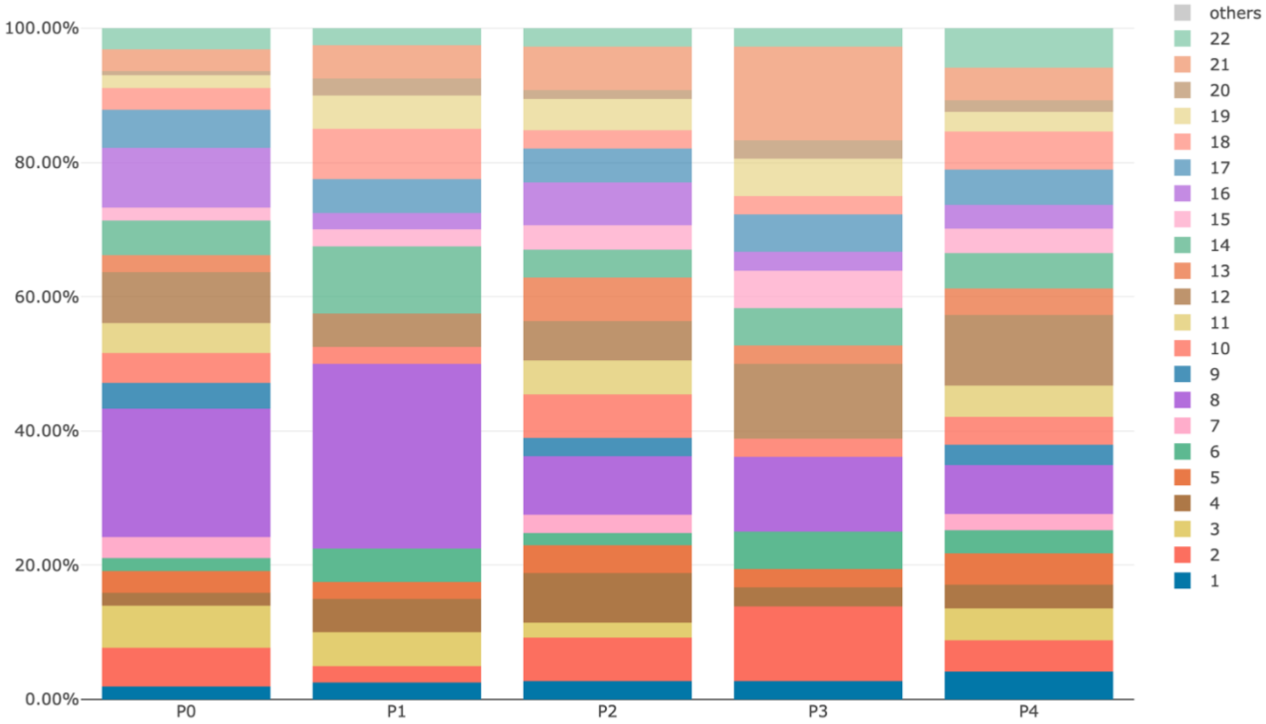


Figure 7.10. MTMvis built on the topic model obtained from the in-text citation contexts contained in the citing entities, shown against the all the periods P0-P4. For each period the visualization plots the topics distribution (e.g., topic 8 is the dominant topic of the periods P0 and P1)

7.2.3 Results and discussion

From a quantitative point of view, while looking at the subject areas of the citing entities we gathered (see Figure 7.1 and Figure 7.2), we noticed an increase in the number of areas involved in time. Indeed, when looking at P0, P2, and P4 (Figure 7.1) the total number of subject areas were 12 in P0 (i.e., before the first partial retraction), while in P2 and P4 we counted 19 and 21 different subject areas respectively. In addition, in P2 and P4 we observed a higher prevalence of non-medical subject areas. Considering the percentage value in P4 with respect to the one in P0, then *social sciences* and *arts and humanities* had increased their percentages of 9.32 and 2.64 times more than those observed in P0, respectively. On the contrary, considering the same periods (P0 and P4), *medicine* and *nursing* had an inverse trend, since their presence decreased by almost 16.8% and 5.62% percent compared with P4, respectively. These figures suggested that the retraction attracted the attention of other subject areas which were not strictly related to the original one of WF-PUB-1998 (i.e., *medicine*).

In addition, we also noticed a continuous increase in the percentage of entities that have explicitly mentioned the retraction of WF-PUB-1998 in their citation contexts over the time. The largest number of citations mentioning the retraction (51%) was reported in the last fifth of P4. A considerable percentage of entities mentioned the retraction even before the full retraction notice (e.g., 25% of entities in the second fifth of P2). So, both the full and partial retractions were acknowledged by the citing entities. Indeed, some acknowledged the retraction after the partial retraction and before the full retraction notice was raised. This aspect might be also related to the kind of the partial retraction (that was “Concerns/Issues About Results” and “Error in Results and/or Conclusions” in WF- PUB-1998) and with the popularity of the particular case in consideration.

Figure 7.3 shows that the intended sentiment carried in the citation contexts of the in-text citations referring to WF-PUB-1998 moved to the negative spectrum over time (from P0 to P4). However, the retraction of WF-PUB-1998 was not always mentioned in these cases. Indeed, as shown in Figure

7.1, in the middle fifth of P2 only 21% of the citing entities mentioned the retraction even if the perceived sentiment in the same year is either negative (for 32.63% of in-text citations) or neutral (for 67.4% of in-text citations).

The distribution of the citation intents annotated in the in-text citations during P0, P2 and P4 showed an increment in the use of generic intents such as *discusses*. This could be related with increasing popularity of the retraction of WF-PUB-1998 in the non-medical subject areas (as already stated previously). Probably, the entities that are part of the non-medical subject areas cited WF-PUB-1998 with a generic intent without recalling strictly medical details in their text, which are out of the scope of their research domains.

As shown in Figure 7.11, the set of intents *uses conclusions from*, *updates*, *extends*, *uses data from* and *obtains support from* decreased a lot after P0, probably due to a lesser use of the data and conclusions contained in WF-PUB-1998 after its retraction. Other citation intents, instead, showed a clear increment of their use along the three periods. For instance, the use of *critiques* seemed to be related somehow with the increment of the negative sentiment overall. Instead, *credits* had an important drop. In this case, the citing entities published before the partial retraction of WF-PUB-1998 used it mostly in a neutral way to credit Wakefield and colleagues for their findings. However, in P2 and P4, beside the overall drop, *credits* had a higher percentage of negative citations. This last aspect was also noticed in the intent *cites as evidence*, although its overall usage has increased in time. However, if before the retraction, *cites for evidence* was used neutrally to refer to WF- PUB-1998 to support some statements or conclusions in the citing entities, after the retraction it was actually used to highlight WF-PUB-1998 as a negative scientific example due to its retraction and, more generally, of faulty science.

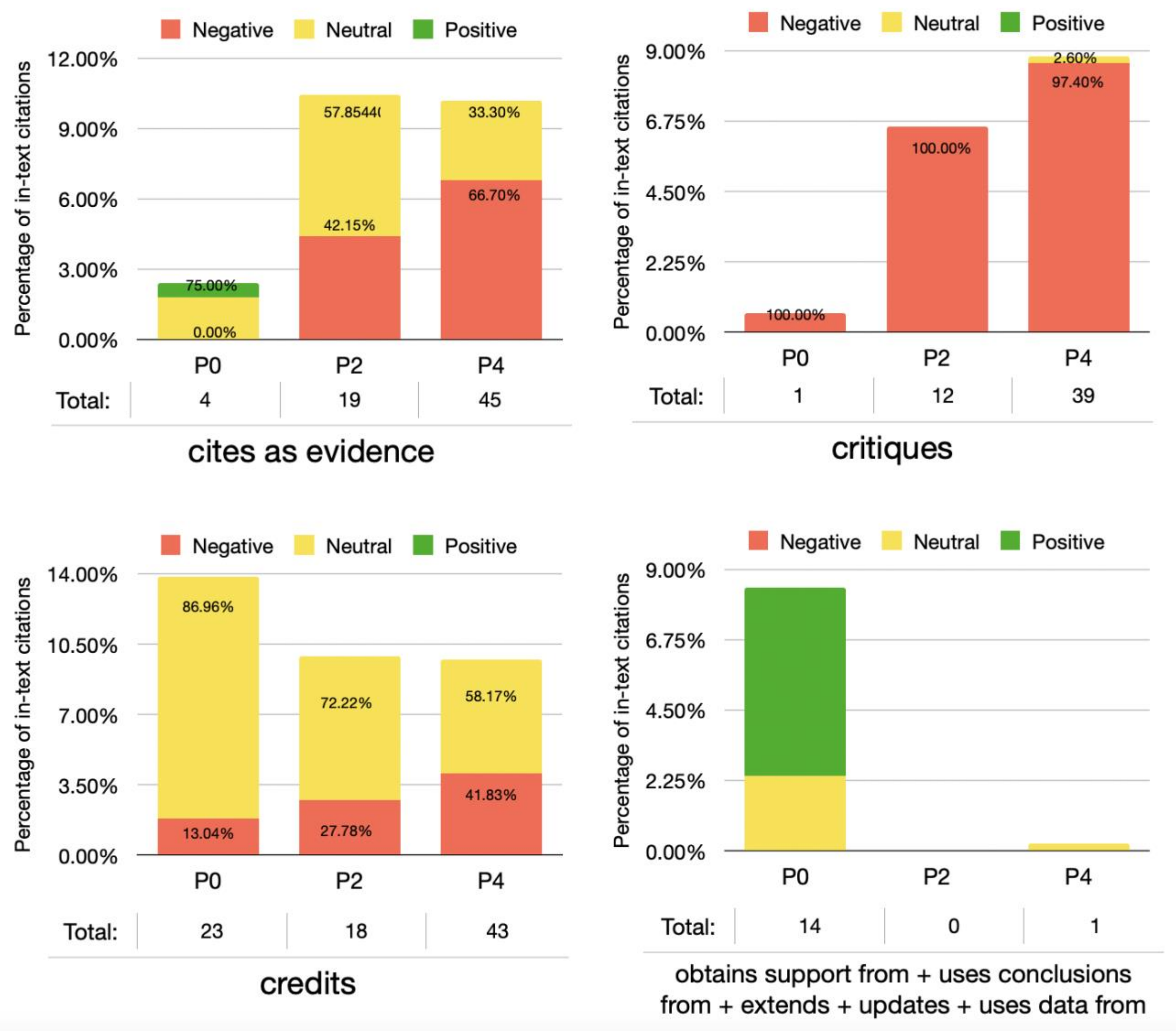


Figure 7.11. The four graphs illustrate the way the use of citation intents changed over time (i.e., the three periods P0, P2 and P4) and according to their perceived sentiment. The citation intents *cites as evidence*, *critiques* and *credits* are illustrated in separated charts, that show an increment in the negative sentiment along the three periods

We investigated the sections of the in-text citations marked as *credits* and *cites as evidence*. On the one hand, the *credits* citations were mostly distributed on descriptive sections – i.e., *introduction, discussion, and background* – during all the three periods. On the other hand, the *cites as evidence* citations appeared also in technical sections – i.e., *results and method*. The sections distribution in P3 for both *credits* and *cites as evidence* followed the overall distribution introduced in Fig. 7.3: the in-text citations have been concentrated in few sections mostly of descriptive type – i.e., *introduction* and *discussion*.

Looking at the retrieved topics in the topic model created using the abstracts of the citing entities, we noticed that topics 1, 2 and 5 were those increasing their presence after the partial retraction (i.e., starting from P2). The themes covered by these topics seemed to refer to discussions on the retraction phenomena (see Table 3 in Appendix) and used a limited number of terms from medical jargon.

A deeper investigation on the evolution of topics 1, 2 and 5 during P0, P2 and P4 on all the subject areas, showed that topics 2 and 5 increased significantly in P4 compared to P2 (3.75% vs 4.92% for topic 2, and 2.5% vs 6.56% for topic 5) while topic 1 has an almost steady trend between P2 and P4 (3.75% vs. 3.28%). This might indicate that topic 1 (and the abstracts linked to it) discussed the retraction phenomena similarly over P2 and P4. In fact, although topic 1 included words that deal with ethical/social issues (see Table 3 in Appendix), it did not include words strongly related to the retraction or having a strongly negative sentiment. The citing entities linked to topic 1 cited WF-PUB-1998 and discuss the case without mentioning the actual retraction of WF- PUB-1998, even after its full retraction (i.e., P4). Using MTMvis we can see that topic 1 is mainly related to the *medicine* subject area (excluding the subject areas with limited number of abstracts). This relation between topic 1 and *medicine* is also interesting: indeed, topic 1 has little engagement with the medical themes, considering its 30 most probable terms. Thus, part of the entities in the *medicine* subject area discussed the retraction of WF-PUB-1998 in non-medical terms as well.

We investigated the distribution of topics 2 and 5 over the subject areas during P2 and P4 and checked if such topics were part of the top five ones of each related subject area. Both the topics 2 and 5 were listed in the top five topics of twelve subject areas. Avoiding considering the subject areas for which we had a small number of abstracts in P2 and P4, we noticed that topics 2 and 5 were highly represented in the *social sciences* subject area with a total percentage of 13% (number of abstract: 10) of all the abstracts in P2 and P4. These considerations suggest that topics 2 and 5 were the ones that better represent and characterize the period after the full retraction (i.e., P4) and that *social sciences* is the subject area that dealt the most with the themes emerged in P4. Contrary to our previous considerations regarding topic 1, in these two topics we found a clear reference to the retraction. The fact that this aspect was manifested in the analysis of the abstracts may indicate that the retraction might have been one of the main subjects discussed in the entities of the abstracts analyzed.

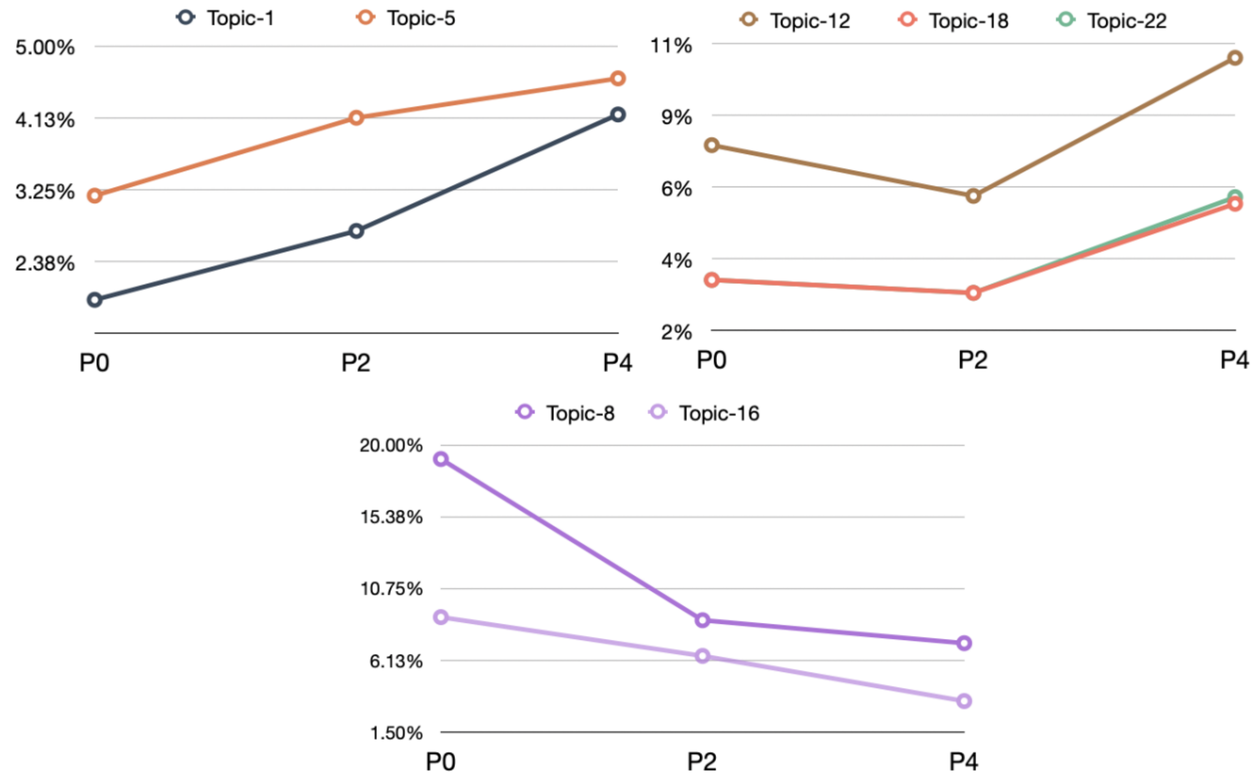


Figure 7.12. The evolution over time of three groups of topics defined from the citation contexts of the in-text citations to WF-PUB-1998

We analyzed also the twenty-two topics we obtained considering the topic model created using the citation contexts of the in-text citations referring to WF-PUB-1998. In particular, as shown in Figure 7.12, we focused on:

1. the topics for which we observed an increasing use over time.
2. the topics which had a huge increment in their use in P4.
3. the topics which had a constant decrease in their use over time.

The topics that increased over P0, P2 and P4 (i.e., topics 1 and 5) included few medical terms and seemed to refer to the controversy of the retraction of WF-PUB-1998 from a mathematical and statistical perspective. A second group of topics (i.e., topics 12, 18 and 22) seemed to refer to WF-PUB-1998 as an example of faulty science, which was acknowledged clearly in P4. The drastic change of these topics in P4 is very significant. Indeed, all the three topics (as shown Table 4 in Appendix) mention the word "retraction" (and its derivatives) along with other words with a strong negative connotation. In other words, it seems that the authors waited the full retraction notice before marking their negative impressions toward WF-PUB-1998 – 22.15% of the in-text citations in P4 are part of this group of topics.

Similar behavior could be noticed also in the citations coming from medical subject areas, since 23.88% of the citations in P4 are coming from *medicine* and *nursing* articles. This suggests that also the entities close to the domain of the retracted article did not hesitate to judge a retracted work done by their colleagues.

The last group of topics (i.e., topics 8 and 16) were mainly related to the medical domain and included some medical themes treated in WF-PUB-1998. The fact that these topics had a clear decrease over time suggests that the most recent citing works provided partial and limited acknowledgement to the conclusions and medical arguments in WF-PUB-1998.

In Figure 7.13, we show the topics that either increased (left panel) or decreased (right panel) their presence over time considering only the citation contexts of the citing entities belonging to the *medicine* subject area. Some of the topics shown in Figure 7.13 are also included in Figure 7.12, although there is an important difference: topic 15 (that concerned the conclusions of WF-PUB-1998 and the controversies arising from it) is not listed in Figure 7.12, even if it seemed relevant when we focus only on the *medicine* subject area. We had a similar situation also with the topics decreasing over time. Indeed, topic 7 (a description of the medical topics in WF-PUB-1998) is not highlighted in Figure 7.12 as well.

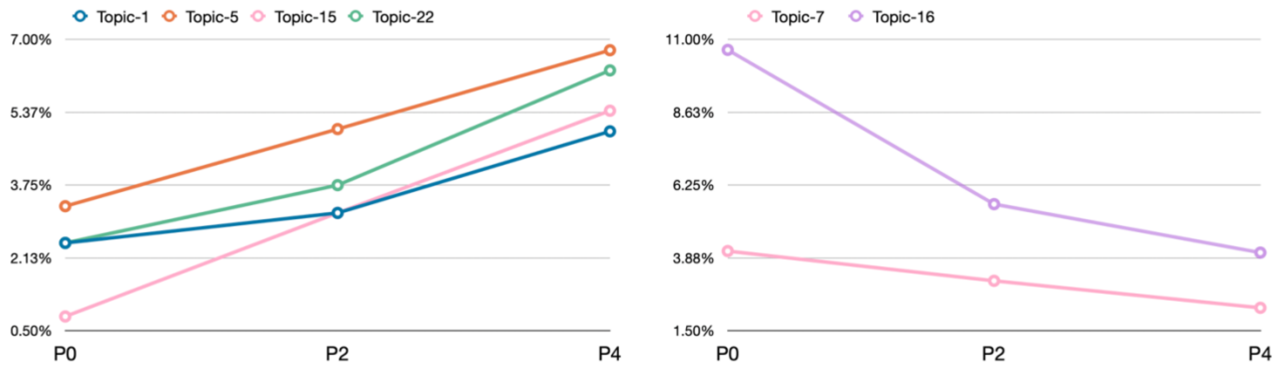


Figure 7.13. The increasing (left) and decreasing (right) topics of the in-text citation topic model, considering only the *medicine* area of study

This scenario suggests that the citing entities in the *medicine* subject area included additional prominent topics when discussing WF-PUB-1998. More precisely, after its final retraction, part of the entities addressed the retraction by pointing out its controversies from a medical perspective. On the other hand, the decreasing relevance of topic 7 indicates that the entities part of the *medicine* area of study addressed less the medical arguments of WF-PUB-1998 and rather focused on citing and discussing the retraction of WF-PUB-1998 without deepening into its actual content (e.g., method and findings).

7.2.3 Differences and similarities with the results obtained on the analysis of retraction in humanities

As for the retraction in humanities, we have reported an incrementing trend in the number of the citing entities after the full retraction of WF-PUB-1998, in addition in both cases the pick was reached on the first fifth of that period (i.e., [-1, -0.61]). This result was also observed in the work of Teixeira da Silva and Dobránszki (2017), who demonstrated this situation with the observation of the number of citations one month after the retraction. In other words, many tend to not wait a lot before citing and talking about the retracted publication, indeed this is happening right after the year of retraction in both STEM and humanities case. We can also speculate on the fact that both STEM and humanities notice the retraction of the publication they are citing in the same way.

Another noteworthy similarity emerges from the distribution of the subject areas of the citing entitles. The original subject area of the retraction case analyzed (medicine and humanities), decreases significantly after the year of the full retraction in favor of a new number of subject areas. In WF-PUB-1998 medicine moved from 52.7% (P0) to 35.9% in the period after the full retraction, while in the humanities case arts and humanities moved from 22.9% (P0) to 18.4% in the period after the full retraction. However, as we can see from these percentages the gap in the percentage values is different and it can be linked to a higher perception and awareness that sensitive subject areas (e.g., medicine) have toward retraction. In addition, in WF-PUB-1998 the social sciences subject area mostly dealt with the ethical arguments emerged in P4, this was not the case in humanities (social sciences decreased significantly in P4). This results, may be highly related to the reason of retraction and the overall popularity of the retraction case among the social scientists, which tend to discuss more the ethical issues around a sever scientific misconduct behavior (i.e., WF-PUB-1998).

The distribution of the citation intents is another interesting aspect. On one hand, in WF-PUB-1998 we noticed an increment in the citing entities discussing (i.e., labeled with the intent *discusses*) the cited retractions in the post-retraction period, this was mainly due to the arise in the number of non-medical subject areas. On the other hand, in humanities the overall intent did not change throughout all the periods, citations have been mainly done for general purposes (e.g., citing for information or for obtaining background to their argumentations). We can speculate that not only the perception of the retraction in STEM is higher from a quantitative point of view, also the behavior of the citing entities changes after the retraction and is more prone to dynamic changes, which is not the case in the retractions of humanities.

Although the citation intent *discusses* in the humanities case did not get the same positive increment as it is in the WF-PUB-1998 case, yet if we consider only the topics and argumentations related to the retraction phenomenon (mentioning the retraction in their citation contexts), then in both cases these have been addressed by the citing entities for an open discussion (i.e., labeled as *discusses*). In other words, in both the STEM case and humanities the retraction event triggered a higher discussion from the citing entities. Therefore, when the citing entities talk about the retraction of the retracted publications that they cite, they usually do that in order to discuss some aspects rather than just mentioning it or describing general facts around the case.

Beside the above discussion, other features which have been also investigated by the past studies, followed the behavior already discussed in the analysis of L1 (Section 7.1). In general, the values of these common features for the WF-PUB-1998 case are in line with the dynamics and characteristics we have observed while dealing with popular retraction cases. More precisely, the percentage of entities which have acknowledged the retraction was significantly high. In addition, the negative sentiment of the citing entities has been continually increasing starting from the year of its retraction, in some cases for some specific fifths the majority of the in-text citations have been labeled as negative. Both these aspects are in line with the results of Suelzer et al. (2019). These facts might indicate that the entities that cite retractions in STEM are more prone to express a negative judgment and stronger opinions (i.e., the intent *critiques* is highly represented in P4 in the analysis of WF-PUB-1998) toward the retracted publication, in contrast to what we have observed in the analysis of the retraction in humanities.

Chapter 8

Conclusions

This thesis started with a general definition and background on retraction, some numbers and statistics on it, and a discussion toward the approaches that could be/had been adopted to study such phenomenon. Then we showed the situation in the humanities domain which is the focus of our work. A review of the past studies and the fact that we wanted to further enrich our knowledge around this phenomenon with qualitative aspects (which gathered less attention compared to the quantitative aspects), helped us motivate our chosen approach for the analysis of retraction in the humanities domain which is based on the concepts introduced by Science of Science (SciSci). SciSci is a transdisciplinary field that uses large datasets to study the dynamics of the science production. Our analysis was mainly focused on the citations of the retracted publications in the humanities domain. The aim was uncovering the quantitative and qualitative aspects following the citing dynamics. The methodology we have designed and presented in the second part of the thesis follows the principles of open science and the methods adopted by SciSci to foster a quantitative and qualitative citation analysis of an established retraction case.

The three research questions we have presented in the introduction of the thesis (i.e., RQ1 – RQ3) have been addressed through a step-by-step application of the methodology in the analysis of two retraction cases: (a) a set of retracted publications from the humanities domain, and (b) the Wakefield et al. retraction case (representing a popular retracted article in STEM). On one hand, the analysis of case (a) helped us address RQ1 and RQ2 strictly related to the dynamics of retraction in the humanities domain. On the other hand, through our analysis of (b) and the past studies involving retractions in STEM we managed to address the last research question (i.e., RQ3).

This chapter gives a summary and highlights the main findings learned from the outcomes of our experiments that allowed us to answer the research questions (Section 8.1), and discusses some (technical and conceptual) limits of the analysis presented in this thesis, providing suggestions on how to address such limits (Section 8.2). Finally, the last section discusses the possible future works that might enrich our knowledge around the retraction phenomenon and help the scientific community benefit from.

8.1 Research questions

The aim of this thesis was answering the three research questions (i.e., RQ1 – RQ3) defined in the introduction chapter. These research questions have been conceptualized around one main macro-objective: understanding the retraction phenomenon in the humanities domain. Each research question played a crucial part for inferring and uncovering new insights around such topic. In this section we revisit the main outcomes and facts we have learned.

The first two research questions (i.e., RQ1 and RQ2) concern retraction in the humanities. We answered these questions based on the outcomes obtained after the application of the methodology on a collection of retracted publications from the humanities domain.

We noticed how on average the retractions in humanities does not have a drop in the number of citations after the year of the full retraction. However, generally, we reported a low negative manifestation from the citing entities toward the retracted publications they have used and cited in the body of their works. Although this event happened on few occasions, it was in part correlated with the affinity of the citing entities to the humanities domain, i.e., citing entities with a low affinity to the humanities domain are those that have manifested most this negative sentiment.

The retraction event in humanities triggered a higher discussion from the citing entities, i.e., entities used to cite the retracted publications to discuss them rather than just obtaining general information and background. In addition, the retraction event had an impact on the subject areas of the citing entities. Subject areas close to the humanities domain such as *social sciences* and *arts and humanities* decreased their presence. On STEM disciplines (especially related to health sciences) such as *psychology* and *medicine* this trend was evident. We suggested that this might indicate a higher concern of these subject areas toward the citation of retracted articles in humanities (or generally to retractions of other domains. This aspect was later investigated while addressing RQ3).

When observing the individual disciplines in the arts and humanities domain, two opposite behaviors came to light after the retraction concerning the citing entities belonging to the *arts and humanities* subject area: (a) a drop in favor of STEM areas (mainly *medicine* and *nursing*), and (b) an increment at the expense of subject areas such as *psychology*. For (a) this was the case when the retracted items belonged to *arts*, while (b) was mainly reported in the retractions of *philosophy*. Although in different ways, both the behaviors are a demonstration of a high concern toward retraction. In both cases this happened mainly thanks to the fact that the retracted publications from the humanities domain included other topics close to STEM disciplines in their studies.

The subject area *social sciences* is a rare non-STEM area that manifested the same behavior of *psychology* in case (b). This fact occurred when dealing with the retractions of *history* and *religion*. Although *social sciences* is not a STEM discipline, the arguments treated were close to sensitive and non-humanities topics such as drugs/alcohol and psychological diseases. This fact explains why its behavior is like the one of the citing entities in STEM, which confirms our previous observations regarding the two behaviors (a) and (b).

To study the differences in the retraction dynamics between STEM and humanities we presented an answer based on two levels of analysis: (L1) a comparison with the outcomes of other past studies of retraction in STEM, and (L2) a comparison with the results obtained after the analysis of a popular retraction case in STEM (Wakefield et al., case) using our methodology.

The retraction of the humanities publications did not imply a drop of citations on the post retraction period. On the other hand, a more prominent positive impact on the post-retraction citation trend was observed in the past while analyzing retractions in general, i.e., without any domain restriction (Bar-Ilan & Halevi, 2018). However, the opposite trend (decrease in the citations number) was found in the analysis of specific disciplines such as biomedicine (Dinh et al., 2019) and psychology (Yang et al., 2020). Based on these results and others (e.g., Chen et al. (2013)), we have speculated that the potential threats and damage of retracted articles has been perceived more seriously by others (i.e., citing entities) on the post-retraction when the retracted publications dealt with sensitive arguments.

The increasing trend in the citations number has been observed in STEM when the analysis focused on one/limited number of retracted publications. However, also in cases with one/limited number of retractions, other studies reported an opposite trend. This made us think that other factors such as the popularity and the media attention to the retraction case have an important role and it is highly subjective to the specific case taken into consideration.

Some past studies on STEM retractions observed a low percentage of negative citations as well as a low percentage of entities mentioning the retraction in the context of their citation. These facts are in line with our findings on humanities. However, works that have analyzed popular media cases (e.g., the Wakefield case) showed a different situation. This was another evidence to our hypothesis regarding the influence of the case popularity and the media attention on the behavior of the citing works. To wrap up the L1 analysis (i.e., a comparison with past studies), we have suggested that this latter relation with the popularity of the case is the factor which impacts more on the reported statistics (e.g., sentiment, number of citations, etc.) and it represented the main decisive factor to consider for a reliable comparison of STEM and humanities retractions.

In the analysis of L2 (i.e., a comparison with the findings over the analysis of the Wakefield et al. (1998) paper) the comparison shifted toward other attributes less debated by past studies, yet part of our methodology.

We noticed how both STEM and humanities tended to not wait a lot before citing a retracted publication, indeed the highest number of citations was reported right after the year of retraction in both cases, which is also a sign that in both cases the retraction has been noticed similarly. Retraction had a similar effect on the subject areas distribution of the citing entitles - the original subject area of the retracted item (either STEM or humanities), decreased significantly after the retraction in favor of new several subject areas, although, in STEM this was more evident: the original subject area of our STEM case (e.g., medicine) decreased significantly in favor of the *social sciences* subject area which dealt with the ethical arguments emerged from the retraction case. Beside the overall popularity of the retraction case, we have hypothesized that this behavior is also related to the reason of retraction as well.

Citing entities addressed and discussed the STEM retraction form several perspectives and for different reasons after the retraction. In other words, the citation intent was more prone to dynamic changes. This was not the case for the retractions in humanities, whom have been cited mainly for general reasons before and after the retraction. The citation behavior in the humanities case was different only when the citing entities explicitly talked about the retraction of the cited publication in their citation contexts, in this case the citing entities manifested other usages beside focusing only on the general aspects.

The features analyzed by the past studies (i.e., L1 analysis) has been considered also in our methodology. Our findings were very close to the ones of the past studies discussed in L1. The retraction was highly acknowledged in the citation context and the negative sentiment was constantly growing starting from the year of the retraction. More precisely, these aspects are in line with the results manifested by studies concerning singular and popular retractions in STEM (e.g., the work of Suelzer et al. (2019) that also investigated the Wakefield case).

8.2 Limitations

The analysis presented in this thesis is not entirely devoid of limits and possible biases which might occur on the methodological session as well as in the interpretation of the results obtained. In this section we revisit and discuss the noteworthy ones and suggest some possible adaptions to the methodology and analysis we have conducted, to help us overcome such limits.

The second phase of the methodology can be easily implemented by means of a computational approach using scripts and APIs which can query the services we have mentioned, e.g., OpenCitations' APIs. Other services could be adopted as well but we will need to further investigate how to integrate them using additional ad-hoc scripts. Thus, the time of execution of this phase should relatively be low. Instead, marking the citing entities which have been retracted is a higher time-consuming process, this is due to the fact that this step relies on manual querying of the Retraction Watch database using its web search interface. An automatic approach is possible if we can access the complete dataset, and this could be done only by signing an agreement with Retraction Watch and on behalf of an institution. This part has not been considered as a direct part of the methodology, due to its bureaucratical nature.

MAG is one of the services indicated in our methodology for gathering the citations and their metadata. Although after years of self-improving, the data in MAG became very accurate, problems such as the entity recognition and disambiguation is far from being solved also in the data provided by MAG (Wang et al., 2019). Microsoft announced that it will shut down MAG by the end of 2021. It remains to be seen whether and how thus service can and will be replaced. OpenAlex is a new (launched on January 3rd, 2022) free and open catalog of the world's scholarly papers, researchers, journals, and institutions which is based on important data sources such MAG and Crossref

(Hendricks et al., 2021). This new index of scholarly data represents a new potential alternative to MAG; however, it is too early to either give a judge considering the few studies which have already used it, or to state whether OpenAlex will be able to cover all the data that has been provided previously by MAG.

The subject area and category classification step can potentially introduce some biases, especially if we did not manage to find the venues of the articles citing the RET-SET in Scimago. In these cases, indeed, it is needed a manual subjective interpretation from the annotator. Although, based on our experience from the case studies analyzed in this thesis, we found a limited number of venues without a corresponding classification in Scimago.

The third phase is the highest time-consuming phase. It is totally based on manual processing, consequently, it is prone to human errors, e.g., text typing, text translation, the correct definition of the in-text citation context windows, etc. However, these errors could be detected in the second step of the phase while annotating the characteristics of the in-text citations, since the annotator needs to read the extracted text again. An important limit of this phase is related to its dependency on the availability of the full text of the citing entities, which relies on the user having access to paywalled contents. In case of using English as main language for the documents, non-English full texts are also accepted as long as the original text is written in a Latin alphabet language. In this case, the dataset produced by the methodology stores a translated version of the original text done by the annotator or by means of automatic tools such as Google Translate. As a result of this flexibility, the translated text may not accurately represent the original one and we might lose some information. Although, since the topic modeling analysis aim is to infer a collection of meaningful keywords, we think this trade-off is reasonable, since we are not interested in a perfect representation of the text.

From our prior experience, we think that the second step of the second phase, i.e., “Annotating the in-text citations characteristics” is the most delicate one. As direct consequence of the results obtained by several studies conducted in the past on the manual and automatic identification of citation functions, such as the works of Di Iorio et al. (2013) and Ciancarini et al. (2014), we have defined a guiding schema to help the annotator in the definition of the citation intent. Of course, this process could still be biased by the human's subjective judgement while reading and analyzing the in-text citation context. We expect that a future adoption of this guiding schema may give us additional clues to eventually refine and extend the proposed schema with additional constraints to limit possible ambiguities. For instance, a particular ambiguous situation we faced in the analysis of the retraction case of Wakefield arose when trying to annotate the following in-text citation context: "This study revealed that mental health correlated with religiosity ...". In that sentence, the word "This" may be referred to either the article itself (the citing one) or another study cited and discussed in the previous sentence. In these cases, for example, we can decide to always apply the second option, or to force a larger definition of the context window. Another aid could come from formalizing the annotation of the citation sentiment into a decisional schema, similarly to what we have done for the citation intent. This may help the annotator in the choice and reduce his subjective interpretations. For example, a possible rule to include in a potential decisional schema is: if the citing article uses the word "controversial" in reference to the cited retracted article, then the citation sentiment should be marked as *negative*. However, in our methodology, we did not propose any annotation schema for mapping the citation sentiment.

We think that the methodological process should also be published as a stand-alone protocol document in Protocols.io – a platform for developing and sharing reproducible methods (Teytelman et al., 2016). The first edition of the methodology has been published on such platform (Heibi & Peroni, 2020a), yet it needs to be redefined/updated following the methodology definition presented in this thesis. The methodology should also include the description of an automatic process for building the charts we have presented in the phase dedicated to the descriptive statistics. We want to ensure that this activity, though, is done using open services and software to foster the reproducibility of the methodology. The idea is to create a customizable workflow as it has been done in MITAO for what it concerns the fourth phase of the methodology. In addition, a potential improvement would be to convert the static charts into a dynamic format (e.g., dynamically outlined following the definition of some filters).

In the last phase of the methodology (i.e., “Running a topic modeling analysis”), most of the execution is handled by MITAO. However, the preparation of the documents to use as input of the process is up to the user. Our plan is to integrate an additional sub-module in MITAO which takes the annotated dataset as input and automatically generates a collection of documents based on an established criterion. Such extension would allow MITAO to automatically generate the abstracts and in-text citation contexts collections from the dataset generated in the previous phases of the methodology, as well as other documents collections based on different attributes, e.g., the subject area.

The fact that we relied only on open citations repositories (e.g., OpenCitations’ COCI) in order to foster the reproducibility of our analysis might have an impact on the final results especially from a quantitative point of view, since we are aware of the fact that many citing entities have been left out of the analysis.

Our findings and observations on the retraction case of Wakefield et al. provided important insights for the work of this thesis. However, we are aware of particular limitations that may have affected the findings and the interpretations we made.

First, we used the data in COCI to gather all the citations, however since COCI contains citations between entities included in Crossref when they are both identified by DOIs, we did not include in our analysis all the citations that involved entities with no DOIs. Also, we missed the citations from articles published by some publishers, such as Elsevier, that did not share openly their reference lists via Crossref – and that, thus, were not available in the COCI dataset.

For a few citing entities (i.e., 22) involved in the citations we gathered, we could not retrieve their full text due to commercial paywalls, preventing us to analyze the citation contexts and in-text citations they defined. We decided to exclude these citing entities, and their related citations, from our analysis.

On the same line of the Wakefield et al. analysis, also our study toward retractions in the humanities domain has some limitations that may have introduced some biases. First, compared to other fields of study, bibliographic metadata in the humanities have a limited coverage in well-known citation databases (Hammarfelt, 2016). This fact led to some limitations when applying a citation analysis in the humanities domain (Archambault & Larivière, 2010). Pragmatically, for what our study was concerned, we surely collected fewer citing entities than those that had in fact cited the retracted publications. The availability of a larger amount of data could have strengthened and improved the quality of our results.

The selection of the retracted articles was another crucial issue since we faced two major problems: (1) some inconsistencies in the data provided by Retraction Watch, and (2) the presence of retracted articles labeled as humanities that, at a close analysis, actually belonged to a different discipline. The first descriptive statistical results, our manual check, and the definition of the domain affinity score, helped us limit the biases of these two issues. However, we can improve the approach adopted by using additional services such as Elsevier's ScienceDirect – as done in (Bar-Ilan & Halevi, 2018) – and to high up the threshold of the humanities affinity level to exclude border cases.

A citation analysis about retraction in the humanities domain has been rarely discussed in the past (with some exceptions such as Halevi (2020)), therefore the discussion of our results included a comparison with similar works which considered also different domains and different retraction cases. Such works have not addressed the humanities domain or investigated either a single or a limited set of retraction cases. Works which considered other domains did not include most of the features that we have analyzed in this work. e.g., the citation intent, which made the comparison with them difficult. We hope that the work of this thesis and others to be done in this field can favor a comparison and improvement in the understanding of the retraction phenomenon in the humanities domain.

8.3 Future work

In addition to the limits regarding our methods and findings and the adaptions we have presented in the previous sections, there are also other aspects that this thesis did not address if compared with the past approaches. There is for sure a big margin for a future extension of the presented work to strengthen the current methodology along with the introduction of new ways and analysis to infer additional information to help us understand the retraction phenomenon.

In particular, a big improvement to the current analysis would be the generation of a citation network starting from either the retracted article(s) as seed or from its citing entities, as suggested by Van der Vet and Nijveen (2016) who proved the importance of such analysis. Such analysis can enlighten us on the negative/positive outcomes of the error propagation of the retracted research results. A citation network analysis helps us find other relevant additional publications to the retracted ones, to demonstrate, qualitatively, and quantitatively, the connections between publications and authors by creating groups (Martinez-Perez et al., 2020). For instance, the VOSviewer software (Van Eck & Waltman, 2010) relies on citation network analysis. If we consider the retracted publications in RET-SET as seeds in our network, then we can decide to what extent/level we want to explore the citations to such retracted publications. We can decide to consider also the entities which have cited the entities which have used the retracted ones (i.e., two levels of citations).

Although the methodology already implies a check for the citing entities which have been retracted as well, an important aspect to further investigate using a citation network analysis is the "chain reaction" of the retracted publications – *"did the original retraction affect the decision of retraction of the citing article?"*. The answer of this question is crucial, it favors an understanding of the factors that trigger a subsequent retraction of the entities which have used the methods and materials of the retracted publication they have cited, and whether the reason of retraction or the popularity of the case play an important role in the retraction decision.

A feature to consider in the features annotation process of the upcoming versions of the methodology is the journals quartiles – quartiles rank the journals from highest to lowest based on their impact factor or impact index, there are four quartiles: Q1(highest), Q2, Q3 and Q4(lowest). The value of this feature is provided by Scimago. Its annotation for each citing entity as well as for each retracted publication of the RET-SET, can favor an investigation toward the relation / impact that it has on the citation behavior.

The reason for retraction is part of the information provided by Retraction Watch which is extracted from the retraction notice published by the journals. However, apart from its usage in the definition of the case study in the first phase of the methodology, this feature has not been reconsidered in the continuation of the analyzes (e.g., in the generation of the descriptive statistics or as parameter to filter the topics generated after the topic modeling analysis). Although the inclusion of this additional feature might introduce several complications in the interpretations of the results, i.e., all the other values and findings should be represented as a function of the retraction reason, it is no doubt that the integration of this option in the future versions of the methodology can in light this study with interesting new outcomes. Also, a possible future step might involve a text analysis of the retraction notices for a classification of the reasons of retraction, and a comparison of these results with the annotations of the retraction reasons provided by Retraction Watch.

The inclusion of new citations data is another crucial point. The integration of other data retrieved from additional non-open services is an option to be considered in the future. However, it is important to note that such choice might have negative repercussions on the open science principles we have adopted, especially on guaranteeing the reproducibility of the results.

Bibliography

[Al-Hidabi & Teh, 2019] Al-Hidabi, M. D. A., & Teh, P. L. (2019). Multiple Publications: The Main Reason for the Retraction of Papers in Computer Science. In K. Arai, S. Kapoor, & R. Bhatia (Eds.), *Advances in Information and Communication Networks* (Vol. 886, pp. 511–526). Springer International Publishing. https://doi.org/10.1007/978-3-030-03402-3_35

[Aldeen AlRyalat et al., 2020] Aldeen AlRyalat, S., Azzam, M., Massad, A., & Alqatawneh, D. (2020). Retractions of research papers by authors from the Arab region (1998-2018). *European Science Editing*, *46*, e51002. https://doi.org/10.3897/ese.2020.e51002

[Anauati et al., 2016] Anauati, V., Galiani, S., & Gálvez, R. H. (2016). Quantifying the Life Cycle of Scholarly Articles Across Fields of Economic Research. *Economic Inquiry*, *54*(2), 1339–1355. https://doi.org/10.1111/ecin.12292

[Anderson & Lemken, 2020] Anderson, M. H., & Lemken, R. K. (2020). Citation Context Analysis as a Method for Conducting Rigorous and Impactful Literature Reviews. *Organizational Research Methods*. https://doi.org/10.1177/1094428120969905

[Archambault & Larivière, 2010] Archambault, É., & Larivière, V. (2010). The limits of bibliometrics for the analysis of the social sciences and humanities literature. https://ost.openum.ca/files/sites/132/2017/06/WSSR_ArchambaultLariviere.pdf

[Arghode, 2012] Arghode, V. (2012). Qualitative and Quantitative Research: Paradigmatic Differences. *Global Education Journal*, *2012*(4).

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

[Arun et al., 2010] Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murthy, M. N. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. *Advances in Knowledge Discovery and Data Mining* (Vol. 6118, pp. 391–402). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-13657-3_43

[Azoulay et al., 2017] Azoulay, P., Bonatti, A., & Krieger, J. L. (2017). The career effects of scandal: Evidence from scientific retractions. *Research Policy*, *46*(9), 1552–1569. https://doi.org/10.1016/j.respol.2017.07.003

[Azoulay et al., 2015] Azoulay, P., Furman, J. L., Krieger, & Murray, F. (2015). Retractions. *Review of Economics and Statistics*, *97*(5), 1118–1136. https://doi.org/10.1162/REST_a_00469

[Azoulay et al., 2011] Azoulay, P., Graff Zivin, J. S., & Manso, G. (2011). Incentives and creativity: Evidence from the academic life sciences. *The RAND Journal of Economics*, *42*(3), 527–554. https://doi.org/10.1111/j.1756-2171.2011.00140.x

[Bhatt, 2020] Bhatt, B. (2021). A multi-perspective analysis of retractions in life sciences. *Scientometrics*, 126(5), 4039–4054. https://doi.org/10.1007/s11192-021-03907-0

[Bar-Ilan & Halevi, 2017] Bar-Ilan, J., & Halevi, G. (2017). Post retraction citations in context: A case study. *Scientometrics*, *113*(1), 547–565. https://doi.org/10.1007/s11192-017-2242-0

[Bar-Ilan & Halevi, 2018] Bar-Ilan, J., & Halevi, G. (2018). Temporal characteristics of retracted articles. *Scientometrics*, *116*(3), 1771–1783. https://doi.org/10.1007/s11192-018-2802-y

[Bengfort et al., 2018] Bengfort, B., Bilbro, R., & Ojeda, T. (2018). Applied text analysis with Python: Enabling language-aware data products with machine learning (First edition). O'Reilly Media, Inc.

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific american*, *284*(5), 34-43.

[Bizer et al., 2018] Bizer, C., Vidal, M.-E., & Skaf-Molli, H. (2018). Linked Open Data. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of Database Systems* (pp. 2096–2101). Springer New York. https://doi.org/10.1007/978-1-4614-8265-9_80603

[Blei et al., 2003] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

[Boyack et al., 2018] Boyack, K. W., van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics*, *12*(1), 59–73. https://doi.org/10.1016/j.joi.2017.11.005

[Boldt, 2000] Boldt, J. (2000). [RETRACTED] The Good, the Bad, and the Ugly: Should We Completely Banish Human Albumin from Our Intensive Care Units?: *Anesthesia & Analgesia*, *91*(4), 887–895. https://doi.org/10.1097/00000539-200010000-00022

[Bolland et al., 2021] Bolland, M. J., Grey, A., & Avenell, A. (2021). Citation of retracted publications: A challenging problem. *Accountability in Research*, 1–8. https://doi.org/10.1080/08989621.2021.1886933

[Bonilla & Grimmer, 2013] Bonilla, T., & Grimmer, J. (2013). Elevated threat levels and decreased expectations: How democracy handles terrorist threats. *Poetics*, *41*(6), 650–669. https://doi.org/10.1016/j.poetic.2013.06.003

[Bordignon, 2020] Bordignon, F. (2020). Self-correction of science: A comparative study of negative citations and post-publication peer review. *Scientometrics*, *124*(2), 1225–1239. https://doi.org/10.1007/s11192-020-03536-z

[Bornemann-Cimenti et al., 2016] Bornemann-Cimenti, H., Szilagyi, I. S., & Sandner-Kiesling, A. (2016). Perpetuation of Retracted Publications Using the Example of the Scott S. Reuben Case:

Incidences, Reasons and Possible Improvements. *Science and Engineering Ethics*, *22*(4), 1063–1072. https://doi.org/10.1007/s11948-015-9680-y

[Bornmann et al., 2020] Bornmann, L., Wray, K. B., & Haunschild, R. (2020). Citation concept analysis (CCA): A new form of citation analysis revealing the usefulness of concepts for other researchers illustrated by exemplary case studies including classic books by Thomas S. Kuhn and Karl R. Popper. *Scientometrics*, *122*(2), 1051–1074. https://doi.org/10.1007/s11192-019-03326-2

[Boschiero et al., 2021] Boschiero, M. N., Carvalho, T. A., & Marson, F. A. de L. (2021). Retraction in the era of COVID-19 and its influence on evidence-based medicine: Is science in jeopardy? *Pulmonology*, *27*(2), 97–106. https://doi.org/10.1016/j.pulmoe.2020.10.011

[Bordignon, 2020] Bordignon, F. (2020). Self-correction of science: A comparative study of negative citations and post-publication peer review. *Scientometrics*, *124*(2), 1225–1239. https://doi.org/10.1007/s11192-020-03536-z

[Brainard, 2018] Brainard, J. (2018). What a massive database of retracted papers reveals about science publishing's 'death penalty.' *Science*. https://doi.org/10.1126/science.aav8384

[Brownlee, 2019] Brownlee, J. (2019). A Gentle Introduction to the Bag-of-Words Model. 30.

[Budd et al., 2011] Budd, J. M., Coble, Z. C., & Anderson, K. M. (2011). Retracted Publications in Biomedicine: Cause for Concern. ACRL Conference.

[Callon et al., 1986] Callon, M., Law, J., & Rip, A. (1986). Qualitative Scientometrics. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the Dynamics of Science and Technology* (pp. 103–123). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-07408-2_7

[Campos-Varela & Ruano-Raviña, 2019] Campos-Varela, I., & Ruano-Raviña, A. (2019). Misconduct as the main cause for retraction. A descriptive study of retracted publications and their authors. *Gaceta Sanitaria*, *33*(4), 356–360. https://doi.org/10.1016/j.gaceta.2018.01.009

[Campos-Varela et al., 2020] Campos-Varela, I., Villaverde-Castañeda, R., & Ruano-Raviña, A. (2020). Retraction of publications: A study of biomedical journals retracting publications based on impact factor and journal category. *Gaceta Sanitaria*, *34*(5), 430–434. https://doi.org/10.1016/j.gaceta.2019.05.008

[Candal-Pedreira et al., 2020] Candal-Pedreira, C., Ruano-Ravina, A., Fernández, E., Ramos, J., Campos-Varela, I., & Pérez-Ríos, M. (2020). Does retraction after misconduct have an impact on citations? A pre–post study. *BMJ Global Health*, *5*(11), e003719. https://doi.org/10.1136/bmjgh-2020-003719

[Casadevall et al., 2014] Casadevall, A., Steen, R. G., & Fang, F. C. (2014). Sources of error in the retracted scientific literature. *The FASEB Journal*, *28*(9), 3847–3855. https://doi.org/10.1096/fj.14-256735

[Cerejo, 2013] Cerejo, C. (2013). What are the most common reasons for retraction? [Data set]. https://doi.org/10.34193/EI-A-9722

[Chen et al., 2013] Chen, C., Hu, Z., Milbank, J., & Schultz, T. (2013). A visual analytic study of retracted articles in scientific literature. *Journal of the American Society for Information Science and Technology*, *64*(2), 234–253. https://doi.org/10.1002/asi.22755

[Chuang et al., 2012] Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. *Proceedings of the International Working Conference on Advanced Visual Interfaces - AVI '12*, 74. https://doi.org/10.1145/2254556.2254572

[Ciancarini et al., 2014] Ciancarini, P., Di Iorio, A., Nuzzolese, A. G., Peroni, S., & Vitali, F. (2014). Evaluating Citation Functions in CiTO: Cognitive Issues. In V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab, & A. Tordai (Eds.), *The Semantic Web: Trends and Challenges* (Vol. 8465, pp. 580–594). Springer International Publishing. https://doi.org/10.1007/978-3-319-07443-6_39

[COPE Council] COPE Council. (2019). COPE Guidelines: Retraction Guidelines. *Committee on Publication Ethics*. https://doi.org/10.24318/cope.2019.1.4

[Corbyn, 2012] Corbyn, Z. (2012). Misconduct is the main cause of life-sciences retractions. *Nature*, *490*(7418), 21. https://doi.org/10.1038/490021a

[Cokol et al., 2007] Cokol, M., Iossifov, I., Rodriguez-Esteban, R., & Rzhetsky, A. (2007). How many scientific papers should be retracted? *EMBO Reports*, *8*(5), 422–423. https://doi.org/10.1038/sj.embor.7400970

[Corbyn, 2012] Corbyn, Z. (2012). Misconduct is the main cause of life-sciences retractions. *Nature*, *490*(7418), 21–21. https://doi.org/10.1038/490021a

[Crothers et al., 2020] Crothers, C., Bornmann, L., & Haunschild, R. (2020). Citation concept analysis (CCA) of Robert K. Merton's book *Social Theory and Social Structure*: How often are certain concepts from the book cited in subsequent publications? *Quantitative Science Studies*, 1–16. https://doi.org/10.1162/qss_a_00029

[Costas et al., 2015] Costas, R., Zahedi, Z., & Wouters, P. (2015). Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective: Do "Altmetrics" Correlate With Citations? *Journal of the Association for Information Science and Technology*, *66*(10), 2003–2019. https://doi.org/10.1002/asi.23309

[Croidieu & Kim, 2018] Croidieu, G., & Kim, P. H. (2018). Labor of Love: Amateurs and Lay-expertise Legitimation in the Early U.S. Radio Field. *Administrative Science Quarterly*, *63*(1), 1–42. https://doi.org/10.1177/0001839216686531

[Daquino et al., 2020] Daquino, M., Peroni, S., Shotton, D., Colavizza, G., Ghavimi, B., Lauscher, A., Mayr, P., Romanello, M., & Zumstein, P. (2020). The OpenCitations Data Model. In J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, & L. Kagal (Eds.), *The Semantic Web – ISWC 2020* (Vol. 12507, pp. 447–463). Springer International Publishing. https://doi.org/10.1007/978-3-030-62466-8_28

[Degn et al., 2018] Degn, L., Franssen, T., Sørensen, M. P., & de Rijcke, S. (2018). Research groups as communities of practice—A case study of four high-performing research groups. *Higher Education*, *76*(2), 231–246. https://doi.org/10.1007/s10734-017-0205-2

[Dal-Ré & Ayuso, 2021] Dal-Ré, R., & Ayuso, C. (2021). For how long and with what relevance do genetics articles retracted due to research misconduct remain active in the scientific literature. *Accountability in Research*, *28*(5), 280–296. https://doi.org/10.1080/08989621.2020.1835479

[DeWitt & Archer, 2015] DeWitt, J., & Archer, L. (2015). Who Aspires to a Science Career? A comparison of survey responses from primary and secondary school students. *International Journal of Science Education*, *37*(13), 2170–2192. https://doi.org/10.1080/09500693.2015.1071899

[Dhammi & Ul Haq, 2016] Dhammi, I., & Ul Haq, R. (2016). What is plagiarism and how to avoid it? *Indian Journal of Orthopaedics*, *50*(6), 581. https://doi.org/10.4103/0019-5413.193485

[Di Iorio et al., 2013] Di Iorio, A., Nuzzolese, A. G., & Peroni, S. (2013, May). Towards the Automatic Identification of the Nature of Citations. In *SePublica* (pp. 63-74).

[DiMaggio et al., 2013] DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, *41*(6), 570–606. https://doi.org/10.1016/j.poetic.2013.08.004

[Ding et al., 2014] Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis: Content-Based Citation Analysis: The Next Generation of Citation Analysis. *Journal of the Association for Information Science and Technology*, *65*(9), 1820–1833. https://doi.org/10.1002/asi.23256

[Dinh et al, 2019] Dinh, L., Sarol, J., Cheng, Y., Hsiao, T., Parulian, N., & Schneider, J. (2019). Systematic examination of pre- and post-retraction citations. *Proceedings of the Association for Information Science and Technology*, *56*(1), 390–394. https://doi.org/10.1002/pra2.35

[Donthu et al., 2021] Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

[Dougherty, 2020] Dougherty, M. V. (2020). *Disguised Academic Plagiarism: A Typology and Case Studies for Researchers and Editors* (Vol. 8). Springer International Publishing. https://doi.org/10.1007/978-3-030-46711-1

[Du & Wong, 2019] Du, X., & Wong, B. (2019). Science career aspiration and science capital in China and UK: A comparative study using PISA data. *International Journal of Science Education*, *41*(15), 2136–2155. https://doi.org/10.1080/09500693.2019.1662135

[Dusdal & Powell, 2021] Dusdal, J., & Powell, J. J. W. (2021). Benefits, Motivations, and Challenges of International Collaborative Research: A Sociology of Science Case Study. *Science and Public Policy*, *48*(2), 235–245. https://doi.org/10.1093/scipol/scab010

[Duster, 1981] Duster, T. (1981). Intermediate steps between micro-and macro-integration: the case of screening for inherited disorders. *Advances in Social Theory and Methodology, Londres, Routledge & Kegan Paul*, 109-135.

[Elango et al., 2019] Elango, B., Kozak, M., & Rajendran, P. (2019). Analysis of retractions in Indian science. *Scientometrics*, *119*(2), 1081–1094. https://doi.org/10.1007/s11192-019-03079-y

[European Organization For Nuclear Research, & OpenAIRE, 2013] European Organization For Nuclear Research, & OpenAIRE. (2013). *Zenodo: Research. Shared.* https://doi.org/10.25495/7GXK-RD71

[Eysenbach, 2006] Eysenbach, G. (2006). Citation Advantage of Open Access Articles. *PLoS Biology*, *4*(5), e157. https://doi.org/10.1371/journal.pbio.0040157

[Fanelli, 2016] Fanelli, D. (2016). Set up a 'self-retraction' system for honest errors. *Nature*, *531*(7595), 415–415. https://doi.org/10.1038/531415a

[Fang & Casadevall, 2011] Fang, F. C., & Casadevall, A. (2011). Retracted Science and the Retraction Index. *Infection and Immunity*, *79*(10), 3855–3859. https://doi.org/10.1128/IAI.05661-11

[Fang et al., 2012] Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, *109*(42), 17028–17033. https://doi.org/10.1073/pnas.1212247109

[Feng et al., 2020] Feng, L., Yuan, J., & Yang, L. (2020). An observation framework for retracted publications in multiple dimensions. *Scientometrics*, *125*(2), 1445–1457. https://doi.org/10.1007/s11192-020-03702-3

[Ferri et al., 2020] Ferri, P., Heibi, I., Pareschi, L., & Peroni, S. (2020). MITAO: A User Friendly and Modular Software for Topic Modelling. *PuntOorg International Journal*, 5(2), 135–149. https://doi.org/10.19245/25.05.pij.5.2.3

[Flick, 2014] Flick, U. (2014), *An Introduction to Qualitative Research,* London: SAGE.

[Fortunato et al., 2018] Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science*, *359*(6379), eaao0185. https://doi.org/10.1126/science.aao0185

[Foster et al., 2015] Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and Innovation in Scientists' Research Strategies. *American Sociological Review*, *80*(5), 875–908. https://doi.org/10.1177/0003122415601618

[Fry et al., 2011] Fry, L. W., Hannah, S. T., Noel, M., & Walumbwa, F. O. (2011). RETRACTED: Impact of spiritual leadership on unit performance. *The Leadership Quarterly*, *22*(2), 259–270. https://doi.org/10.1016/j.leaqua.2011.02.002

[Fukuhara et al., 2005] Fukuhara, A., Matsuda, M., Nishizawa, M., Segawa, K., Tanaka, M., Kishimoto, K., Matsuki, Y., Murakami, M., Ichisaka, T., Murakami, H., Watanabe, E., Takagi, T., Akiyoshi, M., Ohtsubo, T., Kihara, S., Yamashita, S., Makishima, M., Funahashi, T., Yamanaka, S., … Shimomura, I. (2005). [RETRACTED] Visfatin: A Protein Secreted by Visceral Fat That Mimics the Effects of Insulin. *Science*, *307*(5708), 426–430. https://doi.org/10.1126/science.1097243

[Gamson et al., 1992] Gamson, W. A. (1992). *Talking politics*. Cambridge University Press.

[Gaudino et al, 2021] Gaudino, M., Robinson, N. B., Audisio, K., Rahouma, M., Benedetto, U., Kurlansky, P., & Fremes, S. E. (2021). Trends and Characteristics of Retracted Articles in the

Biomedical Literature, 1971 to 2020. *JAMA Internal Medicine*. https://doi.org/10.1001/jamainternmed.2021.1807

[Gasparyan et al., 2014] Gasparyan, A. Y., Ayvazyan, L., Akazhanov, N. A., & Kitas, G. D. (2014). Self-correction in biomedical publications and the scientific impact. *Croatian Medical Journal*, *55*(1), 61–72. https://doi.org/10.3325/cmj.2014.55.61

[Gasparyan et al., 2018] Gasparyan, A. Y., Yessirkepov, M., Duisenova, A., Trukhachev, V. I., Kostyukova, E. I., & Kitas, G. D. (2018). Researcher and Author Impact Metrics: Variety, Value, and Context. *Journal of Korean Medical Science*, *33*(18), e139. https://doi.org/10.3346/jkms.2018.33.e139

[Glänzel & Moed, 2002] Glänzel, W., & Moed, H. F. (2002). Journal impact measures in bibliometric research. *Scientometrics*, *53*(2), 171-193.

[González-Alcaide et al., 2016] González-Alcaide, G., Llorente, P., & Ramos, J. M. (2016). Bibliometric indicators to identify emerging research fields: Publications on mass gatherings. *Scientometrics*, *109*(2), 1283–1298. https://doi.org/10.1007/s11192-016-2083-2

[Graham et al., 2012] Graham, S., Weingart, S., & Milligan, I. (2012). *Getting started with topic modeling and MALLET*. The Editorial Board of the Programming Historian.

[Gregor & Hevner, 2013] Gregor, S., & Hevner, A. R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, *37*(2), 337–355. https://doi.org/10.25300/MISQ/2013/37.2.01

[Grossarth-Maticek & Eysenck, 1990] Grossarth-Maticek, R., & Eysenck, H. J. (1990). Personality, Stress and Disease: Description and Validation of a New Inventory. *Psychological Reports*, *66*(2), 355–373. https://doi.org/10.2466/pr0.1990.66.2.355

[Guéguen et al., 2014] Guéguen, N., Meineri, S., & Fischer-Lokou, J. (2014). RETRACTED: Men's music ability and attractiveness to women in a real-life courtship context. *Psychology of Music*, *42*(4), 545–549. https://doi.org/10.1177/0305735613482025

[Halevi, 2020] Halevi, G. (2020). Why Articles in Arts and Humanities Are Being Retracted? *Publishing Research Quarterly*, *36*(1), 55–62. https://doi.org/10.1007/s12109-019-09699-9

[Hammarfelt, 2016] Hammarfelt, B. (2016). Beyond Coverage: Toward a Bibliometrics for the Humanities. In M. Ochsner, S. E. Hug, & H.-D. Daniel (Eds.), *Research Assessment in the Humanities* (pp. 115–131). Springer International Publishing. https://doi.org/10.1007/978-3-319-29016-4_10

[Hannigan et al., 2019] Hannigan, T. R., Haans, R. F. J., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., Kaplan, S., & Jennings, P. D. (2019). Topic Modeling in Management Research: Rendering New Theory from Textual Data. *Academy of Management Annals*, *13*(2), 586–632. https://doi.org/10.5465/annals.2017.0099

[Hassan et al., 2017] Hassan, S.-U., Akram, A., & Haddawy, P. (2017). Identifying Important Citations Using Contextual Information from Full Text. *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 1–8. https://doi.org/10.1109/JCDL.2017.7991558

[Heibi & Peroni, 2021] Heibi, I., & Peroni, S. (2021). A qualitative and quantitative analysis of open citations to retracted articles: The Wakefield 1998 et al.'s case. *Scientometrics*. https://doi.org/10.1007/s11192-021-04097-5

[Heibi & Peroni, 2021b] Heibi, I., & Peroni, S. (2021). *LCC and Scimago indexes* [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.4767023

[Heibi & Peroni, 2021c] Heibi, I., & Peroni, S. (2021). *Inputs and results of "A quantitative and qualitative citation analysis to retracted articles in the humanities domain"* [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.5639371

[Heibi & Peroni, 2021d] Heibi, I., & Peroni, S. (2021). A protocol to gather, characterize and analyze incoming citations of retracted articles. *ArXiv:2106.01781 [Cs]*. http://arxiv.org/abs/2106.01781

[Heibi & Peroni, 2021e] Heibi, I., & Peroni, S. (2021). A quantitative and qualitative citation analysis of retracted articles in the humanities. *ArXiv:2111.05223 [Cs]*. http://arxiv.org/abs/2111.05223

[Heibi & Peroni, 2020a] Heibi, I., & Peroni, S. (2020). *A methodology for gathering and annotating the raw-data/characteristics of the documents citing a retracted article v2* [Preprint]. https://doi.org/10.17504/protocols.io.bqqumvww

[Heibi & Peroni, 2020b] Heibi, I, & Peroni, S. (2020). Inputs and results of "A qualitative and quantitative analysis of open citations to retracted articles: the Wakefield 1998 et al.'s case" [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5833466

[Heibi et al., 2019] Heibi, I., Peroni, S., & Shotton, D. (2019). Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. *Scientometrics*, *121*(2), 1213–1228. https://doi.org/10.1007/s11192-019-03217-6

[Heibi et al., 2021] Heibi, I., Peroni, S., Pareschi, L., & Ferri, P. (2021). MITAO: a tool for enabling scholars in the humanities to use Topic Modeling in their studies. AIUCD 2021. https://doi.org/10.6092/UNIBO/AMSACTA/6712

[Hendricks et al., 2021] Hendricks, G., Kramer, B., Maccallum, C. J., Manghi, P., & Neylon, C. (2021). Now is the time to work together toward open infrastructures for scholarly metadata. *Impact of social sciences blog*.

[Hendricks et al., 2020] Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, *1*(1), 414–427. https://doi.org/10.1162/qss_a_00022

[Herteliu et al., 2017] Herteliu, C., Ausloos, M., Ileanu, B., Rotundo, G., & Andrei, T. (2017). Quantitative and Qualitative Analysis of Editor Behavior through Potentially Coercive Citations. *Publications*, *5*(2), 15. https://doi.org/10.3390/publications5020015

[Hilgard & Jamieson, 2017] Hilgard, J., & Jamieson, K. H. (2017). *Science as "Broken" Versus Science as "Self-Correcting"* (K. H. Jamieson, D. M. Kahan, & D. A. Scheufele, Eds.; Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190497620.013.9

[Hjørland, 2013] Hjørland, B. (2013). Facet analysis: The logical approach to knowledge organization. *Information Processing & Management*, *49*(2), 545–557. https://doi.org/10.1016/j.ipm.2012.10.001

[Holman Jones, 2007] Holman Jones, S. (2007). Autoethnography. In G. Ritzer (Ed.), *The Blackwell Encyclopedia of Sociology* (p. wbeosa082). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781405165518.wbeosa082

[Huang & Chang, 2008] Huang, M., & Chang, Y. (2008). Characteristics of research output in social sciences and humanities: From a research evaluation perspective. *Journal of the American Society for Information Science and Technology*, *59*(11), 1819–1828. https://doi.org/10.1002/asi.20885

[Hsiao & Schneider, 2021] Hsiao, T.-K., & Schneider, J. (2021). Continued Use of Retracted Papers: Temporal Trends in Citations and (Lack of) Awareness of Retractions Shown in Citation Contexts in Biomedicine. *Quantitative Science Studies*, 1–53. https://doi.org/10.1162/qss_a_00155

[Jan et al.,2018] Jan, R., Bano, S., Syed, I., Mehraj, M. (2018). Context Analysis of Top Seven Retracted Articles: Should Retraction Watch Revisit the List? *Context*. https://digitalcommons.unl.edu/libphilprac/2016/

[Jacob & Lefgren, 2011] Jacob, B. A., & Lefgren, L. (2011). The impact of research grant funding on scientific productivity. *Journal of Public Economics*, *95*(9–10), 1168–1177. https://doi.org/10.1016/j.jpubeco.2011.05.005

[Jelodar et al., 2019] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

[Jha et al., 2017] Jha, R., Jbara, A.-A., Qazvinian, V., & Radev, D. R. (2017). NLP-driven citation analysis for scientometrics. *Natural Language Engineering*, *23*(1), 93–130. https://doi.org/10.1017/S1351324915000443

[Jockers & Mimno, 2013] Jockers, M. L., & Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics*, *41*(6), 750–769. https://doi.org/10.1016/j.poetic.2013.08.005

[Kalpazidou & Cacace, 2017] Kalpazidou Schmidt, E., & Cacace, M. (2017). Addressing gender inequality in science: The multifaceted challenge of assessing impact. *Research Evaluation*, *26*(2), 102–114. https://doi.org/10.1093/reseval/rvx003

[Kang & Evans, 2020] Kang, D., & Evans, J. (2020). Against method: Exploding the boundary between qualitative and quantitative studies of science. *Quantitative Science Studies*, *1*(3), 930–944. https://doi.org/10.1162/qss_a_00056

[Karabag & Berggren, 2012] Karabag, S. F., & Berggren, C. (2012). Retraction, dishonesty and plagiarism: Analysis of a crucial issue for academic publishing, and the inadequate responses from leading journals in economics and management disciplines. *Journal of Applied Economics and Business Research*, *2*(3), 172-183.

[Kjærnsli & Lie, 2011] Kjærnsli, M., & Lie, S. (2011). Students' Preference for Science Careers: International comparisons based on PISA 2006. *International Journal of Science Education*, *33*(1), 121–144. https://doi.org/10.1080/09500693.2010.518642

[Kuhn, 1962] Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago.

[Krippendorff, 2018] Krippendorff, K. (2018). *Content analysis: An introduction to its methodology.* Sage publications

[Ledford, 2020] Ledford, H. (2020). Safety fears over drug hyped to treat the coronavirus spark global confusion. *Nature*, *582*(7810), 18–19. https://doi.org/10.1038/d41586-020-01599-9

[Ledford & Van Noorden, 2020] Ledford, H., & Van Noorden, R. (2020). High-profile coronavirus retractions raise concerns about data oversight. *Nature*, *582*(7811), 160–160. https://doi.org/10.1038/d41586-020-01695-w

[Leydesdorff & Milojević, 2015] Leydesdorff, L., & Milojević, S. (2015). Scientometrics. In International Encyclopedia of the Social & Behavioral Sciences (pp. 322–327). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.85030-8

[Levy & Franklin, 2014] Levy, K. E. C., & Franklin, M. (2014). Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking Industry. *Social Science Computer Review*, *32*(2), 182–194. https://doi.org/10.1177/0894439313506847

[Li & Agha, 2015] Li, D., & Agha, L. (2015). Big names or big ideas: Do peer-review panels select the best science proposals? *Science*, *348*(6233), 434–438. https://doi.org/10.1126/science.aaa0185

[Lipetz, 1965] Lipetz, B. A. (1965). Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. *American documentation*, *16*(2), 81-90.

[Lu, 2011] Lu, Z. (2011). PubMed and beyond: A survey of web tools for searching biomedical literature. *Database, 2011*. https://doi.org/10.1093/database/baq036

[Lu et al., 2013] Lu, S. F., Jin, G. Z., Uzzi, B., & Jones, B. (2013). The Retraction Penalty: Evidence from the Web of Science. *Scientific Reports*, *3*(1), 3146. https://doi.org/10.1038/srep03146

[Luwel et al., 2018] Luwel, M., & van Eck, N. J. (2018, September). The Schön case: Analyzing in-text citations to papers before and after retraction. In STI 2018 Conference Proceedings (pp. 1025-1030). Centre for Science and Technology Studies (CWTS).

[MacRoberts & MacRoberts, 2018] MacRoberts, M. H., & MacRoberts, B. R. (2018). The mismeasure of science: Citation analysis. *Journal of the Association for Information Science and Technology*, *69*(3), 474–482. https://doi.org/10.1002/asi.23970

[Martinez-Perez et al., 2020] Martinez-Perez, C., Alvarez-Peregrina, C., Villa-Collar, C., & Sánchez-Tena, M. Á. (2020). Current State and Future Trends: A Citation Network Analysis of the Academic Performance Field. *International Journal of Environmental Research and Public Health*, *17*(15), 5352. https://doi.org/10.3390/ijerph17155352

[Matsuyama et al., 2005] Matsuyama, W., Mitsuyama, H., Watanabe, M., Oonakahara, K., Higashimoto, I., Osame, M., & Arimura, K. (2005). RETRACTED: Effects of Omega-3 Polyunsaturated Fatty Acids on Inflammatory Markers in COPD. *Chest*, *128*(6), 3817–3827. https://doi.org/10.1378/chest.128.6.3817

[Marshall, 2013] Marshall, E. A. (2013). Defining population problems: Using topic models for cross-national comparison of disciplinary development. *Poetics*, *41*(6), 701–724. https://doi.org/10.1016/j.poetic.2013.08.001

[McFarland et al., 2013] McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., & Jurafsky, D. (2013). Differentiating language usage through topic models. *Poetics*, *41*(6), 607–625. https://doi.org/10.1016/j.poetic.2013.06.004

[Mehra et al., 2020] Mehra, M. R., Desai, S. S., Ruschitzka, F., & Patel, A. N. (2020). RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *The Lancet*, S0140673620311806. https://doi.org/10.1016/S0140-6736(20)31180-6

[Miller, 2013] Miller, I. M. (2013). Rebellion, crime and violence in Qing China, 1722–1911: A topic modeling approach. *Poetics*, *41*(6), 626–649. https://doi.org/10.1016/j.poetic.2013.06.005

[Moed, 2006] Moed, H. F. (2006). *Citation analysis in research evaluation* (Vol. 9). Springer Science & Business Media.

[Mohr & Bogdanov, 2013] Mohr, J. W., & Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. *Poetics*, *41*(6), 545–569. https://doi.org/10.1016/j.poetic.2013.10.001

[Mongeon & Larivière, 2016] Mongeon, P., & Larivière, V. (2016). Costly collaborations: The impact of scientific fraud on co-authors' careers: Costly Collaborations: The Impact of Scientific Fraud on Co-Authors' Careers. *Journal of the Association for Information Science and Technology*, *67*(3), 535–542. https://doi.org/10.1002/asi.23421

[Mott et al., 2019] Mott, A., Fairhurst, C., & Torgerson, D. (2019). Assessing the impact of retraction on the citation of randomized controlled trial reports: An interrupted time-series analysis. *Journal of Health Services Research & Policy*, *24*(1), 44–51. https://doi.org/10.1177/1355819618797965

[Moylan & Kowalczuk, 2016] Moylan, E. C., & Kowalczuk, M. K. (2016). Why articles are retracted: A retrospective cross-sectional study of retraction notices at BioMed Central. *BMJ Open*, *6*(11). https://doi.org/10.1136/bmjopen-2016-012047

[Mueen & Dhubaib, 2011] Mueen Ahmed, K. K., & Dhubaib, B. E. A. (2011). Zotero: A bibliographic assistant to researcher. *Journal of Pharmacology and Pharmacotherapeutics*, *2*(4), 304–305. https://doi.org/10.4103/0976-500X.85940

[Narayan et al., 2012] Narayan, N., Lee, I. H., Borenstein, R., Sun, J., Wong, R., Tong, G., Fergusson, M. M., Liu, J., Rovira, I. I., Cheng, H.-L., Wang, G., Gucek, M., Lombard, D., Alt, F. W., Sack, M. N., Murphy, E., Cao, L., & Finkel, T. (2012). RETRACTED: The NAD-dependent deacetylase SIRT2 is required for programmed necrosis. *Nature*, *492*(7428), 199–204. https://doi.org/10.1038/nature11700

[Ngah & Goi, 1997] Ngah, Z. A., & Goi, S. S. (1997). Characteristics of citations used by humanities researchers. *Malaysian Journal of Library & Information Science*, *2*(2), 19-36

[Nikpay et al., 2020] Nikpay, F., Ahmad, R., Rouhani, B. D., & Shamshirband, S. (2020). RETRACTED: A systematic review on post-implementation evaluation models of enterprise architecture artefacts. *Information Systems Frontiers*, *22*(3), 789–789. https://doi.org/10.1007/s10796-016-9716-0

[OpenCitations, 2018] OpenCitations. (2018). *COCI CSV dataset of all the citation data* (p. 11568165723 Bytes) [Data set]. figshare. https://doi.org/10.6084/M9.FIGSHARE.6741422.V3

[Oransky & Marcus, 2012] Oransky, I., & Marcus, A. (2012). Retraction watch. *Tracking retractions as a window into the scientific process. Late resveratrol researcher Dipak Das up to*, *20*.

[Packalen & Bhattacharya, 2015] Packalen, M., & Bhattacharya, J. (2015). *Age and the Trying Out of New Ideas*. National Bureau of Economic Research. https://doi.org/10.3386/w20920

[Peroni & Shotton, 2012] Peroni, S., & Shotton, D. (2012). FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. *Journal of Web Semantics*, *17*, 33–43. https://doi.org/10.1016/j.websem.2012.08.001

[Peroni & Shotton, 2018a] Peroni, S., & Shotton, D. (2018). The SPAR Ontologies. In D. Vrandečić, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee, & E. Simperl (Eds.), *The Semantic Web – ISWC 2018* (Vol. 11137, pp. 119–136). Springer International Publishing. https://doi.org/10.1007/978-3-030-00668-6_8

[Peroni & Shotton, 2018b] Peroni, S., & Shotton, D. (2018). *Open Citation: Definition*. https://doi.org/10.6084/M9.FIGSHARE.6683855.V1

[Peroni & Shotton, 2020] Peroni, S., & Shotton, D. (2020). OpenCitations, an infrastructure organization for open scholarship. *Quantitative Science Studies*, *1*(1), 428–444. https://doi.org/10.1162/qss_a_00023

[Price, 1963] Price, D. J. D. S. (1963). *Little Science, Big Science*. Columbia University Press. https://doi.org/10.7312/pric91844

[Price, 1965] Price, D. J. de S. (1965). Networks of Scientific Papers: The pattern of bibliographic references indicates the nature of the scientific research front. *Science*, *149*(3683), 510–515. https://doi.org/10.1126/science.149.3683.510

[Qui et al., 2017] Qiu, J., Zhao, R., Yang, S., & Dong, K. (2017). Methods of Citation Analysis. In J. Qiu, R. Zhao, S. Yang, & K. Dong, *Informetrics* (pp. 207–309). Springer Singapore. https://doi.org/10.1007/978-981-10-4032-0_8

[Ramos, 2003] Ramos, J. (2003). *Using TF-IDF to Determine Word Relevance in Document Queries*.

[Rubbo et al., 2019] Rubbo, P., Helmann, C. L., Bilynkievycz dos Santos, C., & Pilatti, L. A. (2019). Retractions in the Engineering Field: A Study on the Web of Science Database. *Ethics & Behavior*, *29*(2), 141–155. https://doi.org/10.1080/10508422.2017.1390667

[Ribeiro & Vasconcelos, 2018] Ribeiro, M. D., & Vasconcelos, S. M. R. (2018). Retractions covered by Retraction Watch in the 2013–2015 period: Prevalence for the most productive countries. *Scientometrics*, *114*(2), 719–734. https://doi.org/10.1007/s11192-017-2621-6

[Rossetto et al., 2018] Rossetto, D. E., Bernardes, R. C., Borini, F. M., & Gattaz, C. C. (2018). Structure and evolution of innovation research in the last 60 years: Review and future trends in the field of business through the citations and co-citations analysis. *Scientometrics*, *115*(3), 1329–1363. https://doi.org/10.1007/s11192-018-2709-7

[Schneider et al., 2020] Schneider, J., Ye, D., Hill, A. M., & Whitehorn, A. S. (2020). Continued post-retraction citation of a fraudulent clinical trial report, 11 years after it was retracted for falsifying data. *Scientometrics*, *125*(3), 2877–2913. https://doi.org/10.1007/s11192-020-03631-1

[Schmiedel et al., 2019] Schmiedel, T., Müller, O., & vom Brocke, J. (2019). Topic Modeling as a Strategy of Inquiry in Organizational Research: A Tutorial With an Application Example on Organizational Culture. *Organizational Research Methods*, *22*(4), 941–968. https://doi.org/10.1177/1094428118773858

[Sengupta, 1992] Sengupta, I. N. (1992). Bibliometrics, Informetrics, Scientometrics and Librametrics: An Overview. *Libri*, *42*(2). https://doi.org/10.1515/libr.1992.42.2.75

[Shi et al., 2015] Shi, F., Foster, J. G., & Evans, J. A. (2015). Weaving the fabric of science: Dynamic network models of science's unfolding structure. *Social Networks*, *43*, 73–85. https://doi.org/10.1016/j.socnet.2015.02.006

[Sievert & Shirley, 2014] Sievert, C., & Shirley, K. E. (2014). *LDAvis: A method for visualizing and interpreting topics*. https://doi.org/10.13140/2.1.1394.3043

[Shibayama & Wang, 2020] Shibayama, S., & Wang, J. (2020). Measuring originality in science. *Scientometrics*, *122*(1), 409–427. https://doi.org/10.1007/s11192-019-03263-0

[Shuai et al., 2017] Shuai, X., Rollins, J., Moulinier, I., Custis, T., Edmunds, M., & Schilder, F. (2017). A Multidimensional Investigation of the Effects of Publication Retraction on Scholarly Impact. *Journal of the Association for Information Science and Technology*, *68*(9), 2225–2236. https://doi.org/10.1002/asi.23826

[Silverman, 2015] Silverman, D. (2015). *Interpreting qualitative data*. Sage.

[Small, 1978] Small, H. G. (1978). Cited Documents as Concept Symbols. Social Studies of Science, 8(3), 327–340. https://doi.org/10.1177/030631277800800305

[Small, 2018] Small, H. (2018). Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty. Journal of Informetrics, 12(2), 461–480. https://doi.org/10.1016/j.joi.2018.03.007

[Small et al., 2017] Small, H., Tseng, H., & Patek, M. (2017). Discovering discoveries: Identifying biomedical discoveries using citation contexts. Journal of Informetrics, 11(1), 46–62. https://doi.org/10.1016/j.joi.2016.11.001

[Steen et al., 2013] Steen, R. G., Casadevall, A., & Fang, F. C. (2013). Why Has the Number of Scientific Retractions Increased? *PLoS ONE*. https://doi.org/10.1371/journal.pone.0068397

[Stone, 1966] Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

[Suelzer et al., 2019] Suelzer, E. M., Deal, J., Hanus, K. L., Ruggeri, B., Sieracki, R., & Witkowski, E. (2019). Assessment of Citations of the Retracted Article by Wakefield et al With Fraudulent Claims of an Association Between Vaccination and Autism. *JAMA Network Open*, *2*(11). https://doi.org/10.1001/jamanetworkopen.2019.15552

[Suppe, 1998] Suppe, F. (1998). The structure of a scientific paper. *Philosophy of Science*, *65*(3), 381-405.

[Tangherlini & Leonard, 2013] Tangherlini, T. R., & Leonard, P. (2013). Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research. *Poetics*, *41*(6), 725–749. https://doi.org/10.1016/j.poetic.2013.08.002

[Teixeira da Silva & Dobránszki, 2017] Teixeira da Silva, J. A., & Dobránszki, J. (2017). Highly cited retracted papers. *Scientometrics*, *110*(3), 1653–1661. https://doi.org/10.1007/s11192-016-2227-4

[Teixeira da Silva & Bornemann-Cimenti, 2017] Teixeira da Silva, J. A., & Bornemann-Cimenti, H. (2017). Why do some retracted papers continue to be cited? *Scientometrics*, *110*(1), 365–370. https://doi.org/10.1007/s11192-016-2178-9

[Teufel et al., 2006] Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing - EMNLP '06*, 103. https://doi.org/10.3115/1610075.1610091

[Teytelman et al., 2016] Teytelman, L., Stoliartchouk, A., Kindler, L., & Hurwitz, B. L. (2016). Protocols.io: Virtual Communities for Protocol Development and Discussion. *PLOS Biology*, *14*(8). https://doi.org/10.1371/journal.pbio.1002538

[Theis-Mahon & Bakker, 2020] Theis-Mahon, N. R., & Bakker, C. J. (2020). The continued citation of retracted publications in dentistry. *Journal of the Medical Library Association*, *108*(3). https://doi.org/10.5195/jmla.2020.824

[The Retraction Watch Database, 2018] The Retraction Watch Database [Internet]. New York: The Center for Scientific Integrity. 2018. ISSN: 2692-465X. Available from: http://retractiondatabase.org/.

[Thornton et al., 2012] Thornton, P. H., Ocasio, W., & Lounsbury, M. (2012). *The institutional logics perspective: A new approach to culture, structure, and process*. Oxford University Press on Demand.

[Truica et al., 2016] Truica, C.-O., Radulescu, F., & Boicea, A. (2016). Comparing Different Term Weighting Schemas for Topic Modeling. *2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, 307–310. https://doi.org/10.1109/SYNASC.2016.055

[Tuarob et al., 2020] Tuarob, S., Kang, S. W., Wettayakorn, P., Pornprasit, C., Sachati, T., Hassan, S.-U., & Haddawy, P. (2020). Automatic Classification of Algorithm Citation Functions in Scientific Literature. *IEEE Transactions on Knowledge and Data Engineering*, *32*(10), 1881–1896. https://doi.org/10.1109/TKDE.2019.2913376

[Van der Vet & Nijveen, 2016] Van der Vet, P. E., & Nijveen, H. (2016). Propagation of errors in citation networks: A study involving the entire citation network of a widely cited paper published in, and later retracted from, the journal Nature. *Research Integrity and Peer Review*, *1*(1), 3. https://doi.org/10.1186/s41073-016-0008-5

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

[Van Eck & Waltman, 2014] Van Eck, N. J., & Waltman, L. (2014). CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, *8*(4), 802–823. https://doi.org/10.1016/j.joi.2014.07.006

[Vayansky & Kumar, 2020] Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582. https://doi.org/10.1016/j.is.2020.101582

[Vicente-Saez & Martinez-Fuentes, 2018] Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, *88*, 428–436. https://doi.org/10.1016/j.jbusres.2017.12.043

[Vom Brocke et al., 2020] vom Brocke, J., Hevner, A., & Maedche, A. (2020). Introduction to Design Science Research. In J. vom Brocke, A. Hevner, & A. Maedche (Eds.), *Design Science Research. Cases* (pp. 1–13). Springer International Publishing. https://doi.org/10.1007/978-3-030-46781-4_1

[Vuong et al., 2020] Vuong, Q.-H., La, V.-P., Ho, M.-T., Vuong, T.-T., & Ho, M.-T. (2020). Characteristics of retracted articles based on retraction data from online sources through February 2019. *Science Editing*, 7(1), 34–44. https://doi.org/10.6087/kcse.187

[Wakefield et al., 1998] Wakefield, A., Murch, S., Anthony, A., Linnell, J., Casson, D., Malik, M., Berelowitz, M., Dhillon, A., Thomson, M., Harvey, P., Valentine, A., Davies, S., & Walker-Smith, J. (1998). RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, *351*(9103), 637–641. https://doi.org/10.1016/S0140-6736(97)11096-0

[Wang et al., 2020] Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., & Kanakia, A. (2020). Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies*, *1*(1), 396–413. https://doi.org/10.1162/qss_a_00021

[Wang et al., 2019] Wang, K., Shen, Z., Huang, C., Wu, C.-H., Eide, D., Dong, Y., Qian, J., Kanakia, A., Chen, A., & Rogahn, R. (2019). A Review of Microsoft Academic Services for Science of Science Studies. *Frontiers in Big Data*, 2, 45. https://doi.org/10.3389/fdata.2019.00045

[Wang & Barabási, 2021] Wang, D., & Barabási, A.-L. (2021). *The Science of Science* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108610834

[Whelden, 1926] Whelden, C. H. (1926). The Trend-Seasonal Normal in Time Series. *Journal of the American Statistical Association*, *21*(155), 321–329. https://doi.org/10.1080/01621459.1926.10502183

[Whorf, 2012] Whorf, B. L. (2012). Language, thought, and reality: Selected writings of Benjamin Lee Whorf. MIT press.

[Wilkinson et al., 2016] Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

[Wray & Anderson, 2018] Wray, K. B., & Andersen, L. E. (2018). Retractions in Science. *Scientometrics*, *117*(3), 2009–2019. https://doi.org/10.1007/s11192-018-2922-4

[Wu et al., 2017] Wu, L., Wang, D., & Evans, J. A. (2017). Large Teams Have Developed Science and Technology; Small Teams Have Disrupted It. *SSRN Electronic Journal*. [Preprint] https://doi.org/10.2139/ssrn.3034125

[Yeo-The & Tang, 2021] Yeo-Teh, N. S. L., & Tang, B. L. (2021). An alarming retraction rate for scientific publications on Coronavirus Disease 2019 (COVID-19). *Accountability in Research*, *28*(1), 47–53. https://doi.org/10.1080/08989621.2020.1782203

[Yousif et al., 2019] Yousif, A., Niu, Z., Chambua, J., & Khan, Z. Y. (2019). Multi-task learning model based on recurrent convolutional neural networks for citation sentiment and purpose classification. *Neurocomputing*, *335*, 195–205. https://doi.org/10.1016/j.neucom.2019.01.021

[Zavaraqi & Fadaie, 2012] Zavaraqi, R., & Fadaie, G.-R. (2012). Scientometrics or science of science: Quantitative, qualitative or mixed one. *Collnet Journal of Scientometrics and Information Management*, *6*(2), 273–278. https://doi.org/10.1080/09737766.2012.10700939

[Zeng et al., 2017] Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., & Stanley, H. E. (2017). The science of science: From the perspective of complex systems. *Physics Reports*, *714–715*, 1–73. https://doi.org/10.1016/j.physrep.2017.10.001

[Zhao et al., 2015] Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, *16*(S13), S8. https://doi.org/10.1186/1471-2105-16-S13-S8

Appendix

Table 1. The 16 topics of the topic model obtained from the abstracts of the entities citing a retracted publication form the humanities domain. For each topic (row) we mention its proportion percentage in the corpus (column 1), and the 30 most relevant terms (column 2), and we give an interpretation and a title to it (column 3).

Topic (proportion)	**Terms (the 30 most probable terms)**	**Interpretation – Title**
1(7%)	*milaitari, musician, phenotyp, innov, mentor, voic, read, reson, macrofossil, actor, societ, transform, societi, wisdom, colha, strang, peatland, vietnames, vietnam, command, creation, canon, sovereignti, doctrin, union, creat, socialist, labor, sociolog, evolutionari*	Covers some socio-political issues, which might in particular involve Vietnam – Socio-political issues possibly related to Vietnam
2(6%)	*leadership, corpor, style, brand, altruist, reput, servant, bayesian, rmsep, boot, patriot, confirmatori, file, flora, membership, republican, heart, multimedia, dimens, elev, intrins, entit, frequentist, frlt, lwwa, lichen, beta, laissez, tehran, accept*	It talks about leadership from a historical and socio-political point of view – Socio-political issues related to leadership
3(3.2%)	*colon, salvador, slaveri, coloni, tropic, ceram, phillip, glaze, artifact, caryl, columbus, durabl, episod, island, edg, women, breach, guanahani, mostar, spain, lucayan, casuistri, wedg, ethnocrat, saltwat, england, flexur, beam, cross, contracept*	Talks about colonies in history, it might include Columbus as a particular event – Colonial history
4(3.5%)	*border, heritag, deviant, uptak, inject, disadvantag, jewish, gomb, hazard, entangl, anthropomorph, rohmer, marquis, multiscalar, yiddish, inmat, smes, stori, divin, convict, transcrib, zone, patriarch, satisfi, interwoven, warn, count, halesian, laic, dual,*	Discusses historical materials which might involve the Jewish culture – History of the Jewish culture
5(7.5%)	*music, sleep, therapi, intervent, listen, aquina, insomnia, meta, wave, week, qualiti, antagonist, romant, care, cognit, random, conform, relax, grace, qinl, chariti, virtu, bias, treatment, intellect, subgroup, saxophon, holocaust, embas, minut*	Talks about the correlation of music with psychological issues and diseases – Music and psychological diseases
6(17.2%)	*spiritu, workplac, organiz, employe, organ, leadership, commit, work, relationship, mediat, procur, research, studi, satisfact, empir, sens, leader, descart, empower, disast, turnov, inequ, intent, worker, multi, influenc, sector, context, perform, equat,*	Talks about leadership, work organization, and management form a social and political point of view – Leadership organization, and management
7(3.6%)	*fire, consum, massacr, imagin, sociopolit, kymlicka, poisson, vote, empow, circular, moslem, clinician, halal, carpathian, retract, zimbabw, efficaci, redistribut, client, realm, atroc, crime, categori, metabol, hungari, lognorm, compound, voter, romanc, perpetu,*	It has some terms close to the history study domain; it includes also the word “retract" – History in general (might be related to the retraction)
8(3.2%)	*plagiar, pill, ppmvs, mccs, dispers, counsel, spend, psychosomat, counselor, plagiarist, medicin, publish, oral, mysteri, citat, peripato, judgment, multicultur, death, genuin, bone, user, autonomi, disciplin, adherend, serial, advic, successor, sphenoid, anatomist*	Talks about plagiarism with a focus on medicine and health science issues – Plagiarism in health sciences

9(9.5%)	*moral, tunnel, engin, cooper, religion, mosqu, movement, religi, polit, believ, social, econom, hwls, donat, mode, congreg, ethic, burk, less, muslim, basic, prevent, smoke, deceti, abraham, orthot, religionist, caregiv, imam, warfar,*	It clearly discusses historical religion materials – Historical religion issues
10(3.5%)	*virus, creativ, terror, humanitarian, racist, stakehold, tempor, mortal, epistemolog, hack, ornament, angri, bacteria, ethiopia, ingroup, filter, hinduism, counter, theistic, salient, hail, mengistu, mariam, sexi, finer, unexpect, bacteriophag, manifold, microb, sarcoma,*	Talks about viruses and medical diseases – Viruses and medical diseases
11(7.3%)	*parent, depress, children, child, religi, apparit, month, older, group, athlet, associ, ageism, ukrain, hemodialysi, religios, young, sentiment, elder, age, warmth, harsh, buddhist, dictat, optimist, church, younger, baselin, adult, medjugorj, longitudin,*	Discusses social materials related to the family and the religion – Family and religion
12(3%)	*ventil, airflow, spruce, eurosceptic, healer, tribe, activewear, chenopodiacea, fatigu, attitud, picea, purchas, turkish, turkey, arbor, cone, crassifolia, ruqyah, twin, communism, dizygot, marshallian, pigou, node, modest, wind, taxa, complianc, speed, switch,*	Covers some socio-political issues, which might involve turkey – Socio-political issues with a focus on Turkey
13(6.6%)	*classif, classifi, featur, genr, bond, needi, paint, recipi, politi, strengthen, mfcc, song, tumor, signal, ritual, pupil, nsmr, interethn, vector, concret, loyola, nearest, blame, neuro, slab, jslrr, bound, rcbt, polym, ignatian,*	Difficult to interpret, might discuss medical and neuroscience issues – Difficult to interpret, could be related to neuroscience issues
14(4.8%)	*brick, clay, wast, environment, forens, metro, news, green, build, lcas, servic, parametris, brickwork, station, simplifi, decapit, cycl, patient, depart, manufactur, behead, emiss, fine, fhus, resili, immigr, psqi, respond, hospit, agro,*	Covers terms related to the material science in the industry domain, might have a connection with architecture. – Material science and architecture
15(6%)	*vagu, pecham, corg, diplomaci, coerciv, hrqol, disord, yangshao, misconduct, iran, afterlif, quarter, architectur, youth, permiss, studio, anima, tractatus, smolensk, pilgrim, defend, nurs, maker, terrestri, srebrenica, probabil, illeg, libya, pray, intercessori,*	Covers political terms, might describe a particular diplomatic event – Political science issues
16(8%)	*lake, holocen, pollen, veget, plateau, monsoon, tibetan, climat, alpin, forest, plant, water, summer, reconstruct, chang, ecosystem, stepp, proxi, region, meadow, precipit, moistur, record, assemblag, qinghai, sediment, eastern, arid, temperatur, basin*	Covers geographic and climatic issues – Geography and climatic issues

Table 2. The 20 topics of the topic model obtained from the contexts of the in-text citations to retracted publication form the humanities domain. For each topic (row) we mention its proportion percentage in the corpus (column 1), and the 30 most relevant terms (column 2), and we give an interpretation and a title to it (column 3).

Topic (proportion)	**Terms (the 30 most probable terms)**	**Interpretation – Title**
1 (3.5%)	*worldview, florian, aksu, arndt, mcgregor, gurlar, gulu, dalda, taubman, salter, oak, tubb, jona, beatitudin, homini, taiwan, jewish, list, tiesler, royal, request, crise, bono, par, mesoand, quibus, dingwal, charit, jew, chris*	Difficult to interpret, it includes several person names and terms such as Jew, and Jewish which might be part of a social or historical discussion, yet it is hard to establish –Difficult to interpret, might be related to the jewish history

2 (3.1%)	*medjugorj, skrbis, croatian, apparit, socialist, govern, undertaken, nonconform, iranian, pilgrimag, ambiti, propaganda, becam, multilater, soul, iran, korea, grant, revolut, jordan, vietnam, chest, plato, phrene, ethno, smeester, iraq, cancer, coronari, ouimet*	Talks about countries with ongoing or past conflicts from a socio-political point of view –Countries in conflict
3 (7.1%)	*health, mental, threat, religion, cope, psycholog, koenig, distress, pargament, tobacco, conform, tite, existenti, ill, affect, hein, proulx, belief, media, posit, esteem, mediterranean, marathon, alcohol, associ, psychiatri, affirm, drug, martin, stereotyp*	Talks about mental and psychological problems/diseases, which might involve the use of alcohol and drugs. –Drugs/Alcohol and psychological diseases
4 (3.1%)	*week, spend, lipid, soil, chronic, composit, hope, habit, somewhat, resin, occas, dose, rome, player, sphagnum, forward, delimit, hoshow, andean, yablonski, inca, realiz, cumul, expect, desir, faith, heterogen, justifi, kullich, harmat,*	Difficult to interpret, it includes terms related to biology and natural sciences as long as others close to historical studies –Difficult to interpret, treats different domains
5 (5.5%)	*contracept, method, barrier, schimel, nigeria, castano, day, plan, ingroup, uptak, warm, fertil, aristotl, women, oppos, outgroup, zhang, saharan, biomass, sedimentari, geel, defend, multiproxi, ventil, check, campo, studi, target, male, static*	Talks about genders' sociological and medical issues, such as contraceptives for women –Gender social issues
6 (5.3%)	*retract, cardiolog, plagiar, jaha, equiti, ethnic, journal, emiss, editor, skill, workforc, overlap, rhythm, underscor, serbian, ecosystem, session, editori, evolut, familiar, song, divers, heart, medicin, demand, techniqu, articl, race, economi, racial*	Talks about the retraction phenomenon (in particular, plagiarism) and about its main actors (e.g. journals, editors) –The retraction phenomenon
7 (4.2%)	*children, nonreligi, household, alloc, punit, christian, yoosefi, shariff, game, flight, million, upbring, versus, citi, religi, buttress, hosseini, unipolar, less, anonym, money, render, parent, greatest, rais, bipolar, polycentr, partner, biebyck, relate*	Talks about social issues related to the family. Might include and discuss religion in it. –Family and religion
8 (4.7%)	*moreno, portolé, solaz, cabo, softwar, lotka, data, squar, physiqu, bibliographi, invers, prophet, cosar, ozman, morbid, voluntar, templat, optim, fall, tool, environment, center, author, taskin, distanc, logarithm, casuistri, braungart, triandafyllidou, autom*	Difficult to interpret, it includes a bunch of names and a terms close to the engineering domain. –Difficult to interpret, might be related to engineering domain
9 (8.4%)	*lake, pollen, plateau, climat, holocen, qinghai, precipit, record, tibetan, forest, shen, monsoon, declin, mieh, veget, mountain, tree, temperatur, summer, chang, sediment, basin, southern, region, luanhaizi, picea, tibet, schlütz, northeastern, site*	Talks about geographic and climatic issues –Geography and climatic issues
10 (5.5%)	*unterrain, deform, fink, wive, cranial, neolith, rite, poisson, prayer, femal, yellow, skull, head, asia, huntington, accident, ladenhauf, huber, wallner, liebmann, glaze, class, intent, communic, regim, obliqu, stori, clash, cultur, church*	It discusses arguments form the biology and anatomy in relation to social issues –Biology and social sciences
11 (6.9%)	*music, anxieti, bond, sleep, therapi, concret, adhes, trimmer, epoxi, qualiti, khasawneh, intervent, symptom, listen, ignor, surfac, benefici, classif, reduct, dementia, older, strength, genr, improv, machin, täljsten, random, patient, turbul, allevi*	Talks about the correlation of music with psychological issues and diseases –Music and psychological diseases

12 (11.7%)	*spiritu, organiz, leadership, employe, perform, workplac, commit, organ, membership, vision, unit, satisfact, product, duchon, sens, call, work, mather, inner, encourag, motiv, hope, slocum, faith, leader, giacalon, militari, ashmo, relationship, valu*	Talks about leadership, work organisation, and management form a social and political point of view –Leadership organisation, and management
13 (5.5%)	*generos, countri, fix, affili, intend, canada, conclus, yield, model, zero, ideal, item, purchas, dummi, reput, vitucci, revers, dictat, religios, household, tabl, punish, empathi, regress, religi, investig, score, injustic, rich, cloth*	Talks and discusses about social issues related with the religion –Religion
14 (1.2%)	*timbral, compani, aerob, pitch, textur, rhythmic, kuhn, market, fleck, fatigu, featur, ambit, coerciv, refocus, indulg, saddam, turmoil, houdt, donnelli, sketchup, googl, bentler, kanungo, christensen, selfless, foci, menon, chalder, ownership, postinfecti*	Talks about a software for architectural design –Software for architectural design
15 (4.6%)	*polici, guéguen, varieti, guitar, pointless, bandit, twenti, erupt, rope, humil, buhari, act, augustin, profil, carri, whatev, franc, relianc, sensor, icc, horror, walk, helvoort, gülmez, werden, einer, aberr, konflikt, homogenisierung, zwar*	It seems the discussion involves a particular socio-political case, some of the terms might indicate a terroristic attack –War and terrorism
16 (6%)	*avolio, gardner, authent, team, servant, sticker, luthan, share, ethic, genuin, appreci, film, beck, asd, keen, triad, leadership, mode, contrast, reincarn, metric, experienc, dent, bjorl, obeyeseker, wrap, colha, massey, crania, hook*	It talks about leadership and socio-ethic issues and controversies –Leadership and socio-ethic issues
17 (2.7%)	*paint, pigment, tactic, charlton, solicit, sung, invok, eugen, jstor, sensationalist, grate, ceylan, aydin, dri, oil, acupunctur, induct, websit, focuss, kirchmaiera, ussr, began, dealt, korean, soviet, probabil, massacr, languag, excel, preserv*	Discusses socio-political issues, which might be particularly related to historical events such as wars. –War and history
18 (5.7%)	*europ, european, rumford, multipl, biebuyck, uniti, imaginari, plural, closer, tension, singular, rather, empower, ident, media, seen, cosmopolitan, clair, realize, protest, sentiment, exist, mani, implic, walter, disrupt, koger, anyth, micro, broton*	Talks about historical and social issues which might be particularly related to Europe. –History of Europe
19 (3%)	*kantola, intellect, meyer, scotus, book, politi, portion, deepli, suarez, smock, twigg, parayitam, transcendent, bandsuch, tombaugh, extant, chap, sphenoid, megatherm, henri, undocu, footnot, dispers, anarch, actorhood, theoriz, plagiar, object, seven, entir*	Difficult to interpret, it includes several person names and other terms related to letterature and the publishing process –Difficult to interpret, might be related to letterature
20 (2.5%)	*andrea, scène, metteur, rohmer, tribe, fight, flow, genocid, artmak, avelar, fiction, sloth, alexand, enrol, peripher, weidman, campaign, ensembl, auteur, newspap, sorabji, mixtec, totonac, bougarel, soviet, marrouchi, former, film, antiqu, aztec*	Talks about historical events, including army conflicts and wars –War and army conflicts

Table 3. The 13 topics of the topic model obtained from the abstracts of the entities citing WF-PUB-1998. For each topic (row) we mention its proportion percentage in the corpus (column 1), and the 30 most relevant terms (column 2), and we give an interpretation and a title to it (column 3).

Topic (proportion)	**Terms (the 30 most probable terms)**	**Interpretation – Title**
1 (3.5%)	*bias, epidem, philosoph, cell, experi, consum, behavior, protect, even, illich, acetaminophen, scientist, scienc, oxid, call, conserv, long, public, scientif, phone, campaign, occur, experiment, ethic, feed, among, politician, reject, dissent, insignific*	Close to the social studies domain, might take in consideration ethical thematic – Social science studies
2 (4.2%)	*retract, symptom, disabl, student, scienc, perspect, expertis, misconduct, forens, paper, probabl, studi, mandat, cite, diseas, treat, provid, research, replic, vaccin, tripl, librari, qualif, adjuv, might, take, younger, case, media, wakefield*	Includes terms related to the retraction phenomena (study-domain independent), it includes terms used in scientometric analysis. – Bibliometrics and retraction
3 (46%)	*vaccin, parent, children, health, inform, immun, public, decis, review, safeti, studi, measl, evid, articl, research, risk, disord, practic, autist, issu, factor, concern, diseas, report, increas, import, relat, child, effect, base*	This is by far the largest topic out of the 13, it contains common terms, highly frequent in the corpus, and close to the WF-PUB-1998 thematic. – Discussion toward the contents of WF-PUB-1998
4 (3.6%)	*access, nurs, bowel, knowledg, hepat, polici, global, mobil, portfolio, immunis, ask, biblic, newspap, pharmaceut, huge, visitor, time, citizen, percept, symptom, organ, internet, take, model, statist, carer, epoch, golden, scientif, cognit*	Includes terms from the Medical and pharmaceutical field. – Medicine and pharmacology
5 (5.1%)	*vaccin, regress, anti, myth, movement, countri, syndrom, social, incom, diagnost, affect, determin, right, iron, hcws, children, overview, diseas, court, peopl, parent, million, routin, acupunctur, danger, mortal, immun, claim, degrad, intervent*	Some terms are out of the medical field of study and might indicate a possible discussion. – General discussion toward medical and non-medical subjects
6 (3.8%)	*statist, cultur, describ, infant, citat, exampl, articl, symptom, case, american, journal, metabol, combin, parent, disagr, doctor, abl, associ, construction, bordetella, basi, advers, illustr, gliadin, illusori, literatur, indic, unit, eat, rather*	A high number of terms are related to the bibliometrics field of study. All the terms are objective and do not indicate an opinion or a discussion. – Bibliometric materials
7 (4.1%)	*biomark, altmetr, nan, behavior, disord, occur, well, postpon, answer, herd, context, fear, genet, graphic, appeal, interquartil, order, vaccin, gene, sensori, evalu, modul, geneticist, chapter, lymphocyt, abstain, putat, approach, protect, homeopathi*	Includes terms from the biology, pharmacology and genetics field of study. Might also indicate a statistical analysis along with an open discussion. – Discussion toward biology, pharmacology, and genetics subjects
8 (4.9%)	*vaccin, balanc, symptom, risk, reaction, link, subgroup, mump, regress, sphere, aefi, opioid, public, prevent, record, case, food, diseas, chronic, media, claim, allerg, week, resid, advers, children, estim, strain, cobalamin, associ*	Large part of the terms are close to WF-PUB-1998 subjects. Mostly from the medical field of study. – General description of the WF-PUB-1998
9 (5.1%)	*fraud, diseas, narrat, vaccin, complic, health, controversi, comorbid, measl, polici, coliti, bowel, neurolog, travel, inflammatori, movement, trust, ocean, research, retract, attribut, percept, public, futur, preserv, caus, ulcer, case, medic, problem*	Concern the retraction phenomena, followed with some medical expressions. It also includes strong terms such as “fraud”. – General description of the WF-PUB-1998

10 (4.9%)	*vaccin, immun, misconduct, polici, patholog, retract, aefi, result, children, research, disord, qualit, report, caus, development, advanc, case, chang, internet, record, expos, mold, infecti, program, vaer, actor, live, sinc, mani, appli*	General terms related to WF-PUB-1998 treated thematics, some are correlated with a discussion around the retraction phenomena. – Description of the WF-PUB-1998 and its retraction
11 (5.6%)	*vaccin, health, immunis, engag, communic, reason, disord, virus, diagnost, messag, coverag, examin, make, chang, accept, client, diseas, measl, development, research, consid, resist, peopl, public, evid, observ, recent, imag, effect, pervas*	Terms related to the medical subjects and related to the contents of WF-PUB-1998 – General medical background of WF-PUB-1998
12 (3.3%)	*uncertainti, boy, scientif, gynecologist, debat, twitter, semant, ongo, intent, vaccin, disturb, variabl, messag, rhetor, liabil, frame, reddit, percept, content, sourc, gfcf, produc, paediatr, rais, pyridox, guilt, fact, advic, link, chang*	Includes terms close to the computer science lexicon and from the pediatric field of study. – Computer science and pediatric
13 (5.9%)	*vaccin, incid, scientif, erad, measl, frame, viral, literatur, enceph, diseas, controversi, mother, workshop, differ, propos, expert, infect, increas, evalu, genet, dramat, recent, coalit, frequent, communic, current, link, programm, polio, scienc*	Includes a large number of general terms from different fields of study, part of them are correlated with WF-PUB-1998 thematic. – Discussion of WF-PUB-1998 and in general toward medicine

Table 4. The 22 topics available in the topic model generated using the in-text citation contexts contained in the articles citing WF-PUB-1998. For each topic (row) we mention its proportion percentage in the corpus (column 1), and the 30 most relevant terms (column 2), and we give an interpretation and a title to it (column 3).

Topic (proportion)	**Terms (the 30 most probable terms)**	**Interpretation – Title**
1 (2.7%)	*valu, reuter, thomson, worth, retract, would, impact, time, mean, paper, immun, comparison, figur, lancet, base, mening, cerebr, associ, identifi, greater, roach, regress, senior, later, europ, vitamin, viral, assign, consist, campaign*	Includes few terms from the medical domain. Might address the retraction of WF-PUB-1998 from a statistical/mathematical perspective. Some terms consider general info about the paper (metadata). – Statistical analysis of the WF-PUB-1998 retraction
2 (5.2%)	*articl, control, associ, public, signific, affect, research, biopsi, scientif, follow, controversi, natur, patient, case, decad, preserv, evid, report, vaccin, use, multipl, clinic, subsequ, differ, first, cell, follicl, symptom, concern, studi*	General terms to describe the WF-PUB-1998 research – Description of WF-PUB-1998
3 (3.9%)	*articl, case, colon, development, lancet, enterocol, report, assert, associ, public, vaccin, disord, follow, signific, diagnosi, base, group, mumpsrubella, treatment, controversi, parent, scientif, evid, result, symptom, studi, publish, link, reaction, unknown*	General terms to describe the WF-PUB-1998 research – Description of WF-PUB-1998
4 (3.9%)	*articl, immun, dose, claim, lancet, requir, discredit, declin, find, infect, diseas, februari, colleagu, herd, coverag, respons, measl, report, first, sever, vaccin, regress, month, andrew, second, mump, research, countri, detail, parent*	General terms to describe the WF-PUB-1998 research. Might also outline some metadata of the paper (e.g., the venue). – Description of WF-PUB-1998, content and metadata

5 (3.6%)	*vaccin, link, research, suggest, report, appear, measl, focus, scientif, develop, univers, reject, possibl, studi, associ, mump, investig, though, fund, even, continu, around, wakefieldet, newspap, result, doubt, relat, signific, high, evid*	Discusses the controversial aspects around the retraction case of Wakefield et al. – Controversy around the WF-PUB-1998 retraction
6 (2.8%)	*design, studi, expert, consider, parent, bias, receiv, requir, risk, messag, control, occur, point, factor, time, start, howev, report, third, call, know, vaccin, connect, major, qualiti, media, research, school, best, mump*	Does not include any medical term, it rather focuses on other aspects concerning the WF-PUB-1998 – Non-medical discussion of WF-PUB-1998
7 (2.2%)	*diseas, caus, characterist, process, read, inflamm, alarm, report, assert, without, show, ileocolon, bowel, declin, safeti, student, first, exampl, intestin, media, young, detail, autist, adult, possibl, scientif, studi, even, control, wide*	General terms that summarize the WF-PUB-1998 research, mostly from a medical point of view. – Description of the medical topics in WF-PUB-1998
8 (13.2%)	*regress, development, increas, hypothes, causal, vaccin, link, bowel, disord, case, report, measl, problem, symptom, author, associ, system, immun, mump, hypothesi, relationship, peptid, suggest, opioid, diseas, popul, onset, studi, risk, autist*	General terms that summarize the WF-PUB-1998 research, mostly from a medical point of view. – Description toward the medical topics of WF-PUB-1998
9 (2.6%)	*seri, development, parent, report, abnorm, loss, child, autist, associ, acquir, consecut, articl, spectrum, coliti, pervas, symptom, specif, attent, author, count, skill, normal, public, concurr, characterist, claim, abdomin, reduc, enabl, eight*	Many terms are related to medicine domain, yet the correlation with WF-PUB-1998 is less evident. – General discussion in the medicine domain
10 (3.9%)	*mucos, regress, trigger, subtl, food, medic, symptom, development, pattern, extens, lead, clear, result, retract, disord, specif, condit, enterocol, intoler, possibl, research, also, develop, affect, suggest, case, diarrhea, includ, year, caus*	General terms that summarize the medical topics of WF-PUB-1998. Might also mention its retraction – A medical background description of WF-PUB-1998 and its retraction
11 (3.7%)	*understand, paper, period, peer, cohort, development, singl, review, widespread, technic, major, help, studi, rat, scienc, public, littl, interact, relationship, origin, week, outbreak, neurochem, normal, media, articl, sometim, regress, debat, report*	Focuses on technical aspects close to the publishing process and does not include any medical term. – Issues related to the publishing process
12 (10.8%)	*find, retract, uptak, studi, subsequ, media, vaccin, paper, fear, link, controversi, increas, evid, measl, scientif, articl, publish, number, safeti, mump, journal, belief, parent, public, mani, health, mother, concern, claim, lancet*	Discusses the retraction of WF-PUB-1998 from different perspectives, e.g., social impact. Might also highlight the negative impacts. – Negative impacts of the WF-PUB-1998 retraction
13 (3.6%)	*lancet, cost, articl, rhetor, health, paper, public, text, scienc, begin, outbreak, emerg, autist, immedi, also, featur, interpret, acquir, origin, caus, controversi, distress, depart, note, might, languag, debat, measl, behavior, vaccin*	Discusses the controversial aspects around the retraction of WF-PUB-1998 from different perspectives, yet far from the medical domain. – Controversy around the WF-PUB-1998 retraction
14 (5.6%)	*unit, publish, studi, state, bowel, possibl, general, paper, measl, immunis, royal, case, press, diseas, report, free, group, three, kingdom, research, link, controversi, hospit, risk, would, receiv, vaccin, women, journal, earlier*	Discusses the medical conclusions of the WF-PUB-1998 study – Outcomes of WF-PUB-1998

15 (3.3%)	*intestin, report, associ, studi, regress, retract, team, hospit, bowel, sever, find, ileum, royal, behavior, hypothesi, free, subsequ, caus, ethic, altmetr, problem, colleagu, vaccin, consider, research, connect, abnorm, paper, number, british*	Discusses the medical conclusions of the WF-PUB-1998 study. Might address the controversies around the paper. – Outcomes and controversies of WF-PUB-1998
16 (4.8%)	*associ, development, regress, whether, initi, spectrum, vaccin, bowel, trigger, possibl, propos, widespread, autoimmun, autist, disord, specif, environment, diseas, increas, coverag, concern, public, articl, hypothesi, virus, aris, question, base, andrew, sinc*	Talks about the medical issues and outcomes of WF-PUB-1998. – Background and outcomes of WF-PUB-1998
17 (5.4%)	*refer, articl, link, vaccin, skill, normal, histori, mump, acquir, side, bowel, support, concern, studi, public, research, suggest, associ, measl, demonstr, paper, lancet, evid, pediatr, symptom, exist, follow, prove, describ, diseas*	Discusses the conclusions of WF-PUB-1998 and the impacts of such outcomes – Impact of the WF-PUB-1998 outcomes
18 (4.6%)	*receiv, countri, author, health, measl, articl, year, effect, andrew, eight, recent, parent, behavior, case, suggest, vaccin, paper, begin, safeti, british, sinc, studi, develop, short, mump, patient, campaign, publish, report, investig*	An overview of the WF-PUB-1998 retraction, without necessarily analyzing its contents and conclusions. – Overview of the WF-PUB-1998 retraction
19 (3.2%)	*articl, factor, constip, nonspecif, britain, caus, first, autist, measl, compon, year, upon, call, intestin, publish, dquo, retract, make, immun, event, disord, neurolog, unnecessari, permeabl, find, apoptosi, vaccin, describ, scientif, syndrom*	Discusses the WF-PUB-1998 retraction and the controversial outcomes of the paper. – Controversial outcomes of the WF-PUB-1998 retracted paper
20 (1.6%)	*theori, council, nationwid, queri, consequ, profession, medic, continu, ultim, despit, accept, long, case, optim, rat, vitamin, pertussi, myelogenesi, impair, persist, altern, indic, https, prsa, cross, field, coverag, notif, dismiss, collaps*	Discusses WF-PUB-1998 as a case study for a better understanding of the research behavior or more specifically the research in the medicine field of study. – Understanding science from the analysis of WF-PUB-1998 case
21 (5%)	*vaccin, coverag, public, associ, link, articl, suggest, publish, research, media, time, citat, claim, thimeros, across, prove, particular, measl, lancet, lead, whether, parent, paper, evid, subsequ, extens, sinc, mump, around, increas*	Talks about the consequences of the Wakefield et al. retraction case. Might consider bibliometric aspects (e.g., citations) – Bibliometric analysis of the WF-PUB-1998 retraction
22 (4.4%)	*connect, retract, paper, potenti, topic, disord, research, deer, receiv, exampl, studi, lancet, report, claim, elliman, causal, use, development, inflammatori, although, concern, data, type, exist, inform, bowel, publish, base, measl, attent*	Discusses the WF-PUB-1998 retraction from non-medical point of view. Might use WF-PUB-1998 as a case study. – The WF-PUB-1998 retraction case