Freie Universität Bozen
Libera Università di Bolzano
Università Liedia de Bulsan

Fakultät für Informatik
Facoltà di Scienze e Tecnologie informatiche
Faculty of Computer Sciences

Msc in Computational Data Science (Data Analytics)

Msc Thesis

Pending Title

Candidate: Samuele Ceol

Supervisor: Barbara Russo

July 2023

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent quis eros vitae diam faucibus mattis. Donec fringilla ullamcorper turpis. Sed fermentum mauris sapien, et scelerisque erat molestie at. Donec vel eros eget augue aliquet hendrerit eget sit amet augue. Nullam sollicitudin posuere felis, iaculis sodales odio eleifend at. Suspendisse ultrices odio at volutpat ornare. Ut at lectus urna. In ac sollicitudin turpis.

Vestibulum vel fermentum odio. Vivamus condimentum rutrum libero eu vehicula. Nunc volutpat neque quis urna aliquet, et bibendum mauris rutrum. Maecenas in arcu non orci imperdiet lacinia eget vitae erat. Pellentesque pretium, tortor non tincidunt convallis, turpis lectus facilisis ex, eget consectetur nunc arcu eu nulla. Nulla ornare aliquet enim nec efficitur. In aliquam posuere augue. Etiam a ex at nisl vehicula rhoncus eu sed sem.

Maecenas imperdiet eleifend nisi nec vulputate. Vestibulum vestibulum eget nulla eu sodales. Pellentesque id metus vel eros sodales lacinia. Duis aliquet lacus quis accumsan finibus. Mauris sit amet ipsum dignissim, pulvinar sapien vitae, tristique massa. Quisque sit amet metus convallis, vulputate tellus a, gravida lacus. In et euismod libero, quis fermentum elit. Nam sed congue metus. Nullam at felis elit. Pellentesque enim sapien, efficitur ut ipsum non, auctor porttitor justo. Quisque efficitur odio at lectus condimentum, a eleifend tortor finibus.

Donec eget pharetra odio, vel iaculis elit. Suspendisse sagittis lacinia diam, sed maximus eros mollis nec. In molestie arcu in mauris egestas feugiat. Quisque sit amet leo ac sapien malesuada laoreet. Nunc non sodales mi. Mauris rhoncus, turpis vel ultrices condimentum, augue enim accumsan est, et viverra magna lectus eget quam. Proin ac aliquam purus.

Praesent vel iaculis leo. Morbi justo sapien, laoreet non lorem eu, iaculis vulputate eros. Etiam finibus blandit risus, nec cursus mi venenatis eget. Pellentesque quis mollis felis. Proin hendrerit sagittis ex pretium convallis. Morbi turpis leo, consectetur nec arcu vitae, sodales ultricies lorem. Cras lobortis nisl erat, eget ultrices ligula porttitor vel. Pellentesque imperdiet mattis enim a tincidunt. Maecenas a diam eu lacus ultrices vehicula at quis sem. Nunc volutpat semper hendrerit. Praesent molestie, libero id feugiat pellentesque, leo est feugiat tellus, id tempor nunc neque id.

Zusammenfassung

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent quis eros vitae diam faucibus mattis. Donec fringilla ullamcorper turpis. Sed fermentum mauris sapien, et scelerisque erat molestie at. Donec vel eros eget augue aliquet hendrerit eget sit amet augue. Nullam sollicitudin posuere felis, iaculis sodales odio eleifend at. Suspendisse ultrices odio at volutpat ornare. Ut at lectus urna. In ac sollicitudin turpis.

Vestibulum vel fermentum odio. Vivamus condimentum rutrum libero eu vehicula. Nunc volutpat neque quis urna aliquet, et bibendum mauris rutrum. Maecenas in arcu non orci imperdiet lacinia eget vitae erat. Pellentesque pretium, tortor non tincidunt convallis, turpis lectus facilisis ex, eget consectetur nunc arcu eu nulla. Nulla ornare aliquet enim nec efficitur. In aliquam posuere augue. Etiam a ex at nisl vehicula rhoncus eu sed sem.

Maecenas imperdiet eleifend nisi nec vulputate. Vestibulum vestibulum eget nulla eu sodales. Pellentesque id metus vel eros sodales lacinia. Duis aliquet lacus quis accumsan finibus. Mauris sit amet ipsum dignissim, pulvinar sapien vitae, tristique massa. Quisque sit amet metus convallis, vulputate tellus a, gravida lacus. In et euismod libero, quis fermentum elit. Nam sed congue metus. Nullam at felis elit. Pellentesque enim sapien, efficitur ut ipsum non, auctor porttitor justo. Quisque efficitur odio at lectus condimentum, a eleifend tortor finibus.

Donec eget pharetra odio, vel iaculis elit. Suspendisse sagittis lacinia diam, sed maximus eros mollis nec. In molestie arcu in mauris egestas feugiat. Quisque sit amet leo ac sapien malesuada laoreet. Nunc non sodales mi. Mauris rhoncus, turpis vel ultrices condimentum, augue enim accumsan est, et viverra magna lectus eget quam. Proin ac aliquam purus.

Praesent vel iaculis leo. Morbi justo sapien, laoreet non lorem eu, iaculis vulputate eros. Etiam finibus blandit risus, nec cursus mi venenatis eget. Pellentesque quis mollis felis. Proin hendrerit sagittis ex pretium convallis. Morbi turpis leo, consectetur nec arcu vitae, sodales ultricies lorem. Cras lobortis nisl erat, eget ultrices ligula porttitor vel. Pellentesque imperdiet mattis enim a tincidunt. Maecenas a diam eu lacus ultrices vehicula at quis sem. Nunc volutpat semper hendrerit. Praesent molestie, libero id feugiat pellentesque, leo est feugiat tellus, id tempor nunc neque id.

Sommario

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent quis eros vitae diam faucibus mattis. Donec fringilla ullamcorper turpis. Sed fermentum mauris sapien, et scelerisque erat molestie at. Donec vel eros eget augue aliquet hendrerit eget sit amet augue. Nullam sollicitudin posuere felis, iaculis sodales odio eleifend at. Suspendisse ultrices odio at volutpat ornare. Ut at lectus urna. In ac sollicitudin turpis.

Vestibulum vel fermentum odio. Vivamus condimentum rutrum libero eu vehicula. Nunc volutpat neque quis urna aliquet, et bibendum mauris rutrum. Maecenas in arcu non orci imperdiet lacinia eget vitae erat. Pellentesque pretium, tortor non tincidunt convallis, turpis lectus facilisis ex, eget consectetur nunc arcu eu nulla. Nulla ornare aliquet enim nec efficitur. In aliquam posuere augue. Etiam a ex at nisl vehicula rhoncus eu sed sem.

Maecenas imperdiet eleifend nisi nec vulputate. Vestibulum vestibulum eget nulla eu sodales. Pellentesque id metus vel eros sodales lacinia. Duis aliquet lacus quis accumsan finibus. Mauris sit amet ipsum dignissim, pulvinar sapien vitae, tristique massa. Quisque sit amet metus convallis, vulputate tellus a, gravida lacus. In et euismod libero, quis fermentum elit. Nam sed congue metus. Nullam at felis elit. Pellentesque enim sapien, efficitur ut ipsum non, auctor porttitor justo. Quisque efficitur odio at lectus condimentum, a eleifend tortor finibus.

Donec eget pharetra odio, vel iaculis elit. Suspendisse sagittis lacinia diam, sed maximus eros mollis nec. In molestie arcu in mauris egestas feugiat. Quisque sit amet leo ac sapien malesuada laoreet. Nunc non sodales mi. Mauris rhoncus, turpis vel ultrices condimentum, augue enim accumsan est, et viverra magna lectus eget quam. Proin ac aliquam purus.

Praesent vel iaculis leo. Morbi justo sapien, laoreet non lorem eu, iaculis vulputate eros. Etiam finibus blandit risus, nec cursus mi venenatis eget. Pellentesque quis mollis felis. Proin hendrerit sagittis ex pretium convallis. Morbi turpis leo, consectetur nec arcu vitae, sodales ultricies lorem. Cras lobortis nisl erat, eget ultrices ligula porttitor vel. Pellentesque imperdiet mattis enim a tincidunt. Maecenas a diam eu lacus ultrices vehicula at quis sem. Nunc volutpat semper hendrerit. Praesent molestie, libero id feugiat pellentesque, leo est feugiat tellus, id tempor nunc neque id.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1. Opening

1.2. Rationale

Topic labeling has existed as a support activity for topic modeling and as an active research field in the domain of natural language processing and information retrieval for several decades. Despite its longstanding significance, it is possible to notice a distinct lack of comprehensive systematic literature reviews explicitly dedicated to the exploration of state-of-the-art techniques and current trends on this matter. This is particularly surprising given the critical role that topic labeling plays in providing concise and easily understandable interpretations of topics generated by means of the available topic modeling techniques.

Most of the existing literature reviews on topic modeling either avoid mentioning (Chauhan and Shah, 2021) or only briefly touch upon the subject of topic labeling (Churchill and Singh, 2022), while only a few have dedicated sections that delve into the utilization of labeling techniques and activities in the covered work (Silva et al., 2021; Chen et al., 2016). In this context, the fact that topic labeling is generally treated as a secondary activity often leads to a limited understanding of the current techniques used in this field and may result in missed opportunities for further developments and potential avenues for improvements.

One of the main driving factors for conducting this work is therefore the desire to address this lack of existing reviews specifically targeting the subject of topic labeling. Additionally, a literature review is generally considered a useful tool to provide an overview of the existing research in the field and to identify the trends, challenges, and future directions of the current research. Moreover, it is important to mention how a systematic literature review can help to synthesize the findings from various studies, identify the strengths and weaknesses of existing methods, and highlight the gaps in the current literature (Cooper, 1988; Kitchenham, 2004).

All the factors mentioned in this section ultimately play a role in justifying the existence of the presented work.

1.3. Research Questions

Research questions play a central role in the development of systematic literature review as they serve as a general guide to the process of identifying, selecting, and analyzing the relevant literature. They help to narrow the focus of the review at hand on a specific topic and ensure that all the relevant information is captured from the collected work. Research questions are typically formulated based on the review's objectives and are designed to answer specific questions about the topic being studied. They provide a clear direction for the literature search and help to keep the review organized and structured. Kitchenham (2004) goes as far as stating that: *"The critical issue in any systematic* *review is to **ask the right question** [...] one that is meaningful and important to practitioners as well as researchers [...] will lead either to changes [...] or to increased confidence in the value of current practice [...] identifies discrepancies between commonly held beliefs and reality [and] identifies and/or scopes future research activities".*

The definition of relevant research questions is therefore of primary importance in guiding the development of the review at hand. In this context, a systematic literature review can even be described as: *"[...] a means of identifying, evaluating and interpreting all available research relevant to a particular **research question**"* (Brereton and Budgen, 2006).

In the context of this work, the following (four) research questions are formulated and briefly described:

- **RQ1. What are the different approaches for topic labeling, how are they used and in which context?** Conceptually, the task of assigning relevant labels to topics can be achieved in many different ways. From the simple manual assignment performed by means of domain experts to more advanced techniques that can automate the process by exploiting topic hierarchies (Magatti et al., 2009) or that integrate the labeling step in the topic discovery process with the help of ontological concepts (Allahyari and Kochut, 2015).

  In this context, this first research question aims at highlighting which techniques are exploited in the collected work to pair the generated distributions with suitable identifiers and the details related to how such techniques are applied.

- **RQ2. Which underlying topic modelling techniques are used and how do they affect the choice of a topic labeling approach?** Numerous techniques exist to automatically extract and categorize latent topics based on the text content of a given corpus. In this context, techniques such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Latent Semantic Analysis (LSA) (Dumais, 2004) and Probabilistic LSA (PLSA) (Hofmann, 2013) ultimately exists to reach (in different ways) a similar goal.

  This research question aims at understanding which are the prevalent algorithms that are used in the collected research to generate the relevant topic distributions. Additionally, this research questions explores whether the choice of a given topic modeling technique can ultimately influence the approach taken when labeling topics.

- **RQ3. How are candidate labels ultimately selected and how is the quality of the final label assignments evaluated?** Once generated and assigned, topic labels can be evaluated with regards to their ability to provide a meaningful description for the topic at hand. This evaluation can be performed manually (Adhitama et al., 2017) using well-known statistics such as Cohen's kappa (McHugh, 2012) or can even be automated using existing resources such as Wikipedia's article titles (and related content) as starting points to derive appropriate assignments of topic labels (Lau et al., 2011; Bhatia et al., 2016).

  In this regard, this research question aims at defining how the quality of generated labels is assessed. Additionally, in some instances multiple candidate labels might be proposed in order to describe a single topic. This research question also tries to summarise how (in the collected work) a single label is identified among a set of potentially valid candidates.

- **RQ4. Which are the prevalent domains on which topic labeling techniques have been applied to and how are they shaped?** The systematic literature review at hand does not focus on any particular domain. This means that the collected research is likely to belong to a wide array of different areas of research. The aim of this research question is therefore to analyse which are the prevalent domains for which relevant work is identified and to highlight how a given domain might influence the choices made when generating and subsequently labeling the extracted topics. Additionally, this research questions aims at defining how said domains are shaped and described by means of the corpus of documents under scrutiny.

The four research questions presented in this section will be used to guide the review process by providing a structured framework for examining the collected work.

Chapter 2

Theoretical Foundations & Related Work

2.1. Theoretical Foundations

2.1.1. Primary, Secondary & Tertiary studies

Fundamental in the understanding of the scope of Systematic Literature Reviews is the notion of categories of information sources. In fact, having a general idea on how literary work is categorised on the basis of the provided insights in the domain of reference is an important first step in the contextualisation of the role that literature reviews have in summarising the most current evidence available for the chosen topic. In this context, literary publications are commonly organised into three categories (Gosling and Iles, 2022) defined on the basis of the level (or type) of analysis and synthesis they provide. Depending on the relevant category of belonging, papers are (generally) identified as: primary, secondary or tertiary studies.

Primary studies are original research studies that generate new data and findings on a particular research question. Examples of primary studies include observational studies, case studies, surveys and, in a more general sense, describe methodologies applied by researchers to collect novel data rather than relying on information extracted from already existing research. Jupp (2006) defines the main goal of primary studies to be: *"The generation of new data in order to address a specific research question, using either direct methods such as interviews, or indirect methods such as observation. Data are collected specifically for the study at hand, and have not previously been interpreted by a source other than the researcher"*. Primary studies are usually published as research articles in academic journals.

Secondary studies are studies that analyze and synthesize data and findings from multiple primary sources. Examples of secondary studies include systematic reviews, meta-analyses, and scoping reviews. Secondary studies use explicit and rigorous methods to identify, select, and analyse primary studies on a specific (set of) research question(s) or topic(s). The aim of secondary studies is to provide a comprehensive and objective summary of the existing evidence on a particular research matter. Systematic literature reviews (like the one presented in the context of this work) are a type of secondary study, as they involve a systematic and comprehensive search for primary studies, followed by an analysis and synthesis of the findings extracted from the included work. As stated in Kitchenham (2004): *"Individual studies contributing to a systematic review are called primary studies; a systematic review is a form of secondary study"*.

Tertiary studies are studies that analyze and synthesize data and findings from multiple secondary studies. For example, when analysing Systematic literature reviews in the context of software engineering research, Kitchenham et al. (2009) notes that: *"In this case the goal of the review is to assess systematic literature reviews (which are referred to as secondary studies), so this study is categorised as a tertiary literature review"*. Tertiary studies aim to provide an overview of the existing evidence on a particular topic, typically offering a broader scope than secondary studies. In this regard, Khan et al. (2019) states that: *"Tertiary studies synthesize the secondary studies to provide a holistic view of an area"*.

2.1.2. Systematic Literature Review

The term systematic review identifies a literary work (and the related systematic approach) conducted in order to identify, appraise, and synthesize all the available evidence relevant to one or more research questions and topics identified as the core focus of the review process. The primary objective of a systematic review can be associated with the task of providing a comprehensive and unbiased **summary of the existing evidence** in the collected research (Kitchenham, 2004). Systematic reviews involve a thorough and reproducible search process involving one or more information sources (venues and related repositories) and other sources (sometimes falling outside traditional publication channels, like in the case of Multivocal Literature Reviews) to identify all relevant studies, followed by a critical appraisal of the quality of the included research, and a synthesis of the findings. Kitchenham (2004) justifies the existence of Systematic Literature Reviews as a tool to:

> *"[1] summarise the existing evidence concerning a [...] technology. [2] identify any gaps in current research in order to suggest areas for further investigation. [3] provide a framework/background in order to appropriately position new research activities. [4] examine the extent to which empirical evidence supports/contradicts theoretical hypotheses or even [5] assist the generation of new hypotheses"*.

As previously stated, the prelude to a Systematic Review process is generally associated with the identification of a topic (or area) of interest, normally presented together with a statement related to the underlying **Rationale** that justifies the execution of the proposed review process (Section 1.2). This rationale is closely associated with one or more **Research Questions** (Section 1.3) which are laid out in order to formally define the kind of answers the review process is attempting to provide. Once completed, this introductory phase is normally followed by another fundamental step, represented by the development of a **Review Protocol** (Chapter 3), which is a formal definition of the relevant methodologies that will be followed throughout the review process. This includes: Identifying the eligibility criteria imposed on the collected publications (Section 3.1), defining the relevant information sources from which the primary studies will be collected (Section 3.2), highlighting the search strategy used to gather the relevant research (Section 3.3), describing any additional study collection activity (e.g. snowballing, Section 3.4), laying out the identified inclusion and exclusion criteria that will be used in the paper selection process (Section 3.5), presenting the data items that will be considered in the analysis of the individual studies (Section 3.6 and 3.7) and highlighting the analysis and synthesis methods that will be used to ultimately summarise the relevant findings (Section 3.8). The **Analysis and Synthesis** of the relevant findings (Chapter 6) contained in the collected work is presented as a final result to the review process and is ultimately used to answer the presented research questions. The synthesis contained in a literature review can be done either narratively or quantitatively (in the case of a meta-analysis).

Clearly, the general structure described in this Section will tend to vary depending on the specific requirements imposed by the topic under scrutiny. In the context of the presented review, the notions contained in Kitchenham (2004) have been used as a starting point in identifying the general components and characteristics making up a Systematic Literature Review. On the other hand, an adapted version of the SEGRESS checklist contained in Kitchenham et al. (2022) is used to structure the present review into its various sections (and related activities).

2.1.3. Systematic Mapping Study

Systematic Mapping Studies (sometimes referred to as scoping studies or scoping reviews) are a type of literature review that aims to provide an overview of the literature on a particular topic or research area. The primary objective of a mapping study is to identify the breadth and depth of existing research in a given area, rather than to synthesize and analyze the findings found within the collected research (Mayeda and Andrews, 2021). Mapping studies typically involve a systematic search of multiple databases and other sources to identify relevant literature, followed by a descriptive summary of the characteristics of the studies, such as the study design, population, and even research methods. When defining the goal of Systematic Mapping Studies, states that: *"Systematic mapping studies or scoping studies are designed to give an **overview of a research area**"* and, when comparing them to Systematic Literature Reviews, highlights that:

> *"Systematic mapping studies are used to structure a research area, while systematic reviews are focused on gathering and synthesizing evidence"* - Petersen et al. (2015)

Therefore, a successful Systematic Mapping Study is generally able to provide a clear overview of the research area of interest and can even be instrumental in justifying the development of a Systematic Literature Review at later stage (Rožanc and Mernik, 2021)

Given this premise, it stands to reason that Systematic Mapping Studies and Systematic Literature Reviews also differ in terms of the research questions they respectively tend to pose. In fact:

> *"Research questions in a mapping study are exploratory in nature and focus on providing insights into the current state of research and practice about a specific topic. Whereas, an SLR focuses on synthesizing results to address a well-defined [set of] research question[s]"* - (Riaz et al., 2015).

This difference in the nature of the research questions naturally leads to the aforementioned separation of the expected outcomes between mapping studies and systematic reviews. In this context, the structuring of the data resulting from the research collected within a given mapping study will be dependent on such research questions and might ultimately include useful insights on the appropriate methodological process required for a more detailed analysis in a subsequent systematic review.

Evidently, since scoping studies are generally useful in mapping *"key concepts underpinning a research area"* and in assessing the *"types of evidence available"* within a given domain (Arksey and O'Malley, 2005), they tend to encompass a larger portion of the relevant studies and can be found to have a **generally broader scope** when compared with literature reviews which require a more detailed exploration of the evidence found within such documents Perryman (2016). Additionally, Perryman (2016) (whilst paraphrasing Kitchenham et al. (2010)), states that:

> *"Mapping and systematic reviews can overlap if they also include discussion of outcomes or classify included works by research methods used, even extending to assessment of research papers in a subcategory, but similarities end where systematic reviews categorize, appraise, and synthesize the included works".*

2.1.4. A comparative example of SMSs and SLRs

On top of the definitions and general parallels drawn in Subsection 2.1.2 and 2.1.3, to develop a better understanding of the fundamental differences between Systematic Mapping Studies (SMSs) and Systematic Literature Reviews (SLRs) it might also be useful to draw some practical comparisons on the characteristics of the two types of studies in order to provide a more grounded insight on their differences. For this purpose, the work of Pergher and Rossi (2013) and Achimugu et al. (2014) is selected and analysed. The two studies, corresponding to a SMS and a SLR respectively, explore the available research in the domain of software requirements prioritization.

As a first step in this analysis, a comparison is made between the two studies in relation to the kind of **Research Questions** they each pose. On the one hand, it is possible to observe how the questions presented in the mapping study are generally broader, and show the clear desire of the researcher to map the research area under scrutiny, and to provide a summary relating to the kind of studies one might expect to find when examining the associated primary study. More specifically, the (two) research questions are presented as follows:

"[1] Which ***areas*** *in RP have been addressed and which* ***types of articles*** *have been published? [2] Which* ***types of empirical studies*** *have been published within the RP area?"*

On the other hand, the chosen SLR structures its Research Questions in a more targeted way, with the obvious intent of using them to delve deeper into the techniques and processes proposed in the collected research. Additionally, in this case the questions also show the interest of the authors to explore potential limitations in the identified approaches, and even the desire profile their taxonomy. Once again, the (four) Research Questions are shown as follows:

"[1] What are the ***existing techniques*** *used for prioritizing requirements? [2] What are the* ***descriptions and limitations*** *of existing prioritization techniques? [2] What* ***taxonomy*** *of prioritization scales does each technique exhibit? [4] What are the* ***processes*** *involved in software requirements prioritization?"*

Secondly, it might be interesting to observe how the two studies structure their corresponding **search strategy** and subsequent **study selection** process. In this regard, the chosen mapping study defines its search string on the basis of two different aspects, namely: the area of research under scrutiny (requirements prioritisation) and some common techniques typically used in the field. In contrast, the systematic review provides a slightly more detailed description of the process used to generate the query, which also takes into account: relevant terms from the research questions, potentially relevant alternative spellings, synonyms and keywords from prominent studies and books. With the initial collection of documents gathered by means of the proposed search queries, both studies conduct an initial examination based on the associated titles in order to perform an initial coarse grained selection and to remove potential duplicates.

At this point, only the SLR provides a formal definition of the **inclusion and exclusion criteria** taken into consideration during the study selection process, whilst the SMP simply refers to a more general *"Screening of Papers"* carried out on the initial selection of primary studies. Naturally, in both cases researchers ensure that the selected work focuses on the topic of requirements prioritisation in software engineering. However, in the case of the mapping study no further (stricter) criteria is added to this initial requirement. This minimal study selection procedure reinforces what is stated in Subsection 2.1.3 and further underlines the general tendency of SMS to have a broader approach with regards to the selection of the included studies. On the other hand, the literature review adds to this initial requirement further inclusion criteria based on the paper's language and time period (i.e. studies written in English and published in the years 1996-2013) and a further criteria is introduced requiring the collected papers to address at least one of the proposed Research Questions. Additionally, the authors specifically exclude grey literature from the final selection, including only papers from: *"peer-reviewed journals, refereed conference proceedings, workshops, symposiums, book chapters and IEEE bulletins"*. These additional requirements tend to align with the general characteristics of a traditional systematic review, were further measures are taken in order to ensure the collection of studies containing meaningful, and maybe even more importantly, trustworthy insights.

Additionally, the SLR also includes in its selection process a **Quality Assesment (QA)** step to evaluate the credibility, completeness and relevance of each document on the basis of: the clarity of the research goals and associated techniques, the suitability of the experimental design and the used data sets and the value ultimately provided by the related insights.

Surprisingly, the two secondary studies collected, selected and ultimately analysed a similar number of papers (presenting a final selection of 65 and 73 documents in the SMP and SLR respectively). In this regard, it must be noted that the chosen review was conducted by a larger group of authors than the selected mapping study (three reviewers for the SLR against two for the SMP), which might ultimately justify why this information would not match with the formal definitions provided in the previous Subsections.

Finally, the **Research Findings** associated with each study fundamentally match the prior expectation determined by the difference in the nature of the posed research questions. Naturally, in addition to the discussion associated with the relevant findings, the mapping study also proposes some recommendations which can conceptually be used as guidance by researchers seeking to operate in the chosen domain or by future reviewers working on a SLR in the field.

Given the initial definition of SLRs and SMSs and the more practical comparison presented in this Subsection, it might nonetheless be useful to underline that, in many ways, the discourse surrounding the practical distinctions between SMSs and SLRs still remains *"somewhat fuzzy"* (Napoleão et al., 2017), with differences on the characteristics of individual studies which remains in many ways tied to the domain under scrutiny, the goals of the reviewers and the general scope of the presented work.

2.1.5. Topic modeling

Topic modeling is a computational technique used to extract latent semantic structures from a corpus of textual data. It is a statistical modeling approach that identifies patterns of co-occurring words and topics within a collection of documents (Anthes, 2010). Topic modeling assumes that a document is a mixture of several topics, and each topic is characterized by a set of words that frequently co-occur. The objective of topic modeling is to automatically discover these latent topics that underlie the content of a corpus and to assign each document a probability distribution over these topics (Griffiths and Steyvers, 2004).

Topic modeling is most commonly used as an automated mean to index, cluster and search large collections of (generally unlabeled) documents. In this regard, Blei and Lafferty (2009) state that: *"Effectively using such collections requires interacting with them in a more structured way: finding articles similar to those of interest, and exploring the collection through the underlying topics that run through it. The central problem is that this structure [...] is not readily available in most modern collections, and the size and growth rate of these collections preclude us from building it by hand. To develop the necessary tools for exploring and browsing modern digital libraries, we require automated methods of organizing, managing, and delivering their contents".*

Topic modeling approaches have been widely applied in a variety of domains: in information retrieval systems they allow for a more flexible document retrieval process, which enables to capture the user's intent or context rather than simply using the keywords found in the provided query. For example, if a user searches for "machine learning", a search engine that uses topic modeling can identify not only documents that contain the keyword "machine learning", but also documents that discuss related topics such as artificial intelligence, data mining, and deep learning. Similarly, recommendation systems can benefit from these approaches by using topics in order to suggest relevant items to users based on their interests and past interactions. In social network analysis, topic modeling can be used to identify currently trending subjects based on recently posted content and in journalism, policy studies and human sciences, they represent important tools in environments where being able to quickly understand the content and key themes of a given corpus is of central importance. In all of these cases, the popularity of such approaches can be traced back to the fact that:

> *"Topic models allow us to answer big-picture questions quickly, cheaply, and without human intervention. Once trained, they provide a framework for humans to understand document collections [...] by "reading" models or [...] by using topics as input variables for further analysis"* (Boyd-Graber et al., 2017).

General terminology

In addition to the general definition provided within this section, it might also be useful to highlight some of the key concepts (and associated terminology) often found when dealing with topic modeling approaches. In this context, using the nomenclature from Chen et al. (2016) (which is also highlighted in Silva et al. (2021)), the following notions are presented:

- A **word** (or term) w is a sequence (of arbitrary length) of concatenated (individual) characters. Words are not necessarily unique within a given corpus.
- A **document** d is an ordered sequence of terms. Therefore, the i^{th} document d_i found in the corpus and having length N (where N is the number of terms found within the document) is defined as $d_i = \{w_1, ..., w_n\}$.

- A **corpus** (or collection) C is an unordered sequence of documents. Therefore, a corpus C containing n document is defined as $C = \{d_1, ..., d_n\}$.
- A **vocabulary** V is an unordered sequence of m non-repeating (i.e. unique) terms found in a given corpus (i.e. across all documents).
- A **term-document matrix** A is a $m \times n$ matrix whose i^{th}, j^{th} entry represent the weight of term w_i with regards to document d_j. Such weight is represented by an arbitrary weighting function. A simple example of such a weighting function is the term frequency, representing a count of the occurrences of w_i in d_j.
- A **topic** (sometimes also referred to as concept or topic model) z is a vector of size m representing the distribution of probabilities associated with each word in V with regards to z.
- A **document-topic matrix** θ is a n by K matrix (where K represents the total number of generated topics) whose i^{th}, j^{th} entry represents the probability of topic k_j appearing in document d_j.
- A **topic membership vector** θ_d is a vector of size K representing the probability of each topic appearing in document d. In other words, θ_d is a single row extracted from θ
- A **topic-term matrix** Φ is a K by m matrix whose i^{th}, j^{th} entry represents the probability of word w_j in topic z_i. In other words, each row in Φ represent a single topic z_i.

2.1.6. Topic labeling

Providing a formal definition of topic labelling is a somewhat more challenging task than characterising the notion of topic modelling (as presented in the previous Subsection). In fact, only a relatively small portion of primary studies dealing with topic modeling activities ultimately include a topic labeling step. In this context, even fewer publications tend to include Additionally,

As a first step in finding an early definition of Topic Label(ing), the Semantic Scholar repository is queried using the keyword *"topic label*"* and the results are sorted by citation count. Additionally, a date range restriction is imposed on the resulting papers in order to limit the results to work published after the year 2003. This constraint is imposed in an attempt to find more topically relevant studies by filtering out research preceding Blei et al. (2003), which introduced the first introduced the notion of Latent Dirichtlet Allocation in machine learning. After a relatively brief manual inspection process, the work by Mei et al. (2007) is identified as relevant in providing the required definitions.

In its introductory section, the selected paper states that a common challenge for topic models is to correctly comprehend the meaning of each topic. In this context, the interpretation of a topic solely based on the associated multinomial distribution can be a daunting task, especially for users who might not be well-versed with the source collection from which the topic has been generated. In fact, it often might be rather challenging to answer the simple questions of *"What is a topic model about?"* and to understand *"How is one distribution different from another distribution of words?"*. Given this premise, and after having defined the concept of a topic (model) generated from a collection C and described as a distribution $\{p(w|z)\}_{w \in V}$ over a vocabulary set V, the paper formally defines the notion of **Topic Label** as follows:

> *"A topic label, or a "label", l, for a topic model z, is a sequence of words which is semantically meaningful and covers the latent meaning of z. Words, phrases, and sentences are all valid labels under this definition".*

Starting from this definition, the authors also provide a description of the activity of Topic Model Labeling (or simply **Topic Labeling**):

> *"Given a topic model z extracted from a text collection, the problem of single topic model labeling is to (1) identify a set of candidate labels $L = \{l_1, ..., l_m\}$, and (2) design a relevance scoring function $s(l_i, z)$ [to measure the semantic similarity between the label and the topic model]. With L and s, we can then select a subset of n labels with the highest relevance scores $L_z = \{l_{z,1}, ..., l_{z,n}\}$ for z.*
> *This definition can be generalized to label multiple topics. Let $Z = \{z_1, ..., z_k\}$ be a set of k*

topic models, and $L = \{l_1, ..., l_m\}$ *be a set of candidate topic labels. The problem of multiple topic model labeling is to select a subset of* n_i *labels,* $L_i = \{l_{i,1}, ..., l_{i,n_i}\}$*, for each topic model* z_i*".*

Notice that in Mei et al. (2007) the θ symbol is used to identify a single topic. In this section, the label has been changed to z in order to align it to the terminology definition found in Chen et al. (2016) and presented in Subsection 2.1.5.

Following this definition, a general description of what a suitable (set of) topic label(s) should look like is provided. In this context, it is stated that a label should be: *"(1) understandable, (2) semantically relevant, (3) covering the whole topic well, and (4) discriminative across topics"*. Additionally, it is underlined how in most cases the relevant candidate labels might not be immediately accessible to the researchers. In this regard, it is appropriate to assume that such candidates should be obtained from the same corpus used to generate the topics. Then, a relevant scoring function should be identified and used to rank (and ultimately assign) each label w.r.t the generated topics. Still, it is noted how the process of identifying such candidate labels (especially with limited domain knowledge) will often be quite a challenging task, in large part due to the significant difference in semantics between topic distributions and potential labels.

2.2. Related Work

2.2.1. Existing literature reviews

2.2.2. Existing Meta-Analyses

Chapter 3

Review Protocol (Methods)

3.1. Eligibility Criteria

Eligibility criteria play a crucial role in ensuring the rigor and comprehensiveness of a systematic literature review. In a general sense, eligibility criteria are used as an initial tool to broadly determine which studies can be included in the review. Firstly, eligibility criteria help to establish the time period that will be the focus of the review, which is highly dependent on the number of reviewers and available time and resources. Secondly, they define any existing language limitations that might be imposed on the collected work. Finally, they ensure that studies are selected from appropriate venues, such as reputable journals or conferences. By using eligibility criteria to filter out studies from unreliable or low-quality sources, systematic reviews can increase the quality and reliability of the evidence base they rely on.

3.1.1. Search start & end dates

Establishing eligibility criteria is a crucial aspect of conducting a systematic literature review, and one of the key steps in this process is determining the time-frame of the search. By setting a specific start and end period related to the publication date of the collected research, reviewers can effectively focus their search and ensure that they are collecting only the most up-to-date and relevant literature. The choice of the time frame can be influenced by several factors, such as the scope of the review, the research question, and the availability of resources.

For instance, Silva et al. (2021) justify their decision to limit the search to the last twelve years by explaining that it allows them to focus on more recent and mature work in software engineering research. This approach helps to ensure that the review is based on a comprehensive and up-to-date selection of studies that reflect the latest trends and advancements in the field.

Similarly, for the proposed review, the focus has been narrowed to cover the years **2017-2022**. This choice enables the collection of the latest state-of-the-art research while also ensuring that the set of publications to be processed is manageable within the available time and resources. By setting a specific time frame, the review can avoid the inclusion of outdated or irrelevant studies, which may not contribute to the proposed research questions. The chosen period of 2017-2022 is recent enough to gather the most current research findings and insights while providing a sufficient range of publications for a comprehensive and reliable review.

3.1.2. Language limitations

This review will only consider papers written in the **English language**. This choice ensures that the described research (and corresponding results) are fully understandable to the reviewers. Additionally, it has been decided to analyse only work where the executed topic modeling and labeling techniques have been applied on a corpus made up of documents written (at least partially) in the

English language or where full English translation have been provided for the generated labels (and related topics).

3.1.3. Venue selection process

Gathering an initial set of candidate venues (Journals and Conferences) establishing the basis of the venue selection process represents one of the most important steps in fully defining the eligibility criteria of the covered literature. In order to perform a grounded selection, the repositories of a set of five major scholarly research publishers are queried and the resulting papers aggregated by venue and counted. The consulted **repositories** are: IEEE Xplore; ACM Digital Library; SpringerLink; ScienceDirect (Elsevier) and ACL Anthology.

In order to perform the initial exploratory search, three different queries are formulated. The three queries all relate to the notion of "topic labeling" or "topic modeling," but differ in their structure and meaning:

1. `"topic label*" OR "topic model*"`: This query uses the Boolean operator `OR` to search for any documents that contain either of the root terms "topic label" or "topic model". This means that the search will return documents that contain either phrase, but may also include different formulations of the root terms, such as "topic labeling" or "topic modeling."
2. `"topic label*"`: This query searches for documents that contain the root terms "topic label". This leads to a narrower search compared to the first query, as it does not retrieve any documents containing the root term "topic model".
3. `"topic label*" OR ("topic model*" AND "label*")`: This query uses both the Boolean operator `OR` and parentheses to group two different sub-queries together. The first sub-query is the same as the second query, searching for documents that contain the root term "topic label". The second sub-query uses the Boolean operator `AND` to search for documents that contain both the root term "topic model" and any words that start with "label." The two sub-queries are then combined using "OR" meaning that the search will return any documents that match either sub-query.

The three queries are executed on the selected repositories for the period 2017-2022. The aggregated results stemming from this execution are highlighted in Figure 3.1.

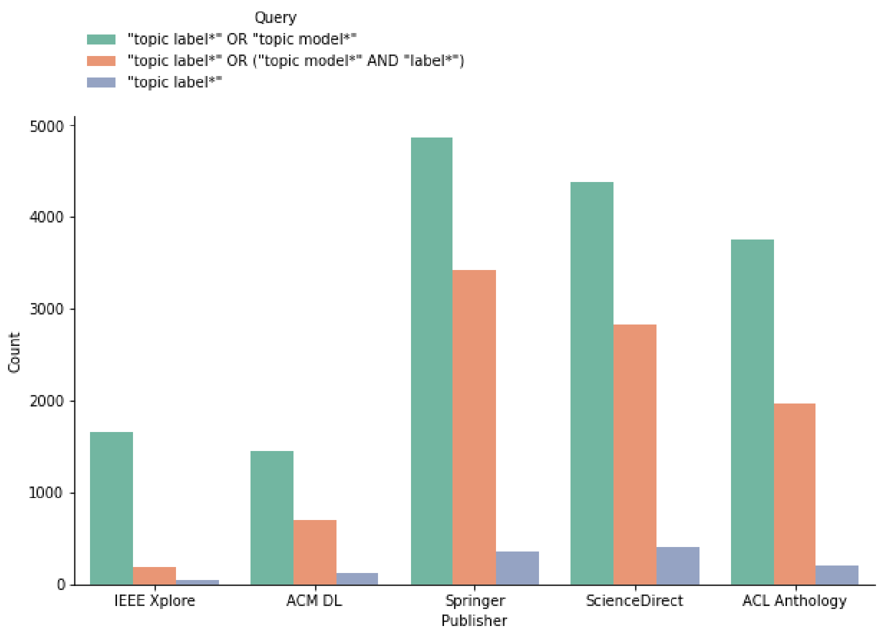


Figure 3.1: Exploratory search results

Ultimately, the venue selection is performed on each repository based on the aggregated results obtained using the query: `"topic label*" OR ("topic model*" AND "label*")`.

The choice of using this query stems from the fact that it represents a middle ground between the narrow query that selects only papers that explicitly mention topic labeling (or topic labels, topic label, etc...) and the broad query that also retrieves all work containing terms related to "topic model". The chosen query ensures that the collected papers that do not explicitly mention topic labeling have at least some mention of terms related to the root word "label". At the same time, it avoids retrieving a significant amount of potentially irrelevant papers lacking any mentions of any labeling activities.

The results obtained from the chosen query are evaluated and the venues selected by taking into account both the **nr. of retrieved papers** and the **rating** issued by the chosen reference bodies. In this regard, the Scimago Journal & Country Rank (SJR, 2023) score averaged in the time span 2017-2021 is considered for journals and the Computing Research and Education Association of Australasia (CORE, 2023) and the The GII-GRIN-SCIE (GGS, 2023) Conference ratings are considered for conferences.

All repositories are queried considering **all metadata** information available for a given candidate result (title, abstract, content, tags, etc...).

The only notable exception to this fact is represented by the ACL Anthology, which does not directly allow to obtain aggregated data from query results. The results are instead obtained solely from the full bibliography with abstracts. Unlike the other repositories, this complete bibliography only offers two useful search fields for each paper (title and abstract). This limitation is taken into account when analysing the results related to the candidate venues published in the ACL Anthology.

Finally, all the selected journals publish original research, ensuring that the sources used in the review provide valuable insights and contribute to the advancement of the field.

Note on the SJR(2) score

As previously mentioned, the obtained candidate journals are ranked using the SJR(2) score (Guerrero-Bote and Moya-Anegón, 2012) averaged in the period 2017-2021. The SJR2 indicator measures the prestige of a scientific journal over a journal citation network. In this network, nodes represent journals and edges the citation relationships between journals. In this context, the score computation is divided into two phases.

At the start of the **first phase**, each journal in the network is assigned a starting prestige value of $\frac{1}{N}$. The starting value is updated iteratively using the following formula:

$$PSJR2_i = \overbrace{\frac{(1-d-e)}{N}}^{1} + \overbrace{e \cdot \frac{Art_i}{\sum_{j=1}^{N} Art_j}}^{2} + \overbrace{\frac{d}{PSJR2D} \cdot \left[\sum_{j=1}^{N} Coef_{ji} \cdot PSJR2_j \right]}^{3} \tag{3.1}$$

Where: $d = 0.9$, 0.0999, N is the number of journals in the repository and Art_i represents the number of citable primary documents in journal i. Additionally, the coefficient $Coef_{ji}$ is computed (before the beginning of each iteration) as follows:

$$Coef_{ji} = \frac{(Cos_{ji} \cdot C_{ji})}{\sum_{h=1}^{N} (Cos_{jh} \cdot C_{jh})} \tag{3.2}$$

Where: C_{ji} represents the numbers of citations from journal j to journal i and Cos_{ji} is the cosine between co-citation profiles of journals j and i. In this context, co-citation is defined as: *"the frequency with which two documents are cited together"* (Small, 1973).

The cosine of the co-citation profiles is expressed as follows:

$$Cos_{ij} = \frac{\sum_{h=1,h\neq i,h\neq j}^{N} Cocit_{ih} Cocit_{hj}}{\sqrt{\sum_{h=1,h\neq i,h\neq j}^{N} (Cocit_{ih})^2} \sqrt{\sum_{h=1,h\neq i,h\neq j}^{N} (Cocit_{jh})^2}} \tag{3.3}$$

Where: $Cocit_{ji}$ is the co-citation of journals j and i.

Finally, the factor $PSJR2D$ representing the total prestige distributed in the current iteration is expressed by the formula:

$$PSJR2D = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{(Cos_{ji} \cdot C_{ji})}{\sum_{h=1}^{N} (Cos_{jh} \cdot C_{jh})} \cdot PSJR2_j \tag{3.4}$$

The individual contributions to the prestige indicator can therefore be identified in the parts 1,2 and 3 of equation 3.1. These parts represent respectively:

1. A base prestige value derived from being part of the SJR repository.
2. The prestige determined by the number of articles included in the journal.
3. The prestige related to the number, “importance”, and “closeness” of citations received by the journal.

During **phase 2**, the final prestige value is computed. Since the `PSJR` value computed in phase 1 is a size-dependent metric (i.e. larger journals will end up having higher prestige values), the final result is obtained by dividing the Phase 1 PSJR score by the ratio of citable documents of each journal:

$$SJR2_i = \frac{PSJR2_i}{(Art_i / \sum_{j=1}^{N} Art_j)} = \frac{PSJR2_i}{Art_i} \cdot \sum_{j=1}^{N} Art_j \tag{3.5}$$

The decision to utilise the SJR(2) score as a reference metric in the evaluation of the gathered journals stems from the overall popularity of the metric which is commonly regarded as a robust way to numerically represent the prestige of a given venue. This fact is further reinforced by the utilisation of this value in citation databases such as Scopus, where SJR appears as one of the three journal-level metrics (together with CiteScore and SNIP).

Details of the CORE rankings

The main goal of the CORE ranking is identified as providing: "guidance to early career researchers, or those entering a new field, regarding likely sources of quality conference papers to become familiar with, and venues to aim to publish in and attend" (CORE, 2023a).

In this context, major conferences are evaluated and ranked periodically by the CORE Executive Committee and assigned a corresponding category (A* to C). In the evaluation process, committee members take into account:

- The percentage of impactful papers (in a given area) based on the corresponding citation data. Citation data sources include: Google Scholar (h5), Elsevier citation (considered by individual FoR codes), ACM (average) citations per paper, Aminer number of citation for the most cited paper belonging to the venue.
- The size of the network of strong researchers publishing in the conference. In this context, strong authors are identified on the basis of the h-index relative to other researchers in the same area.
- The strength of the review process based on the h-index and overall experience of Program Committee members.

The assignment of an A* rating is generally motivated by a particularly strong visibility of the particular conference (even beyond the particular area of research).

Detailed information on the latest evaluation process at the time of writing this review can be found in CORE (2023b).

Details of the GGS rankings

GGS is an initiative is sponsored by GII (Group of Italian Professors of Computer Engineering), GRIN (Group of Italian Professors of Computer Science), and SCIE (Spanish Computer-Science Society) which aims to provide appropriate ratings for computer science conferences by assigning a letter-based rating (from A++ to B-).

The simple rating algorithm (GGS, 2023) used by GGS relies on three distinct data sources from which ratings are obtained, converted to letter values and averaged in order to obtain the final evaluation found in the GGS website. In this regard, the three base data sources are:

- The previously described CORE (2021) rating.
- Microsoft Academics.
- LiveSHINE.

At the time of writing, the latest data used by GGS was downloaded from the corresponding sources on June, 1st 2021.

In the context of this review, the decision to use GGS is justified by the desire to introduce an evaluation metric which is widely used in literary works produced in Italy when needing to asses the quality of a conference.

Methodology for final venue selection

The final venue selection is performed by accounting for both the number of associated publications resulting from the issued query, and the prestige as defined by either the SJR score (average 2017-2021) for journals or the CORE and GGS ratings for conferences.

For journals, the cut-off value related to the number of retrieved publications is set to **30 documents** (i.e. all journals with fewer than 30 articles retrieved from the issued query are discarded). This threshold allows for an initial selection of **23 candidate journals**. The generated candidate set is sorted in descending order according to the SJR metric and the **first quartile** (Q1) representing the top 25% of best scoring journals is selected. This selection, which is visually highlighted in Figure 3.2, ultimately generates the set of **6 journals** used in the context of this review. Said journals are: Pattern Recognition, Journal of Informetrics, Information Sciences, Decision Support Systems, Knowledge-Based Systems, Expert Systems with Applications.

On the other hand, the minimum number of publication for a given candidate conference is set to **20 documents** for all conferences except the ones published in the ACL Anthology were the cut-off value is set to 10. As previously mentioned, this choice is justified by the limitation imposed by the ACL Anthology repository which limited the scope of the search process to the papers titles and abstracts (as opposed to the other repositories which also included other metadata information such as the full document content, tags, etc.). The imposed threshold leads to the initial selection of **11 candidate conferences** represented in Table 3.1.

Conference	CORE	GGS
SIGIR	A*	A++
KDD	A*	A++
ACL	A*	A++
CIKM	A	A+
EMNLP	A	A+
NAACL	A	A+
ECML PKDD	A	A
COLING	A	A
EACL	A	A
ECIR	A	A-
PAKDD	A	B

Table 3.1: Set of (11) candidate conferences

From this initial selection, all conferences having at least a **"A" CORE rating** and a **"A-" GGS rating** are selected from the candidate pool of conferences meeting the initial requirements. This second filtering leads to the final set of **10 conferences** represented in the following list:

- International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)

- ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)
- Annual Meeting of the Association for Computational Linguistics (ACL)
- ACM Conference on Information and Knowledge Management (CIKM)
- Conference on Empirical Methods in Natural Language Processing (EMNLP)
- North American Chapter of the Association for Computational Linguistics (NAACL)
- Machine Learning and Knowledge Discovery in Databases (ECML PKDD)
- International Conference on Computational Linguistics (COLING)
- Conference of the European Chapter of the Association for Computational Linguistics (EACL)
- Advances in Information Retrieval (ECIR)

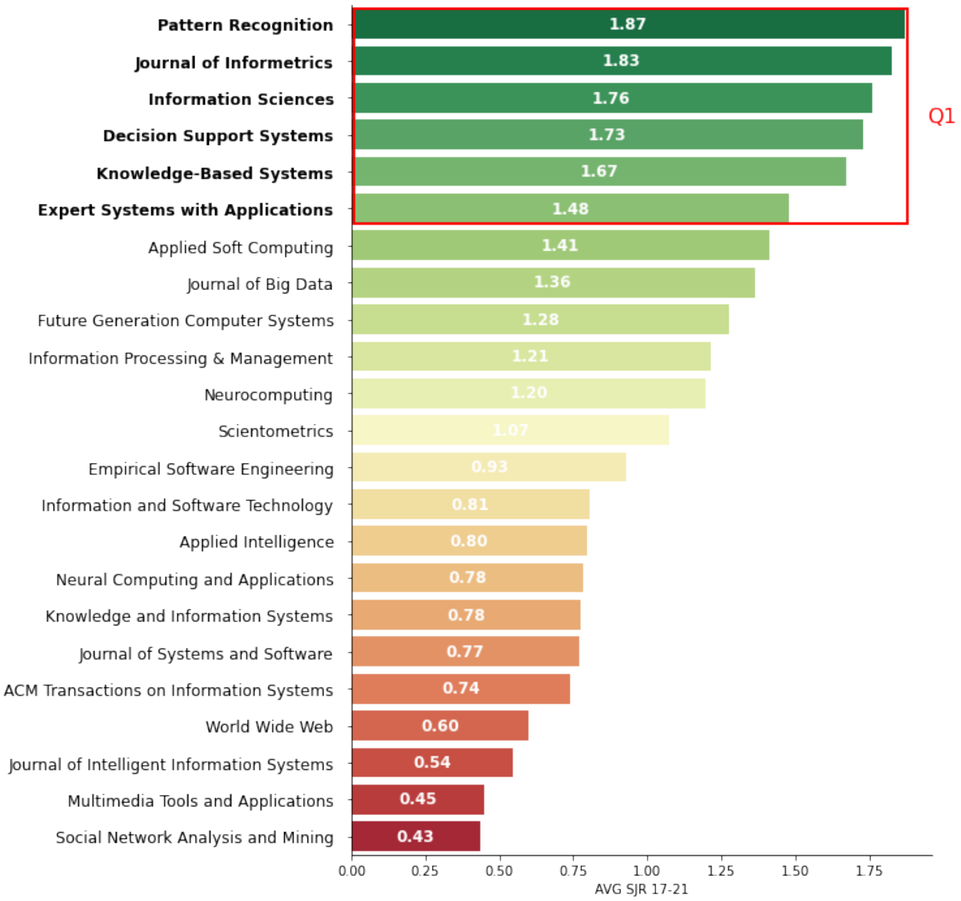


Figure 3.2: Set of (23) candidate journals

3.2. Information Sources

In the context of a systematic literature review, information sources refer to the venues and related repositories searched to identify and gather the relevant studies and other sources of information that will be included in the review. For a comprehensive review it is important to have a clear and comprehensive strategy for identifying relevant information sources, in order to ensure that the overall review process is conducted in a transparent and reproducible manner.

Based on the list of selected venues defined in Subsection 3.1.3, the repositories on which the literature search process is performed are listed below:

- ACL Anthology (ACL, EMNLP, NAACL, COLING, EACL)

- ACM Digital Library (SIGIR, SIGKDD, CIKM)
- Springer (ECML PKDD, ECIR)
- ScienceDirect (Pattern Recognition, Journal of Informetrics, Information Sciences, Decision Support Systems, Knowledge-Based Systems, Expert Systems with Applications)

Additionally, a quantitative summary of such sources is provided in Table 3.2, where each row represents a different publisher/association and its corresponding repositories, conferences, journals, and total number of sources. Additionally, the last column indicates the total number of information sources associated with each repository.

Publisher/Association	Repository	Conferences	Journals	Total
ACL	ACL Anthology	5		5
ACM	ACM Digital Library	3		3
Springer	SpringerLink	2		2
ScienceDirect	Elsevier		6	6
		10	**6**	**16**

Table 3.2: Information sources

3.3. Search Strategy

In the context of a systematic literature review, a search strategy is a well-defined and structured approach that specifies the process for identifying the initial set of relevant research studies on a particular topic. It involves determining the key terms (and related wildcard operators) that are relevant to the research questions, as well as identifying an appropriate query to use in the search process. The search strategy is designed to ensure that all potentially relevant literature is identified and included in the review, while minimizing the risk of missing important studies. It typically involves a combination of search terms and Boolean operators (such as AND, OR, and NOT) that are used to construct complex search queries (Aromataris and Riitano, 2014). The quality and comprehensiveness of the search strategy can have a significant impact on the results of the review, as a poorly designed strategy may result in the exclusion of relevant studies or the inclusion of irrelevant ones. Additionally, providing a formal definition of the utilised search strategy increases the replicability of the presented study.

Starting from the notions introduced in Subsection 3.1.3, the query `"topic label*" OR ("topic model*" AND "label*")` applied on the time period established in Subsection 3.1.1 (2017-2022) is used as a starting point to define the chosen search strategy. The following subsection highlights how the chosen query is further refined to better accommodate the time and resource constraints associated with the presented work.

3.3.1. Refining the chosen query

Given the large amount of papers related to the query spanning the time period 2017-2022 (as highlighted in Figure 3.1), a further filtering step is proposed on the initially established query by imposing a **proximity constraint** (using the `NEAR` operator) between the root terms `label*` and `topic*`. This choice is justified by the assumption that those sentences for which the root term `label` does not appear in close proximity with the root term `topic` are unlikely to be relevant for the proposed review and should therefore not be flagged as positives during the search process.

Since proximity operators are not natively supported by most the chosen repositories (with the exception of SpringerLink), it is decided to first gather the initial set of papers using the more relaxed query `"topic label*" OR ("topic model*" AND "label*")` and then to further filter them locally by imposing the proposed proximity constraint.

This two step process is made possible by the fact that imposing the proximity constraint:

$$\texttt{"label*" NEAR "topic*"} \tag{3.6}$$

On the papers gathered from the query:

$$\texttt{"topic label*" OR ("topic model*" AND "label*")} \tag{3.7}$$

Ultimately leads to the same set of results that would be obtained by directly executing, on the selected repositories, the query:

$$\texttt{"topic label*" OR ("topic model*" AND ("label*" NEAR "topic*"))} \tag{3.8}$$

In fact, the query containing the proximity operator is simply a stricter version of the one used to gather the initial set of papers.

The filtering on local files is performed on the FoxTrot Professional Search tool using, in the indexed folders, the following query specified in using the specific tool's syntax:

$$\texttt{[\{20\} "topic*" "label*"]} \tag{3.9}$$

Note that the value `{20}` in the query 3.9 indicates the strictness of the proximity constraint. In this context, the maximum distance (i.e. the nr. of words) between the two root terms is 20. For readability purposes, for the rest of the chapter the syntax `"label*" NEAR "topic*"` will be used instead in order to indicate the query imposing the proximity constraint used within the FoxTrot search tool.

In this context, all the papers gathered from the first part of the original query (`’’topic label*’’`) are kept. This is due to the fact that in the interested papers the two root terms always appear in direct proximity to one another. On the other hand, for the second part of the original query (`"topic model*" AND "label*"`) only those papers meeting the newly imposed proximity constraint are kept. In other words, the initially gathered papers mentioning topic model(ing) and label(s) are kept only if the proximity constraint between the two root terms (`label` and `topic`) is satisfied.

Formally, if P is defined as the set of papers contained in the selected venues for the chosen time period, and:

- Q_0 as the query `"topic*" NEAR "label*"` (Query 3.6)
- Q_1 as the query `"topic label*" OR ("topic model*" AND "label*")` (Query 3.7)
- Q_2 as the query `"topic label*" OR ("topic model*" AND ("label*" NEAR "topic*"))` (Query 3.8)

Then:

- $Q_1(P)$ represents the set of results obtained by executing Q_1 on P
- $Q_2(P)$ represents the set of results obtained by executing Q_2 on P

In this context, it follows that:

$$Q_2(P) \subseteq Q_1(P) \tag{3.10}$$

Additionally, it follows that:

$$Q_0(Q_1(P)) \equiv Q_2(P) \tag{3.11}$$

3.4. Snowballing Methods

Snowballing, also known as citation chaining (Jobin et al., 2019) or citation chasing (Cooper et al., 2017), is a research method used in literature reviews to identify additional relevant sources of information by exploiting the references and citations in the already selected research studies. Formally, it is defined as a: *"literature-searching technique that involves locating a relevant source paper and then using this paper as a starting point for either working back from its references or for conducting additional citation searches"* (Booth et al., 2012). The purpose of snowballing is to ensure that all relevant studies on a given topic are included in a literature review. By starting with a small number of studies and following the references and citation tied to those studies, researchers can identify additional work that may not have been initially found through database searches or other traditional methods. Snowballing can be particularly useful when (like in the context of this work) strict limitations have been imposed on the consulted repositories and related source venues. In this context, this technique allows to broaden the scope of the presented work by exploiting an already defined set of initially selected studies. Additionally, snowballing can help researchers identify seminal studies or key researchers in a given field, as well as track the development of ideas or concepts over time. This particular possibility is further explored in Chapter 5 where a set of literature graphs are built and analysed from the full set of gathered papers.

Overall, snowballing is a valuable tool for conducting comprehensive literature reviews and ensuring that all relevant research is taken into consideration in the research process. In order to highlight the methodological process related to the snowballing activities tied to this work, the current section will define the guidelines followed in order to conduct the backward and forward snowballing activities, together with the tools used to support them.

3.4.1. Backward Snowballing

Backward snowballing refers to the activity of extracting relevant work from the references appearing in the initial set of selected papers. In other words, this means finding yet undiscovered (referenced) work that should be included in the final review. In this context, the activities involved with backward snowballing can be summarised in the following steps:

1. Traversing the selected list of papers and, for each document, extracting the related list of references (see Subsection 3.4.3).
2. Filtering the extracted references with regards to the imposed time constraints (time frame 2017-2022).
3. Gathering (and storing) the initial set of publications resulting from this reference extraction process.
4. (Locally) applying the previously constructed search query (`"topic label*" OR ("topic model*" AND ("label*" NEAR "topic*"))`) to filter down the gathered publications.
5. Inspecting the content of each remaining document and determining its relevance by applying the same set of inclusion/exclusion criteria utilised for the main set of papers (Section 3.5).

In the same way as the process described in Subsection 3.3.1, the filtering on local files performed in the context of step 4 is achieved via the FoxTrot Professional Search tool.

For the papers gathered from forward snowballing, no quality-based venue filtering procedure is enforced. This choice is justified by the fact that, in a general sense, the quality constraints imposed on the initial selection (and described in Section 3.1.3) are expected to carry over to the referenced work. In this context, it is assumed that an high quality paper will tend to reference only work that is qualitatively on par (or superior) to itself, making the formulation of a further filtering step unneeded.

3.4.2. Forward Snowballing

In a similar fashion, the process of forward snowballing is executed on the publications belonging to the initial selection in order to extract the related (relevant) citations. The steps required to complete this procedure are similar to the ones described for the backward snowballing counterpart (steps 1-5 in Subsection 3.4.1) with the major differentiating aspect being the fact that step 2 can be skipped entirely. In fact, since all papers that are part of the initial selection belong to the time frame 2017-2022, all the related citations are bound to respect the same time constraint. Additionally (and unlike backward snowballing), a further filtering procedure is applied before inspecting the collected work (i.e. before executing step 5). This filtering procedure is guided by the same metrics defined in Subsection 3.1.3 and its results are described in details in the Results Subsection 4.2.2. The reason for adding this additional step (which is entirely absent in the methods of Subsection 3.4.1) stems from the fact that the gathered citations have no associated intrinsic guarantee of quality, and may represent sources that do not meet the qualitative standards enforced for the other publications gathered in the context of this review. In fact, any set of authors, independently from the quality of the work they have produced, can freely cite any of the papers selected in the context of this work. Therefore, not implementing a quality-based filtering procedure for the gathered citations could have an impact on the overall quality of the presented review.

3.4.3. Semantic Scholar API

Manually gathering references and citations from a given piece of work is a laborious and time consuming process, especially for highly cited papers or for those publications where references are not organized in a standard format. Additionally, such a manual process executed on a larger scale tends to be error-prone, which would ultimately lead to missing or incorrect information or incomplete references. Automating the process of gathering references and citations with an API can help to address these issues. Additionally, an API can ensure that references are organized in a consistent manner, regardless of the citation style used by the author.

In order to automate the process of gathering references and citations tied to a given piece of work, the Semantic Scholar URL-based Academic Graph API is used. The Academic Graph API allows to easily automate the extraction of references and citation from a given paper by generating a request URL for each of the publications belonging to the initial selection. In this context, a request URL looks as follows:

```
api.semanticscholar.org/graph/v1/paper/PAPER_ID?fields=references,citations
```

And is composed by a base URL (`api.semanticscholar.org/graph/v1/paper/`), a 40-character alphanumeric identifier (`PAPER_ID`) which uniquely identifies a given paper in the Semantic Scholar database and (optionally) the set of fields that need to be retrieved for the paper at hand (`?fields=references,citations`). In this case, only the `references` and `citations` fields are requested for inclusion as part of the API response.

A JSON response from the Academic Graph API looks as follows:

```
{
    "paperId": "29890a936639862f45cb9a987dd599dce9759bf5",
    "citations": [
        {
            "paperId": "66e54f3dce8c4c156d8c90327d4557d240fe16bd",
            "title": "Predictive maintenance..."
        },
        {
            "paperId": "c51ccc4ebb9631248748a72939fee74aaf39c385",
            "title": "Twitter as a predictive..."
        },
        ...
    ],
    "references": [
        {
            "paperId": "8b6e751d10d557e27b3fe373f41cb573259b9d24",
            "title": "An empirical study of..."
        },
        {
            "paperId": "6f2e6f655e8a08010c72826d048401d61df29ef4",
            "title": "Getting Evidence into..."
        },
        ...
    ]
}
```

Listing 3.1: Academic Graph API response

Starting from such a response, the identifier (`paperID`) of each citation and reference is extracted and a new request is made for each of the newly identified papers in order to retrieve the publication's digital object identifier (DOI) or any equivalent identifier that can be used to universally describe the document. All such identifiers are saved under the field `externalIds` in Semantic Scholar and can be obtained by simply modifying the `?fields` part of the presented URL.

A standard response for a retrieved reference or citation looks as follows:

```
{
    "paperId": "ae202083ed5759d70f0fffbfde7543f7663eb4c0",
    "externalIds": {
        "ACL": "2021.eacl-main.121",
        "DBLP": "conf/eacl/PopaR21",
        "DOI": "10.18653/v1/2021.eacl-main.121",
        "CorpusId": 233189658
    }
}
```

Listing 3.2: External Ids tied to a given paper

3.5. Selection Process (Inclusion / Exclusion Criteria)

Inclusion criteria are characteristics that studies must meet in order to be included in the systematic review. Similarly, exclusion criteria define those characteristics that need to be absent in a given piece of work in order for it to be considered valid for the review at hand. As stated in Meline (2006) (starting from the notion of relevance and acceptability introduced in Robey and Dalebout (1998)): *"Prospective studies for systematic reviews are evaluated for eligibility on the basis of relevance and acceptability [...]. Systematic reviewers ask: Is the study relevant to the review's purpose? Is the study acceptable for review? Systematic reviewers then formulate inclusion and exclusion criteria to answer these questions"*. In this context, it stands to reason that inclusion and exclusion criteria should also be defined by taking into consideration the established research questions (Kitchenham, 2004).

In order to be selected for this review, a paper should propose or actively apply, either with a primary or secondary focus a topic labeling technique. Papers appearing in the selected research do not necessarily need to describe the implementation of a novel labeling approach, but it is important that they do not meet any of the following **exclusion criteria**:

- **EC1.** The paper does not actively apply any topic labeling techniques.
- **EC2.** The labels do not posses descriptive properties with regards to the specifics of the topics' content.
- **EC3.** All the described labeling approaches are taken from existing work and re-proposed as-is (on the same corpus and set of topics).
- **EC4.** The paper and/or the analysed corpus do not match the imposed language restrictions.
- **EC5.** The paper is a systematic review (secondary/tertiary study).

With regards to **EC2**, some examples of papers that would match the exclusion criteria are presented as follows:

- Tang et al. (2019), associates (binary) sentiment labels to topics.
- Bahrainian et al. (2018), extracts topics from a corpus of 28 years of scholarly articles divided into one year time slices. A binary label (continued/not-continued) is assigned to each topic to indicate whether the given topic is also included in the subsequent time slice.
- Figueiredo and Jorge (2019), uses a SVM classifier to assign binary labels to LDA topics in order to classify tweets as either "relevant" or "irrelevant".

These three studies involve different techniques for topic labeling, but they share a common feature: the labels are binary and are not related to the content of the topics themselves. Instead, the labels reflect some other characteristic of the topics (such as sentiment or relevance).

A note on grey literature and MLRs

Grey literature refers to information that is produced and distributed outside of traditional commercial publishing channels, and is often produced by non-commercial entities such as government agencies, academic institutions, or industry organizations (Farace and TransAtlantic, 1997). This can include reports, working papers, conference proceedings, and other documents that are not formally published in peer-reviewed journals or other commercial publications.

In this context, a Multivocal Literature Review (MLR) is a type of systematic literature review that goes beyond traditional published academic sources to include a broader range of materials, including grey literature such as blog posts and white papers (Garousi et al., 2019). By incorporating a wider range of sources, MLRs might be able to provide a more comprehensive understanding of a particular topic. In essence, MLRs are tied to grey literature because they actively seek out and incorporate this type of information in order to provide a more nuanced and inclusive view of the literature on a particular topic.

In the context of this work, it has been decided **not to include grey literature** among the analysed work. Although such an inclusion might ultimately enhance the information carried over by the review process, the effort required seek out and individually asses the trustworthiness of work produced outside of traditional distribution channels has been deemed to be excessive with regards to the available time and resources.

3.6. Data Items

In the context of a literature review, **data items** refer to the specific pieces of information that researchers are looking for or recording from the sources they are reviewing. These may include key findings, research questions, study design, sample size, data analysis methods, and other relevant information related to the research topic. The definition of a set of data (collection) items allows to neatly summarise the individual pieces of information that are required to be extracted from each of the collected papers.

A **data extraction form**, on the other hand, is a tool used to systematically collect and organize the data items identified during the literature review process. Kitchenham (2004) states that: *"The data extraction forms must be designed to collect all the information needed to address the review questions and the study quality criteria"* and that *"The objective [...] is to design data extraction forms to accurately record the information researchers obtain from the primary studies"*. It typically includes a list of data items, together with a short description and the research question to which a given data item relates to. The purpose of a data extraction form is to facilitate the organization and analysis of data from multiple sources, making it easier to identify patterns, themes, and gaps in the literature. It also helps ensure consistency and accuracy in data collection, allowing researchers to more easily compare and synthesize findings across studies.

The data extraction form used in this study is presented in Table 3.3. The presented form is organised into 23 distinct data items. All data items, with the exception of 1-4 (which are used to store general descriptor for a given paper), are associated with one of the four research questions introduced in Section 1.3. The data items are grouped together (and sorted) on the basis of the research question they belong to. In addition to the item's number, name, description and research question, a checkmark symbol (✓) is used to indicate mandatory data items (i.e. those data items for which associated data will be present in the data extraction forms of all the selected studies) and a cross symbol (✗) highlights those data items for which data might be absent in some of the analysed papers.

Notice that data item 6 is used to define whether the paper under scrutiny possesses a primary or secondary focus with regards to the topic covered by this study (topic labeling). In this context, a paper is identified to have a primary focus on topic labeling if the mention of labeling activities is made in the paper's title, abstract or introductory section.

The number of data items associated with activities related to topic labeling (5-11 and 15-18) reflects the main focus of this review. On the other hand, only a relatively generic description of the topic modeling task(s) performed for a given piece of work is provided (items 12-14). This was a deliberate choice made in order to avoid shifting the focus of the presented work away from the chosen area of research and towards the more commonly covered task of topic modeling.

The last set of five data items (19-23) offers a broader view of the work under review. In this context, together with information related to the research domain (item 19) and to the corpus being examined (items 21-23), data item 20 introduces the general problem statement associated with the work at hand. This summary item is included to facilitate the contextualization of data collection entries when reviewing the data at the end of the data collection phase.

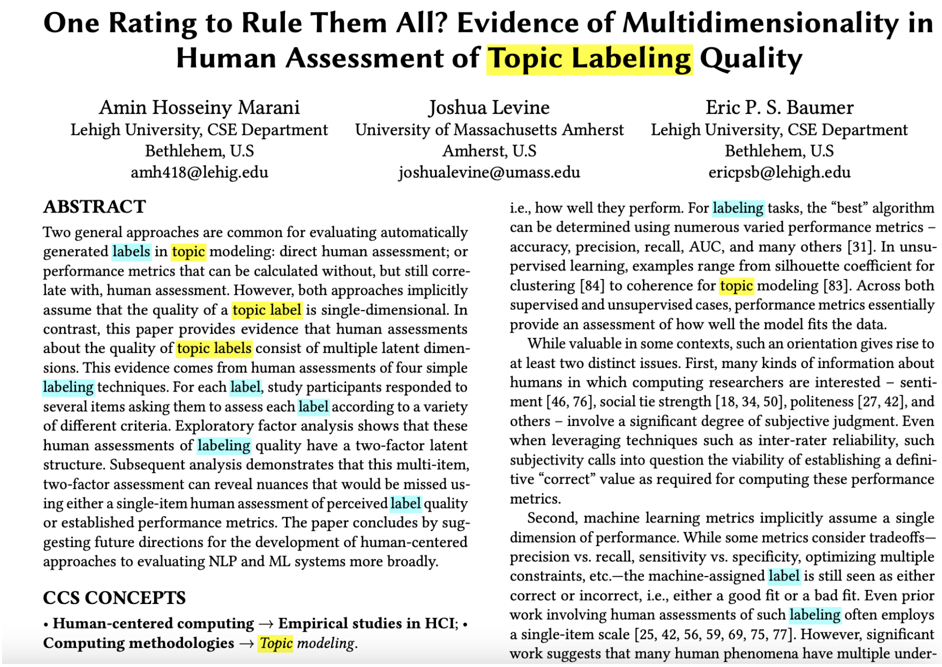
Item nr	Item	Description	RQ	Mandatory
1	Year	Publication year	-	✓
2	Author(s)	Publication author(s)	-	✓
3	Title	Publication title	-	✓
4	Venue	Publication venue	-	✓
5	Topic labeling	Topic labeling approach(es)	RQ1	✓
6	Focus	Primary / Secondary focus on topic labeling	RQ1	✓
7	Type of contribution	Established / Novel approach for topic labeling	RQ1	✓
8	Underlying technique	Technique / Algorithm on which the topic labeling approach is based	RQ1	✓
9	Topic labeling par.	Parameter used for topic labeling	RQ1	✗
10	Label generation	Label generation process	RQ1	✗
11	Motivation	Motivation for applying a labeling step	RQ1	✗
12	Topic modeling	Topic modeling approach(es)	RQ2	✓
13	Topic modeling par.	Parameter used for topic modeling	RQ2	✓
14	Nr. of topics	Nr of topics generated from the corpus	RQ2	✓
15	Label	Label description (e.g. single- multi-word)	RQ3	✓
16	Label selection	Nr of candidates per topic & approach(es) for selection	RQ3	✗
17	Label quality evaluation	Quality metric(s) for label evaluation	RQ3	✗
18	Assessors	Number and details of the assessors involved in the selection and evaluation	RQ3	✗
19	Domain	Domain(s) of interest	RQ4	✓
20	Problem statement	Summary of problem statement	RQ4	✓
21	Corpus	Origin, format, shape and content of the corpus	RQ4	✓
22	Document	Format of individual documents in the corpus	RQ4	✗
23	Pre-processing	Pre-processing steps performed on documents	RQ4	✓

Table 3.3: Data extraction form

3.7. Data Collection Process

In the context of a literature review, the data collection process refers to the methodological approach taken to extract relevant information from the collected studies. Kitchenham et al. (2022) states that the data collection process should: *"Specify the method used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and, if applicable, details of automation tools used in the process"*. In other words, if the data extraction form (described in Section 3.3) answers to the question of: "**What** pieces of information are extracted from a primary study?", then the description of the data collection process answers to the question of: "**How** are those pieces of information extracted?". In this context, defining a data collection process is important in describing the resources allocated for the analysis of the collected work, together with the tools that were used to support it. Additionally, clarifying the details of the employed collection procedure allows to more easily justify the limitations imposed during the search process with regards to the start & end dates (Subsection 3.1.1) and the venue selection process (Subsection 3.1.3).

For this systematic review, the data for the previously described data extraction forms is collected by a single reviewer working independently. The reviewer is (at the time of collecting the data) a 26-year-old Data Science Msc. student at the Free University of Bozen-Bolzano. In order to facilitate the data extraction process, a tool called PDF Search (conceptually similar to the previously described FoxTrot Professional Search) is used to search and highlight salient words in the collected work. This highlighting functionality (an example of which is provided in Figure 3.3) can help in quickly identifying relevant portions of a given piece of work when scrutinizing it. Conceptually, this allows the reviewer to skim over the content of a given publication and to fixate only on those paragraphs that appear as being strongly highlighted by the search tool. This approach helps to easily visualise which sections of a given publication are more likely to contain information that can be useful in the data extraction process and consequently lowers the average time required for the analysis of a single study. Naturally, more traditional techniques for analysing papers such as reading the title and full abstract, using the paragraph titles to guide the search process, etc. are used in conjunction with the chosen tool. The query `topic* label*` is used in the selected tool to highlight words belonging to the root terms: topic, label and topic label.



One Rating to Rule Them All? Evidence of Multidimensionality in Human Assessment of Topic Labeling Quality

Amin Hosseiny Marani
Lehigh University, CSE Department
Bethlehem, U.S
amh418@lehig.edu

Joshua Levine
University of Massachusetts Amherst
Amherst, U.S
joshualevine@umass.edu

Eric P. S. Baumer
Lehigh University, CSE Department
Bethlehem, U.S
ericpsb@lehigh.edu

ABSTRACT

Two general approaches are common for evaluating automatically generated labels in topic modeling: direct human assessment; or performance metrics that can be calculated without, but still correlate with, human assessment. However, both approaches implicitly assume that the quality of a topic label is single-dimensional. In contrast, this paper provides evidence that human assessments about the quality of topic labels consist of multiple latent dimensions. This evidence comes from human assessments of four simple labeling techniques. For each label, study participants responded to several items asking them to assess each label according to a variety of different criteria. Exploratory factor analysis shows that these human assessments of labeling quality have a two-factor latent structure. Subsequent analysis demonstrates that this multi-item, two-factor assessment can reveal nuances that would be missed using either a single-item human assessment of perceived label quality or established performance metrics. The paper concludes by suggesting future directions for the development of human-centered approaches to evaluating NLP and ML systems more broadly.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Topic modeling*.

i.e., how well they perform. For labeling tasks, the "best" algorithm can be determined using numerous varied performance metrics – accuracy, precision, recall, AUC, and many others [31]. In unsupervised learning, examples range from silhouette coefficient for clustering [84] to coherence for topic modeling [83]. Across both supervised and unsupervised cases, performance metrics essentially provide an assessment of how well the model fits the data.

While valuable in some contexts, such an orientation gives rise to at least two distinct issues. First, many kinds of information about humans in which computing researchers are interested – sentiment [46, 76], social tie strength [18, 34, 50], politeness [27, 42], and others – involve a significant degree of subjective judgment. Even when leveraging techniques such as inter-rater reliability, such subjectivity calls into question the viability of establishing a definitive "correct" value as required for computing these performance metrics.

Second, machine learning metrics implicitly assume a single dimension of performance. While some metrics consider tradeoffs—precision vs. recall, sensitivity vs. specificity, optimizing multiple constraints, etc.—the machine-assigned label is still seen as either correct or incorrect, i.e., either a good fit or a bad fit. Even prior work involving human assessments of such labeling often employs a single-item scale [25, 42, 56, 59, 69, 75, 77]. However, significant work suggests that many human phenomena have multiple under-

Figure 3.3: Example of text highlighting

3.8. Analysis and Synthesis Methods

Defining a clear methodology for the analysis and synthesis of the collected studies is an important step in enabling the review process to thoroughly synthesize the relevant results into a coherent narrative. Such methodology involves the integration of findings from the various studies into a broad overview of the topic under review. This involves aggregating the data and identifying similarities and differences across studies and enables researchers to identify relevant patterns, gaps, and inconsistencies in the collected literature, and should ultimately lead to new insights into the research topic. This synthesis can be descriptive (or non-quantitative) and can often be integrated by using statistical techniques in order to extrapolate quantitative findings from the primary studies. A review heavily centered around providing a quantitative research synthesis is called a meta-analysis.

Meta-analyses summarise the key findings by: *"encoding the magnitude and direction of each relevant statistical relationship in a collection of studies"* and allow for: *"an analytically precise examination of the relationships between study findings"* (Lipsey and Wilson, 2001). Independently from the type of review being conducted, and in order to facilitate the synthesis process, the analysis can be organised into distinct categories corresponding to the key insights the review is ultimately trying to provide. Kitchenham (2004) states that: *"Extracted information about the studies (i.e. intervention, population, context, sample sizes, outcomes, study quality) should be tabulated in a manner consistent with the review question [...] should be structured to highlight similarities and difference between study outcomes [and should] identify whether results from studies are consistent one with another (i.e. homogeneous) or inconsistent (e.g. heterogeneous)"*

Considering the fact that the main goal of a Systematic Literature Review is to answer the proposed Research Questions as thoroughly as possible with regards to the evidence contained in the collected work, it naturally follows that the synthesis tasks would be structured accordingly. Additionally, it might also be useful to extended the more trivial analysis conducted on the information extracted from the items associated with each question by adding some additional categories, or "points of view", each representing a different approach taken in the summarisation of the collected data. By doing so, the systematic review should be able to provide a more thorough overview of the research area under scrutiny. These different frames of reference, organised into three distinct points of view, are structured to provide a complete synthesis of the gathered insights in terms of: (1) What is found in the body of research (**the evidence**), (2) what is not found in the body of research (**the gaps**) and (3) what are the noteworthy techniques and approaches found in the individual studies (**the insights**). In other words, we can describe each element of this analysis structure as follows:

1. **Synthesis of the primary studies**: This category represent the more traditional approach to synthesising results in the context of a literature review. Here, data extracted from multiple studies is combined, compared and presented. General information, observations and insights relating to the given research question and the corresponding data items are provided. In other words, data from the collected research is presented **as a whole** and different papers are **compared** with regards to their methodology, content and findings. Given its very systematic nature, this particular category of results is presented by strictly adhering to the considered data items. In other words, a summary of the collected information is highlighted for each Research Question and related data items.
2. **Gaps in the research**: Since the first point deals with insights that are found in the collected research (i.e. What is there?), this second category deals with **gaps** in the literature (i.e. What is NOT there?). Here the objective is to try to identify those areas of the topic under review that are generally lackluster given the information contained in the collected work. Once again (and in a similar fashion to the first point), this step requires comparing the content of multiple studies. In this context, it is also important to note that the relevant gaps are recognised by analysing the insights provided in (1) but that they not necessarily tied to a single data item or even to an individual Research Question. This means that the more structured analysis deriving from (1) serves as a starting point in the gap identification process. More details on the type of analysis carried out in order correctly profile a gap in the collected literature is provided in Subsection 3.8.1.
3. **Insights from individual studies**: Often times, interesting (and more specific) insights can be extracted from an **analysis of the individual papers**. On top of the comparative analysis of present and missing evidence carried out by point (1) and (2), an Systematic Literature Review should also ideally summarise specific (and potentially novel) approaches for future readers that might want to get familiar with the most current state-of-the-art approaches and technologies. Therefore, this category deals with all the relevant information that do not pertain to an holistic analysis of the research. Additionally, insights are classified based on their level of reproducibility and presented accordingly. This means that an individual insight should not only be topically relevant, but also practically applicable by future researchers that might be interested in integrating it into their own research. Details on the classification performed for the relevant insights is presented in Subsection 3.8.2.

As previously mentioned, the general synthesis generated in (1) is conducted within the context of each of the presented Research Questions. Specifically, the individual data items found in the data extraction form presented in Table 3.3 and associated with such questions are used as the individual building blocks within the analysis process. Additionally, the information carried over by the items that are missing a Research Question association (Data items 1 to 4) are used to briefly summarise the shape of the analysed work.

3.8.1. Profiling a gap

3.8.2. Evaluating insights from individual studies

In a general sense, the goal of a Systematic Literature Overview needs not necessarily be limited to answering the presented Research Questions. In fact, and as stated in the introduction to this Section, the synthesis of primary studies can be integrated by an identification of potential gaps in the collected work and by a more detailed analysis of the potentially relevant insights identified in some of the primary studies. This last activity, whilst having the chance of providing a summary of the useful tools and methodologies that might be of interest to perspective readers, runs the risk of becoming a daunting task for the reviewer tasked with collecting and presenting an informative review of such techniques. Additionally, when preparing these summaries, one also needs to keep into consideration that not all insights have the capability of being equally applicable in a practical context. In this regard, two characteristics bound to the nature of the presented approach and to the characteristics of the study (and venue) in which the approach is found are considered for this classification. Such characteristics are: (1) the **level of accessibility of the primary study** and (2) the **reusability of the proposed approach**.

With regards to point (1), the main distinction associated with the degree of accessibility of the document is made with regards to the notion of **Open Access (OA)**. Traditionally, scientific literature is published within the boundaries of subscription-based journals, requiring readers (or, more commonly, the institutions in which they operate) to pay subscription fees to access the hosted literature. Unfortunately, the prices associated with such fees have been shown to be steadily increasing, often even outpacing inflation (Creaser and White, 2008). Naturally, this creates a situation where, depending on the economical availability of individual institutions, perspective readers might find very difficult (or even impossible) to gain access to specific publications. In order to partially account for said difficulty, the past decades have seen a rise in the modes of OA publishing, allowing to go beyond traditional *"toll-access"* publishing (Langham-Putrow et al., 2020). In this context, an initial distinction is made between documents belonging to subscription-based venues and OA ones. Clearly, insights extracted from OA sources will be prioritised in the efforts associated with their description, given their more easily accessible nature. Additionally, it is also important to specify that the specific by which OA is implemented gives the rise to distinct OA sub-categories (Neuhaus, 2019; Solomon, 2013), which will also be taken into consideration in this classification. The following list contains the two categories considered in the context of this review to classify documents associated with the identified insights on the basis of their access characteristics. Additionally, the specific OA sub-categories (which are also taken into account) are presented and briefly described:

- **Subscription-based access**: The document requires payment of a subscription fee to a specific venue or repository in order to be accessed.
- **Open Access**:
  - **Gold & Hybrid OA**: Work presented within gold journals, meaning venues where the content is made immediately available for free upon publication. Sometimes, this category is also referred to as Hybrid Open Access (Paid Open Access) if article-processing charges are imposed to the authors. In this context, only articles paid upfront are made to be OA. Since the difference between these two approaches cannot be perceived by the end reader, they are grouped together in the context of this classification.
  - **Green OA**: Refers to work that, upon completion of the peer review process associated with the submission to a subscription-based venue, is made available within a freely ac-

cessible repositories of choice. Depending on the venue's policies, the authors might be able to upload in the OA repositories only a specific version of the article (e.g. manuscript, pre-print, post-peer review, etc.).

In relation to point (2), one of the determining factors associated with the usefulness of a potential insight is dictated by its level of reusability for perspective readers. In fact, a relevant methodology or technique should not only be effective in its ultimate goal, but also properly documented and (even more importantly) reproducible to third parties. Therefore, together with the accessibility of the individual study, the reusability aspect of the insight and of the required associated **resources** (e.g. software repositories, pre-trained models, relevant software tools, etc,), is taken into consideration when describing insightful methodologies. In this context, *"The **FAIR** Guiding Principles for scientific data management and stewardship"* (Wilkinson et al., 2016) can be used as a starting point in further evaluating the identified insights. The principles described in this document represent a collection of best practices designed to improve the reuse capabilities of scholarly data on the basis of their level of: **F**indability, **A**ccessibility, **I**nteroperability and **R**eusability. Additionally, it must be noted that in recent years further efforts have been made to adapt the principles described in the original FAIR document (which is mainly designed around data, especially in the domain of life sciences) to information that is more technical in nature. In this context, the work of Lamprecht et al. (2020) (*"Towards FAIR principles for research software"*), which re-formulates the FAIR statements in the context of research software, is used in conjunction with the original FAIR document as guidance to formulate relevant questions posed to evaluate the reusability of the encountered insights and associated resources. The questions formulated from the considerations of Lamprecht et al. (2020) and used to evaluate the reusability of the encountered insights are presented, organised into the four FAIR categories, as follows:

- **Findabilty**
  - Are the relevant resources made available by the authors?
    - If that is the case, are they linked from the primary study or from the repository in which the study is hosted?
    - If that is not the case, is the resource unambiguously identifiable by querying specialised registries or dedicated repositories?
  - Is the provided metadata sufficient to uniquely tie the methodology described in the primary study with a specific version of the resource?
- **Accessibility**
  - Can the resource be freely accessed?
  - Is the hosted resource still retrievable (i.e. inspectable and downloadable)? If that is not the case, has the associated metadata being persisted?
  - Additionally, are there specific requirements in place for execution (High-performance machinery, paid registration, proprietary software etc.)?
  - Are specific data samples required for execution? If so, are such data samples made available with the resource?
- **Interoperability**
  - Are required dependencies (and associated versions), input and output data formats, communication interfaces and deployment options specified in the provided metadata?
  - Are there concerns associated with portability (e.g. specific OS requirements)?
- **Reusability**
  - Can the resource be used with data distinct from the test one? Can it be adapted to use cases distinct from the ones described in the primary study?

- To this end, is proper documentation provided?
- Is it conceptually possible to extend the provided resource? Is the provided usage license compatible with this possibility?

Naturally, it is reasonable to assume that some of the identified insights will be **non-technical in nature** and will not require external resources in order to be made usable to third parties. In this context, the interested insights will be evaluated for inclusion in the review only on the condition that the description provided in the associated primary study is sufficiently detailed to render the insight reusable to perspective readers.

In conclusion, describing insights extracted from the primary studies on the basis of these two metrics should conceptually help to: (1) Alleviate the workload posed on the reviewers by allowing to place a primary focus on the summarisation of only the most accessible and replicable techniques and (2) help the reader to easily find those approaches that are not only topically relevant, but also easily exploitable in their future research endeavors.

Chapter 4

Study Selection

Following the search strategy defined in Section 3.3, the selected repositories and related venues (Section 3.2) are searched. In this context, the initial query `"topic label*" OR ("topic model*" AND "label*")` (Section 3.3) is utilised in order to retrieve relevant studies in the time period 2017-2022 (Subsection 3.1.1). This initial gathering process resulted in the selection of **388** conference papers and **549** journal papers, for a final collection comprising of a total of **937** distinct publications. The papers were retrieved and stored locally on the 2nd of November 2022. As discussed in Subsection 3.3.1, a further filtering step is imposed on the original collection by means of a proximity constraint. In this regard, multiple threshold values for such constraints are tested (20, 10 and 5 terms) and the aggregated results are compared with the baseline retrieval (performed without proximity constraints). The result of this analysis are presented in Table 4.1 and 4.2 and show the effects of proximity constraints on the number of articles gathered from the various journals and conferences. Each table is structured into six columns, where the first column lists the venue names, and the remaining four show the number of studies retrieved in each venue under different proximity constraints. The column named "No constraint" lists the results obtained when no root term proximity restriction is imposed. Additionally, the same data is visually highlighted in Figure 4.1 (for jorunals) and in Figure 4.2 (for conferences).

	Proximity constraint			
Journal	**No constraint**	**20 terms**	**10 terms**	**5 terms**
Decision Support Systems	34	14	11	8
Expert Systems with Applications	206	84	71	59
Information Sciences	93	41	36	30
Journal of Infometrics	30	19	17	14
Knowledge-based Systems	140	54	44	38
Pattern recognition	47	11	8	7

Table 4.1: Effects of proximity constraints on journals

Ultimately, the results obtained from the laxer 20 term constraint are selected and used as a starting point in the selection process (the methodology of which is described in detail in Section 3.5). Additionally, a further justification on the chosen constraint value, together with a discussion on its impact on the final collection of primary studies, can be found in Section 4.1).

The remaining set of 424 papers obtained from the proposed query and filtered using the 20 term proximity operator is manually inspected and a final selection of **64 papers** is generated on the basis of the established inclusion and exclusion criteria. This set of 64 papers (**34** of which are retrieved from the selected journals and **30** from conferences) fundamentally represent the core selection of studies stemming from the presented review process. In other words, the content related to all of the publications belonging to this initial set (which are listed in Appendix A) is analysed and summarised as part of the literature review process. Additionally, such publications are used as the starting point

	Proximity constraint			
Conference	**No constraint**	**20 terms**	**10 terms**	**5 terms**
ACL	70	40	33	30
CIKM	44	21	20	18
COLING	32	19	17	16
EACL	16	8	8	5
ECIR	40	17	16	11
ECML PKDD	21	11	11	10
KDD	25	15	11	9
EMNLP	69	35	31	24
NAACL	41	24	24	18
SIGIR	30	11	10	10

Table 4.2: Effects of proximity constraints on conferences

for the secondary gathering process stemming from the described snowballing activities (the results of which are presented in Section 4.2).

A note on EMNLP 2022

As discussed in the previous section, the paper retrieval procedure is carried out at the end of 2022 in order to locally store an initial set of primary studies which are later filtered down and further analysed. After the conclusion of such retrieval process, it is noted that the 2022 instance of the Conference on Empirical Methods in Natural Language Processing (EMNLP) has not yet taken place. Given the fact that EMNLP appears among the selected conferences (Section 3.1.3) and that its content might potentially be relevant for the proposed work, it is decided to conduct a second retrieval process upon conclusion of EMNLP'22. This choice is further justified by the desire of including all selected conferences (up to their latest iteration at the time of writing) in the presented review. In this context, the interested conference took place in the days between the 7th and the 11th of December 2022. The search is conducted following the same modalities described in Section 3.3 and implemented for the other venues as described at the beginning of this Chapter and is ultimately conducted on the 20th of December 2022 (well after the conclusion of the conference and the publication of the related papers on the ACL repository). From this secondary retrieval task conducted for the 2022 instance of the EMNLP conference a set of 828 papers is collected. To this initial set of papers, the previously presented query with proximity operator is applied, resulting in a subset of 20 articles (29 before application of the proximity constraint). This subset of publication is manually inspected for detailed evaluation. In this regard, after the application of the agreed upon inclusion and exclusion criteria (Section 3.5), no paper is deemed relevant for the presented review and therefore no changes are made to the original set of selected publications.

Despite the fact that this additional research activity ultimately led to no perceivable changes to the content of the presented work, it was still deemed necessary to present this task as a side note related to the process of generating final study selection. In a general sense, this subsection underlines the fact that a single study retrieval activity might not always be sufficient to fully satisfy the requirements of a given review, and that the presence of upcoming conferences should be generally taken into account when inspecting the selected venues in order to avoid (as much as possible) the creation of reviews potentially missing some information related to the state-of-the-art with regards to the presented topic.

4.1. Impact of the proximity constraint

This Subsection offers a brief overview on the impact of proximity constraint on the selected papers. The choice of introducing a proximity constraint between the root terms `topic*` and `label*` on the original query results has been justified in Subsection 3.3.1 by the desire of filtering out those papers containing the root terms `topic model*` and `label*` that were unlikely to carry information that would be useful in the context of the proposed review. The reasoning behind this choice stems from the fact that if the root term `label*` appears in isolation (i.e. not in the vicinity of `topic*`) it is likely not being used to describe a topic labeling activity. Some examples of such cases are presented in the following list:

- Wu et al. (2019) proposes an approach to model opinion targets and their sentiment. In this context, the `label*` root term is often used to refer to the assignment of opinion targets to snippet of text:
  - *"It's labor-intensive to manually **label** opinion targets in each domain"*
  - *"... we propose an unsupervised method to identify opinion targets automatically, which solves the difficulty of **labeling** opinion targets manually in different domains"*
- Demszky et al. (2019) proposes an NLP framework to analyse political polarisation in social media. In this case, the root term `label*` can be seen used to describe the activity of attaching a political affiliation to a given user:
  - *"We begin by quantifying polarization [...] between the language of users **labeled** Democrats and Republicans"*
  - *"We **label** a user as a Democrat if they followed more Democratic than Republican politicians in November 2017"*

Adding a proximity constraint generally allows to filter out these unwanted instances by only flagging sentences (within documents) where the two terms are used in near conjunction with one another. Some examples of such sentences, all extracted from papers retrieved during the search process, are provided below:

- *"**Topics** can be more readily interpretable when they are assigned semantically meaningful **labels**"* (Hosseiny Marani et al., 2022)
- *"Various methods have been proposed to assign concise **labels** to **topics** to improve interpretability"* (Zosa et al., 2022)
- *"An interpretable **topic** is one that can be easily **labeled**. How easily a **topic** could be **labeled** ..."* (Doogan and Buntine, 2021)

The choice of using 20 terms as a (somewhat broader) constraint can be justified by the information found in Griffies et al. (2013) which states that: *"The average length of sentences in scientific writing is only about **12-17 words**"*. In this context, the chosen proximity constraint should generally be able to account for paper containing instances of the two root terms appearing in the same sentence.

4.1.1. Lowering the threshold

Another potentially useful insight can be gathered by observing the influence of stricter proximity constraints over the set of selected papers. In this context, it might be interesting to take a closer look at those studies that would be filtered out by lowering the maximum distance between key terms in the proximity constraint. Changing such constraint from the initial 20 terms to a lower distance of 5 terms filters out two of the selected studies (Huang et al., 2020; Yang et al., 2017). Further lowering this threshold to 3 terms eliminates another two publications (Kim et al., 2020; Mukherjee et al., 2020). In other words, by moving the proximity constraint down to a maximum distance of three terms (from the original value of 20) it is possible to observe how a total of 4 out of 64 papers (6% of the full selection) are excluded. By taking a closer look at the four interested papers, the following observations can be drawn:

- Huang et al. (2020) proposes a method to perform seed-guided topical taxonomy construction. Within the constructed taxonomy, each node (topic) is represented by a concept name and a cluster of coherent terms. Unlike the topically similar paper by Zhang et al. (2018a) (also published in KDD), this paper does not directly refer to concept names as topic labels. The sentence allowing for the retrieval of this paper is the following: *"Then we use majority votes to **label** the pairs and use all the true parent-children pairs from different methods to construct a gold standard taxonomy. Since each **topic** is represented by a cluster of words, ..."*.

- Yang et al. (2017) proposes the use of tree priors (using tree LDA) to improve interpretability of topics. In this context, concepts appearing in multi levelled tree priors can be used to assign topic names. Even though the paper does not refer to single word topic identifiers as "labels", the paper is retrieved due to the following sentence: *"All the results are averaged across five-fold cross-validation using 20 **topics** with hyper-parameters $\alpha = \beta = 0.01$. For 20NewsGroups classification, a post's newsgroup is its **label**"*.
- Kim et al. (2020) proposes an LSA model (based on Word2vec and Spherical k-means clustering) to perform a trend analysis on blockchain research. In this context: *"the name of the cluster is defined by considering the characteristics of the words assigned to the cluster, and it is considered as a topic"*. Notice that in this paper the word cluster (or topic) name is not referred to as a label. Instead, the word "topic" is used to refer to both the identifier and the terms contained in the cluster. The paper is retrieved due to the following sentence: *"For measuring the accuracy of allocated **topics** to the documents, existing studies have used the data for text classification, which was already categorized or assigned to the **topic**. Since there are no exact **labels** for our data, we propose a quantitative evaluation method"*.
- In Mukherjee et al. (2020), the manual labeling step is captured by the laxer proximity constraint thanks to the following sentence: *"each extracted **topic** is manually interpreted by looking at its representative words and assigned a genuine aspect **label**"*.

With the exception of one publication, a common characteristic of the papers that would be excluded by a stricter proximity constraint resides in the fact that they do not refer to topic identifier as "labels". In fact, the terms "concept name", "concept" or simply "topic" are used instead. Despite this fact, the highlighted work is kept in the final selection due to its relevance with regards to the proposed survey.

4.2. Snowballing Results

The results related to the two distinct snowballing activities carried out in the context of this systematic review, the methodologies of which are presented in Section 3.4 (Subsection 3.4.1 and 3.4.2 for backward and forward snowballing respectively), are presented in this Section. As mentioned previously, the proposed snowballing phases are put into place to broaden the scope of the review process which, up until this point, has been fairly limited by the strict constraints imposed on the initial selection of venues (Subsection 3.1.3 and related information sources (Section 3.2). In fact, although the process of narrowingly focusing the initial retrieval efforts to very specific subset of potential avenues is required in order to allow for a more manageable (in terms of time and resources) and (maybe more importantly) reproducible review, it is undeniable how such a choice might have substantial impacts on the completeness of the information contained in the final work. Therefore, the hope is that the activities described in this Section may help in identifying relevant studies belonging to previously unexplored information channels, allowing for a more complete overview of the topic at hand. In this context, a short summary highlighting the shape of the collected set of additional studies is presented for both snowballing activities, together with specific mentions of papers belonging to the original selection that were re-retrieved as part of this process. Additionally, the previously unexamined venues to which the newly collected publications belong to are also presented, together with a short mention of their respective domains. Finally, the Subsection related to Forward Snowballing is integrated with a detailed explanation of the quality-based venue selection process (the introduction of which is justified in Subsection 3.4.2) conducted in order to further reduce the number of collected papers based on the prestige of the related source venues. In this Subsection, it is also possible to find a brief note related to a subset of papers that were not originally retrievable by means of the tools utilised for the other studies, together with an explanation of how this particular issue was addressed and ultimately solved.

4.2.1. Backward Snowballing

As mentioned in Subsection 3.4.1, the backward snowballing process is conducted as a first snowballing activity starting from the full set of references belonging to the initial selection of primary studies identified in the context of this review and extracted in a semi-automated way using the Academic Graph API (the technical details of which are presented in 3.4.3). In this regard, the same time-related constraints that are used for the core selection (i.e. work released from 2017 onward) is applied as a first filtering mechanism. On the other hand, no limitation on the origin venue (journal / conference) of the extracted work is applied. Using this selection criteria, an initial corpus of **738** items is obtained. Based on the assumption that the initial set of papers is able to provide a sufficient guarantee of quality for the work collected at this stage, no filtering step is applied with regards to the venue to which the extracted work belongs to. On this broad selection, the previously described query containing the 20 terms proximity operator (`"topic label*" OR ("topic model*" AND ("label*" NEAR "topic*"))`) is applied using the FoxTrot Professional Search tool. The corpus resulting from this filtering step returns a total of **160** items (231 if the proximity constraint is removed). In the same way as the initial study selection process (Chapter 4), all papers contained in the corpus resulting from the initial reference extraction (and filtering) phase are individually inspected and evaluated against the established selection criteria. From the initial set of 160 papers, a final selection of **28 studies** is found to meet the imposed requirements and is therefore added to the systematic review. Of this 28 publications, 21 originate from journals and 7 from conferences. With the exception of one paper, only one relevant study is found for each newly encountered venue. In other words, the 28 papers are found to belong to a set of 27 different journals and conferences. The full set of studies selected at this stage (and their related venues) can be found in Appendix B.

This new venues can be grouped into the following domains:

- **Computer Science**: International Journal of Advanced Computer Science and Applications, International Conference on Informatics and Computational Sciences (ICICoS)
- **Information Systems**: Brazilian Symposium on Information Systems (SBSI), Journal of Information Processing Systems, Social Network Analysis and Mining
- **Natural Language Processing**: Natural Language Processing and Information Systems, Transactions of the Association for Computational Linguistics
- **Artificial Intelligence & Machine Learning**: Applied Artificial Intelligence, The Journal of Machine Learning Research, International Conference on Humanized Computing and Communication (HCC)
- **Information Science**: Journal of the Association for Information Science and Technology, Journal of Information Science, International Journal of Information Management
- **Economics**: Ecological Economics, Journal of Comparative Economics
- **Political Science**: Political Science Research and Methods
- **Social Science**: Social Forces, Media and Communication, Research & Politics
- **Database Systems**: International Conference on Database Systems for Advanced Applications (DASFAA)
- **Bioinformatics**: ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB)
- **Data Science**: Data Science and Advanced Analytics (DSAA)
- **Marketing**: Journal of Advertising
- **Management**: Management Science
- **Communication**: Communication Methods and Measures
- **Agricultural sciences**: Small-scale Forestry
- **Transportation**: Transportation Research Part C
- **Tourism & Hospitality management**: Tourism Management

During the presently described snowballing activity, a total of **15** papers belonging to the initial selection are also re-retrieved. This fact indicates that there exists some level of interconnection between the 60 studies selected during the first phase of the presented work. The nature of this links among previously gathered publications is further explored as part of the literature network analysis offered in Chapter 5.

4.2.2. Forward snowballing

After having obtained the relevant (outgoing) references from the selected set of papers following the backward snowballing activity described in the previous section, the similar process of forward snowballing is carried out in order to capture the relevant (incoming) citations stemming from the initial selection. Executing forward snowballing on the set of 64 initially selected papers results in a total of **1147** extracted citations. At this point, it can be noticed how this resulting set is substantially larger than the one initially obtained during backward snowballing (where a total of 738 studies where retrieved using the Semnatic Scholar API). Once again, the query (and proximity operator) are applied to the extracted citations. This step generates a set of **358** studies (590 if the proximity operator is not applied) which need to be manually evaluated with regards to the inclusion and exclusion criteria used in the previous activities of the proposed review. Similarly to what is seen for backward snowballing in Subsection 4.2.1, a total of **18** papers that were already part of the initial selection of publications are retrieved once again following the forward snowballing procedure. Additionally, one (1) paper (Béchara et al., 2021) is retrieved and immediately discarded due to the fact that it had already been selected in the context of the backward snowballing phase.

As previously mentioned, the full set of retrieved papers is manually inspected and the chosen inclusion and exclusion criteria are applied. From this analysis, a collection of **58 studies** is ultimately selected and prepared for further analysis. As anticipated in the introduction to this Section, such analysis (and subsequent filtering step) is based on the quality of the respective source venues as established by the metrics introduced in Subsection 3.1.3 and is presented following a short note on non-retrievable papers.

A note on non-retrievable papers

Throughout the presented work, papers have been retrieved and locally stored in .pdf format from their respective repositories. This is made possible by the access provided by the Free University of Bozen-Bolzano which has proved itself to be sufficient for the collection of studies identified for the initial selection and for the first snowballing task. Despite this fact, in the context of forward snowballing a set of 13 papers (out of the total 1147) is found to fall outside of the coverage offered by the university's access. Additionally, the interested publications are not found to be freely accessible by any other means (for instance, they are not available for download on the arXiv repostiory or in any of the respective author's websites). This issue can be attributed to the large amount of studies that are collected in the context of this work and, in a more general sense, can be expected to present itself in most systematic literature review having a sufficiently extensive scope. Additionally, it must be pointed out that the arisal of this issue is also highly conditional on the resources available to the reviewers. In this regard, a review conducted by a single author (like in the case of the presented work) or by multiple authors belonging to the same institution (and that are therefore generally expected to have access to the same resources) can be considered more likely to encounter this issue than a review work stemming from the collaborations of individuals having a more varied background. Independently from it's cause, not having access to a subset of publications that the imposed search strategy has flagged to be relevant can be potentially threatening to the validity of the review process, which might ultimately be impacted by the lack of information provided by the missing papers. In fact, even though the interested studies represent only a very small percentage of the total collection, it is impossible to establish a-priori the usefulness of the data they contain. In other words, failing to find alternative way to collect this publications might ultimately lead to lackluster information on the state-of-the-art on topic labeling. Because of this reason, the inter-library loan and document delivery service offered the UniBZ's library is utilised in order attempt fetching the relevant documents from other institutions that might have access to the interested venues. Using this service, all of the thirteen missing publications are successfully collected. Due to the university specific limitations governing the provision of external documents, all the interested studies are collected in paper format. In this context, the documents are scanned and converted into .pdf format.

Additionally, the OCR functionality of the PDF Search tool is used to maintain the same highlighting functionality utilised throughout the review process and visually presented in Figure 3.3. At this point, the 13 papers are manually inspected and a total of three publications is found to be relevant and added to the collection. Notice that the 61 studies mentioned at the beginning of this section already takes into account the research gathered using the inter-library document delivery service.

Venue-based quality evaluation

When carrying out the backward snowballing activities, a subsequent check on the quality of the collected papers was not deemed necessary due to the fact that, in a general sense, it was assumed that the quality constraints imposed on the work from which the references were extracted (i.e. the core selection of studies) would carry over to the newly collected set. In other words, the expectation was that papers identified to be of a relatively high quality (based on the evaluation metrics associated with their respective source venues) would tend to consider only other studies roughly meeting the same quality threshold for inclusion among the respective references. In this context, the same reasoning cannot be applied to the collection stemming from the forward snowballing phase. In fact, when dealing with citations, no assumptions can be made about the studies collected at this stage based on the work they are obtained from. Additionally, knowing that the original selection of publications is extracted from a set of venues respecting certain quality characteristics makes it generally more likely for the papers contained withing them to be found, and consequently, used as reference material. Because of this reason, upon completion of the F.S. phase it is decided to carry out a **quality-based analysis** similar to the one proposed for the initial selection in Subsection 3.1.3. In this regard, the interested set of 61 publications is analysed and the respective venues are extracted.

For journals, the respective **SJR Scores** (averaged from 2017 to 2021) are extracted and the venues are sorted accordingly. The sorted list is divided into (quality-based) quartiles. At this point, it is important to remember that: (1) The number of papers associated with each venue is known and that; (2) All 61 papers have already been inspected with regards to the established inclusion and exclusion criteria. Given this premise, the **first two quartiles** (Q1 and Q2) corresponding to the top 50% of the extracted conferences are selected. This selection encapsulates all venues having at least a value of 1 related to the average SJR score and allows for the inclusion of **25** studies found in 19 distinct journals. In this regard, 3 journals (Scientometrics, Cognitive Computation and International Journal of Hospitality Management) are found to contain more than one relevant study (5, 2 and 2 papers respectively). At this point, an additional journal (**IEEE Access**) falling outside of the selected quartiles (and found in Q3) is also selected due to the topical relevance of the **2** publications it contains (Truică and Apostol, 2021; He et al., 2019). Such relevance had previously been established when manually inspecting the publications during the selection process. The set of 37 journals found in the four SJR-based quartiles is visually highlighted in Figure 4.3. In this context, notice that: (1) The selected venues are highlighted in bold and that; (2) The four journals missing an SJR evaluation (Journal of Computational Social Science, Social Sciences & Humanities Open, SSRN Electronic Journal, COLLNET Journal of Scientometrics and Information Management) and for which papers where retrieved during this phase are not ultimately included in the visualisation.

When dealing with conferences, a similar approach based on the CORE and GGS ratings is taken in order to evaluate the 10 gathered venues. Unfortunately, such an evaluation reveals that most of the conferences appearing among the papers collected during forward snowballing are listed as either being “Unranked” or as “Work in Progress” in the CORE / GGS database. Nonetheless, the resulting ranking for conferences is presented in Table 4.3. In this context, notice that the Australasian (B) rating found for the ALTA conference is defined in the CORE website as representing: *"... a conference for which the audience is primarily Australians and New Zealanders (these may be Australasian B or Australasian C)"*. Additionally, notice that ICIS has been positioned higher in the ranking than the ALTA conference because it also presents a formal evaluation provided in the form of a GGS rating, which is instead missing for ALTA. Following the collected evaluations, it is decided to keep in the presented review all papers linked to conferences that present **at least one rating** among the two considered metrics. This choice allows for the selection of a total of **4** conferences together with an equal number of related studies.

Conference	CORE rating	GGS rating
AACL-IJCNLP	B	A-
IJCNN	B	A-
ICIS	C	A-
ALTA	Australasian B	Work in Progress
SAICSIT	UNRANKED	Work in Progress
ITHET	UNRANKED	Work in Progress
MLN	UNRANKED	UNRANKED
SDP	UNRANKED	UNRANKED
ASEW	UNRANKED	UNRANKED
HNICEM	UNRANKED	UNRANKED

Table 4.3: Set of (10) FS conferences. Selected conferences in bold.

In summary, starting from an initial set of 61 publications collected in the context of the forward snowballing task, a venue-based quality evaluation activity is performed in order to find a suitable subset of studies to be included in the systematic review. In this regard, the final selection contains:

1. 19 Journals appearing in the first two quartiles (Q1 and Q2) based on the respective SJR scores averaged in the period 2017-2021 (i.e. all journals with a score above 1).
2. 1 Journal added to the selection due to the topical relevance of the (two) papers it contained.
3. 4 Conferences having at least one CORE or GGS rating.

This selection process leads to a new set containing a total of **31 publications** which are listed in Appendix C. Additionally, the following list contains the previously unexplored venues to which the work selected by the forward snowballing task belongs to, together with their respective domains:

- **Computer Science**: Scientometrics, IEEE Access, Computers & Education, Neurocomputing, Cognitive Computation, European Journal of Operational Research, Journal of the Association for Information Science and Technology, Journal of Big Data
- **Tourism & Hospitality management**: Tourism Management, Annals of Tourism Research, Journal of Travel & Tourism Marketing, International Journal of Hospitality Management, International Journal of Contemporary Hospitality Management
- **Management**: Journal of Retailing and Consumer Services, Journal of Manufacturing Technology Management, Electronic Commerce Research and Applications
- **Natural Language Processing**: Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP), Annual Workshop of the Australasian Language Technology Association (ALTA)
- **Political Science**: Government Information Quarterly
- **Information Systems**: International Conference on Information Systems (ICIS)
- **Artificial Intelligence & Machine Learning**: International Joint Conference on Neural Networks (IJCNN)
- **Health informatics**: Journal of Medical Internet Research
- **Food Science**: Food Quality and Preference
- **Transportation**: Transportation Research Part D

4.3. A note on predatory venues

Conceptually, the literary content of any systematic literature review is highly dependent on the chosen venue selection process and on similar filtering steps applied throughout the various snowballing phases. In this regard, it must be noted that the rigorous selection process conducted in Section 3.1.3 is not necessarily common practice in most literature reviews, where often times the inclusion of specific venues tends to be influenced primarily by their topical relevance to the subject under review rather than being solely based on established evaluation metrics. For example, whilst still taking into account openly available venue evaluation metrics, Silva et al. (2021) primarily selects the relevant information sources based on their topical affinity to the software engineering domain. Clearly, the decision of not limiting the review process to a specific domain (like in the case of the presented work) forces the reviewers to pose a more central focus on the aforementioned evaluation metrics during the venue selection procedure. On the other hand, the notion that some reviewers might purposefully decide not to impose strict constraints when making a decision on the inclusion of specific journals and conferences in their work and the idea of extending the search process beyond traditional publication channels stands at the very basis of the creation of Multivocal Literature Reviews, where (as mentioned in Section 3.5) extending the scope of the queried resources can ultimately lead to a more complete overview on the state-of-the-art. A the same time, this laxer approach might open up the review process to potential threats related to the validity of the collected work. In this context, one of such threats comes in the form of **predatory venues**.

Such notion, for which the term "predatory" was first introduced in Beall (2010), is generally used to identify all those venues aiming to: *"deceive authors to publish for a fee without providing robust peer-review or editorial services, thereby putting profit over trustworthy and dependable science"* (Elmore and Weston, 2020). Given this general definition, and as stated in Byard (2016), the clear consequence stemming from the existance of such venues is: *"that nowadays, for a price, anyone can have almost anything published"*. In this context: *"even the most bizarre theories with inadequate or no scientific validation could be published. [...] These papers would appear to be no different to those published in legitimate journals, and without a clear knowledge of a particular journal's reputation and process, may be difficult to exclude"*. Furthermore: *"Predatory journals may be used [...] to legitimize fringe theories and to validate bogus experts"*.

In the context of this work, a central focus has been placed on assessing the prestige of the retrieved venues in order to use it as a proxy in the establishment of the quality of the collected work. Because of this reason, the risk of inclusion of predatory venues during the study collection phase is drastically reduced. In fact, one can immagine that it would be highly unlikely to find untrustworthy venues having high ratings with regards to the chosen evaluation metrics. Clearly, the notable exception to this fact is represented by papers collected during the backward snowballing procedure, where no particular venue-based quality filtering step was applied with regards to the collected papers. Despite this fact (and as previously noted), the quality of such subset of studies should be somewhat guaranteed by the prestige of the publications from which they are retrieved (and where they were originally found as references).

Nonetheless, for the presented review it was still decided to formally include a check on predatory venues among the collected work. In this context, potentially predatory journals are checked using the Beall's List repository. The repository is organised into two separate list: The original list redacted by library scientist Jeffrey Beall (last updated on January 9, 2017 at the time of writing) and an newer list including an additional set of potentially predatory venues (last updated December 8, 2021 at the time of writing). The two lists are manually inspected and as a result **none of the journals** related to the selected work is found to be potentially predatory. Therefore, no work is removed from the presented review.

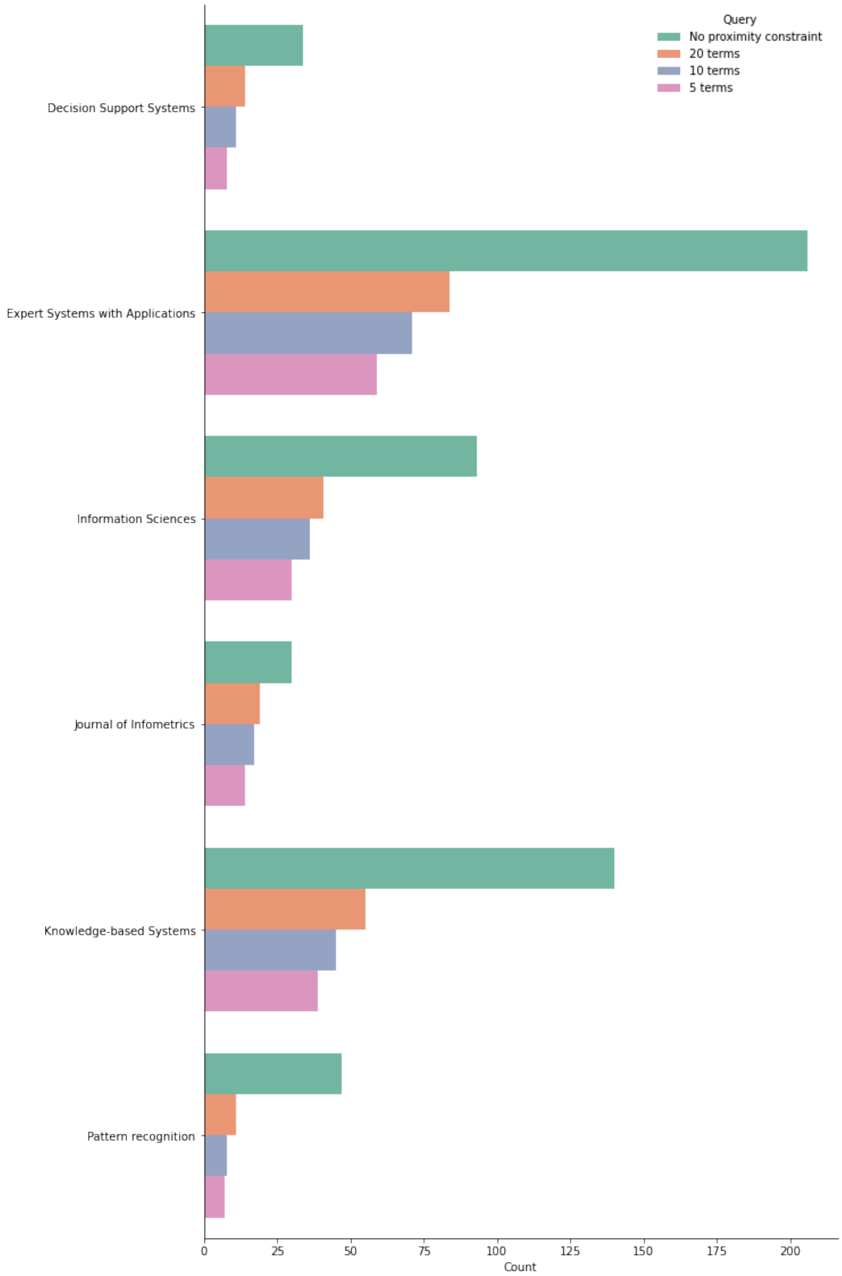


Figure 4.1: Effects of proximity constraints on journals

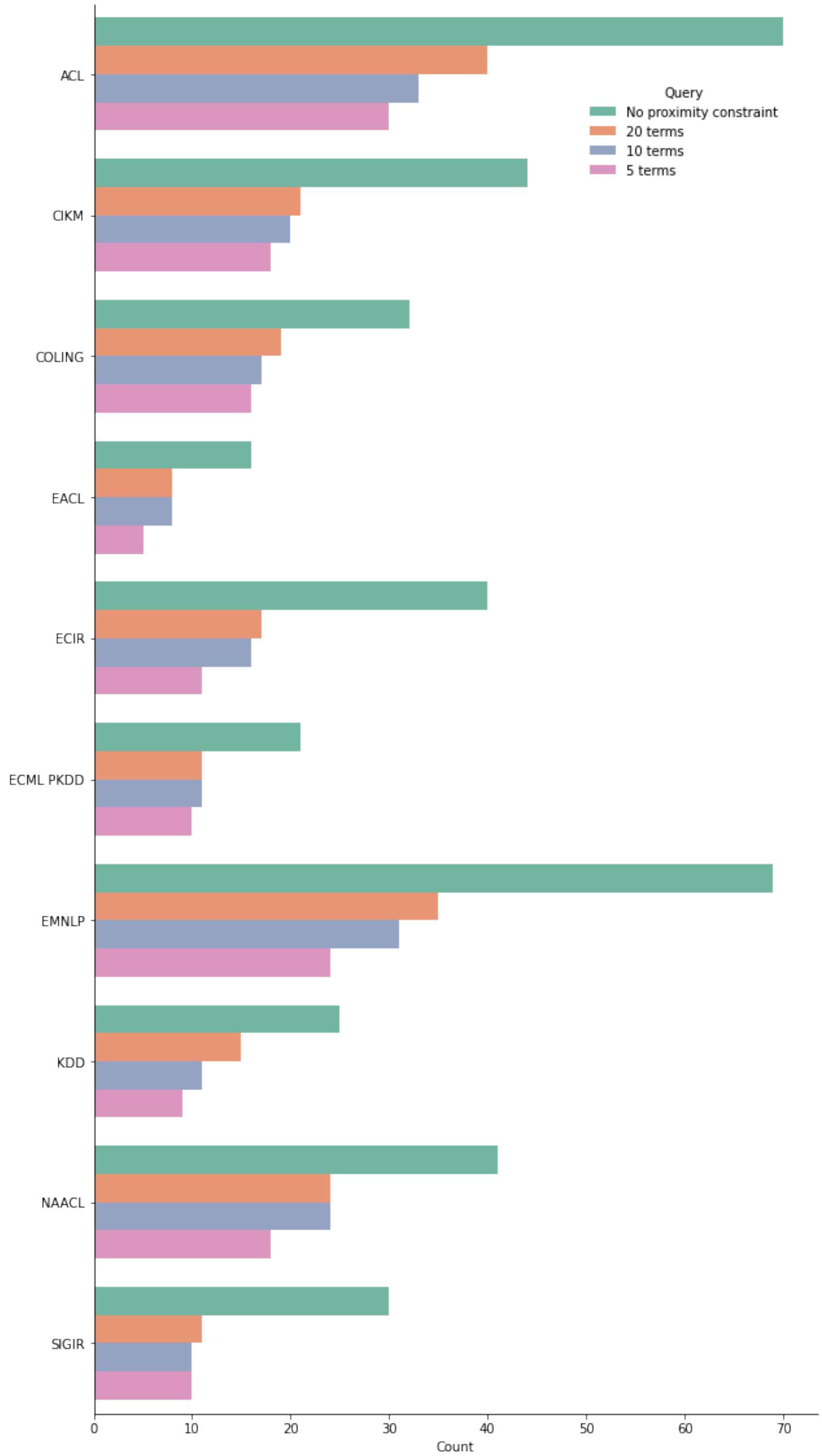


Figure 4.2: Effects of proximity constraints on conferences

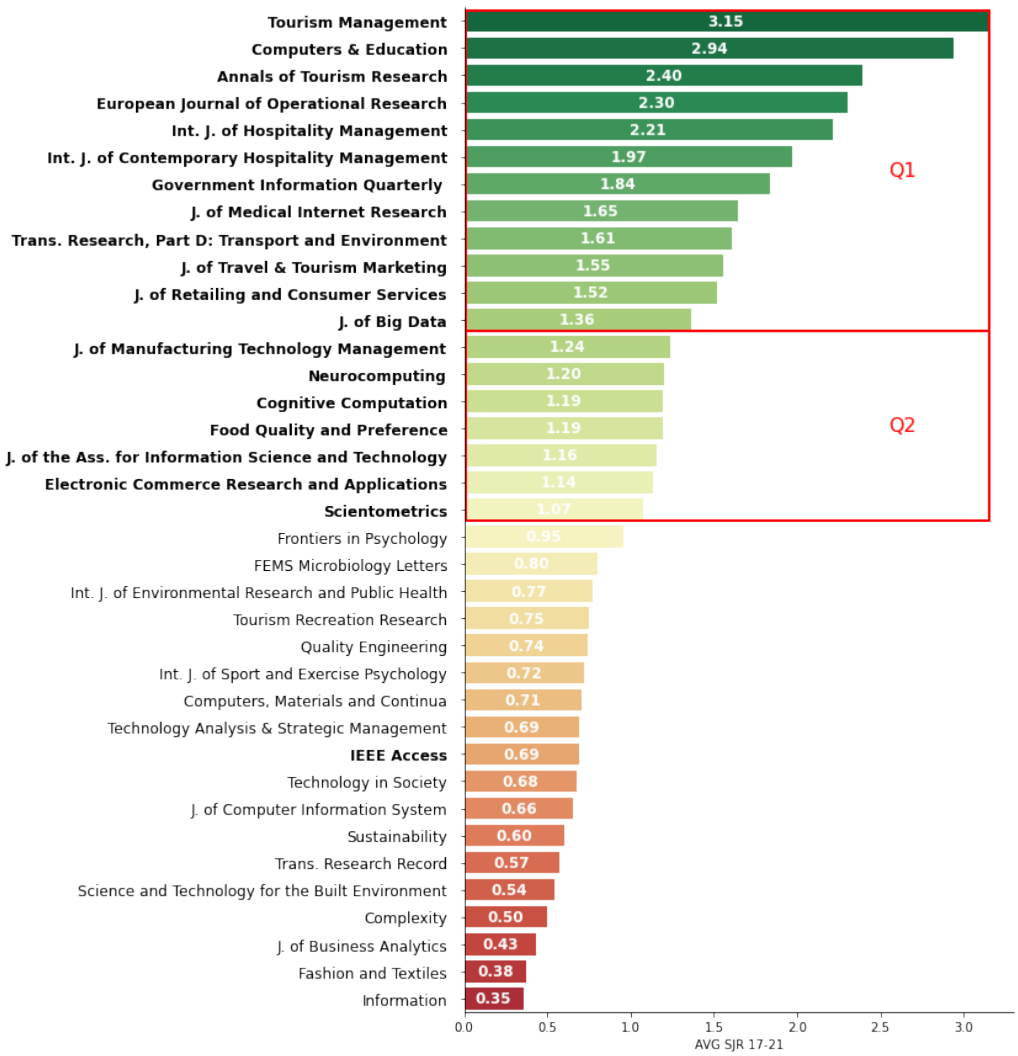


Figure 4.3: Set of (38) FS journals. Unranked journals are excluded.

Chapter 5

Literature network analysis

5.1. Methodology

The graphical representations of the literature network are generated, in an automated manner, starting from the *.bib* files built following the study collection and selection process performed for the presented review. In this context, the initial paper selection described at the beginning of Chapter 4 results in the creation of two files named `Selection_Conferences_2017-2022.bib` and `Selection_Journals_2017-2022.bib` and containing the *BibTeX* entries related to the initial selection of relevant studies identified in the selected conferences and journals respectively and used as a starting point in the execution of the two snowballing phases. Within these two files, the key of each entry (i.e. study) is represented in a standardised format in order to improve readability of the collected data and to ultimately ensure a consistent and clear representation of the individual nodes appearing within the generated networks (were such citation keys are used as labels in the representation of the individual nodes).

In this context, the key format `SURNAME_YEAR_TITLE` is used to express the citation key for a given paper. In case of multiple authors, only the surname of the first author is used in the related key. Additionally, punctuation and special characters are removed from titles and words are concatenated using the underscore symbol. Additionally, all words that take part in a given citation key are lower cased.

During the execution of the backward and forward snowballing activities described in Section 4.2, two additional *.bib* files (one for each snowballing task) are generated for each publication found in the initial paper selection. Each one of these files is named following the citation key of the study from which it originates and contains all the gathered citations (forward snowballing) or references (backward snowballing) that survived the search strategy imposed by the query described in Section 3.3. Further details on this file structure are presented in Section 5.2.

Conceptually, if one of the initially selected studies represents a single node in the generated networks, papers drawn as a result of the backward and forward snowballing activities allow for the creation of the respective outgoing and incoming edges. In other words (and in a more general sense), the out-degree of a node appearing in the presented networks is determined by the number of corresponding (relevant) references. Similarly, the in-degree of a node represents the number of (relevant) citations. Naturally, what can be considered as a "relevant" reference or citation strictly depends on the literature network at hand. For instance, Section 5.3 describes a paper graph that only contains nodes representing those studies that were ultimately selected (and therefore included) in the literature review. On the other hand, Section 5.4 seeks to provide a more holistic view of the gathered work, were the full collection of studies stemming from the two snowballing stages are presented in single extensive graph.

At this point it must be noted that whilst indispensable, the information inferable from the generated *.bib* files is not (on its own) sufficient to draw a complete picture of the analysed literature. In fact, the graphs that would be generated by representing nodes and edges using solely the collected

BibTeX entries would simply be a visual summary of the already described study collection results and would lack any information on the additional interdependence that might exists between the presented research (e.g. Are papers from the initial selection connected to one another? Do some of the papers gathered from the snowballing activities reference other work that is part of the network?). Therefore, another important step in providing a complete representation of the relevant literature networks is the identification of these additional edges.

A general overview of the techniques and tools employed to exploit the characteristics of the generated files in order to create the required literature networks is presented in the following subsection.

5.2. Tools

In this Section, the relevant libraries and an high level description of the logic used to generate the literature networks is presented. As previously mentioned, in order to create the proposed visual representation for the collected studies, the available *.bib* files are used as the initial data sources. In this context, the relevant files are organised on the local storage into two folders (`BS` and `FS`), each containing the data associated with the respective snowballing phase. Each folder is organised into two sub-directories (`Conferences` and `Journals`) referencing the different kinds of explored venue types. Finally, each venue folder contains a *.bib* file named using the key of the paper from which it originates.

For example, the two *.bib* files associated with the paper *"BART-TL: Weakly-Supervised Topic Label Generation "* (Popa and Rebedea, 2021) (originally published in the EACL conference) are found under `BS/Conferences/EACL` and `FS/Conferences/EACL`. In both cases the files are named `popa_2021_bart_tl_weakly_supervised_topic_label_generation.bib`. In this regard, the *.bib* file found in the root directory `BS` contains all the relevant references extracted from Popa and Rebedea (2021). On the other hand, the one found under `FS` contains the gathered citations.

A partial view of the described file structure is visually highlighted as follows:

```
Snowballing
├─BS
│ ├─Conferences
│ │ ├─ACL
│ │ │ ├─maiti_2019_spatial_aggregation_....bib
│ │ │ └─...
│ │ ├─CIKM
│ │ └─...
│ └─Journals
│   ├─Decision Support Systems
│   │ ├─goldberg_2022_sourcing_product_innovation_....bib
│   │ └─...
│   ├─Expert Systems with Applications
│   └─...
└─FS
  └─...
```

Starting from the presented file structure, the Python programming language is used to prepare a series of methods in order to automate the graph generation process. As a starting point, the functionalities offered by the NetworkX Library are used to easily generate the network's structure. The Python NetworkX library is an open-source package for the creation, manipulation, and study of complex networks or graphs. It provides a flexible and user-friendly interface to work with graphs and allows easy integration with other Python scientific libraries. It has extensive tools and functionalities for network analysis, visualization, and modeling.

To build the initial NetworkX `DiGraph` (i.e. directed graph) object, the `build_graph()` method is presented. The method accepts the `path` of the root folder containing the relevant *.bib* files (`BS` or `FS` folders), a `selection` list containing the keys of the selected studies and a boolean variable `selection_only`, which indicates whether only the selected studies should be included in the generated graph. After having initialised the `DiGraph` object (here called `G`), the method iterates over each `venue_type` folder (Conferences and Journals), each `venue` folder and ultimately over each `paper` (i.e. *.bib* file) and calls the `add_node()` method to add the current `paper` to `G`. Then, it parses the BibTeX entries found within each file (representing either citations or references associated with a paper from the initial selection) and, for each entry: (1) builds a valid citation key (`cite_key`) by passing the author, year and title information of the current entry to the `build_cite_key()` method; (2) Adds the node corresponding to the current entry using the `add_node()` method and; (3) Generates the corresponding directed edge (`add_edge()` method) going from `node` to `cite_key` (in the case of backward snowballing) or vice versa (in the case of forward snowballing).

Here, it is important to notice that steps described in (2) and (3) are always executed if the `selection_only` parameter is set to `False` (i.e. the full graph is being generated). On the other hand, if `selection_only` is set to `True`, the generated key needs to be part of the `selection` list in order for the node (and corresponding edge) to be added to the graph.

Additionally, notice that the `build_cite_key()` method simply executes on the provided parameters the processing steps required to generate a key that conforms to the format described in Section 5.1 (i.e. extracting the surname of the first author, connecting the three parameters using underscores and lower casing the final string).

```
def build_graph(path, selection = [], selection_only = True):
 G = nx.DiGraph() # Initialise the directed graph

  for venue_type in path:
   for venue in venue_type: # Each venue is a folder
    for paper in venue: # Each paper is a .bib file
     G.add_node(
      paper,
      width=2, height=2,
      # Additional style info...
     )
     conn_papers = biblib.parse(paper).get_entries() # Get cit. or ref.

     for conn_paper in conn_papers:
      cite_key = build_cite_key(ent['author'], ent['year'], ent['title'])
      if(selection_only == False or cite_key in selection):
       G.add_node(
        cite_key,
        width=2, height=2,
        # Additional style info...
       )
       if(path.contains('BS'):
        G.add_edge(node, cite_key)
       else:
        G.add_edge(cite_key, node)

 return resize_nodes(G)
```

In order to visually account for the "importance" of each paper with regards to the citation it receives, the `resize_nodes()` method is proposed. This method simply adjusts the `width` and `height` values of each node in the network based on the number of incoming edges. In this regard, the size increase S_{inc} of each interested node N with in-degree N_{in} is computed as follows:

$$S_{inc} = log_e(N_{in}) + 1 \tag{5.1}$$

The size increase is then multiplied by the default `width` and `height` values used in the `build_graph()` method (which is equal to 2). Since the natural logarithm of 1 equals to zero, an actual size increase is imposed only on those papers having at least two incoming edges. This is done in order to avoid having the nodes associated with the studies gathered from the backward snowballing phase being, on average, larger than the other nodes (since they always an in-degree which is equal or larger than one).

```
def resize_nodes(G):
 for node in G.nodes(): # Iterate through all nodes
  if(G.in_degree(node) > 0): # Node in-degree is not 0
   size_increase = math.log(G.in_degree(node)) + 1
   G.nodes[node]['width'] = 2 * size_increase
   G.nodes[node]['height'] = 2 * size_increase
 return G
```

As mentioned in the previous section, basing the network generation process solely on the content of the generated *.bib* files would not be sufficient to allow for the creation of a sufficiently interconnected graph. Because of this reason, the `add_additional_edges()` method is proposed in order to automate the collection of the required additional edges. These method accepts a NetworkX graph `G` and a `glossary` variable (which simply represent the content of a `Glossary.bib` file, containing the BibTex entries of all studies encountered during the course of the presented review) and uses the Academic Graph API (called by means of a Python library) in order to collect the required information. In this context, for each node in `G`, the method: (1) Retrieves the associated BibTex entry from the `glossary`; (2) Fetches the paper from SemanticScholar (using the paper's DOI or title); (3) Collects the associated references; (4) Builds, for each reference, the citation key using the `build_cite_key()` method; (5) Searches `G` for the node associated with the generated key and, if the node is found to be present in the graph; (6) Adds the missing edge.

```
def add_additional_edges(G, glossary):
 sch = SemanticScholar()

 for node in G.nodes():
  cur_paper = glossary[node]

  if('doi' in cur_paper): # DOI search
   sch_paper = sch.get_paper(cur_paper['doi'])
  else: # Title search
   res_list = sch.search_paper(cur_paper['title'])
   if(len(res_list) == 0):
    print(f"Paper not found: {cur_paper_key}")
    continue
   else:
    sch_paper = res_list[0]

   if(sch_paper.references): # Explore references
    for reference in sch_paper.references:
     # Build key for reference
     cite_key = build_cite_key(
      reference.authors[0].name,
      reference.year,
      reference.title
     )

     # Check if generated key is found in graph
     if(G.has_node(cite_key)):
      G.add_edge(node, cite_key) # Generate missing edge
 return resize_nodes(G)
```

At this point, all the required logic is in place to allow for the generation of the relevant (complete) graphs. As a first step, the two network relating to the backward and forward snowballing activities are generated separately using the `build_graph()` method. Notice that together with the path of the relevant folder (where the required *.bib* files are located), the `bs_selection` and `fs_selection` are also passed as parameters. These two variables are simply lists containing the key of the studies that were ultimately selected and included in the review. The content of such lists is used when the boolean parameter `selection_only` is set to `True`.

Once generated, the networks can then be merged using the `compose()` method of the NetworkX library, which allows for the nodes (and related edges) to be represented in a single `DiGraph` element.

Clearly, this method is designed to work even in situations where the node sets and edges sets are not disjoint. The composed graph is then passed as a parameter to the `add_additional_edges()` method, which fetches the remaining connections from the SemanticScholar repository. Finally, the complete network is written on file using the `write_dot()` method, which takes the complete graph and writes it to file using the Graphviz .dot syntax.

```
BS_Graph = build_graph('../path/to/bs/folder', bs_selection)
FS_Graph = build_graph('../path/to/fs/folder', fs_selection)
FULL_Graph = add_additional_edges(nx.compose(BS,FS), glossary)

# Persist the graph on file
drawing.nx_pydot.write_dot(FULL_Graph, "FULL_Graph.dot")
```

In a similar fashion, the larger network containing all studies gathered from the two snowballing activities (and not only the selected ones) can be generated as follows:

```
BS_Graph = build_graph('../path/to/bs/folder', selection_only = False)
FS_Graph = build_graph('../path/to/fs/folder', selection_only = False)
FULL_Graph = add_additional_edges(nx.compose(BS,FS), glossary)

# Persist the graph on file
drawing.nx_pydot.write_dot(FULL_Graph, "FULL_Graph.dot")
```

Ultimately, the network visualisation is produced via the Netgraph library, which facilitates the generation of highly customisable graphs (ultimately represented as Matplotlib objects) starting from the existing NetworkX representations. In this context, the `draw_diGraph()` method accepts a `path` pointing to the graph object saved in .dot format and an `output_name` variable containing the path and name indicating where the resulting visualisation should be saved (in .pdf format). Using these information, the method reads the .dot file using the `read_dot()` function and generates the corresponding NetworkX `DiGraph` object. Then, the method iterates over all nodes contained within the graph and saves, for each node, the relevant color, height and label information. Such information are stored in dictionaries and ultimately passed as parameters for the creation of a Netgraph `Graph` object (`plot_instance`). The resulting Matplotlib plot is ultimately printed and saved as a .pdf file.

```
def draw_diGraph(path, output_name):
 diGraph = nx.DiGraph(nx.nx_pydot.read_dot(path))

 node_color = dict()
 node_size = dict()
 node_labels = dict()

 for node in diGraph:
  if 'color' in diGraph.nodes[node]:
   node_color[node] = diGraph.nodes[node]['color']
  else:
   node_color[node] = 'white' # Default color

  if 'height' in diGraph.nodes[node]:
   node_size[node] = float(diGraph.nodes[node]['height'])
  else:
   node_size[node] = 1 # Default size

  if 'label' in diGraph.nodes[node]:
   node_labels[node] = diGraph.nodes[node]['label']

 fig, ax = plt.subplots(figsize=(300,300))
 plot_instance = Graph(diGraph,
  node_labels=node_labels, node_size=node_size, node_color=node_color,
  edge_layout='bundled',
  arrows=True, edge_width = 0.5
 )

 fig.canvas.draw()
 fig.savefig(f'{output_name}.pdf', dpi=100)
```

5.3. Analysis I - Papers selection

Using the tools and methodologies described in Section 5.2 and 5.1, a literature network pertaining to the paper included in the presented review is generated and an analysis of its shape and content is presented in this Section. In this context, the resulting graph only shows those studies part of the final selection that posses at least one outgoing or incoming edge (i.e nodes with an in- or out-degree different from zero). In other words, isolates (i.e. nodes with no neighbors) are not included in the visualisation and subsequent analysis. Such isolates are the documents from the initial selection from which no additional studies where included in the review during one of the two snowballing activities.

The generated graph is found to have a total of **98 nodes**. This value indicates that a total of [...] papers, which were ultimately not represented in the visualisation, are isolates. The 98 nodes are organised into a set of **14 subgraphs**. These elements (which in the case of an undirected graph would be defined as *components*) are distinct directed graphs which share no common vertex with one another. Given this fact, it might be conceptually interesting to understand whether the studies associated with each of identified subgraph share some discernible characteristics that, in a way, might allow to better characterise them and that might ultimately provide an initial **thematic overview** of the work included in the review. When examining the size of the individual subgraphs, it is possible to notice that most of the nodes in the produced visualisation (64 out of 98) are organised into two large elements, made up of 34 and 30 documents respectively. The remaining studies are instead found in three medium sized subgraphs (made up of 9, 4 and 3 documents) and 9 elements containing only two nodes.

Rather than a fully fledged topic modeling approach, which is deemed unnecessary considering the fact that all papers contained in the presented graph are fully examined in the context of the review process, a relatively simpler analysis on the most frequently occurring words (and n-grams) in the titles and abstracts associated with each study is performed. This analysis serves the purpose of providing an initial (general) idea of the thematic content of each subgraph, and provides some starting data allowing for further discussion. In this context, the content of the titles and abstract associated with each study is extracted. A few pre-processing steps are conducted (by means of the tools provided by the Python's NLTK Library (2023)), namely: Removal of stop words, lemmatization, identification of 2- and 3- grams, tokenization. Additionally, the entries associated with the root terms *"topic*"* and *"model*"*, which are likely to be found in virtually all documents in the collection, are removed from the generated results.

Given this premise, a first analysis is offered for the largest of the identified components. In this context, Figure 5.1 visually highlights this first subgraph. As mentioned in Section 5.1, the size of a given node is directly tied to its in-degree. Additionally, notice that the different colors identify the specific phase in which the paper associated with a specific node has been collected. In this regard, the *"skyblue"* color identifier is used to highglight documents belonging to the initial selection, whilst the *"mistyrose"* and *"lightgoldenrodyellow"* colors are used for the bakward and forward snowballing papers respectively. Finally, the displayed label for each node is generated using the first two elements (author surname and year) found in the corresponding citation key.

Performing the previously described process to identify the most prominent n-grams leads to the set of results highlighted in Figure 5.2, where the 15 most prominent n-grams are presented. Here, the terms directly associated with labeling are presented in bold.

In this context, the prominence of the terms: *"label"* and *"labeling"* and of the 2-grams *"topic label"* and *"topic labeling"* are strong indicator of the presence of work focusing on topic labeling activities. Additionally, the terms *"approach"* and *"method"* might indicate a tendency of the interested studies to focus on the description of specific (potentially novel) methodologies. Finally, the term *"automatic"* might already suggests that a substantial portion of the related studies refers to techniques tied to the process of automated topic labeling. Naturally, this initial assumptions need to be verified by further exploring the related data.

Further inspection reveals that 19 out of the 34 papers present the root term *"label*"* in the examined fields. Of the interested studies, all but one study use the term in the context of labeling topics. This insights suggests that roughly half of the documents that are part of this subnetwork

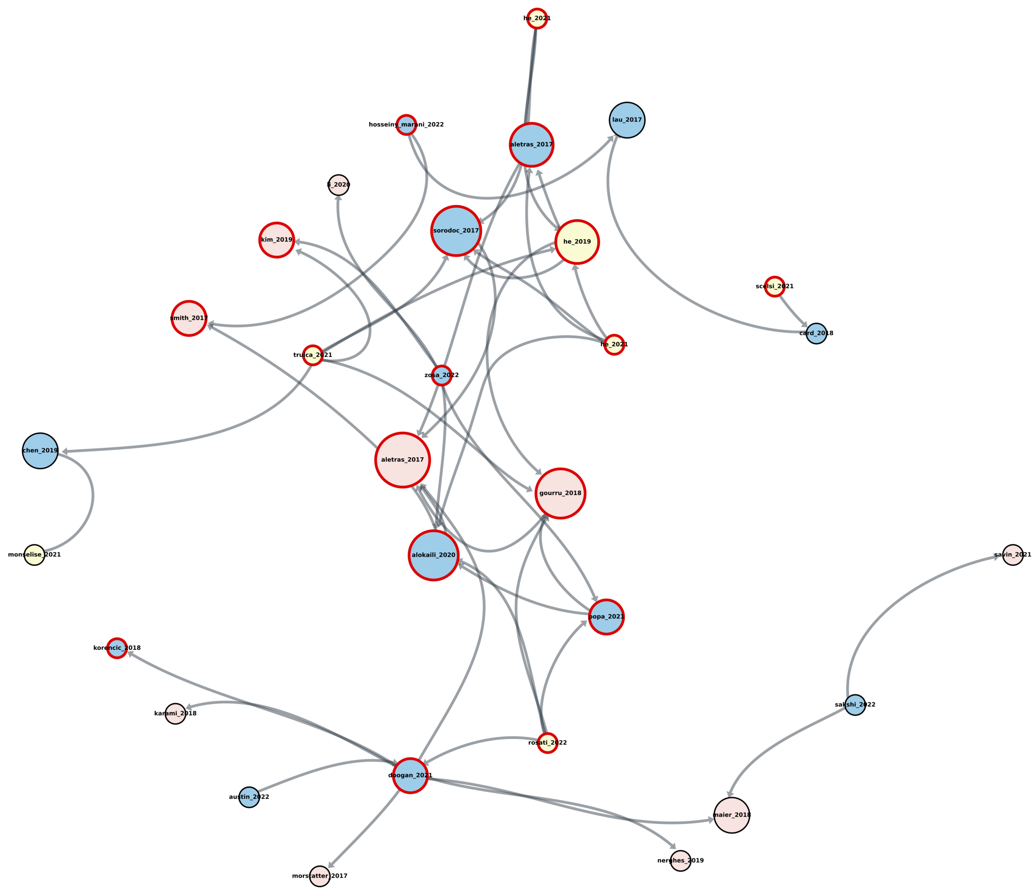


Figure 5.1: Largest subgraph

pose at least a somewhat primary focus on the activity of topic labeling. The interested studies are represented with a red border in Figure 5.1. In this context, the term *"automatic"* can be exploited to further characterise these subgroup. In fact, 9 of the 10 total appearances of this word are found within the documents that also contain the root term *"label*"*. Inspecting the associated titles and abstracts reveals that among these nine appearances, the term is always used to refer to the activity of automatic topic labeling, suggesting that roughly half of these primary studies show at least some interest in the discussion of such techniques. A further observation that might potentially be useful in this analysis is the distribution of the in-degree (i.e. the metric controlling the size of each node) across the documents in the subgraph, especially with regards to how the references (and corresponding citations) are organised among the highlighted work. In this regard, ...

In this subset of studies, the terms *"approach"* and *"method"* (appearing in 8 and 9 papers respectively) are predominantly used to refer to newly proposed topic labeling approaches.

This same analysis is not presented for components having three of fewer nodes, given that the number of associated studies is likely not sufficient to provide meaningful insights. Instead, the content of this smaller subgraphs is manually inspected

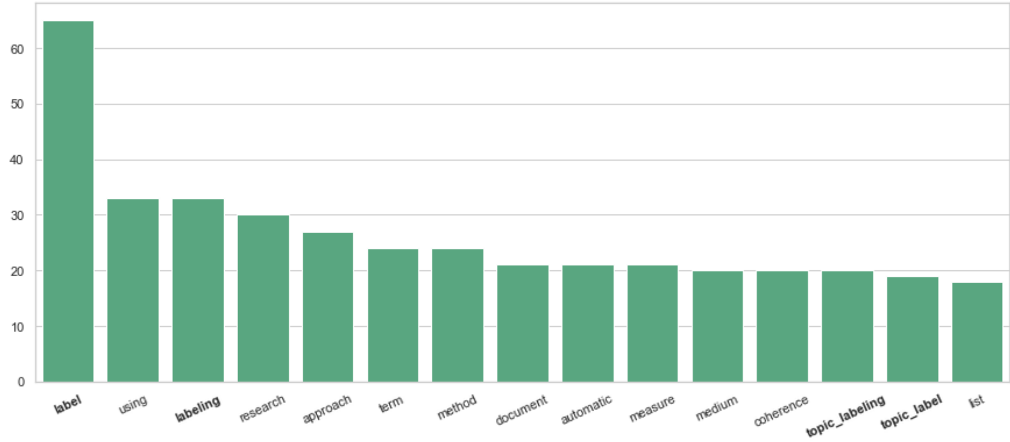


Figure 5.2: Top terms of largest subgraph.

5.4. Analysis II - Complete literature network

5.5. Analysis III - K-cores

5.6. Analysis IV - Main path analysis

Chapter 6

Results of Analyses and Syntheses

6.1. Synthesis of the primary studies

6.1.1. Overview of the included work

As mentioned in Section 3.8, the relatively straightforward process of synthesising the content of the selected primary studies is fundamentally conducted following the general order of the agreed upon data items (presented in Section 3.6). In this context, the relevant information contained within such items are grouped together and synthesised with regards to the Research Question they belong to. When observing the Data Extraction Form (Table 3.3) used within the data collection process, it is possible to notice how **Data items 1-4** do not belong to any specific Research Question. Instead, such items are used to store generic descriptors (publication year, authors, title and venue) of the collected work which are not directly tied to any of the specific questions the presented review is trying to answer but that, nonetheless, allow to easily describe the general shape of the included studies. Naturally, Data Items 1-4 are fully presented in Appendix A, B & C and, in a general sense, author's surnames and publication years are used to represent the referenced material throughout the entirety of this document. Additionally, further figures relating to Data Item 4 (the included venues) are highlighted in the introduction to Chapter 4 and in the subsequent Section 4.2.

Despite this fact, and before going further into the exploration of the data pertaining to the presented Research Questions, it might still be interesting to present, in a tabular format, a general overview of the body of research under scrutiny. As a first step in this exploration, Table 6.1 shows a division of the initial set of studies by venue and year. Venues are sorted by total number of selected studies and further organised alphabetically if they present the same totals. In this initial overview, it is possible to observe how the venue from which the largest amount of papers was ultimately selected is *Expert Systems with Applications*, with a total of 14 studies included in the review. This is probably due to the fact that the interested journal is fairly large, containing a total of 4208 documents published in the period from 2017-2021. Despite this fact, this factor only provides a partial explanation on the relevance of this venue. In fact, it is possible to observe that the only other journal having a similar size (Information Sciences with 4397 documents in the same time period) saw only a total of three selected studies after the application of the exclusion and inclusion criteria. Additionally, in the same table it is possible to observe that one of the initially selected journals (Pattern recognition) was ultimately found to contain no relevant publications.

At this point, it might also be interesting the present a year-by-year outline of all the selected primary studies including, on top of the already presented data, the work collected for the conducted snowballing activities. For this purpose, Table 6.2 provides a representation of the aggregated figures (the subtotals) associated with each distinct data collection activity, together with a further division relating to documents collected from conferences and journals. Finally, the table presents a row containing the totals (across all activities) for each year. As expected, it is possible to observe that most of the work selected during the backward snowballing phase is gathered from the early years of the covered time period. In fact, the year 2017 saw the highest number of studies collected during this

	Year						
Venue	2017	2018	2019	2020	2021	2022	**Total**
Expert Systems with Applications	1	2	1	3	5	2	14
Journal of Infometrics	1	1	0	2	1	2	7
Knowledge-based Systems	2	2	1	1	0	0	6
ACL	1	2	1	1	0	0	5
COLING	0	1	0	1	0	2	4
Decision Support Systems	0	0	2	0	2	0	4
EMNLP	1	0	0	2	1	0	4
KDD	0	1	0	2	0	1	4
EACL	1	0	0	0	2	0	3
Information Sciences	1	0	0	0	1	1	3
CIKM	0	0	1	0	0	1	2
ECIR	1	0	0	0	0	1	2
NAACL	0	0	0	0	2	0	2
SIGIR	1	0	0	2	0	0	2
ECML PKDD	0	0	0	1	0	0	1
Pattern recognition	0	0	0	0	0	0	0
Total	10	9	6	15	14	10	63

Table 6.1: Initial selection papers by venue and year

phase (with a total of 10 papers included) and a constant decline is observed when moving forward in time (with 0 papers selected in 2022). On the other hand, the opposite trend can be observed with regards to the forward snowballing activity, where most of the additional studies belong to the later years. Here, it is possible to see that 2017 and 2018 yielded no additional work. In a general sense, one can expect to see roughly the same patterns in most systematic review where citations and references from the originally collected work are used to extended the scope of the review process. When observing the yearly trend of the initial selection and of the last row referencing the totals, no overly strong pattern is identifiable. The value range from a minimum of 16 to a maximum of 26 documents and the years 2020 and 2021 show the highest paper count (with a value of 25 and 26 respectively).

		Year						
	Venue Type	2017	2018	2019	2020	2021	2022	**Total**
Initial sel.	Conf.	5	4	2	9	5	5	30
	Journal	5	5	4	6	9	5	34
	Subtotal	10	9	6	15	14	10	63
Backward S.	Conf.	2	1	0	4	0	0	7
	Journal	8	6	6	0	1	0	21
	Subtotal	10	7	6	4	1	0	28
Forward S.	Conf.	0	0	1	0	3	1	5
	Journal	0	0	5	6	8	7	26
	Subtotal	0	0	6	6	11	8	31
	Total	20	16	18	25	26	18	123

Table 6.2: Selected papers by activity, venue type and year

6.2. Identified Gaps in the research

6.3. Insights from the individual studies

Chapter 7

Discussion

Chapter 8

Conclusion

8.1. Availability of Data, Code and Other Materials

Appendices

Appendix A

Papers Reviewed - Initial selection

Year	Venue	Title	Reference
2020	SIGIR	Automatic Generation of Topic Labels	(Alokaili et al., 2020)
2017	SIGIR	TOTEM: Personal Tweets Summarization on Mobile Devices	(Chin et al., 2017)
2020	SIGIR	Read what you need: Controllable Aspect-based Opinion Summarization of Tourist Reviews	(Mukherjee et al., 2020)
2021	NAACL	Topic Model or Topic Twaddle? Re-evaluating Semantic Interpretability Measures	(Doogan and Buntine, 2021)
2021	NAACL	A Disentangled Adversarial Neural Topic Model for Separating Opinions from Plots in User Reviews	(Pergola et al., 2021)
2018	KDD	TaxoGen: Unsupervised Topic Taxonomy Construction by Adaptive Term Embedding and Clustering	(Zhang et al., 2018a)
2020	KDD	Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding	(Meng et al., 2020)
2022	KDD	Automatic Phenotyping by a Seed-Guided Topic Model	(Song et al., 2022)
2020	KDD	CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring	(Huang et al., 2020)
2020	EMNLP	Neural Topic Modeling with Cycle-Consistent Adversarial Training	(Hu et al., 2020)
2020	EMNLP	Condolence and Empathy in Online Communities	(Zhou and Jurgens, 2020)
2021	EMNLP	Phrase-BERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration	(Wang et al., 2021)
2017	EMNLP	Adapting Topic Models using Lexical Associations with Tree Priors	(Yang et al., 2017)
2022	ECIR	Multilingual Topic Labelling of News Topics Using Ontological Mapping	(Zosa et al., 2022)
2017	ECIR	Labeling Topics with Images Using a Neural Network	(Aletras and Mittal, 2017)
2017	EACL	Multimodal Topic Labelling	(Sorodoc et al., 2017)
2021	EACL	BART-TL: Weakly-Supervised Topic Label Generation	(Popa and Rebedea, 2021)
2021	EACL	Adversarial Learning of Poisson Factorisation Model for Gauging Brand Sentiment in User Reviews	(Zhao et al., 2021)
2022	COLING	Community Topic: Topic model inference by consecutive word community discovery	(Austin et al., 2022)
2020	COLING	Mining Crowdsourcing Problems from Discussion Forums of Workers	(Nouri et al., 2020)

Year	Venue	Title	Reference
2022	COLING	Improving Deep Embedded Clustering via Learning Cluster-level Representations	(Yin et al., 2022)
2018	COLING	Model-Free Context-Aware Word Composition	(An et al., 2018)
2022	CIKM	One Rating to Rule Them All? Evidence of Multidimensionality in Human Assessment of Topic Labeling Quality	(Hosseiny Marani et al., 2022)
2019	CIKM	ConCET: Entity-Aware Topic Classification for Open-Domain Conversational Agents	(Ahmadvand et al., 2019)
2019	ACL	Spatial Aggregation Facilitates Discovery of Spatial Topics	(Maiti and Vucetic, 2019)
2018	ACL	PhraseCTM: Correlated Topic Modeling on Phrases within Markov Random Fields	(Huang, 2018)
2018	ACL	Neural Models for Documents with Metadata	(Card et al., 2018)
2019	Decision Support Systems	A text analytics approach for online retailing service improvement: Evidence from Twitter	(Ibrahim and Wang, 2019)
2022	Expert Systems With Applications	Recent trends in mathematical expressions recognition: An LDA-based analysis	(Sakshi and Kukreja, 2023)
2022	Expert Systems With Applications	Providing recommendations for communities of learners in MOOCs ecosystems	(Campos et al., 2022)
2022	Journal of Informetrics	Is it all bafflegab? – Linguistic and meta characteristics of research articles in prestigious economics journals	(Amon and Hornik, 2022)
2019	Knowledge-Based Systems	Experimental explorations on short text topic mining between LDA and NMF based Schemes	(Chen et al., 2019)
2020	Journal of Informetrics	Application of machine learning techniques to assess the trends and alignment of the funded research output	(Ebadi et al., 2020)
2022	Journal of Informetrics	Developing a topic-driven method for interdisciplinarity analysis	(Kim et al., 2022b)
2020	Journal of Informetrics	Topic-linked innovation paths in science and technology	(Xu et al., 2020)
2018	Journal of Informetrics	Does deep learning help topic extraction? A kernel k-means clustering method with word embedding	(Zhang et al., 2018b)
2022	Journal of Informetrics	Exploring scientific trajectories of a large-scale dataset using topic-integrated path extraction	(Kim et al., 2022a)
2018	Knowledge-Based Systems	Identifying topical influencers on twitter based on user behavior and network topology	(Alp and Öğüdücü, 2018)
2018	Knowledge-Based Systems	Detecting and predicting the topic change of Knowledge-based Systems: A topic-based bibliometric analysis from 1991 to 2016	(Zhang et al., 2017)
2021	Expert Systems With Applications	Supporting digital content marketing and messaging through topic modelling and decision trees	(Gregoriades et al., 2021)
2018	Expert Systems With Applications	Document-based topic coherence measures for news media text	(Korenčić et al., 2018)
2022	Expert Systems With Applications	Large scale analysis of open MOOC reviews to support learners' course selection	(Gomez et al., 2022)

Year	Venue	Title	Reference
2018	Expert Systems With Applications	W2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis	(García-Pablos et al., 2018)
2021	Expert Systems With Applications	Criteria determination of analytic hierarchy process using a topic model	(Fang and Partovi, 2021)
2019	Expert Systems With Applications	Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews	(Korfiatis et al., 2019)
2019	Expert Systems With Applications	Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints	(Bastani et al., 2019)
2022	Expert Systems With Applications	The climate change Twitter dataset	(Effrosynidis et al., 2022)
2020	Expert Systems With Applications	Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis	(Kim et al., 2020)
2022	Decision Support Systems	Data-driven decision-making in credit risk management: The information value of analyst reports	(Roeder et al., 2022)
2022	Decision Support Systems	Sourcing product innovation intelligence from online reviews	(Goldberg and Abrahams, 2022)
2020	Decision Support Systems	How do consumers in the sharing economy value sharing? Evidence from online reviews	(Xu, 2020)
2022	Expert Systems With Applications	Topic2Labels: A framework to annotate and classify the social media data through LDA topics and deep learning models for crisis response	(Wahid et al., 2022)
2022	Expert Systems With Applications	Preliminary exploration of topic modelling representations for Electronic Health Records coding according to the International Classification of Diseases in Spanish	(Lebeña et al., 2022)
2022	Expert Systems With Applications	Social media analysis by innovative hybrid algorithms with label propagation	(Altınel, 2022)
2019	Knowledge-Based Systems	Learning document representation via topic-enhanced LSTM model	(Zhang et al., 2019)
2020	Knowledge-Based Systems	A topic-sensitive trust evaluation approach for users in online communities	(Chen et al., 2020a)
2017	ACL	Topically Driven Neural Language Model	(Lau et al., 2017)

Appendix B

Papers Reviewed - Backward Snowballing

Year	Venue	Title	Reference
2017	Journal of Advertising	An Investigation of Brand-Related User-Generated Content on Twitter	(Liu et al., 2017)
2017	Management Science	Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach	(Huang et al., 2017)
2020	Lecture Notes in Computer Science	What Are MOOCs Learners' Concerns? Text Analysis of Reviews for Computer Science Courses	(Chen et al., 2020c)
2019	Journal of Comparative Economics	Toward understanding 17th century English culture: A structural topic model of Francis Bacon's ideas	(Grajzl and Murrell, 2019)
2017	Tourism Management	A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism	(Xiang et al., 2017)
2019	Journal of Information Science	Twitter speaks: A case of national disaster situational awareness	(Karami et al., 2019)
2018	International Journal of Information Management	Characterizing diabetes, diet, exercise, and obesity comments on Twitter	(Karami et al., 2018)
2017	Social Forces	Managing the Boundaries of Taste: Culture, Valuation, and Computational Social Science	(Light and Odden, 2017)
2021	Ecological Economics	Free associations of citizens and scientists with economic and green growth: A computational-linguistics analysis	(Savin et al., 2021)
2018	Transportation Research Part C	Using structural topic modeling to identify latent topics and trends in aviation incident reports	(Kuhn, 2018)

Year	Venue	Title	Reference
2017	DSAA	Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation	(Syed and Spruit, 2017)
2019	Media and Communication	Narratives of the Refugee Crisis: A Comparative Study of Mainstream-Media and Twitter	(Nerghes and Lee, 2019)
2019	Journal of Information Processing Systems	An Ontology-Based Labeling of Influential Topics Using Topic Network Analysis	(Kim and Rhee, 2019)
2021	Research & Politics	Transfer Topic Labeling with Domain-Specific Knowledge Base: An Analysis of UK House of Commons Speeches 1935–2014	(Béchara et al., 2021)
2020	SBSI	Recommendation System for Knowledge Acquisition in MOOCs Ecosystems	(Campos et al., 2020)
2020	Communication Methods and Measures	Applying LDA topic modeling in communication research: Toward a valid and reliable methodology	(Maier et al., 2018)
2017	ICICoS	Topic Labeling Towards News Document Collection Based on Latent Dirichlet Allocation and Ontology	(Adhitama et al., 2017)
2017	Natural Language Processing and Information Systems	United We Stand: Using Multiple Strategies for Topic Labeling	(Gourru et al., 2018)
2019	Social Network Analysis and Mining	Topic modeling and sentiment analysis of global climate change tweets	(Dahal et al., 2019)
2019	Small-scale Forestry	Modelling Research Topic Trends in Community Forestry	(Clare and Hickey, 2019)

Appendix C

Papers Reviewed - Forward Snowballing

Year	Venue	Title	Reference
2020	Journal of Manufacturing Technology Management	A topic-based patent analytics approach for exploring technological trends in smart manufacturing	(Wang and Hsu, 2020)
2022	Journal of Big Data	Modeling the public attitude towards organic foods: a big data and text mining approach	(Singh and Glińska-Neweś, 2022)
2020	Journal of Travel & Tourism Marketing	Topic modelling for theme park online reviews: analysis of Disneyland	(Luo et al., 2020)
2021	International Journal of Contemporary Hospitality Management	Revealing industry challenge and business response to Covid-19: a text mining approach	(Yang and Han, 2021)
2021	Scientometrics	The use of citation context to detect the evolution of research topics: a large-scale analysis	(Jebari et al., 2021)
2020	Cognitive Computation	A Structural Topic Modeling-Based Bibliometric Study of Sentiment Analysis Literature	(Chen and Xie, 2020)
2021	Journal of Medical Internet Research	Topics and Sentiments of Public Concerns Regarding COVID-19 Vaccines: Social Media Trend Analysis	(Monselise et al., 2021)
2022	International Journal of Quality & Reliability Management	Quality 4.0: big data analytics to explore service quality attributes and their relation to user sentiment in Airbnb reviews	(Amat-Lefort et al., 2022)
2019	SAICSIT	A Computational Analysis of News Media Bias: A South African Case Study	(Cornelissen et al., 2019)
2020	Computers & Education	Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education	(Chen et al., 2020b)

Year	Venue	Title	Reference
2021	Scientometrics	Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing	(Ebadi et al., 2021)
2020	International Journal of Hospitality Management	Employing structural topic modelling to explore perceived service quality attributes in Airbnb accommodation	(Ding et al., 2020)

Bibliography

Academic Graph API. Academic Graph API's website. https://www.semanticscholar.org/product/api, 2023. [Online; accessed October 2022 - June 2023].

P. Achimugu, A. Selamat, R. Ibrahim, and M. N. Mahrin. A systematic literature review of software requirements prioritization research. *Information and Software Technology*, 56(6):568–585, 2014. ISSN 0950-5849. doi: https://doi.org/10.1016/j.infsof.2014.02.001. URL https://www.sciencedirect.com/science/article/pii/S0950584914000354.

ACL Anthology. ACL Anthology Website. https://aclanthology.org/, 2023. [Online; accessed October 2022 - June 2023].

ACM Digital Library. ACM Digital Library Website. https://dl.acm.org, 2023. [Online; accessed October 2022 - June 2023].

R. Adhitama, R. Kusumaningrum, and R. Gernowo. Topic labeling towards news document collection based on latent dirichlet allocation and ontology. In *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*, pages 247–252, 2017. doi: 10.1109/ICICOS.2017.8276370. URL https://ieeexplore.ieee.org/document/8276370.

A. Ahmadvand, H. Sahijwani, J. I. Choi, and E. Agichtein. Concet: Entity-aware topic classification for open-domain conversational agents. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1371–1380, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369763. doi: 10.1145/3357384.3358048. URL https://doi.org/10.1145/3357384.3358048.

N. Aletras and A. Mittal. Labeling topics with images using a neural network. In J. M. Jose, C. Hauff, I. S. Altıngovde, D. Song, D. Albakour, S. Watt, and J. Tait, editors, *Advances in Information Retrieval*, pages 500–505, Cham, 2017. Springer International Publishing. ISBN 978-3-319-56608-5.

M. Allahyari and K. Kochut. Automatic topic labeling using ontology-based topic models. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 259–264, 2015. doi: 10.1109/ICMLA.2015.88. URL https://ieeexplore.ieee.org/abstract/document/7424318.

A. Alokaili, N. Aletras, and M. Stevenson. Automatic generation of topic labels. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1965–1968, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401185. URL https://doi.org/10.1145/3397271.3401185.

Z. Z. Alp and Ş. G. Öğüdücü. Identifying topical influencers on twitter based on user behavior and network topology. *Knowledge-Based Systems*, 141:211–221, 2018. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2017.11.021. URL https://www.sciencedirect.com/science/article/pii/S0950705117305439.

A. B. Altınel. Social media analysis by innovative hybrid algorithms with label propagation. *Expert Systems with Applications*, 210:118606, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2022.118606. URL https://www.sciencedirect.com/science/article/pii/S095741742201658X.

N. Amat-Lefort, F. Barravecchia, and L. Mastrogiacomo. Quality 4.0: big data analytics to explore service quality attributes and their relation to user sentiment in airbnb reviews. *International Journal of Quality & Reliability Management*, 40(4):990–1008, sep 2022. doi: 10.1108/ijqrm-01-2022-0024. URL https://doi.org/10.1108%2Fijqrm-01-2022-0024.

J. Amon and K. Hornik. Is it all bafflegab? – linguistic and meta characteristics of research articles in prestigious economics journals. *Journal of Informetrics*, 16(2):101284, 2022. ISSN 1751-1577. doi: https://doi.org/10.1016/j.joi.2022.101284. URL https://www.sciencedirect.com/science/article/pii/S1751157722000360.

B. An, X. Han, and L. Sun. Model-free context-aware word composition. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2834–2845, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1240.

G. Anthes. Topic models vs. unstructured data. *Commun. ACM*, 53(12):16–18, dec 2010. ISSN 0001-0782. doi: 10.1145/1859204.1859210. URL https://doi.org/10.1145/1859204.1859210.

H. Arksey and L. O'Malley. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*, 8(1):19–32, feb 2005. doi: 10.1080/1364557032000119616. URL https://doi.org/10.1080%2F1364557032000119616.

E. Aromataris and D. Riitano. Constructing a search strategy and searching for evidence. a guide to the literature search for a systematic review. *Am J Nurs*, 114(5):49–56, May 2014. ISSN 1538-7488 (Electronic); 0002-936X (Linking). doi: 10.1097/01.NAJ.0000446779.99522.f6.

E. Austin, O. R. Zaïane, and C. Largeron. Community topic: Topic model inference by consecutive word community discovery. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 971–983, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.81.

S. A. Bahrainian, I. Mele, and F. Crestani. Predicting topics in scholarly papers. In G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, editors, *Advances in Information Retrieval*, pages 16–28, Cham, 2018. Springer International Publishing. ISBN 978-3-319-76941-7.

K. Bastani, H. Namavari, and J. Shaffer. Latent dirichlet allocation (lda) for topic modeling of the cfpb consumer complaints. *Expert Systems with Applications*, 127:256–271, 2019. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2019.03.001. URL https://www.sciencedirect.com/science/article/pii/S095741741930154X.

J. Beall. "predatory" open-access scholarly publishers. *The Charleston Advisor*, 11(4):10–17, 2010. URL https://core.ac.uk/download/pdf/11886760.pdf.

Beall's List. Beall's List. https://beallslist.net/, 2023. [Online; accessed October 2022 - June 2023].

H. Béchara, A. Herzog, S. Jankin, and P. John. Transfer learning for topic labeling: Analysis of the UK house of commons speeches 1935–2014. *Research & Politics*, 8(2):205316802110222, apr 2021. doi: 10.1177/20531680211022206. URL https://doi.org/10.1177%2F20531680211022206.

S. Bhatia, J. H. Lau, and T. Baldwin. Automatic labelling of topics with neural embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 953–963, Osaka, Japan, Dec. 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1091.

D. M. Blei and J. D. Lafferty. Topic models. In *Text Mining*, pages 101–124. Chapman and Hall/CRC, jun 2009. doi: 10.1201/9781420059458-12. URL https://doi.org/10.1201%2F9781420059458-12.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, mar 2003. ISSN 1532-4435.

A. Booth, D. Papaioannou, and A. Sutton. *Systematic Approaches to a Successful Literature Review*. Sage Publications, 01 2012. URL https://www.researchgate.net/publication/235930866_Systematic_Approaches_to_a_Successful_Literature_Review.

J. Boyd-Graber, Y. Hu, and D. Mimno. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296, 2017. doi: 10.1561/1500000030. URL https://doi.org/10.1561%2F1500000030.

P. Brereton and D. Budgen. Performing systematic literature reviews in software engineering. In *Software Engineering, International Conference on*, pages 1051–1052, Los Alamitos, CA, USA, may 2006. IEEE Computer Society. doi: 10.1145/1134285.1134500. URL https://doi.ieeecomputersociety.org/10.1145/1134285.1134500.

R. W. Byard. The forensic implications of predatory publishing. *Forensic Science, Medicine, and Pathology*, 12(4):391–393, apr 2016. doi: 10.1007/s12024-016-9771-3. URL https://doi.org/10.1007%2Fs12024-016-9771-3.

R. Campos, R. Santos, and J. Oliveira. Recommendation system for knowledge acquisition in moocs ecosystems. In *Anais Estendidos do XVI Simpósio Brasileiro de Sistemas de Informação*, pages 93–108, Porto Alegre, RS, Brasil, 2020. SBC. doi: 10.5753/sbsi.2020.13132. URL https://sol.sbc.org.br/index.php/sbsi_estendido/article/view/13132.

R. Campos, R. P. dos Santos, and J. Oliveira. Providing recommendations for communities of learners in moocs ecosystems. *Expert Systems with Applications*, 205:117510, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2022.117510. URL https://www.sciencedirect.com/science/article/pii/S0957417422008375.

D. Card, C. Tan, and N. A. Smith. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2040, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1189. URL https://aclanthology.org/P18-1189.

U. Chauhan and A. Shah. Topic modeling using latent dirichlet allocation: A survey. *ACM Comput. Surv.*, 54(7), sep 2021. ISSN 0360-0300. doi: 10.1145/3462478. URL https://doi.org/10.1145/3462478.

T.-H. Chen, S. W. Thomas, and A. E. Hassan. A survey on the use of topic models when mining software repositories. *Empirical Software Engineering*, 21(5):1843–1919, 2016. doi: 10.1007/s10664-015-9402-8. URL https://doi.org/10.1007/s10664-015-9402-8.

X. Chen and H. Xie. A structural topic modeling-based bibliometric study of sentiment analysis literature. *Cognitive Computation*, 12(6):1097–1129, Nov 2020. ISSN 1866-9964. doi: 10.1007/s12559-020-09745-1. URL https://link.springer.com/content/pdf/10.1007/s12559-020-09745-1.pdf.

X. Chen, Y. Yuan, M. A. Orgun, and L. Lu. A topic-sensitive trust evaluation approach for users in online communities. *Knowledge-Based Systems*, 194:105546, 2020a. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2020.105546. URL https://www.sciencedirect.com/science/article/pii/S0950705120300435.

X. Chen, D. Zou, G. Cheng, and H. Xie. Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of computers & education. *Computers & Education*, 151:103855, 2020b. ISSN 0360-1315. doi: https://doi.org/10.1016/j.compedu.2020.103855. URL https://www.sciencedirect.com/science/article/pii/S0360131520300555.

X. Chen, D. Zou, H. Xie, and G. Cheng. What are MOOCs learners' concerns? text analysis of reviews for computer science courses. In *Lecture Notes in Computer Science*, pages 73–79. Springer International Publishing, 2020c. doi: 10.1007/978-3-030-59413-8_6. URL https://doi.org/10.1007%2F978-3-030-59413-8_6.

Y. Chen, H. Zhang, R. Liu, Z. Ye, and J. Lin. Experimental explorations on short text topic mining between lda and nmf based schemes. *Knowledge-Based Systems*, 163:1–13, 2019. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2018.08.011. URL https://www.sciencedirect.com/science/article/pii/S0950705118304076.

J. Y. Chin, S. S. Bhowmick, and A. Jatowt. Totem: Personal tweets summarization on mobile devices. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1305–1308, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450350228. doi: 10.1145/3077136.3084138. URL https://doi.org/10.1145/3077136.3084138.

R. Churchill and L. Singh. The evolution of topic modeling. *ACM Comput. Surv.*, 54(10s), nov 2022. ISSN 0360-0300. doi: 10.1145/3507900. URL https://doi.org/10.1145/3507900.

S. M. Clare and G. M. Hickey. Modelling research topic trends in community forestry. *Small-scale Forestry*, 18(2):149–163, Jun 2019. ISSN 1873-7854. doi: 10.1007/s11842-018-9411-8. URL https://link.springer.com/content/pdf/10.1007/s11842-018-9411-8.pdf.

C. Cooper, A. Booth, N. Britten, and R. Garside. A comparison of results of empirical studies of supplementary search techniques and recommendations in review methodology handbooks: a methodological review. *Systematic Reviews*, 6(1), nov 2017. doi: 10.1186/s13643-017-0625-1. URL https://doi.org/10.1186%2Fs13643-017-0625-1.

H. M. Cooper. Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society*, 1(1):104, 1988. doi: 10.1007/BF03177550. URL https://doi.org/10.1007/BF03177550.

CORE. Details of data used in CORE Rankings and response to some concerns. https://drive.google.com/file/d/19q8UdRB0sKMVpz70Oc5VG5MN-sUBwMyJ/view, 2023a. [Online; accessed October 2022 - June 2023].

CORE. CORE conference rankings 2021: Process followed and data considered. https://drive.google.com/file/d/1bKa40nheaQ3zfuXu3jSpKIw5TnhK9USR/view, 2023b. [Online; accessed October 2022 - June 2023].

CORE. Computing Research and Education Association of Australasia (CORE) Website. https://www.core.edu.au/, 2023. [Online; accessed October 2022 - June 2023].

L. A. Cornelissen, L. I. Daly, Q. Sinandile, H. de Lange, and R. J. Barnett. A computational analysis of news media bias. In *Proceedings of the South African Institute of Computer Scientists and Information Technologists 2019*. ACM, sep 2019. doi: 10.1145/3351108.3351134. URL https://doi.org/10.1145%2F3351108.3351134.

C. Creaser and S. White. Trends in journal prices: an analysis of selected journals, 2000–2006. *Learned Publishing*, 21(3):214–224, 2008. doi: https://doi.org/10.1087/095315108X323866. URL https://onlinelibrary.wiley.com/doi/abs/10.1087/095315108X323866.

B. Dahal, S. A. P. Kumar, and Z. Li. Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining*, 9(1):24, Jun 2019. ISSN 1869-5469. doi: 10.1007/s13278-019-0568-8. URL https://link.springer.com/content/pdf/10.1007/s13278-019-0568-8.pdf.

D. Demszky, N. Garg, R. Voigt, J. Zou, J. Shapiro, M. Gentzkow, and D. Jurafsky. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2970–3005, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1304. URL https://aclanthology.org/N19-1304.

K. Ding, W. C. Choo, K. Y. Ng, and S. I. Ng. Employing structural topic modelling to explore perceived service quality attributes in airbnb accommodation. *International Journal of Hospitality Management*, 91:102676, 2020. ISSN 0278-4319. doi: https://doi.org/10.1016/j.ijhm.2020.102676. URL https://www.sciencedirect.com/science/article/pii/S0278431920302280.

C. Doogan and W. Buntine. Topic model or topic twaddle? re-evaluating semantic interpretability measures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3824–3848, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.300. URL https://aclanthology.org/2021.naacl-main.300.

S. T. Dumais. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230, 2004. doi: https://doi.org/10.1002/aris.1440380105. URL https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/aris.1440380105.

A. Ebadi, S. Tremblay, C. Goutte, and A. Schiffauerova. Application of machine learning techniques to assess the trends and alignment of the funded research output. *Journal of Informetrics*, 14(2):101018, 2020. ISSN 1751-1577. doi: https://doi.org/10.1016/j.joi.2020.101018. URL https://www.sciencedirect.com/science/article/pii/S1751157718301901.

A. Ebadi, P. Xi, S. Tremblay, B. Spencer, R. Pall, and A. Wong. Understanding the temporal evolution of covid-19 research through machine learning and natural language processing. *Scientometrics*, 126(1):725–739, Jan 2021. ISSN 1588-2861. doi: 10.1007/s11192-020-03744-7. URL https://link.springer.com/content/pdf/10.1007/s11192-020-03744-7.pdf.

D. Effrosynidis, A. I. Karasakalidis, G. Sylaios, and A. Arampatzis. The climate change twitter dataset. *Expert Systems with Applications*, 204:117541, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2022.117541. URL https://www.sciencedirect.com/science/article/pii/S0957417422008624.

S. A. Elmore and E. H. Weston. Predatory journals: What they are and how to avoid them. *Toxicologic Pathology*, 48(4):607–610, apr 2020. doi: 10.1177/0192623320920209. URL https://doi.org/10.1177%2F0192623320920209.

J. Fang and F. Y. Partovi. Criteria determination of analytic hierarchy process using a topic model. *Expert Systems with Applications*, 169:114306, 2021. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2020.114306. URL https://www.sciencedirect.com/science/article/pii/S0957417420310046.

D. Farace and A. N. TransAtlantic. *GL'97 Proceedings: Third International Conference on Grey Literature: Perspectives on the Design and Transfer of Scientific and Technical Information*. GL conference series. TransAtlantic, 1997. ISBN 9789074854177. URL https://books.google.it/books?id=Qx1rwgEACAAJ.

F. Figueiredo and A. Jorge. Identifying topic relevant hashtags in twitter streams. *Information Sciences*, 505:65–83, 2019. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2019.07.062. URL https://www.sciencedirect.com/science/article/pii/S0020025519306668.

FoxTrot Professional Search. FoxTrot Professional Search. https://foxtrot-search.com/foxtrot-professional.html, 2023. [Online; accessed October 2022 - June 2023].

A. García-Pablos, M. Cuadros, and G. Rigau. W2vlda: Almost unsupervised system for aspect based sentiment analysis. *Expert Systems with Applications*, 91:127–137, 2018. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2017.08.049. URL https://www.sciencedirect.com/science/article/pii/S0957417417305961.

V. Garousi, M. Felderer, and M. V. Mäntylä. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and Software Technology*, 106: 101–121, 2019. ISSN 0950-5849. doi: https://doi.org/10.1016/j.infsof.2018.09.006. URL https://www.sciencedirect.com/science/article/pii/S0950584918301939.

GGS. GGS Rating data sources. https://scie.lcc.uma.es:8443/conferenceRating.jsf, 2023. [Online; accessed October 2022 - June 2023].

GGS. The GII-GRIN-SCIE (GGS) Conference Rating Website. https://scie.lcc.uma.es:8443/, 2023. [Online; accessed October 2022 - June 2023].

D. M. Goldberg and A. S. Abrahams. Sourcing product innovation intelligence from online reviews. *Decision Support Systems*, 157:113751, 2022. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2022.113751. URL https://www.sciencedirect.com/science/article/pii/S0167923622000227.

M. J. Gomez, M. Calderón, V. Sánchez, F. J. G. Clemente, and J. A. Ruipérez-Valiente. Large scale analysis of open mooc reviews to support learners' course selection. *Expert Systems with Applications*, 210:118400, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2022.118400. URL https://www.sciencedirect.com/science/article/pii/S0957417422015081.

C. M. Gosling and R. Iles. Sourcing the available evidence (primary, secondary and tertiary evidence). *Introducing, Designing and Conducting Research for Paramedics*, page 39, 2022.

A. Gourru, J. Velcin, M. Roche, C. Gravier, and P. Poncelet. United we stand: Using multiple strategies for topic labeling. In *NLDB: Natural Language Processing and Information Systems*, volume LNCS, pages 352–363, Paris, France, June 2018. doi: 10.1007/978-3-319-91947-8_37. URL https://hal-lirmm.ccsd.cnrs.fr/lirmm-01910614.

P. Grajzl and P. Murrell. Toward understanding 17th century english culture: A structural topic model of francis bacon's ideas. *Journal of Comparative Economics*, 47(1):111–135, 2019. ISSN 0147-5967. doi: https://doi.org/10.1016/j.jce.2018.10.004. URL https://www.sciencedirect.com/science/article/pii/S0147596718304426.

A. Gregoriades, M. Pampaka, H. Herodotou, and E. Christodoulou. Supporting digital content marketing and messaging through topic modelling and decision trees. *Expert Systems with Applications*, 184:115546, 2021. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2021.115546. URL https://www.sciencedirect.com/science/article/pii/S0957417421009532.

S. M. Griffies, W. A. Perrie, and G. Hull. Publishing connect: Elements of style for writing scientific journal articles. *Elsevier*, 2013. URL https://www.gfdl.noaa.gov/wp-content/uploads/2018/08/Elements_of_Style.pdf.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl_1):5228–5235, apr 2004. doi: 10.1073/pnas.0307752101. URL https://doi.org/10.1073%2Fpnas.0307752101.

V. P. Guerrero-Bote and F. Moya-Anegón. A further step forward in measuring journals' scientific prestige: The sjr2 indicator. *Journal of Informetrics*, 6(4):674–688, 2012. ISSN 1751-1577. doi: https://doi.org/10.1016/j.joi.2012.07.001. URL https://www.sciencedirect.com/science/article/pii/S1751157712000521.

D. He, M. Wang, A. M. Khattak, L. Zhang, and W. Gao. Automatic labeling of topic models using graph-based ranking. *IEEE Access*, 7:131593–131608, 2019. doi: 10.1109/ACCESS.2019.2940516.

T. Hofmann. Probabilistic latent semantic analysis, 2013. URL https://arxiv.org/abs/1301.6705.

A. Hosseiny Marani, J. Levine, and E. P. Baumer. One rating to rule them all? evidence of multidimensionality in human assessment of topic labeling quality. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 768–779, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392365. doi: 10.1145/3511808.3557410. URL https://doi.org/10.1145/3511808.3557410.

X. Hu, R. Wang, D. Zhou, and Y. Xiong. Neural topic modeling with cycle-consistent adversarial training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9018–9030, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.725. URL https://aclanthology.org/2020.emnlp-main.725.

A. Huang, R. Lehavy, A. Zang, and R. Zheng. Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science*, 64, 06 2017. doi: 10.1287/mnsc.2017.2751.

J. Huang, Y. Xie, Y. Meng, Y. Zhang, and J. Han. Corel: Seed-guided topical taxonomy construction by concept learning and relation transferring. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1928–1936, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403244. URL https://doi.org/10.1145/3394486.3403244.

W. Huang. PhraseCTM: Correlated topic modeling on phrases within Markov random fields. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 521–526, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2083. URL https://aclanthology.org/P18-2083.

N. F. Ibrahim and X. Wang. A text analytics approach for online retailing service improvement: Evidence from twitter. *Decision Support Systems*, 121:37–50, 2019. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2019.03.002. URL https://www.sciencedirect.com/science/article/pii/S0167923619300405.

IEEE Xplore. IEEE Xplore Website. https://ieeexplore.ieee.org, 2023. [Online; accessed October 2022 - June 2023].

C. Jebari, E. Herrera-Viedma, and M. J. Cobo. The use of citation context to detect the evolution of research topics: a large-scale analysis. *Scientometrics*, 126(4):2971–2989, Apr 2021. ISSN 1588-2861. doi: 10.1007/s11192-020-03858-y. URL https://link.springer.com/content/pdf/10.1007/s11192-020-03858-y.pdf.

A. Jobin, M. Ienca, and E. Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, sep 2019. doi: 10.1038/s42256-019-0088-2. URL https://doi.org/10.1038%2Fs42256-019-0088-2.

V. Jupp. *The SAGE Dictionary of Social Research Methods*. SAGE Publications, Ltd, 2006. doi: 10.4135/9780857020116. URL https://doi.org/10.4135%2F9780857020116.

A. Karami, A. A. Dahl, G. Turner-McGrievy, H. Kharrazi, and G. Shaw. Characterizing diabetes, diet, exercise, and obesity comments on twitter. *International Journal of Information Management*, 38 (1):1–6, 2018. ISSN 0268-4012. doi: https://doi.org/10.1016/j.ijinfomgt.2017.08.002. URL https://www.sciencedirect.com/science/article/pii/S0268401217306126.

A. Karami, V. Shah, R. Vaezi, and A. Bansal. Twitter speaks: A case of national disaster situational awareness. *Journal of Information Science*, 46(3):313–324, mar 2019. doi: 10.1177/0165551519828620. URL https://doi.org/10.1177%2F0165551519828620.

M. U. Khan, S. Sherin, M. Z. Iqbal, and R. Zahid. Landscaping systematic mapping studies in software engineering: A tertiary study. *Journal of Systems and Software*, 149:396–436, 2019. ISSN 0164-1212. doi: https://doi.org/10.1016/j.jss.2018.12.018. URL https://www.sciencedirect.com/science/article/pii/S0164121218302784.

E. H. Kim, Y. K. Jeong, Y. Kim, and M. Song. Exploring scientific trajectories of a large-scale dataset using topic-integrated path extraction. *Journal of Informetrics*, 16(1):101242, 2022a. ISSN 1751-1577. doi: https://doi.org/10.1016/j.joi.2021.101242. URL https://www.sciencedirect.com/science/article/pii/S1751157721001139.

H. Kim, H. Park, and M. Song. Developing a topic-driven method for interdisciplinarity analysis. *Journal of Informetrics*, 16(2):101255, 2022b. ISSN 1751-1577. doi: https://doi.org/10.1016/j.joi.2022.101255. URL https://www.sciencedirect.com/science/article/pii/S1751157722000074.

H. H. Kim and H. Y. Rhee. An ontology-based labeling of influential topics using topic network analysis. *J. Inf. Process. Syst.*, 15:1096–1107, 2019.

S. Kim, H. Park, and J. Lee. Word2vec-based latent semantic analysis (w2v-lsa) for topic modeling: A study on blockchain technology trend analysis. *Expert Systems with Applications*, 152:113401, 2020. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2020.113401. URL https://www.sciencedirect.com/science/article/pii/S0957417420302256.

B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman. Systematic literature reviews in software engineering – a systematic literature review. *Information and Software Technology*, 51(1):7–15, 2009. ISSN 0950-5849. doi: https://doi.org/10.1016/j.infsof.2008.09.009. URL https://www.sciencedirect.com/science/article/pii/S0950584908001390. Special Section - Most Cited Articles in 2002 and Regular Research Papers.

B. Kitchenham, P. Brereton, and D. Budgen. The educational value of mapping studies of software engineering literature. In *2010 ACM/IEEE 32nd International Conference on Software Engineering*, volume 1, pages 589–598, 2010. doi: 10.1145/1806799.1806887.

B. Kitchenham, L. Madeyski, and D. Budgen. Segress: Software engineering guidelines for reporting secondary studies. *IEEE Transactions on Software Engineering*, PP:1–1, 01 2022. doi: 10.1109/TSE.2022.3174092.

B. A. Kitchenham. Procedures for performing systematic reviews. Technical report, Keele University, Department of Computer Science, Keele University, Kelee, UK, 07 2004. URL https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf.

D. Korenčić, S. Ristov, and J. Šnajder. Document-based topic coherence measures for news media text. *Expert Systems with Applications*, 114:357–373, 2018. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2018.07.063. URL https://www.sciencedirect.com/science/article/pii/S0957417418304883.

N. Korfiatis, P. Stamolampros, P. Kourouthanassis, and V. Sagiadinos. Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. *Expert Systems with Applications*, 116:472–486, 2019. ISSN 0957-4174. doi: https://doi.org/10.1016/j.esw a.2018.09.037. URL https://www.sciencedirect.com/science/article/pii/S0957417418306146.

K. D. Kuhn. Using structural topic modeling to identify latent topics and trends in aviation incident reports. *Transportation Research Part C: Emerging Technologies*, 87:105–122, 2018. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2017.12.018. URL https://www.sciencedirect.com/science/article/pii/S0968090X17303881.

A.-L. Lamprecht, L. Garcia, M. Kuzak, C. Martinez, R. Arcila, E. M. D. Pico, V. D. D. Angel, S. van de Sandt, J. Ison, P. A. Martinez, P. McQuilton, A. Valencia, J. Harrow, F. Psomopoulos, J. L. Gelpi, N. C. Hong, C. Goble, and S. Capella-Gutierrez. Towards FAIR principles for research software. *Data Science*, 3(1):37–59, jun 2020. doi: 10.3233/ds-190026. URL https://doi.org/10.3233%2Fds-190026.

A. Langham-Putrow, C. J. Bakker, and A. Riegelman. Is the open access citation advantage real? a systematic review of the citation of open access and subscription-based articles. *PLoS ONE*, 16, 2020.

J. H. Lau, K. Grieser, D. Newman, and T. Baldwin. Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1536–1545, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1154.

J. H. Lau, T. Baldwin, and T. Cohn. Topically driven neural language model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 355–365, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1033. URL https://aclanthology.org/P17-1033.

N. Lebeña, A. Blanco, A. Pérez, and A. Casillas. Preliminary exploration of topic modelling representations for electronic health records coding according to the international classification of diseases in spanish. *Expert Systems with Applications*, 204:117303, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2022.117303. URL https://www.sciencedirect.com/science/article/pii/S0957417422006662.

R. Light and C. Odden. Managing the Boundaries of Taste: Culture, Valuation, and Computational Social Science. *Social Forces*, 96(2):877–908, 08 2017. ISSN 0037-7732. doi: 10.1093/sf/sox055. URL https://doi.org/10.1093/sf/sox055.

M. Lipsey and D. Wilson. *Practical Meta-Analysis*. Applied Social Research Methods. SAGE Publications, 2001. ISBN 9780761921684. URL https://books.google.it/books?id=G-PnRSMxdIoC.

X. Liu, A. C. Burns, and Y. Hou. An investigation of brand-related user-generated content on twitter. *Journal of Advertising*, 46(2):236–247, mar 2017. doi: 10.1080/00913367.2017.1297273. URL https://doi.org/10.1080%2F00913367.2017.1297273.

J. M. Luo, H. Q. Vu, G. Li, and R. Law. Topic modelling for theme park online reviews: analysis of disneyland. *Journal of Travel & Tourism Marketing*, 37(2):272–285, feb 2020. doi: 10.1080/10548408.2020.1740138. URL https://doi.org/10.1080%2F10548408.2020.1740138.

D. Magatti, S. Calegari, D. Ciucci, and F. Stella. Automatic labeling of topics. In *2009 Ninth International Conference on Intelligent Systems Design and Applications*, pages 1227–1232, 2009. doi: 10.1109/ISDA.2009.165. URL https://ieeexplore.ieee.org/abstract/document/5364126.

D. Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri, and S. Adam. Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2-3):93–118, feb 2018. doi: 10.1080/19312458.2018.1430754. URL https://doi.org/10.1080%2F19312458.2018.1430754.

A. Maiti and S. Vucetic. Spatial aggregation facilitates discovery of spatial topics. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 252–262, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1025. URL https://aclanthology.org/P19-1025.

M. Mayeda and A. Andrews. Chapter two - evaluating software testing techniques: A systematic mapping study. volume 123 of *Advances in Computers*, pages 41–114. Elsevier, 2021. doi: https://doi.org/10.1016/bs.adcom.2021.01.002. URL https://www.sciencedirect.com/science/article/pii/S0065245821000279.

M. L. McHugh. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22(3):276–282, 2012. ISSN 1330-0962 (Print); 1846-7482 (Electronic); 1330-0962 (Linking).

Q. Mei, X. Shen, and C. Zhai. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, page 490–499, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595936097. doi: 10.1145/1281192.1281246. URL https://doi.org/10.1145/1281192.1281246.

T. Meline. Selecting studies for systemic review: Inclusion and exclusion criteria. *Contemporary Issues in Communication Science and Disorders*, 33(Spring):21–27, mar 2006. doi: 10.1044/cicsd_33_s_21. URL https://doi.org/10.1044%2Fcicsd_33_s_21.

Y. Meng, Y. Zhang, J. Huang, Y. Zhang, C. Zhang, and J. Han. Hierarchical topic mining via joint spherical tree and text embedding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1908–1917, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403242. URL https://doi.org/10.1145/3394486.3403242.

M. Monselise, C.-H. Chang, G. Ferreira, R. Yang, and C. C. Yang. Topics and sentiments of public concerns regarding COVID-19 vaccines: Social media trend analysis. *Journal of Medical Internet Research*, 23(10):e30765, oct 2021. doi: 10.2196/30765. URL https://doi.org/10.2196%2F30765.

R. Mukherjee, H. C. Peruri, U. Vishnu, P. Goyal, S. Bhattacharya, and N. Ganguly. Read what you need: Controllable aspect-based opinion summarization of tourist reviews. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1825–1828, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401269. URL https://doi.org/10.1145/3397271.3401269.

B. M. Napoleão, K. R. Felizardo, É. F. de Souza, and N. L. Vijaykumar. Practical similarities and differences between systematic literature reviews and systematic mappings: a tertiary study. In *International Conference on Software Engineering and Knowledge Engineering*, 2017.

A. Nerghes and J.-S. Lee. Narratives of the refugee crisis: A comparative study of mainstream-media and twitter. *Media and Communication*, 7, 09 2019. doi: 10.17645/mac.v7i2.1983.

NetworkX Library. NetworkX Library. https://networkx.org/, 2023. [Online; accessed October 2022 - June 2023].

S. Neuhaus. Research guides: Scholar commons guide: Types of open access. 2019.

NLTK Library. NLTK Library. https://www.nltk.org/, 2023. [Online; accessed October 2022 - June 2023].

Z. Nouri, H. Wachsmuth, and G. Engels. Mining crowdsourcing problems from discussion forums of workers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6264–6276, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.551. URL https://aclanthology.org/2020.coling-main.551.

PDF Search. PDF Search. https://pdfsearch.app/, 2023. [Online; accessed October 2022 - June 2023].

M. Pergher and B. Rossi. Requirements prioritization in software engineering: A systematic mapping study. In *2013 3rd International Workshop on Empirical Requirements Engineering (EmpiRE)*, pages 40–44, 2013. doi: 10.1109/EmpiRE.2013.6615215.

G. Pergola, L. Gui, and Y. He. A disentangled adversarial neural topic model for separating opinions from plots in user reviews. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2870–2883, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.228. URL https://aclanthology.org/2021.naacl-main.228.

C. L. Perryman. Mapping studies. *J Med Libr Assoc*, 104(1):79–82, jan 2016. doi: https://doi.org/10.3163/1536-5050.104.1.014.

K. Petersen, S. Vakkalanka, and L. Kuzniarz. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64:1–18, 2015. ISSN 0950-5849. doi: https://doi.org/10.1016/j.infsof.2015.03.007. URL https://www.sciencedirect.com/science/article/pii/S0950584915000646.

C. Popa and T. Rebedea. BART-TL: Weakly-supervised topic label generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1418–1425, Online, Apr. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.121. URL https://aclanthology.org/2021.eacl-main.121.

M. Riaz, T. Breaux, and L. Williams. How have we evaluated software pattern application? a systematic mapping study of research design practices. *Information and Software Technology*, 65:14–38, 2015. ISSN 0950-5849. doi: https://doi.org/10.1016/j.infsof.2015.04.002. URL https://www.sciencedirect.com/science/article/pii/S0950584915000774.

R. R. Robey and S. D. Dalebout. A tutorial on conducting meta-analyses of clinical outcome research. *Journal of Speech, Language, and Hearing Research*, 41(6):1227–1241, dec 1998. doi: 10.1044/jslhr.4106.1227. URL https://doi.org/10.1044%2Fjslhr.4106.1227.

J. Roeder, M. Palmer, and J. Muntermann. Data-driven decision-making in credit risk management: The information value of analyst reports. *Decision Support Systems*, 158:113770, 2022. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2022.113770. URL https://www.sciencedirect.com/science/article/pii/S0167923622000410.

I. Rožanc and M. Mernik. Chapter three - the screening phase in systematic reviews: Can we speed up the process? volume 123 of *Advances in Computers*, pages 115–191. Elsevier, 2021. doi: https://doi.org/10.1016/bs.adcom.2021.01.006. URL https://www.sciencedirect.com/science/article/pii/S0065245821000310.

Sakshi and V. Kukreja. Recent trends in mathematical expressions recognition: An lda-based analysis. *Expert Systems with Applications*, 213:119028, 2023. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2022.119028. URL https://www.sciencedirect.com/science/article/pii/S0957417422020462.

I. Savin, S. Drews, and J. van den Bergh. Free associations of citizens and scientists with economic and green growth: A computational-linguistics analysis. *Ecological Economics*, 180:106878, 2021. ISSN 0921-8009. doi: https://doi.org/10.1016/j.ecolecon.2020.106878. URL https://www.sciencedirect.com/science/article/pii/S0921800920309484.

ScienceDirect (Elsevier). ScienceDirect (Elsevier) Website. https://www.sciencedirect.com/, 2023. [Online; accessed October 2022 - June 2023].

C. C. Silva, M. Galster, and F. Gilson. Topic modeling in software engineering research. *Empirical Software Engineering*, 26(6):120, 2021. doi: 10.1007/s10664-021-10026-0. URL https://doi.org/10.1007/s10664-021-10026-0.

A. Singh and A. Glińska-Neweś. Modeling the public attitude towards organic foods: a big data and text mining approach. *Journal of Big Data*, 9(1), jan 2022. doi: 10.1186/s40537-021-00551-6. URL https://doi.org/10.1186%2Fs40537-021-00551-6.

SJR. Scimago Journal & Country Rank Website. https://www.scimagojr.com/, 2023. [Online; accessed October 2022 - June 2023].

H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269, jul 1973. doi: 10.1002/asi.4630240406. URL https://doi.org/10.1002%2Fasi.4630240406.

D. Solomon. Types of open access publishers in scopus. *Publications*, 1:16–26, 06 2013. doi: 10.3390/publications1010016.

Z. Song, Y. Hu, A. Verma, D. L. Buckeridge, and Y. Li. Automatic phenotyping by a seed-guided topic model. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 4713–4723, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3542675. URL https://doi.org/10.1145/3534678.3542675.

I. Sorodoc, J. H. Lau, N. Aletras, and T. Baldwin. Multimodal topic labelling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 701–706, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2111.

SpringerLink. Springer Website. https://link.springer.com/, 2023. [Online; accessed October 2022 - June 2023].

S. Syed and M. Spruit. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 165–174, 2017. doi: 10.1109/DSAA.2017.61.

F. Tang, L. Fu, B. Yao, and W. Xu. Aspect based fine-grained sentiment analysis for online reviews. *Information Sciences*, 488:190–204, 2019. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2019.02.064. URL https://www.sciencedirect.com/science/article/pii/S0020025519301872.

C.-O. Truică and E. S. Apostol. Tlatr: Automatic topic labeling using automatic (domain-specific) term recognition. *IEEE Access*, 9:76624–76641, 2021.

J. A. Wahid, L. Shi, Y. Gao, B. Yang, L. Wei, Y. Tao, S. Hussain, M. Ayoub, and I. Yagoub. Topic2labels: A framework to annotate and classify the social media data through lda topics and deep learning models for crisis response. *Expert Systems with Applications*, 195:116562, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2022.116562. URL https://www.sciencedirect.com/science/article/pii/S0957417422000604.

J. Wang and C.-C. Hsu. A topic-based patent analytics approach for exploring technological trends in smart manufacturing. *Journal of Manufacturing Technology Management*, 32(1):110–135, sep 2020. doi: 10.1108/jmtm-03-2020-0106. URL https://doi.org/10.1108%2Fjmtm-03-2020-0106.

S. Wang, L. Thompson, and M. Iyyer. Phrase-BERT: Improved phrase embeddings from BERT with an application to corpus exploration. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10837–10851, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.846. URL https://aclanthology.org/2021.emnlp-main.846.

M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), mar 2016. doi: 10.1038/sdata.2016.18. URL https://doi.org/10.1038%2Fsdata.2016.18.

S. Wu, F. Wu, Y. Chang, C. Wu, and Y. Huang. Automatic construction of target-specific sentiment lexicon. *Expert Systems with Applications*, 116:285–298, 2019. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2018.09.024. URL https://www.sciencedirect.com/science/article/pii/S0957417418306018.

Z. Xiang, Q. Du, Y. Ma, and W. Fan. A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, 58:51–65, 2017. ISSN 0261-5177. doi: https://doi.org/10.1016/j.tourman.2016.10.001. URL https://www.sciencedirect.com/science/article/pii/S0261517716301807.

H. Xu, J. Winnink, Z. Yue, Z. Liu, and G. Yuan. Topic-linked innovation paths in science and technology. *Journal of Informetrics*, 14(2):101014, 2020. ISSN 1751-1577. doi: https://doi.org/10.1016/j.joi.2020.101014. URL https://www.sciencedirect.com/science/article/pii/S175115771930210X.

X. Xu. How do consumers in the sharing economy value sharing? evidence from online reviews. *Decision Support Systems*, 128:113162, 2020. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2019.113162. URL https://www.sciencedirect.com/science/article/pii/S0167923619301915.

M. Yang and C. Han. Revealing industry challenge and business response to covid-19: a text mining approach. *International Journal of Contemporary Hospitality Management*, 33(4):1230–1248, mar 2021. doi: 10.1108/ijchm-08-2020-0920. URL https://doi.org/10.1108%2Fijchm-08-2020-0920.

W. Yang, J. Boyd-Graber, and P. Resnik. Adapting topic models using lexical associations with tree priors. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1901–1906, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1203. URL https://aclanthology.org/D17-1203.

Q. Yin, Z. Wang, Y. Song, Y. Xu, S. Niu, L. Bai, Y. Guo, and X. Yang. Improving deep embedded clustering via learning cluster-level representations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2226–2236, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.195.

C. Zhang, F. Tao, X. Chen, J. Shen, M. Jiang, B. Sadler, M. Vanni, and J. Han. Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2701–2709, New York, NY, USA, 2018a. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3220064. URL https://doi.org/10.1145/3219819.3220064.

W. Zhang, Y. Li, and S. Wang. Learning document representation via topic-enhanced lstm model. *Knowledge-Based Systems*, 174:194–204, 2019. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2019.03.007. URL https://www.sciencedirect.com/science/article/pii/S0950705119301182.

Y. Zhang, H. Chen, J. Lu, and G. Zhang. Detecting and predicting the topic change of knowledge-based systems: A topic-based bibliometric analysis from 1991 to 2016. *Knowledge-Based Systems*, 133:255–268, 2017. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2017.07.011. URL https://www.sciencedirect.com/science/article/pii/S0950705117303271.

Y. Zhang, J. Lu, F. Liu, Q. Liu, A. Porter, H. Chen, and G. Zhang. Does deep learning help topic extraction? a kernel k-means clustering method with word embedding. *Journal of Informetrics*, 12(4):1099–1117, 2018b. ISSN 1751-1577. doi: https://doi.org/10.1016/j.joi.2018.09.004. URL https://www.sciencedirect.com/science/article/pii/S1751157718300257.

R. Zhao, L. Gui, G. Pergola, and Y. He. Adversarial learning of Poisson factorisation model for gauging brand sentiment in user reviews. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2341–2351, Online, Apr. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.199. URL https://aclanthology.org/2021.eacl-main.199.

N. Zhou and D. Jurgens. Condolence and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 609–626, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.45. URL https://aclanthology.org/2020.emnlp-main.45.

E. Zosa, L. Pivovarova, M. Boggia, and S. Ivanova. Multilingual topic labelling of news topics using ontological mapping. In M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, and V. Setty, editors, *Advances in Information Retrieval*, pages 248–256, Cham, 2022. Springer International Publishing. ISBN 978-3-030-99739-7.