unibz
Fakultät für Ingenieurwesen
Facoltà di Ingegneria
Faculty of Engineering

Msc in Computational Data Science (Data Analytics)

Msc Thesis

Enhancing the methodological approach to Systematic Literature Reviews: A practical application on Topic Labeling research

Candidate: Samuele Ceol

Supervisor: Barbara Russo

July 2023

Abstract

The creation of a Systematic Literature Review (SLR) is often a challenging endeavor which requires both a deep understanding of the topic under scrutiny and the ability to satisfactorily summarise the insights found in the collected research whilst, at the same time, being able to identify any potential gaps that might surface from the conducted analysis. The difficulty associated with this task can be especially prominent when dealing with relatively underexplored domains, where the lack of dedicated secondary studies might further complicate the process of defining the appropriate methodological structure required to devise a grounded review procedure.

This work provides what the authors believe to be the first SLR entirely dedicated to exploring research tied to the task of Topic Labeling. In addition to the straightforward contribution associated with providing a comprehensive summary of the insights derived from the relevant primary studies in the period 2017-2022, the many considerations made throughout the review process allowed to formulate a set of methodological contributions capable of enhancing established guidelines for conducting secondary studies and which are believed to be generally applicable by authors operating outside of the specific boundaries imposed by the domain analysed within the context of this work. These contributions, ranging from the use of existing quality metrics in the Venue Selection and Snowballing phases, to the introduction of suitable proximity operators in the employed Search Strategy, should all play a role in aiding prospective researchers in the development of their respective Systematic Reviews.

Ultimately, the findings and observations derived from the included 108 primary studies are broken down and analysed with regards to the four posed Research Questions, which seek to explore the current state-of-the-art associated with Topic Labeling research on the basis of: (1) The identified macro-categories of labeling approaches and the specific techniques characterising each category, together with the motivational context provided by authors for justifying the employment a labeling step. (2) The topic modeling activities conducted to generate the underlying topic distributions. (3) The structure of the generated identifiers and the implemented label selection and quality evaluation procedures. (4) The categories of corpora (and related document types) exploited as the basis for the encountered analyses. Additionally, data associated with this initial synthesis is used as a means to identify notable shortcomings arising from the collected studies, which are ultimately found to take the shape of three distinct Research Gaps. Lastly, a few insights derived from individual documents and selected on the basis of their accessibility and reproducibility are highlighted as a way to guide readers in the task of identifying pre-existing labeling techniques that could conceptually be utilised in their respective research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1. 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 (Chen et al., 2016; Boyd-Graber et al., 2017; Silva et al., 2021). 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 and to provide a strucured overview of the existing research in the field capable of synthesizing findings derived from the various studies, identifying the strengths and weaknesses of existing methods, and highlighting the gaps in the current literature (Cooper, 1988; Kitchenham, 2004).

Additionally, the strong **focus on methodological rigor** adopted throughout the presented work, initially stemming from the general desire of the authors of properly grounding each decision made throughout the review process, ultimately led to the creation of a set of contributions which are considered to be generally applicable outside the boundaries of the explored research and the specific domain under scrutiny. Therefore, a further motivation for conducting this work lies in the provision of a suitable set of methodological pointers that prospective researchers might be able to employ in the context of supporting their respective work, especially in scenarios (such as the one described in this document) lacking previously existing Systematic Reviews which (in more a traditional setting) might act as the initial basis for the produced work.

1.2. Objectives (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 stud-

ied. 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. How are the underlying topic modelling phases characterised and how do they relate to the chosen topic labeling approaches?** 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 the encountered labels generally structured, how are candidates 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 corpora 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 of interest. This means that the collected research is likely to conduct the relevant labeling activities on topics generated starting from a vastly heterogeneous set of corpora. The aim of this research question is therefore to analyse which are the prevalent categories of corpora (and associated documents) in the included work and how such categories might reflect on the subsequent labeling activities.

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.

1.3. Main contributions

In this Section, the principal (novel) contributions tied to the work conducted throughout this review are briefly summarised and described. Notice that the items included in the following list are representative of those activities that were specifically devised for this work and that are generally found to present substantial differences (or to be fundamentally absent) from already existing (comparable) studies. In other words, common activities that can be expected to be present in most literature reviews (e.g. the definition of inclusion and exclusion criteria, the establishment of start & end dates and of language limitations for the collected work) will not be considered in this regard. The main contributions associated with the presented work can therefore be summarised in the following activities:

1. Adapting the SEGRESS structure for secondary studies proposed in Kitchenham et al. (2022) to better suit the requirements tied to exploring existing research on topic labeling.
2. Establishing a grounded **Venue Selection Approach** capable of accounting for:
    - Data collected from an initial exploratory search conducted on selected repositories using multiple specifically defined queries.
    - Established Venue quality metrics.
3. Defining a detailed multi-step **Search Strategy** exploiting advanced queries with wildcards and proximity operators tied to topically relevant terms.
    - Formally highlighting the impact of the imposed proximity constraint on the set of selected studies.
4. Detailing the methodology tied to the two **(Backward and Forward) Snowballing** phases via a five step procedure taking into account existing inclusion constraints and aspects tied to previously defined search criteria.
    - Including a further quality-based filtering task for documents extracted from the Forward Snowballing activities as a mean to prevent potential losses in quality in the final study selection.
5. Verifying the presence of potentially **predatory venues** among the collection of included studies.
6. Proposing a 20-item **Data Extraction Form** highlighting the relevant set of information (Data Items) tied to each Research Question.
7. Enhancing the **Data Collection Process** by means of external tools exploited to highlight salient portions (i.e. keywords) of the inspected work.
8. Structuring the **Analysis and Synthesis** process into a three-tier architecture where data collected from the included documents is explored in terms of:
    - The findings inferable from the information tied to the defined Data Items (which are ultimately used to address the posed **Research Questions**).
    - The shortcomings such findings are capable of highlighting and which are ultimately expressed as a series of relevant **Research Gaps**.
    - Individual **Insights** (and methodologies) extracted from a subset of selected studies.
        - The provision of such insights is further contextualised with regards to the level of **accessibility** of the interested documents and the **reproducibility** of the associated methodologies.
9. Setting up an initial methodological structure for producing literature networks tied to the collected studies.

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 (SE) 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 exist-

ing 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 1.1). This rationale is closely associated with one or more **Research Questions** (Section 1.2) 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 5) 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 SE. 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. Guidelines for conducting Secondary Studies

In a general sense, a valuable aid in establishing the general structure for a given Systematic Literature Review can come in the form of SLR guideline documents, which are studies capable of offering a set of pointers with the purpose of clarifying how a given secondary study should be reported and what should be the relevant content tied to each of the highlighted sections making up the final report. As a general premise, one should understand that these documents are (by definition) generic in nature, and their guidelines should be adapted to the individual requirements tied to the specifics of the domain under scrutiny. Nonetheless, the existence of well-structured instructions for producing secondary studies can be especially useful in scenarios (such as the one presented here) where previously existing Review work associated with a given topic is lackluster or completely absent.

As previously stated, an initial understanding tied to the structural content of the proposed secondary study was developed following the suggestions found in Kitchenham (2004) - *"Procedures for Performing Systematic Reviews"*. In this article, B. Kitchenham collects the guidelines found in three distinct documents (Collaboration, 2003; Health et al., 2000; Dissemination, 2001) stemming from the medical domain and adapts them with the purpose of increasing their relevance for SE researchers. In this context, after underlining the role of SLRs in providing exhaustive summaries of the existing evidence (see Subsection 2.1.2), the author organises the general structure of the review process into three phases. The detail associated with each phase is described as follows:

- **Planning** the Review refers to the set of activities associated with identifying the need (i.e. rationale) for the secondary study (Subsection 1.1) and with the development of a suitable Review Protocol (Chapter 3) highlighting the detailed methodological shape characterising the given work and introducing the Research Questions that the Review needs to ultimately answer.
- **Conducting** the Review fundamentally means executing (in a practical manner) the various activities methodologically defined within the Review Protocol. This includes gathering the relevant documents by means of the defined Search Strategy, filtering them via the established inclusion and exclusion criteria, populating the relevant Data Extraction Forms by inspecting the individual (selected) primary studies and synthesizing the collected information with regards to the posed Research Questions.
- **Reporting** the Review ties to the notion of structuring (and presenting) the collected information within the context of the final document. Here, Kitchenham proposes a 13-item structure for reporting secondary studies.

In the context of this work, the description of the various elements associated with the Planning and Conducting phases offered valuable insights exploited for the initial organisation of the work associated with the presented review process. On the other hand, the specific report structure provided at the end of the document has been (for the most part) ignored and the structural guidelines provided in a subsequent publication (described below) have been used instead.

In a practical sense, the organisation of the presented report followed in large part the structure (and related descriptions) found within the **SEGRESS checklist** provided in Kitchenham et al. (2022) - *"SEGRESS: Software Engineering Guidelines for REporting Secondary Studies"*. SEGRESS was developed as a way to address common criticisms brought forward by recent tertiary studies with regards to SLRs conducted in the SE domain. Here, the provided checklist can be seen as an extension of the one found within the **PRISMA** (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement (Page et al., 2021), which represents the de facto standard for reporting secondary studies in the healthcare domain. The highly rigorous development process that characterised the creation of PRISMA and the fact that it allows to employ ***"terminology aligned with other empirical disciplines"*** are both used as a justification for choosing it as the basis for the proposed guidelines. Additionally, in this regard the authors state that: *"Since the original SE guidelines for software engineering systematic reviews (Kitchenham, 2004) were based on healthcare guidelines, it seems plausible that PRISMA 2020 could be of use to SE researchers"*. Ultimately, whilst sharing the same set of overarching checklist items, the need to propose SEGRESS as an extension of the original PRISMA statement stems from the fact that PRISMA is conceptually designed to address Quantitative Reviews, Mixed-Method studies and Meta-Analyses, but its suitability for **Qualitative Reviews** and **Systematic Mapping Studies** is not taken into consideration. To address this shortcoming, the authors incorporate relevant guidelines for Mapping Studies (Tricco et al., 2018) and for Qualitative Reviews (Tong et al., 2012; Wong et al., 2013) in the description of each PRISMA Item characterising the SEGRESS checklist. Additionally, it is noted that the PRISMA 2020 statement might not be intuitive to understand for researchers operating outside of the medical domain. In this context, SEGRESS provides a more readable checklist for non-healthcare practitioners and further supplements it with examples (found in the Supplementary Material) extracted from SE literature.

Ultimately, whilst recognising that the proposed Systematic Literature Review does not strictly pertain to the domain of Software Engineering, it is believed that the pointers provided within the individual items of SEGRESS mostly maintain their relevance in the explored sub-domain and, in any case, offer a more easily interpretable set of guidelines than the (healthcare-centric) ones found within PRISMA 2020. Furthermore, the previously stated advantages tied to utilising a PRISMA-derived checklist, together with the (mostly) qualitative nature of the presented work, lead the authors to believe that SEGRESS does indeed represent a suitable option for structuring the presented work.

2.2.2. Secondary Studies on Topic Labeling

In this section, existing secondary studies describing potentially relevant research tied to the presented Review are referenced and briefly summarised. As initially stated in Section 1.1, one should note that Topic Labeling still represents somewhat of a niche activity within the broader boundaries defined by Topic Modeling research and that, to the best of the authors knowledge, no prior Systematic Literature Review solely focusing on this broad category of tasks exists at the present time. On the one hand, this notion further validates the potential significance of this work in providing a (first) systematic analysis on the available state-of-the-art characterising this particular sub-domain. Nonetheless, this fact also makes the activity of collecting information from existing secondary studies somewhat more involved, as one is not able rely on dedicated publications as a mean of deriving useful information for assessing the latest methodologies and techniques. In this regard, the search conducted for the purpose of finding relevant Related Work takes into consideration reviews dedicated to describing different facets of Topic Modeling research which also offer (by means of dedicated Sections) some attention with regards to how the included work deals with the task of naming the generated topic distributions.

As a starting point in this exploration, the secondary study by Boyd-Graber et al. (2017) offers a

broad overview of the latest (academic and industrial) applications of topic modeling approaches and provides a few considerations on the current state of labeling techniques. Here, (automated) labeling is organised into two broad categories identified on the basis of the information exploited during the label generation procedure. In this context, the authors broadly identify two families of methodologies: Those using solely internal information (derived directly from the topic model) and those relying on external resources. Ultimately, this initial observation leads to the further definition of four (more granular) categories which recognise:

- **Internal labeling** approaches that use information tied to the topics (i.e. terms prominence) and additional context given by relevant documents as means for extracting suitable (phrase) labels from a given distribution. Examples of such internal labeling procedures are found in Mei et al. (2007) and Mao et al. (2012).
- **Labeling with Supervised Labels**, referring to those (supervised) approaches were weights are learned starting from a labeled datasets containing topics associated with pre-existing gold-standard identifiers. The generated models are then capable of tying unseen distributions to relevant topic labels observed during training.
- **Labeling with Knowledge Bases**, which exploit information contained in existing structures such as ontologies or graphs as the source for labeling topics.
- The **use of Labeled Documents** to devise topic modeling approaches, such as Labeled LDA (Ramage et al., 2009), capable of incorporating the identifiers into the output of the modeling phase.

In the presented review, the techniques derived from the collected research are similarly divided into an initial set of coarse-grained categories (Subsection 5.2.1) and then further explored by means of the respective underlying techniques characterising them (Subsection 5.2.4).

A second set of relevant insights is provided by the secondary study of Chen et al. (2016), where existing research is surveyed with a focus on the use of topic models in SE research. Here, topic labeling methodologies are explored in the chapter dedicated to common pitfalls identified among the collected studies. In this context, several documents are found to provide comparisons between automatically generated and **manually produced identifiers** for the purpose of labeling code artifacts (De Lucia et al., 2012, 2014) or execution traces (Medini et al., 2012). A similar kind of analysis is conducted for documents taking part in the presented Review and is ultimately highlighted in Subsection 5.4.2, where the observed quality evaluation procedures are described in detail and are often found to associate automatically produced topic labels with their human-generated counterpart. Additionally, when examining the included work, Chen et al. (2016) finds evidence in support of the idea that conducting labeling activities on the generated topics represents a valid way to make them more easily interpretable. In this context, Hindle et al. (2012) states that: *"One should use [...]* ***domain experts*** *to label topics. [In fact,] we found that non-experts have questionable labelling accuracy [and] respondents with the most familiarity gave the most relevant topic labels"*. The same observation tied to the use of domain experts in the labeling procedure is given as a way to address the second Research Gap described in Section 6.2. Ultimately, Chen et al. (2016) concludes the chapter by arguing that: *"labelling and interpreting topics can be difficult and subjective, and may require much human effort. Future studies should explore ways to apply different approaches to automatically label the topics"*.

Finally, the last of the identified secondary studies including a relevant Section on labeling approaches is represented by the work of Silva et al. (2021) which, in a similar fashion to Chen et al. (2016) also focuses on topic modeling research in the domain of SE. Here, one of the four posed research questions tries to answer how the generated topics were named in the documents making up the collected research. As a starting point in this endeavor, the authors are able to broadly identify three distinct ways for labeling topic distributions, and organise them into the following categories:

- **Automated** techniques capable of naming distributions without any intervention from human beings.
- **Manual** approaches where the label is entirely generated by assessors inspecting the topic terms.

- **Manual & Automated** where manual labeling is applied for generating the training data ultimately used to build the (automated) classifiers.

A very similar observation is made for the documents inspected within the presented Review, which ultimately led to the identification of the three macro-categories of approaches described in Subsection 5.2.1. Here, one can observe a fundamental difference between the Manual & Automated scenario of Silva et al. (2021) and the Partially Automated category defined in the presented Review, which does not ground its characterisation on the conducted training and testing activities, but instead identifies approaches in this (middle-ground) category as: *"all those techniques that are able to partially offset the labeling efforts required by the traditional manual labeling approaches, but that are still dependent (to various degree) on human intervention"*. When reviewing the collected research, the authors of Silva et al. (2021) are able to observe that most of the interested documents (focusing on topic modeling) do not conduct any activity for generating topic names. Additionally, one documents ties the decision of using a manual naming approach to a **lack of existing approaches** capable of satisfying the same requirement automatically. For this exact reason, Chapter 7 tries to highlight some (reproducible) topic labeling techniques that prospective authors might be able to use to extend their work. A sizable number of papers also highlight potential limitations of manual labeling on the basis of the additional effort required to conduct this task and the potential threats to validity brought upon the inclusion of a **human component** in the process. Similar considerations tied to the risk of bias and subjectivity when dealing with human assessors are also presented in this work (in the context of the identified quality evaluation procedures) in Section 6.2 - *"Considerations on human ratings"*. Ultimately, Silva et al. (2021) concludes the Section dedicated to the results tied to topic naming by highlighting the **recommendations** that one document (Hindle et al., 2015) gives with regards to activities conducted in this regard. These recommendations can be summarised as follows:

- Not all topics will always be relevant to the conducted research. Therefore, one should name only the ones that ultimately seem to contribute to the posed Research Questions.
- The involvement of domain experts is crucial in the production of meaningfuil identifiers.
- The topic labeling procedure should have strong ties with the data originally used to generate the topics. In fact: *"the content of the topic can be interpreted in many different ways and LDA does not look for the same patterns that people do"*.

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.

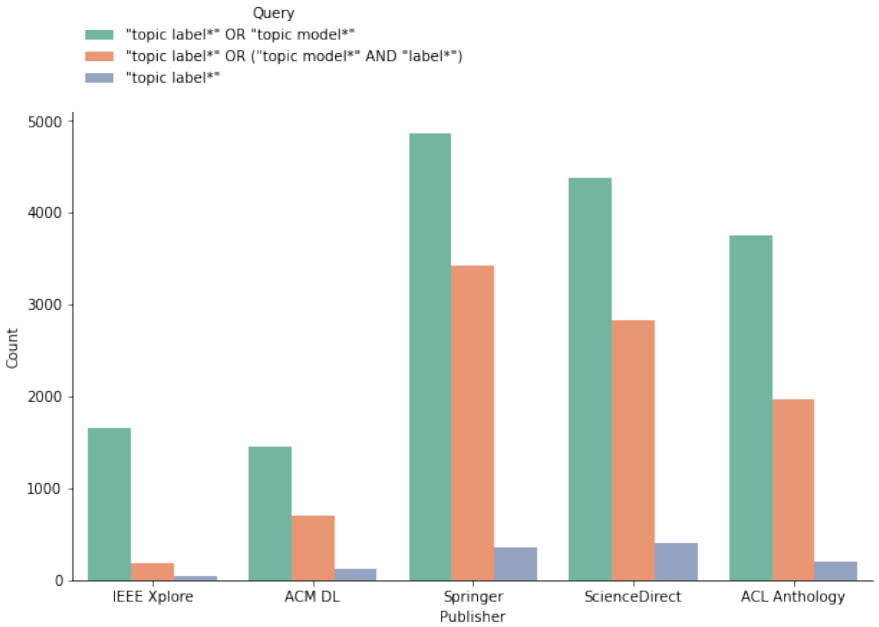


Figure 3.1: Exploratory search results

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)

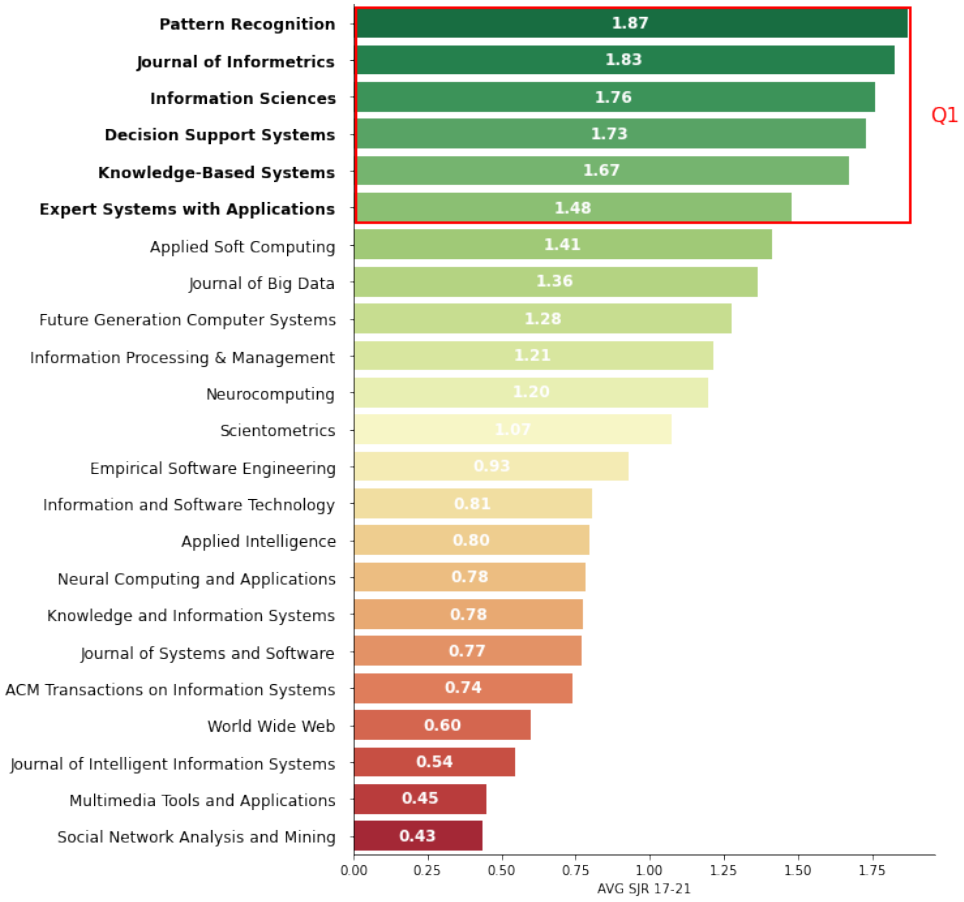


Figure 3.2: Set of (23) candidate journals

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:

```
"label*" NEAR "topic*"
```
(3.6)

On the papers gathered from the query:

```
"topic label*" OR ("topic model*" AND "label*")
```
(3.7)

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

```
"topic label*" OR ("topic model*" AND ("label*" NEAR "topic*"))
```
(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:

```
[{20} "topic*" "label*"]
```
(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 D 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. (2019a), 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 20 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 1.2. 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 14-17) 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 12-13). 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 (18-20) offers a broader view of the work under review. In this context, such items store the information related to the encountered corpora and documents and to the conducted pre-processing activities.

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	Context for applying a labeling step	RQ1	✗
12	Topic modeling	Topic modeling approach(es)	RQ2	✓
13	Nr. of topics	Nr of topics generated from the corpus	RQ2	✓
14	Label	Label structure (e.g. n-gram, sentence)	RQ3	✓
15	Label selection	Nr of candidates per topic & approach(es) for selection	RQ3	✗
16	Label quality evaluation	Quality metric(s) for label evaluation	RQ3	✗
17	Assessors	Number and details of the assessors involved in the selection and evaluation	RQ3	✗
18	Corpus	Type, Origin and size of the corpus	RQ4	✓
19	Document	Format of individual documents in the corpus	RQ4	✗
20	Pre-processing	Pre-processing steps performed on documents	RQ4	✓

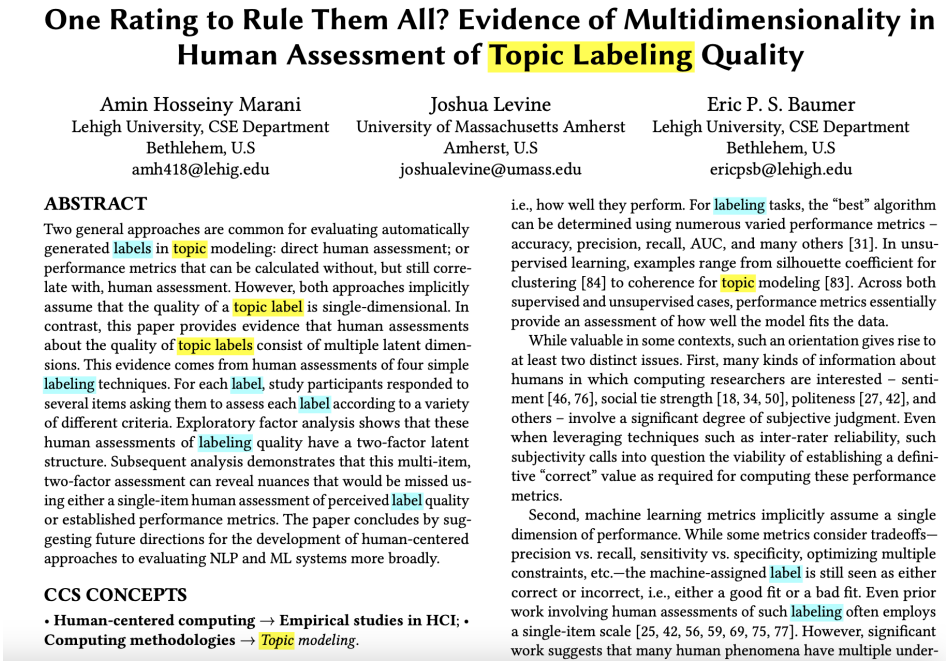
Table 3.3: Data extraction form

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 consis-*

tent 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. Structuring the identification of research gaps

One of the natural consequence of carrying out the tasks of reviewing and subsequently synthesising primary studies is the identification of potential voids in the scrutinised research. Such gaps might not necessarily adhere to individual data items used to summarise the content of the collected research and (as previously stated) might even present a level of abstraction over the overarching Research Question. Nonetheless, the identification and successive description of these gaps is a central step in providing a comprehensive description of the area of research under review. To this point, Robinson et al. (2011) states that presenting such gaps is an expected **output** of all literature reviews, and underlines how they play an important role for prospective authors aiming to conduct further research in the field. In relation to defining the origin of research gaps, the study states that:

> *"The consideration of existing evidence often highlights important areas where deficiencies in information limit our ability to make decisions. When the ability of the systematic reviewer to draw conclusions is limited, we have [...] a 'research gap.' "*

Nevertheless, whilst helpful to draw a clearer picture of what research gaps represent and how they generally arise, this insight alone is not sufficient to structure a rigorous gap analysis. In a general sense, a further characterisation of such research gaps and a more structured set of **best practices** are necessary to better understand how a reviewer might approach the process of research gap identification and extraction from the data associated by the analysis and synthesis of the collected studies. To address this issue, and by integrating notions found in Robinson et al. (2011) and Jacobs (2011), Müller-Bloch and Kranz (2015) (in their work titled *"A framework for rigorously identifying research gaps in qualitative literature reviews"*) proposes a **framework** for identifying research gaps in the context of qualitative reviews. This framework, built on the basis of data extracted from 40 literature reviews, is structured into the following **four activities** (or components):

> Localisation, Characterisation, Verification and Presentation.

The first component, called **Localisation**, naturally refers to the activities associated with the first identification of research gaps. In this context, the authors argue that such localisation should already commence during the synthesis phase, when the identified research is being scrutinised for the first time. Here, the study synthesis and research identification activities should be seen as being **intertwined**. This is made possible by the fact that:

> *"When researchers start to chart the reviewed literature according to various concepts [data items], they already perceive blank fields in the chart [data collection form], which may indicate research gaps."*

Therefore, one should realise that the process of identifying relevant gaps should start as soon as possible during the analysis and synthesis phase, and reviewers should be aware that profiling gaps should not happen (in its entirety) after the selected papers have been fully scrutinised.

In close relation to localisation, the notion of **Characterisation** refers to the idea that gaps exists within a set of pre-defined **classes**, which are defined on the basis of the individual gap's originating factors. The rationale behind providing such classes is that, once identified, they should provide useful guidance with regards to the specifics of the **localisation strategy** that should be employed for identifying gaps belonging to the different categories. In this regard, it is easy to understand how these first two tasks are highly connected with one another. In fact:

> *"Different types of research gaps require different localization strategies [...]. Thus, the concept of characterization is capable of enhancing the process of localizing research gaps. When trying to localize research gaps, researchers can constantly refer to [it]. [...] The two concepts [...] are also related vice versa: The characterization of research gaps might enable scholars to specify what kind of research is required to resolve the respective [...] gap."*

Starting from the notions of Jacobs (2011), which offer a classification of gaps (referred to in the document as *"research problems"*) into six categories, Müller-Bloch and Kranz (2015) provides a set of suitable localisation strategies associated to each of the respective categories (Table 3.4). In this context, such categories can be briefly defined as follows:

- A **methodological conflict** refers to those instances where the chosen methodology has an influence on the produced results.
- **Contradictory evidence** occurs when individual studies allow for conclusions that reveal themselves to be contradictory when explored from a more abstract point of view.
- **Knowledge voids** are simply research findings that do not yet exists.
- An **action-knowledge conflict** is realised when the advocated practices, methodologies or behaviors of authors differs from the actual behavior.
- **Evaluation voids** are research results (findings or propositions) that have not been empirically verified.
- A **theory application void** verifies when theoretical concepts are not applied to specific research issues that could lead to new insights.

Type of Research Gap	Localisation Strategy
Methodological conflict	Scrutinize if findings on a certain topic are inconclusive with regard to applied research methods.
Contradictory evidence	Synthesize key findings and determine contradictions.
Knowledge void	Analyze literature with regard to theoretical concepts and look for specific gaps or under-researched areas of research.
Action-knowledge conflict	Collect information about action and relate this information to the knowledge base.
Evaluation void	Analyze if research findings have been evaluated and empirically verified.
Theory application void	Analyze the theories that have been employed to explain certain phenomena and identify further theories that might contribute to the knowledge base as well.

Table 3.4: Localisation strategy associated with each type of gap

Here, notice that research gaps' classes tend to be relatively coarse grained and that the associated localisation strategies are only briefly defined. To this end, one should keep in mind that these serve only as a general support to the task of localisation and that they do not exclude a-priori the existence of other kinds of research gaps.

The third task, defined as **Verification**, refers to the activity of verifying that a given gap does indeed exist. Naturally, limitations imposed by the methodological process followed for the systematic review (like the venue selection presented in Subsection 3.1.3) might create a situation where the collected primary studies suggest the existence of gaps that reveal themselves to be already addressed (or not present at all) when examining the broader literature. In this context, a potential first solution for this issue is conducting a **forward search** starting from the documents where the gap has been identified. The rationale behind this suggestion stems from the assumption that subsequent studies that have addressed (and solved) a potential research gap are likely to cite the paper from which it originated. Additionally, it is also suggested that a **general search** in texbooks or repositories using relevant keywords might yield a similar result. Naturally, these searches might reveal solution for the gap at hand or allow the reviewers to find other studies where the gap has previously been described.

The last component, called **Presentation**, tries to provide some general details with regards to how an identified research gap should be presented within the review document. In this context, the two approaches of sequential and parallel presentation are identified. **Parallel presentation** refers to the activity of describing gaps when presenting the result of the analyses and syntheses. Here, gaps can be presented together with the sets of information from which they stem, allowing the reviewers

to more easily present why a given gap is in place and how it ties to the overall synthesis. On the other hand, when performing **Sequential presentation**, reviewers describe gaps in a separate section after the synthesis of primary studies (which, in this document, would be Chapter 6). This approach offers an easier way for readers to immediately locate the identified gaps in the document, but might make it more difficult to tie them to specific originating factors. Ultimately, Müller-Bloch and Kranz (2015) suggests that, for a given review, these two approaches should be applied in conjunction with one another. In fact, parallel and sequential presentation should generally **complement each other**, by highlighting the origin of the identified gaps and allowing them to be presented in an easily readable way.

The four components of the presented framework are visually highlighted in Figure 3.4 and will be used as a points of reference in this review to guide the process of identifying, extracting and presenting the relevant research gaps. To this end, it must be stated that:

> *"this procedure [of gaps identification] is often creative, implicit, and informal"* - Müller-Bloch and Kranz (2015)

Therefore, the presented concepts should be interpreted as general guidelines and best practices rather than strict methodologies to which reviewers need to adhere to in an absolute manner.

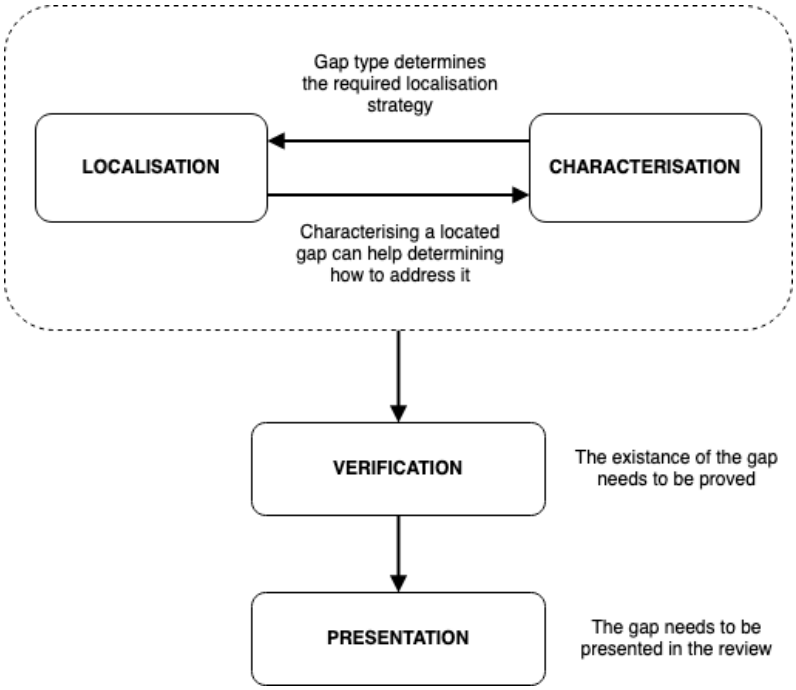


Figure 3.4: Structure of the gap identification framework

3.8.2. Evaluating insights from selected 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 prospective 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, prospective 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 accessible 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 prospective 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?
- **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.)?
- **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?
  - Is it conceptually possible to extend the provided resource? Is the provided usage license compatible with this possibility?

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 **55 papers** is generated on the basis of the established inclusion and exclusion criteria. This set of 55 papers (**31** of which are retrieved from the selected journals and **24** 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 55 papers (7% 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 **23 studies** is found to meet the imposed requirements and is therefore added to the systematic review. The 23 papers are found to belong each one to a different journal or conference. 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
- **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
- **Social Science**: Social Forces, Media and Communication, Research & Politics
- **Database Systems**: International Conference on Database Systems for Advanced Applications (DASFAA)
- **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 D.

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 55 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 58 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 58 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 58 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 **24** 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 (4, 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

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.

(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.

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 **30 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.

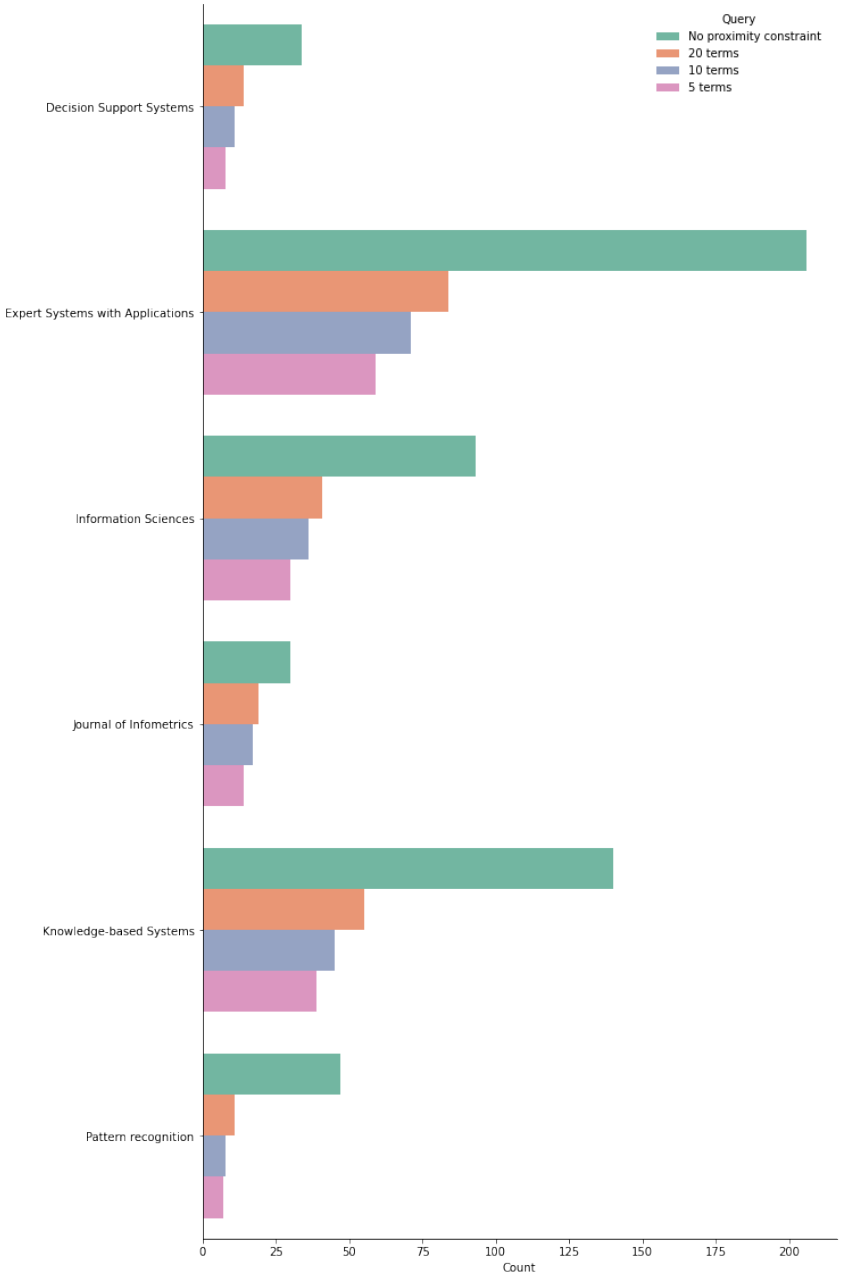


Figure 4.1: Effects of proximity constraints on journals

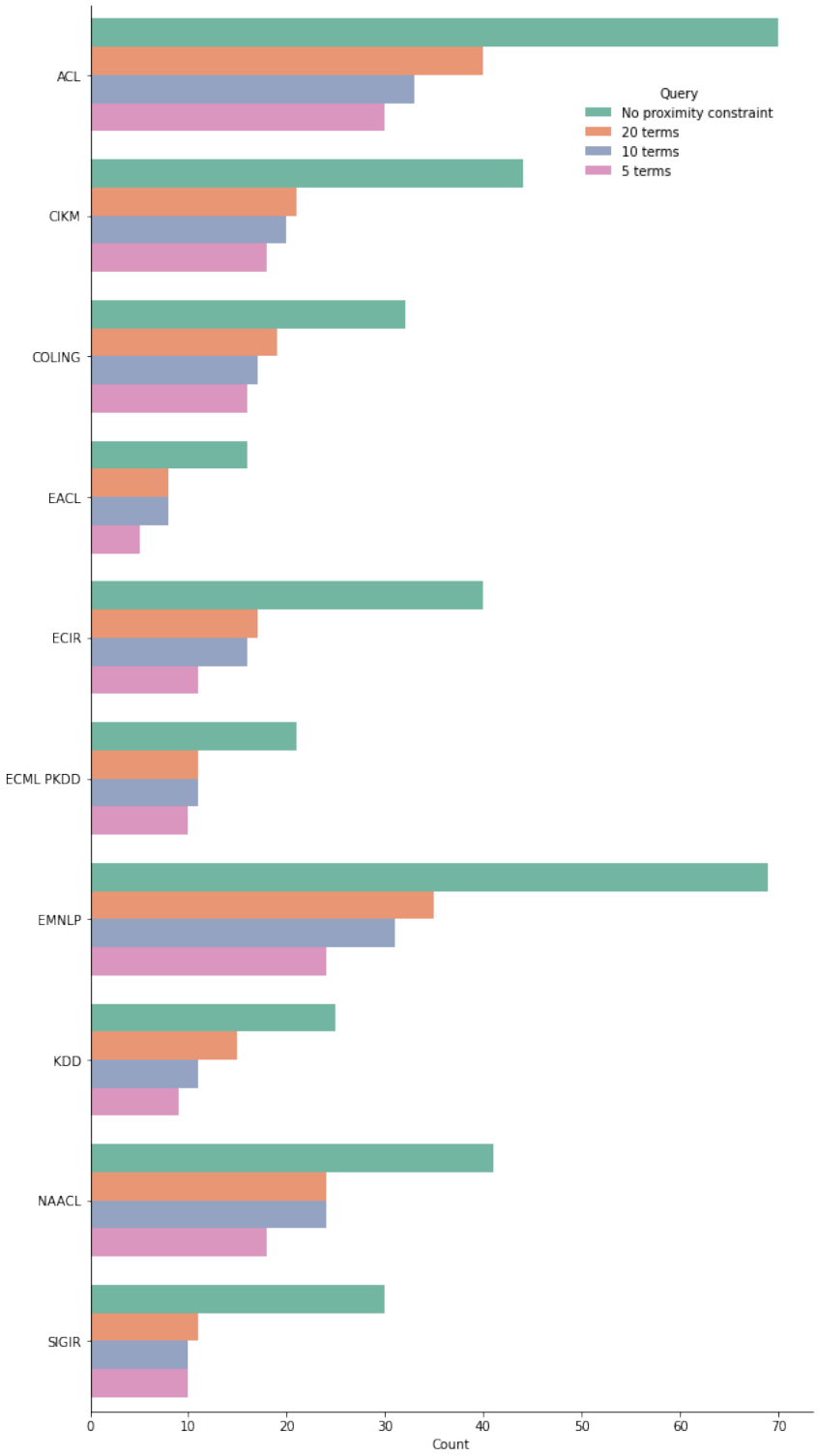


Figure 4.2: Effects of proximity constraints on conferences

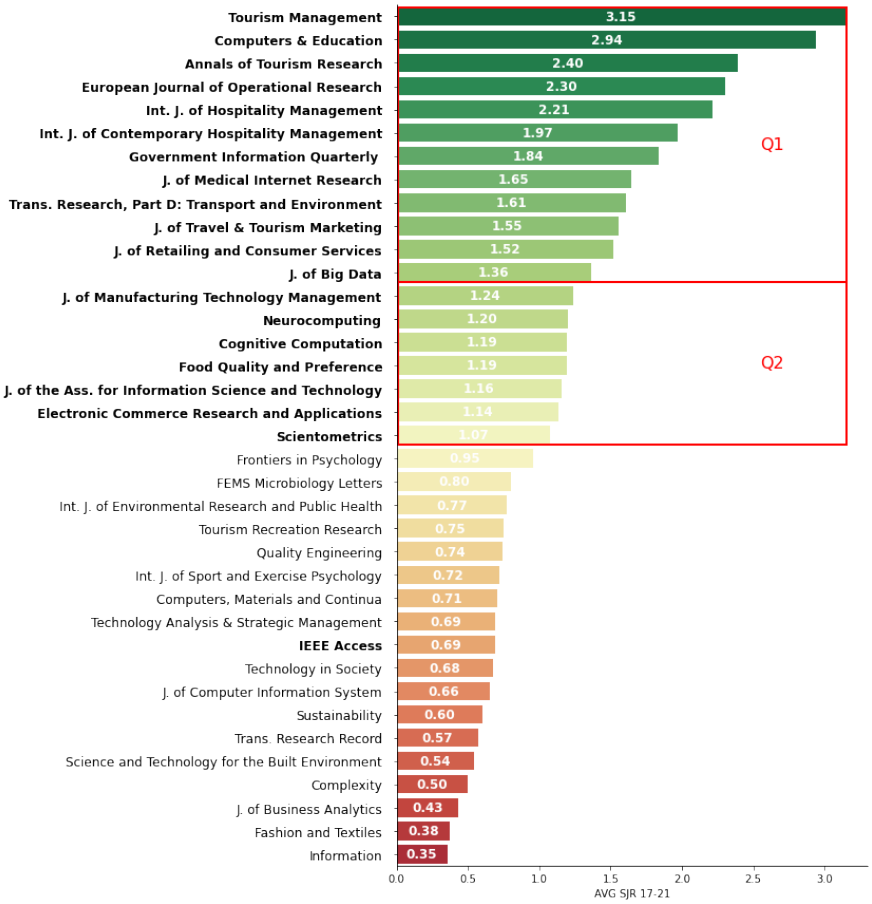


Figure 4.3: Set of (38) FS journals. Unranked journals are excluded.

Chapter 5

Results of Analyses and Syntheses

5.1. 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 5.1 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 three of the initially selected venues (ECML PKDD, Information Sciences and Pattern recognition) were 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 5.2 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
Decision Support Systems	0	0	2	0	2	0	4
EMNLP	1	0	0	2	1	0	4
ACL	1	1	1	0	0	0	3
COLING	0	1	0	1	0	1	3
EACL	1	0	0	0	2	0	3
KDD	0	1	0	2	0	0	3
SIGIR	1	0	0	2	0	0	3
ECIR	1	0	0	0	0	1	2
NAACL	0	0	0	0	2	0	2
CIKM	0	0	0	0	0	1	1
ECML PKDD	0	0	0	0	0	0	0
Information Sciences	0	0	0	0	0	0	0
Pattern recognition	0	0	0	0	0	0	0
Total	9	8	5	13	13	7	55

Table 5.1: Initial selection of papers by venue and year

phase (with a total of 9 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 13 to a maximum of 26 documents and the year 2021 shows the highest paper count (with a value 26 documents).

		Year						
	Venue Type	2017	2018	2019	2020	2021	2022	**Total**
Initial sel.	Conf.	5	3	1	7	5	3	24
	Journal	4	5	4	6	8	4	31
	Subtotal	9	8	5	13	13	7	55
Backward S.	Conf.	2	1	0	1	0	0	4
	Journal	7	3	6	2	1	0	19
	Subtotal	9	4	6	3	1	0	23
Forward S.	Conf.	0	0	1	0	3	1	5
	Journal	0	0	3	6	9	7	26
	Subtotal	0	0	4	6	12	8	30
	Total	18	12	15	22	26	15	108

Table 5.2: Selected papers by activity, venue type and year

5.2. RQ1 - Summary of topic labeling approaches

As stated in the introductory chapter to this work, the first Research Question associated with the presented review naturally revolves around the summarisation of the various topic labeling ap-

proaches utilised in the selected work throughout the chosen time period. In this regard, a total of seven data items (5 to 11 as presented in Table 3.3) are tied to such question and seek to provide not only a complete overview of the encountered methodologies (data item 5), but also an insight on various additional metrics characterising, in a broader sense, how these different approaches tie into the primary studies at hand. On this note, data item 6 highlights the role that such methodologies play within the context of each individual document by defining a first divide between work having a primary focus on (the evaluation or proposal of) topic labeling techniques and the remaining studies treating the technique as a somewhat secondary step exploited in pursue of a broader research objective. Continuing in this vein, data item 7 helps to further describe the collected research by drawing a clearer picture with regards to type of contribution associated with the implemented labeling step. In particular, a distinction is made between the proposal of novel approaches and the use of previously established ones. Data items 8, 9 & 10 delve further into the characterisation of the relevant approaches by providing an overview of the utilised techniques, the underlying parameters and the details of the label generation procedures. Especially for this last three items, it is important to underline that the main objective of this Section is to address the Research Questions by providing an holistic view of the analysed research. To do so, the focus is mostly oriented towards revealing overarching characteristics common to multiple studies in order to address Research Questions that are, by definition, broad in scope and the answers of which cannot be inferred from any individual publication. Therefore, one should expect a better (and especially more detailed) characterization of the individual approaches in Chapter 7, where the broad strokes of the more relevant techniques are presented and contextualised with regards to their level of accessibility and reusability. Finally, data item 11 explores the motivators (or lack thereof) that ultimately led researches to include a labeling step in their work. In this introduction, it must also be noted that the analyses performed in the context of this Subsection (and the subsequent ones) are foundational in identifying the gaps in the research that will be ultimately presented and discussed in Chapter 6.

5.2.1. Categories of topic labeling

With regards to the types of **topic labeling approaches**, this review proposes to organise such approaches into three coarse grained categories. Notice that at this point no further categorisation is provided with regards to the specific techniques characterising each approach, which are instead profiled in more details in the context of data items (8) to (10). The three identified categories referring the different approaches encountered when analysing the selected studies are: Manual, Partially Automated and Fully Automated labeling.

As one might expect, **Manual** approaches refer to the more traditional implementations of labeling activities which fully rely on human annotators that, by inspecting relevant information associated with the topic model at hand (generally the most prominent words associated a given topic and the documents where such topic is more prevalent), produce a meaningful label to characterise each topic.

On the other hand, **Partially Automated** approaches are more difficult to formally define, and represent a sort of middle ground between the aforementioned manual labeling and other techniques that are capable to fully automate the process. In the context of this review, partial automation is recognised in all those techniques that are able to partially offset the labeling efforts required by the traditional manual labeling approaches, but that are still dependent (to various degree) on human intervention. This not only includes (1) automated techniques requiring as input an already existing (potentially partial) label structure, but also (2) manual approaches guided by pre-existing label sets or (3) aided by other forms of automatically generated metadata useful in the generation or assignment of definitive labels (e.g. automatic generation of a summary sentence for each topic). Clearly, the choice of including (2) and (3) in this category might seem somewhat controversial, given the fact that the the label assignment is still ultimately being determined by human annotators. Despite this fact, it was still deemed to be necessary to formally highlight those studies where some effort was expended to simplify or optimise the traditional labeling procedure by introducing some level of automation in the process.

Finally, **Fully Automated** approaches refer to all those applications of topic labeling that do not

require any human intervention throughout the entirety of the labeling procedure. These can be generally identified as all those approaches that take as input the set of topics (i.e. the full probability distribution of terms over a dictionary or the top words associated with each topic) and output a descriptive label for each of the proposed distributions. Often times, these approaches are associated with advanced underlying machine learning algorithms or natural language processing techniques and are capable of automatically analysing the structure of the generated topics and to assign relevant labels on the basis of their textual features, without any input from human annotators.

A visual summary associated with the number of documents related to each of the approaches is highlighted in Figure 5.1. Starting from this initial representation, it is immediately noticeable how the most prominent category identified in the examined research is the Manual Labeling approach with a total of **67** studies presenting some form of manual labeling step (roughly half of the selected research). This is a somewhat expected result, considering that manual labeling does not strictly require any overly specific technique or prior experience in order to be applied and can be generally carried out directly by the respective authors or by domain experts especially selected for the task. In fact, the only practical requirement associated with manual labeling is a general knowledge of the domain of interest to which the generated topics belong to.

With regards to this first insight, it is also important to point out that one should not assume that studies associated with this somewhat simpler implementation of topic labeling hold no value for the presented work. In fact, the decision to have such a granular organisation of items proposed in the utilised Data Extraction Forms exists with the clear purpose of extracting all potentially relevant insights associated with the encountered labeling activities. Whilst certainly interesting (and clearly important in defining the nature of each individual study), the chosen labeling approach (together with the associated underlying technique and label generation process explored in the later items) is not the only determining factor shaping the collected research.

Moving to the Partially Automated approaches, it is possible to notice that they represent the smallest minority in the collected research and are found in only **14** studies. Once again, this result is not completely unexpected, and can likely be attributed to the fact that researchers wanting to move away from traditional manual labeling techniques might ultimately be more likely to end up choosing techniques that completely automate the process rather than relying only on partial automation approaches. This potential explanation will be further evaluated when exploring the motivating factors highlighted within the context of data item 11. Finally, it must also be pointed out that a total of **7** studies (García-Pablos et al., 2018; Liu et al., 2017; Karami et al., 2019; Maier et al., 2018; Amat-Lefort et al., 2022; Ding et al., 2020; Jebari et al., 2021) include both Manual and Partially automated approaches.

Finally, various techniques to fully automate the topic labeling task are observed in **36** distnct publications. Generally, one can assume that the techniques described (or utilised) within studies making use of such approaches are of particular interest to this review, since they represent a radically different approach to what is canonically associated with the traditional approach to topic labeling. Additionally, it must be pointed out that these studies also carry the bulk of the content associated with the individual insights that are ultimately presented in Chapter 7. As mentioned, techniques that fully automate the labeling task are generally insightful and (depending on their level of reusability) often worth exploring in more details with regards to how they are structured and ultimately implemented in their respective studies. Here, it must also be pointed out that one study (Gomez et al., 2022) uses a fully automated approach for the topics generated by one topic model (where each topic is simply labeled by its three most important keywords) and partially automated approach for the remaining model (where labels are taken from existing content categories present in the corpus). Additionally, in another paper (Smith et al., 2017) it is possible to find an implementation of both manual and fully automated approaches.

In summary, the most **frequently encountered** approach in the analysed research is the one associated with manual labeling (encountered in a total of 67 studies). This is followed by fully automated topic labeling approaches, observed in 36 distinct documents. Finally, a total of 14 observations of partially automated approaches is made across the analysed research.

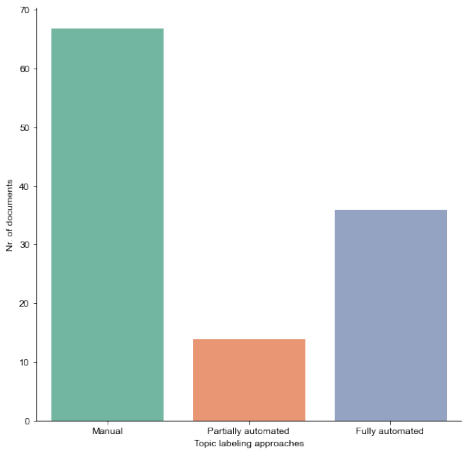


Figure 5.1: Topic labeling approaches

5.2.2. Focus of the research

To further contextualise the information carried over by the preceding data item, the focus of the analysed research is considered in relation to the utilised labeling approaches. As previously mentioned, topic labeling is often considered as a somewhat secondary activity, and is generally used to support and better contextualise the result obtained from topic modeling tasks. Because of this reason, it is fairly common to encounter primary studies where the main focus of the proposed research activity is not directly tied with the labeling task. Because of this reason, studies showing a primary interest on topic labeling have the capacity to be especially relevant in the context of this review, in particular when it comes to identifying novel and insightful approaches. Starting from this notion, papers are identified as having either a Primary or Secondary Focus on topic labeling. In general, this distinction between primary and secondary focus might not be immediately discernible. Therefore, without a clear identification of the metrics required to be taken into account to consider a study as having a primary focus on topic labeling, it might be difficult to objectively classify each study. To address this concern, in the context of this review it is decided to employ a very simple approach to make such a classification on the analysed papers. In this context, the **title**, **abstract** and **introductory section** are used to draw this distinction On this note, a study is labeled as having a **Primary Focus** on topic labeling if it is found to have a mention of the conducted labeling activities in the interested sections. When that is not the case, the interested document is labeled as having a **Secondary Focus** on topic labeling.

Whilst understanding that this represents a very simplistic methodology for partitioning the collected research, it is believed that it should provide a sufficient level of precision in addressing the goal of the associated data item. In fact, one can safely assume that (for a given study) whichever activity lacks any mention of being conducted in the abstract, tile or introductory chapter is unlikely to be considered of primary focus by the respective authors. Nonetheless (and similarly to the mention made on the potential usefulness of manual approaches in the context of the previous data item), it is important to realise that a secondary focus on topic labeling does not automatically invalidate the relevance of a study for the presented review. Additionally, it must be noted that the methodology used for distinguishing these two kind of studies has the useful property of being immediately discernible to the reviewers and is easily justifiable to prospective readers.

As one could expect from the premise given at the beginning of this Subsection, when examining the data associated with the relevant research it is possible to observe that a majority of the studies pose a secondary focus on the activity of topic labeling. In fact, data shows that a total of **81** documents show a non-primary attention in this regard. This result can be tied back to the decision of applying the chosen search query (Section 3.3) to the entirety of the documents' content when searching the selected repositories. Indeed, the strategy of looking for relevant keywords in the body

of each individual study (in addition to the title and abstract sections) allowed (in the first place) for the collection of this category of documents. Additionally, no exclusion criteria is in place explicitly demanding for documents to solely focus on topic labeling tasks. To this end, it must be stressed that these decisions were made consciously when defining the Review Protocol associated with this study and are believed to be necessary for a comprehensive review in the chosen time period.

Finally, in order to link these information to the data presented in the previous Subsection, it might also be worth observing how this division of focus affects the individual categories representing the identified labeling approaches. Here, it is interesting to see how the ratio of studies primarily focusing on topic labeling steadily grows as one moves to more involved categories. More specifically, of all the analysed studies containing manual labeling approaches, only **9%** show a primary focus on topic labeling. On the other hand, this metric is found to be substantially higher for papers referencing partially automated approaches (where primary focus studies make up **14%** of the collection) and represents more than half of the collection interested by fully automated approaches (where **58%** of the documents primarily focus on topic labeling). This growth is visually highlighted in Figure 5.2.

The main takeaway from the analysis performed on the data associated with this Subsection is that the proportion of studies showing a primary research focus on topic labeling grows as one moves from manual labeling to the generally more advanced techniques associated with the partially automated and fully automated approaches.

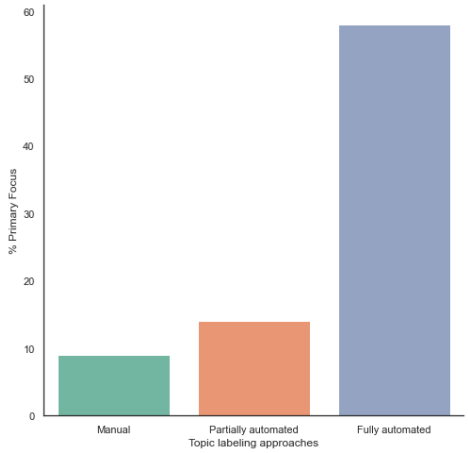


Figure 5.2: Focus on topic labeling for each approach

5.2.3. Novelty of the proposed approaches

Following the identification of the categories of approaches observed in the analysed studies and the attention that such approaches received in the context of the individual documents, it might also be relevant to identify to which degree authors decided to implement original techniques, representing novel approaches in the context of topic labeling. Clearly, a **novel approach** refers to a process for topic labeling that does not reflect (i.e. is not related to) already existing findings associated with previous studies. In other words, if the authors of a given study did not take inspiration from existing research to devise the labeling procedure, the approach is considered to be novel.

Naturally, this general idea comes with a few exception that need to be considered for a correct overview associated with the current data item. In fact, many of the studies implementing manual labeling approaches do not cite any previous work making use of the same methodology. Independently form this fact, it is implicitly clear that the labeling approaches brought forward by such papers do not constitute a novel idea, and therefore should not be treated as such in the presented review. At the same time, it must be pointed out that the utilisation of a manual approach (Data Item 5) does not automatically exclude the possibility of a given study including a novel contribution. In

fact, there might exist some specific characteristics related to how the (manual) labeling process is carried out that might justify the the identification of such a process as a novel endeavor. For example, Ojo and Rizun (2021) proposes a multi-step manual labeling process involving two experts that: (1) Independently label the generated distributions by looking at the top scoring words; (2) Compare and discuss the generated labels; (3) Individually refine the labels by integrating the previous analyses with the top scoring documents associated with each topic and; (4) Discuss again the results, reaching a final label solution.

Clearly, whenever such a condition does not hold, or whenever the authors explicitly state that a specific methodology is taken from existing research, the study is classified has possessing an already **established approach** to topic labeling.

Examining the data collected from the generated Data Extraction Forms reveal that **70** of the 108 analysed studies refer back to already existing approaches. In other words, this figure shows that the authors of more than half of the analysed studies decided to rely on previous research (or on simple manual labeling techniques) in order to assign labels to the generated term distributions. Additionally, one paper (Hosseiny Marani et al., 2022) makes use of both established and novel techniques. In particular, the authors explore a total of four different approaches, two of which are taken from existing research, whilst the remaining ones propose novel labeling processes using the salience score metric introduced in Chuang et al. (2012).

By itself, these notions are not particularly informative and it should be further contextualised with regards to the already explored data items. In this context, the information associated with the novelty of the proposed techniques is compared to (1) the types of labeling approaches and (2) the focus shown in the individual papers with regards to the conducted labeling activities.

With regards to (1), one can observe a trend similar to the one highlighted in the previous Subsection with regards to the three identified categories of labeling approaches. In fact, in the analysed studies there exists a constant growth in the percentage of documents introducing novel topic labeling techniques as one moves from manual approaches, to partially and finally to fully automated ones. In this context, the included research shows a percentage of novel insights of **13%**, **33%** and **75%** for manual, partially and fully automated approaches respectively.

At the same time, it is possible to observe that in the context of (2), most studies having a primary focus on topic labeling (as identified in Data Item 6) also include the proposal of novel approaches. More specifically, one can observe that this fact is true for **78%** of such studies. On the other hand, the same figure is only equal to **21%** for the collected work presenting a secondary attention on the matter.

In summary, the insights obtained from Data Item 7 allow to extract the following information relating to the analysed research:

- The authors in slightly more than half of the analysed work **reference existing approaches** for topic labeling rather than providing their own novel technique for supporting such activity.
- Studies with a **primary focus** on topic labeling are much more likely to introduce a novel approach. To support this information, one can observe that only a relatively small percentage (roughly 22%) of the collected documents showing a secondary focus on the labeling activity ultimately include a novel contribution for the labeling activity. This is opposed to the work having a primary focus in this regard, where more than three quarter of the scrutinised documents show some form of novel technique.
- Finally, and in a similar fashion to what is observed in Subsection 5.2.2 (*"Focus of the research"*) with regards to studies showing a primary focus on topic labeling, one can see a clear growth pattern in the percentage of novel approaches when moving from the simple manual labeling techniques to the more advanced partially and fully automated approaches. To this end, we have that roughly 75% of studies employing **fully automated approaches** do so in the context of a novel technique.

Nr.	Technique	Papers	Total
1	Ontology-based	Zosa et al. (2022), Kim and Rhee (2019), Adhitama et al. (2017), Allahyari et al. (2017)	4
2	Transformer-based	Pergola et al. (2021), Wang et al. (2021), Popa and Rebedea (2021), He et al. (2021a)	4
3	Frequency- & rank- based term(s) extraction	Zhang et al. (2018a), Campos et al. (2020), Scelsi et al. (2021)	3
4	Graph-based	Chin et al. (2017), He et al. (2019), He et al. (2021b)	3
5	Supervised Topic modeling	Zhang et al. (2019), Hu et al. (2020), Lau et al. (2017)	3
6	Cosine similarity-based	Scelsi et al. (2021) Zhang et al. (2021)	2
7	Deep Neural Network	Aletras and Mittal (2017), Sorodoc et al. (2017)	2
8	Probabilistic approach	Gourru et al. (2018), Hosseiny Marani et al. (2022)	2
9	Automatic Term Reconignition	Truică and Apostol (2021)	1
10	Deep embedded clustering	Yin et al. (2022)	1
11	Long short-term memory model	Lau et al. (2017)	1
12	Multi-document summarization	Rosati (2022)	1
13	Sequence-to-sequence model	Alokaili et al. (2020)	1
14	Transfer topic labeling	Béchara et al. (2021)	1

Table 5.3: Fully automated novel labeling techniques

5.2.4. Encountered labeling techniques

In this Subsection, a general overview is given with regards to the various topic labeling techniques encountered throughout the analysed research. In this context, the relevant data is presented in a tabular format and structured following (1) the three categories of topic labeling approaches (defined in Subsection 5.2.1 - *"Categories of topic labeling"*) and (2) further divided into novel and established approaches (as presented in Subsection 5.2.3 - *"Novelty of the proposed approaches"*). Additionally (and when required), a brief comment is provided on the general details associated with the individual techniques and on noteworthy characteristics associated with the individual papers.

Firstly, Table 5.3 and 5.4 highlight the techniques encountered within the context of papers presenting **Fully Automated** topic labeling approaches (Data Item 5). The two Tables contain references to the individual (relevant) studies regarding Novel and Established approaches (Data Item 7) respectively.

In total, a set of 14 overarching techniques are identified across the documents introducing **Novel** approaches. In this context, Frequency- & rank- based term(s) extraction methods (3) use metrics such as Term Frequency (Zhang et al., 2018a; Campos et al., 2020), Term Frequency - Inverse Document Frequency (Scelsi et al., 2021) and and Okapi BM250 (Zhang et al., 2018a) to rank candidate

Nr.	Technique	Original proposal	Papers	Total
1	Top-terms selection	Sievert and Shirley (2014), Sievert and Shirley (2014),	Cassi et al. (2017), Hagen et al. (2019), Gomez et al. (2022), Hosseiny Marani et al. (2022)	4
2	Graph-based	Aletras and Stevenson (2013) Wang et al. (2014)	Aletras et al. (2017), Huang et al. (2022)	2
3	Supervised Topic modeling	Ramage et al. (2009) Ramage et al. (2011)	Chen et al. (2020a), Lebeña et al. (2022)	2
4	Cosine similarity -based	Lau et al. (2011)	Smith et al. (2017)	1
5	Frequency- & rank- based term(s) extraction	Campos et al. (2020)	Campos et al. (2022)	1
6	Probabilistic approach	Mei et al. (2007)	Hosseiny Marani et al. (2022)	1
7	Support Vector Regression	Lau et al. (2011)	Aletras et al. (2017)	1

Table 5.4: Fully automated established labeling techniques

terms based on their relevance to the given topic distributions. (9) Automatic Term Recognition (Trucă and Apostol, 2021) refers to a labeling technique utilising (domain- specific) term recognition to extract (and subsequently score) candidate n-grams from relevant documents belonging to a topic using the C-Value metric. The encountered Deep embedded clustering method (10) (Yin et al., 2022) refers to an Embedded Topic Model (ETM) which learns topics together with the clusters they belong to. Here, clusters are associated with ground-truth labels obtained from the underlying dataset. The observed Multi-document summarisation (MDS) approach (12) uses MDS models based on the longformer architecture (Beltagy et al., 2020) to generate natural language summaries.

On the other hand, 6 distinct Fully Automated techniques based on already **Established** work are encountered across 12 documents. Here, Table 5.4 also provide a reference to the primary studies in which the interested techniques were first encountered. An exception to this detail is Gomez et al. (2022), for which an original proposal paper is not presented. This is due to the fact that the proposed approach simply build a topic label by selecting the top scoring words for the topic and merging them using an underscore symbol. Whilst not referencing any particular previous work, the simplicity of this technique prevents it from being classified as a novel contribution.

In Table 5.5, all studies containing **Partially Automated** approaches are listed and divided between Novel and Established approaches. In this last category, only one overarching technique is found and it refers to all those studies that use as guidance in their labeling procedure a predetermined set of existing labels (1). In terms of Novel techniques, (3) (Xu et al., 2020) refers to the automatic generation of topic summaries using LexPageRank (Erkan and Radev, 2004) in order to aid experts in the labeling process. Similarly, in (4) Roeder et al. (2022) uses automatically generated metrics (of topic-document prevalence) to further contextualise each distribution and to offer guidance in the labeling procedure. Finally, Maximum entropy modeling (5) refers to the W2VLDA model proposed by García-Pablos et al. (2018) that employs a topic modelling approach combined with continuous word embeddings and a Maximum Entropy classifier.

Finally, Table 5.6 lists papers showcasing **Manual Approaches**. For both Novel and Established processes the two categories refer to "traditional" Manual Labeling (where labelers simply inspect a subset of prevalent terms associated with a topic) and to Manual Labelling assisted by associated documents (where the individual documents from the corpus are also used as further guidance to contextualise a given topic). In this context, notice that whenever authors decided not to give detailed information with regards to the labeling process, this was assumed to be based solely on the (top) words associated with each distribution.

With regards to the Novel approaches identified in the context of Manual Labeling techniques, it might be necessary to state that these approaches were classified as such because they present some meaningful (original) differences with regard to the way in which the corresponding techniques are

Type	Nr.	Technique	Papers	Total
Established	1	Manual labeling assisted by pre-existing label set	Liu et al. (2017), Yang et al. (2017), Karami et al. (2018), Karami et al. (2019), Ding et al. (2020), Fang and Partovi (2021), Jebari et al. (2021), Amat-Lefort et al. (2022), Gomez et al. (2022)	9
Novel	2	Transformer-based	Maier et al. (2018), Huang et al. (2020)	2
	3	Manual labeling guided by generated summaries	Xu et al. (2020)	1
	4	Manual labeling guided by topic prevalence information	Roeder et al. (2022)	1
	5	Maximum entropy modeling	García-Pablos et al. (2018)	1

Table 5.5: Partially automated techniques

generally presented. For example, Kuhn (2018) distinguishes its manual labeling procedure by using three distinct metrics in order to inspect the top ranked words for each topic. Such top scoring words where obtained by probability of occurrence conditional on the topic (Prob), by lift (Lift), and using the FREX metric (FREX). Another example could be provided with regards to Chen et al. (2019), where a two round procedure is performed by first classifying topics as either "coherent" or "not coherent" and then assigning a label only to those topics that were marked as coherent by both judges during the initial phase.

A note on label generation & labeling parameters

In the context of the proposed Data Extraction Form (Table 3.3), Data Items 9 & 10 are put into place to allow (for each one of the analysed studies) the gathering of detailed information associated with the label generation procedures and of the values tied to the relevant parameters influencing such processes. In other words, these items are generally exploited to obtain a more granular description of the individual documents to allow (when needed) to highlight the specifics related to how the different authors decided to implement their topic labeling activities. Because of the level of detail characterising the information associated with these Data Items, they do not reveal themselves particularly useful in the context of this Section. In fact, since the specifics of the label generation procedure (and associated parameters) are generally strictly tied to the individual study to which they refer to, it is somewhat difficult to draw an holistic analysis associated with these information.

On the other hand, these Data Items reveal themselves to be fundamental in the presentation of the **insights from individual studies** presented in the context of Chapter 7. Because of this reason, the publications identified to be relevant on the basis of the metrics outlined in Subsection 3.8.2 will be described (in large part) using the information found in these items.

5.2.5. Factors motivating the labeling procedure

In a general sense, a multitude of factors can lead researchers to implement a labeling step in the presented research. Although often non-mandatory, the activity of topic labeling can often be an instrumental tool in enhancing and further refining the results obtained by the underlying topic modeling procedure. In this context, Data Item 11 tries to capture the various motivation that led authors of the analysed research to include a labeling activity in their work. In this context, Table

Type	Technique	Papers	Total
Established	Manual labeling	Huang et al. (2017), Light and Odden (2017), Smith et al. (2017), Xiang et al. (2017), Zhang et al. (2017),Alp and Öğüdücü (2018), An et al. (2018), Huang (2018), Karami et al. (2018), García-Pablos et al. (2018), Alp and Öğüdücü (2019), Bastani et al. (2019), Clare and Hickey (2019), Cornelissen et al. (2019), Maiti and Vucetic (2019), Chen et al. (2020c), Ebadi et al. (2020), Kim et al. (2020), Luo et al. (2020), Mukherjee et al. (2020), Nouri et al. (2020), Stamolampros et al. (2020), Xu (2020), Zhou and Jurgens (2020), Aman et al. (2021), Doogan and Buntine (2021), Ebadi et al. (2021), Monselise et al. (2021), Zhao et al. (2021), Amon and Hornik (2022), Chung et al. (2022), Effrosynidis et al. (2022), Goldberg and Abrahams (2022), Kim et al. (2022b), Kim et al. (2022a), Meena and Kumar (2022), Sakshi and Kukreja (2023), Singh and Glińska-Neweś (2022), Wahid et al. (2022), Yoo et al. (2022)	40
	Manual labeling assisted by associated documents	Syed and Spruit (2017), Maier et al. (2018), Dahal et al. (2019), Grajzl and Murrell (2019), Korfiatis et al. (2019), Nerghes and Lee (2019), Stamolampros et al. (2019), Chen and Xie (2020), Chen et al. (2020b), Ding et al. (2020), Wang and Hsu (2020), Gregoriades et al. (2021), Savin et al. (2021), Symitsi et al. (2021), Yang and Han (2021), Amat-Lefort et al. (2022), Altınel (2022), Chen et al. (2022)	18
Novel	Manual labeling	Kuhn (2018), Zhang et al. (2018b), Chen et al. (2019), Ibrahim and Wang (2019), Meng et al. (2020), Ojo and Rizun (2021)	6
	Manual labeling assisted by associated documents	Korenčić et al. (2018), Ojo and Rizun (2021)	2

Table 5.6: Manual techniques

5.7 summarises the **10** reasons identified for conducting such labeling activities. Additionally, for the numerical counts associated with each motivation a further division is made between studies explicitly justifying their reasons and studies where such motivation needed to be inferred starting from contextual clues.

As previously stated, it is generally well accepted that one of the main reasons for associating descriptive labels to generated term distributions is the objective of (2) **producing more easily interpretable topics** with the goal of making such distribution more immediately understandable to human readers. An example of this motivation can be found in Chin et al. (2017), where a topic-modeling based tweets summarisation engine (TOTEM) capable of generating recent timeline summaries for specific Twitter users is proposed. Here, topic labels are generated to provide more concise and easily readable topic summaries. Sometimes, studies carrying out topic modeling tasks do so in order to shed light on specific domains of interest. In these scenarios, labels are generally attributed to (1) **better characterise the underlying corpus** in terms of the generated topics. For example, in Sakshi and Kukreja (2023) the labels play the role of identifiers for the different (main) research areas found in the documents making up the collected research (which in this case refers to the mathematical expression recognition domain.) Ultimately, the motivations associated with (1) and (2) both refer to a desire of the authors to create more intuitively understandable topics. Here,

Nr	Motivation	Excplicitly stated	Inferred	Total
1	Characterising the corpus	18	31	49
2	Producing more easily interpretable topics	17	10	27
3	Adressing the overhead of manual labeling	11	4	15
4	Proxy to measure topic interpretability / coherence	8	0	8
5	Simplifying / Enhancing topic visualisations	2	2	4
6	Proxy to evaluate topic representation approaches	2	0	2
7	Complementing textual labels	1	0	1
8	Identifying non-relevant topics	1	0	1
9	Providing a basis for the analysis of human judgment of topic label quality	1	0	1
10	Verifying the robustness of coherence metrics	1	0	1

Table 5.7: Motivations for topic labeling

the main difference lays in the fact that documents that were found to be referring to (1) present a generally closer focus on the collected documents (e.g. by seeking to provide a complete analysis of the underlying corpus), whilst studies associated with (1) tended to concentrate on the proposed methodologies. A good portion of the analysed studies are also found to present topic labeling approaches as a way to (3) **address the overhead of topic labeling** activities. As one can imagine, this particular motivation is given (or inferred) from some of the studies proposing fully or partially automated approaches. Next, for a total of 8 documents the motivation given for executing a labeling step is the desire to (4) use labeling as a **proxy to measure the coherence and/or interpretability** of the generated distributions. In this context, Effrosynidis et al. (2022) states that: *"One sign of a healthy outcome of a topic modeling procedure is when the top words that define a topic can easily guide a human to provide this topic's title"*. Another example can be found in Huang (2018), where the time required to human annotators to label topics is used as a metric to compare the results of the two utilised topic models. In a total of four publications, labels are used as a tool to (5) **simplify and/or enhance the presented visualisations**. For instance, in Clare and Hickey (2019) labels are used as node names in the representation of the generated topic correlation map. In the collected research, labels are also used as a (6) **proxy to evaluate existing approaches for representing topics**. In this context, Smith et al. (2017) evaluate differences in ease of labeling for topics represented using: word lists, word lists with bar graphs, word clouds and network graphs. On the other hand, Aletras et al. (2017) use labels as part of a comparison drawn between topic represented as: word lists, phrase labels and images. In Aletras and Mittal (2017), an automated approach for producing image labels is proposed in order to (7) offer a language independent representation of topics that **complements textual labels**. Nerghes and Lee (2019) uses the proposed manual labeling step to (8) **identify** (and subsequently discard) **non-relevant topic** which are not meaningful in the representation of the underlying corpus. For (9), Hosseiny Marani et al. (2022) the generation of topic labels represents a requirement for executing the proposed **analysis on human judgments of topic label quality**, where human ratings are analyzed using exploratory factor analysis. Finally, in Doogan and Buntine (2021) the (10) **robustness of established topic coherence metrics** is (also) evaluated using topic labels, starting from the assumption that: *"An interpretable topic is one that can be easily labeled."* and *"An interpretable topic has high agreement on labels."*.

5.2.6. Addressing RQ1

Ultimately, the collection of information associated with Data Items 5 to 11 presented in this Section all build towards addressing the first Research Question proposed (at the beginning of this document) in Section 1.2. As a reminder, the question reads as follows:

> ***"What are the different approaches for topic labeling, how are they used and in which context?"***

In order to answer such a question, a brief summary of the most important information gathered from the relevant Data Items is presented.

When analysing the collected work, a total of **three macro-categories** of labeling approaches have been identified. These are: Manual, Partially Automated, and Fully Automated approaches. As one could expect, the most prominent category is the one associated with manual labeling processes, which are found in 62% of studies. Following the manual approaches, the second most common category is represented by fully automated approaches, identified in 33% of the included documents. Finally, the middle ground between manual and fully automated techniques represented by partially automated approaches is encountered only in around 12% of the relevant documents.

As a subsequent step, these categories (and the associated documents) have been further contextualised with regards to: (1) the **focus** posed on the labeling step and (2) the **novelty** of the proposed approach. In this context, it was found that the ratio of studies showing a primary focus on topic labeling and the proposal of novel approaches grows as one moves from studies including manual labeling approaches, to work proposing the generally more advanced techniques associated with partially automated and fully automated approaches. Here, one can observe that only 9% of papers relating to manual labeling show a primary focus on topic labeling. Additionally, novel techniques are found only in about 13% of these studies. On the other hand, the same metrics are shown to be higher (14% and 33%) for documents relating to partially automated approaches and ultimately increase even more drastically in the case of fully automated ones, where they are found respectively in 58% and 75% of the interested studies.

In an attempt to delve more deeply into the specific **(categories of) techniques** associated with each of the identified approaches, the individual documents have been analysed and a set of tabular results has been presented.

With regards to novel endeavors associated with fully automated approaches, a total of 12 distinct techniques have been identified (Table 5.3). Among such techniques, the two most popular ones are found to be the Transformer- and Ontology-based ones. In terms of established approaches (Table 5.4), a total of 6 different techniques has been recognised, the most popular of which corresponds to probabilistic approaches.

For Partially Automated labeling, a total of 5 (4 novel and 1 established) approaches are found to be present in the research (Table 5.5). Here, the single established approach refers the the technique of manual labeling assisted by a pre-existing label set.

Finally, Manual techniques (Table 5.6 are divided into two macro-categories corresponding to: (Traditional) Manual labeling and Manual labeling assisted by associated documents.

As a last step in addressing this first Research Question, the (provided or inferred) **motivational context** behind the choice of applying a topic labeling step have been explored. In this regard, a total of 10 different justifications have been recognised (Table 5.7). As one could expect, it is discovered that most studies use labels to increase the interpretability of the generated distributions and to provide a better characterisation of the corpus under scrutiny (e.g. identifying relevant research areas in a collection of scientific documents). Additionally, the proposal of automated labeling techniques is often supported by the desire of the authors to reduce the effort (and cognitive load) generally associated to manual labeling activities. Surprisingly, in quite a few instances (8 studies) it was shown that labels were used as an evaluation metric in order to assess the interpretability (and coherence) of the produced topics.

5.3. RQ2 - Topic modeling approaches and nr. of generated topics

In this second Subsection, the collected research is analysed within the context of the encountered topic modeling approaches. Here, Data Items 12 & 13 serve as the starting point for structuring such an analysis and for drawing the relevant observation in relation to the already presented data. More specifically, a general summary of all the topic modeling techniques used throughout the included documents is provided following the information found in Data Item 12. Then, such approaches are further contextualised by including a quick overview associated with the number of generated topic (Data Item 14). Naturally (and when possible), comments are made on the re-

lation that the items associated with this second Research Question might have on the information presented in Subsection 5.2 with regards to the encountered topic labeling approaches.

5.3.1. Approaches for generating topics

Considering the fact that the topic labeling activity (representing the main focus of this review) can be seen as a somewhat secondary step to topic modeling, it might be interesting to explore which are the topic modeling techniques that characterised the research that is included in this work. To this end, Table 5.8 highlights the **16** distinct techniques that were identified throughout the synthesis process together with a reference to the work they were first proposed in and a count of the total number of studies in which such approaches were found.

As one could expect, (1) Latent Dirichlet allocation (LDA) Blei et al. (2003) reveals itself to be the most widely used techniques for generating topic distribution in the analysed work. This is not surprising, considering the fact that LDA can generally be regarded as one of the most popular techniques for conducting a topic modeling task. Interestingly, a total of five distinct techniques extending the LDA model (identified as "LDA-based techniques" in Table 5.8 were also found to be present in an equal number of documents.

Following LDA, the second most commonly identified approach to generating topics is (2) Structural Topic Modeling (STM) (Roberts et al., 2013, 2019) which builds on the concepts of traditional probabilistic topic models, such as LDA and Correlated Topic Modeling (CTM) (Lafferty and Blei, 2005) to propose a model capable of incorporating document metadata (e.g. authors, publisher, date, etc.) into the generative process. [EXPAND HERE?]

Starting from (3), the number of documents associated with each of the identified studies starts to fall drastically, with only 8 documents associated with the statistical method of Non-negative Matrix Factorization (NMF) (Paatero and Tapper, 1994). Here, approaches based on (4) K-means clustering (Lloyd, 1982), (5) Latent Semantic Analysis (Dumais, 2004) and (6) Neural Topic Models (**?**) are all found in 4 documents. Additionally, (7) Dynamic Topic Models (Blei and Lafferty, 2006) (DTMs), which are generally used to analyse topical evolution in a corpus of sequential documents, are found in only three distinct studies. Finally, all the remaining topic modeling techniques (8 to 17) are each found in only one document.

To briefly characterise further the interested research with regards to the information provided in Table 5.8, it might be interesting to observe how prevalent the (three) most popular topic modeling techniques (LDA, STM & NMF) are with regards to the three **topic labeling approaches** presented in the context of Data Item 5 (Subsection 5.1 - "Categories of topic labeling"). In this context, this comparison could in theory clarify whether any clear correlation exists between the choice of a specific type of labeling approach and the underlying technique used to generate the term distributions.

Starting from LDA, it is possible to observe that this technique represents the topic modeling approach of choice for roughly 52%, 73% and 58% of documents making use of Manual, Partially Automated and Fully Automated labeling techniques respectively. Given this set of information, one can observe that, whilst generally popular among all three categories, no overly clear pattern can be found underlining a clear preference for any of the labeling approaches. Here, it should also be pointed out that the slightly higher figure found for Partially Automated approaches should be contextualised with regards to the notion that such a labeling category is represented by a total of only 14 documents. Therefore, metrics associated with this category can be generally considered to be the most susceptible to changes among the three categories.

With regards to STM, the same metric is equal to 28% and 13% for Manual and Partially automated approaches and is found to be completely absent in all studies presenty Fully Automated techniques. This is somewhat interesting, as it highlights that the use of **Structural Topic Models** is particularly relevant for documents found to include Manual Topic Labeling techniques. In the analysed collection, this observation can be attributed to the fact that STM documents presenting Manual Labeling approaches are found to be generally focused on providing a detailed analysis of the corpus under scrutiny (rather than describing an elaborate labeling procedure). In this context, STM allows the respective authors to incorporate additional (potentially relevant) document metadata into the topic modeling task. To further support this evidence, the documents making use of

Nr.	Technique	Proposed In	Count
1	Latent Dirichlet allocation	Blei et al. (2003)	64
LDA-based techniques	KB-LDA	Movshovitz-Attias and Cohen (2015)	1
	MetaLDA	Zhao et al. (2017)	1
	PLDA	Ramage et al. (2011)	1
	sLDA	Blei and McAuliffe (2010)	1
	tLDA	Boyd-Graber et al. (2007)	1
	W2VLDA	García-Pablos et al. (2018)	1
2	Structural Topic Model	Roberts et al. (2013)	20
3	Non-negative matrix factorization	Paatero and Tapper (1994)	8
4	K-means clustering method	Lloyd (1982)	4
5	Latent Semantic Analysis	Dumais (2004)	4
6	Neural topic model		4
7	Dynamic topic model	Blei and Lafferty (2006)	3
8	Attention-based Aspect Extraction	He et al. (2017)	1
9	Community detection based on Fast Unfolding	Blondel et al. (2008)	1
10	Dirichlet-Multinomial regression	Mimno and McCallum (2012)	1
11	Embedded topic model	Yin et al. (2022)	1
12	Joint Spherical tree and text embedding model	Meng et al. (2020)	1
13	LSTM model	Hochreiter and Schmidhuber (1997)	1
14	Sequence-to-sequence model	Sutskever et al. (2014)	1
15	Correlated Topic Model	Lafferty and Blei (2005)	1
16	Transformer model	Vaswani et al. (2017)	1

Table 5.8: Topic modeling techniques

STM are further contextualised with regards to the underlying **motivation** (initially presented in Table 5.7) driving the subsequent labeling step. Here, it is found that 86% of the interested studies are associated with the motivation titled "Characterising the corpus". This finding further supports the statement made above with regards to the interested studies.

Lastly, the NMF topic modeling technique represents 4% and 5% of papers containing Manual and Fully automated approaches and is found to have no appearances in documents referencing Partially Automated ones. Here, the generally limited sample of publications found to use this modeling technique makes it hard to draw any conclusions on the significance of the associated figures tied to the labeling approaches.

As a final representation, Table 5.9 highlights which topic labeling techniques (described in Subsection 5.2.4 - *"Encountered labeling techniques"*) if found to be present for each one of the observed topic modeling approaches.

5.3.2. Number of generated topics

Following the analysis associated with the encountered topic modeling approaches provided within the context of the previous Subsection, it could also be interesting to explore (in a general sense) how many topics were ultimately generated by such models within the documents that are part of the analysed research. Naturally, if the data associated with this metric (the number of generated topics) were to be presented in isolation it would carry little informational value, especially within the boundaries of the proposed review. In fact, the real insightfulness of this Data Item stems from its ability to determine whether the choice of a specific **labeling approach** has any influence over the number of topics generated during the underlying topic modeling phase. In this context, it is reasonable to expect that techniques that can partially or fully automate the labeling task should

	LDA	STM	NMF	K-means	LSA	NTM	DTM	ABAE	CDFU	DMR	ETM	JSTTE	LSTM	S2S	CTM	Trans.
Automatic Term Reconignition (ATR)	✓		✓													
Cosine similarity -based	✓			✓			✓									
Deep embedded clustering											✓					
Deep Neural Network	✓															
Frequency- & rank- based term extraction	✓		✓	✓			✓									
Graph-based	✓								✓							
Manual l.	✓	✓	✓	✓	✓			✓		✓		✓			✓	
Manual labeling assisted by associated documents	✓	✓														
Manual labeling assisted by pre existing label set	✓	✓					✓									
Manual labeling guided by generated summaries	✓															
Manual labeling guided by topic prevalence information	✓															
Maximum entropy modeling	✓															
Multi-document summarization (MDS)															✓	
Ontology-based	✓															
Probabilistic approach	✓															
Probabilistic approach	✓															
Sequence-to-sequence model														✓		
Supervised Topic modeling	✓					✓							✓			
Support Vector Regression	✓															
Top-terms selection	✓															
Transfer topic labeling							✓									
Transformer-based	✓					✓										✓

Table 5.9: Topic labeling techniques encountered for each topic modeling approach

encourage researcher to generate an higher number of topics when compared to those studies using Manual Labeling techniques (which are generally assumed to be substantially more time consuming). Additionally, here it is important to state that the number of topics collected from each of the analysed papers refers to the figure highlighted in the corresponding result sections. In this regard, one must understand that this fact does not automatically exclude the possibility that authors (especially of studies making use of Partially and Fully automated techniques) might have experimented with even larger collection of topics which were ultimately reduced in size on the basis of the results observed from established topic coherence or label quality metrics. Given this premise, the data relating to the number of generated topics (Data Item 13) is associated with the three identified labeling approaches (Data Item 5) and visually highlighted by means of a boxplot in Figure 5.3. Looking at the median value for each box, it can be determined that the previously stated assumption holds for the analysed data. In fact, it is possible to observe that the lowest median value (20) for the number of generated topics is tied to those studies including Manual Labeling approaches. This metric then increases as one moves to Partially Automated approaches (with a median of 30 topics) and is found to have the highest value (48) for Fully Automated techniques. Additionally, it is possible to see that both Partially Automated and Fully Automated approaches also present generally higher results for the minimum and maximum values and for the upper (Q3) and lower (Q1) quartiles when compared to the corresponding metrics found in the box associated with Manual Approaches.

This last approach also shows a much lower dispersion (with an interquartile range value equal to 30) when compared to Partially and Fully automated approaches (having ranges of 90 and 75 respectively). The fact that such observations tend to be more tightly clustered around the median value of 20 topics might additionally suggests that authors employing Manual Labeling processes might be less willing to experiment with the generation of an higher number of topics given the inherent bottleneck that such a manual approach would imply during the labeling phase.

Although these information show a general tendency towards the generation of more topics for partially and especially fully automated techniques, such an increase is not as pronounced as one could expect, especially within the context of fully automated approaches. In fact, if one imagines that a fully automated topic labeling approach should completely exonerate human annotators from the labeling procedure, it is somewhat surprising to observe that the values associated with such approaches are generally alligned with the one observed for partially automated techniques.

In this context, it might be once again worth mentioning that the high spread and the positive skew found in the box associated with Partially Automated approaches should be contextualised with regards to the relatively low number of documents (14) associated with this particular category of topic labeling.

5.3.3. Addressing RQ2

The inclusion of Data Items 12 & 13 serves the general purpose of addressing Research Question 2, which is formulated as follows:

> ***How are the underlying topic modelling phases characterised and how do they relate to the chosen topic labeling approaches?***

Here, a brief summary relating to the information presented in this Section is provided as an answer to this question.

The analysis of the literature included in this studies revealed a total of 16 distinct **approaches to topic modeling** used for generating the relevant term distributions. In this context, **LDA** revealed itself to be the most popular approach, characterising a total of 58 distinct studies. Additionally, five different techniques (KB-LDA, MetaLDA, PLDA, sLDA, tLDA & W2VLDA) stemming from the extension of the underlying model structure proposed by LDA are found in an equal number of documents. Additionally, no distinctive pattern could be observed between the use of LDA and the choice of a specific topic labeling approach. In fact, this first modeling technique was found to be fairly popular among papers across all labeling approaches. Following LDA (and its related models), **STM** is

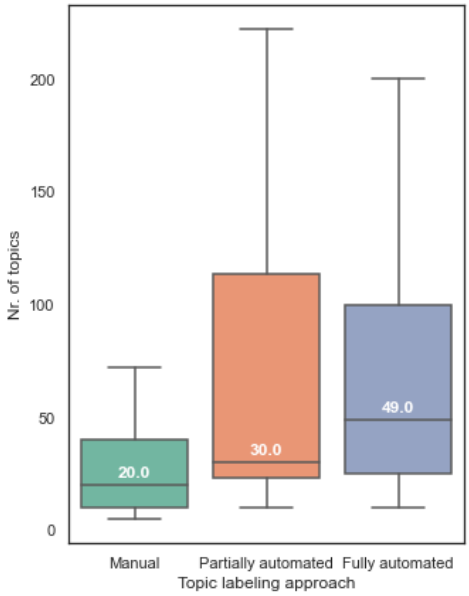


Figure 5.3: Nr of topics generated for each labeling approach

found to be the second most commonly encountered technique for generating topics (with 20 associated documents). Here, data associated with the collected research suggests that this specific technique is more likely to be found in studies utilising Manual Labeling processes. In fact, its popularity is seen to drastically decrease for Partially Automated approaches and the technique is completely absent in documents using Fully Automated ones. In this context, it is hypothesised that the use of STM could be tied to studies where the main goal is a thorough exploration of the underlying corpus and where the labeling phase is approached as a somewhat secondary activity. In this regard, STM allows author to incorporate additional document metadata into the modeling procedure. This thesis seems to be supported by the fact that almost all (86%) documents implementing STM justify the labeling approach as a necessary step to *"Characterise the corpus"*. The third modeling approach in terms of associated documents is **NMF** (found in 8 studies). Here, the limited number of publications makes it impossible to draw any definitive correlation with the associated labeling approaches.

To further characterise the conducted modeling phases, the **number of topics** generated within the context of each study is recorded and the resulting distributions are drawn with regards to the three labeling approaches. As one could expect, a general increase in the number of generated topics is observed as one moves from Manual Labeling techniques (where the associated documents have a median value of 20 generated topics) to Partially and finally to Fully Automated ones (showing values of 30 and 49 respectively for the same metric). Additionally, the distributions associated with the number of topics generated for these last two labeling approaches show a much higher dispersion (around the median value) when compared to the one associated to Manual Labeling. This further suggests that such manual approaches might act as a sort of bottleneck with regards to the acceptable range of (labeled) topics one is realistically able to generate when employing this kind of topic labeling techniques.

5.4. RQ3 - Label structure, candidate selection & quality evaluation

Shifting once again the focus towards the conducted labeling activities, the information associated with Data Items 14 to 17 is analysed as a mean to address the third Research Question. In this context, this Section starts by providing a general overview with regards to the different kinds of la-

bels encountered throughout the inspected research (Data Item 14) and then moves to presenting the different procedures for candidate selection (Data Item 15) and for label quality evaluation (Data Item 16). Additionally, these last two items are further contextualised with information associated with the assessors (Data Item 17) involved in the selection and evaluation processes.

5.4.1. Label structures

Whilst sharing the same overarching goal of enriching the output of the conducted topic modeling phase, the descriptive labels assigned to the individual topic distributions achieve this task by taking many different shapes, which generally reflect the type of characterisation the respective authors want to attribute to the topics when conducting the labeling step. Because of this reason, the data associated with Item 14 is explored in order to shed light on the encountered label structures. For this purpose, Table 5.10 summarises the number of encounters associated with the **four** general structures identified during the synthesis phase for each of the three labeling approaches. As one could expect, the most commonly recorded structure utilised to build topic labels is the one corresponding to **n-gram identifiers**. In this context, an individual label is simply represented as a contiguous sequence of (n) terms capturing the content of the associated topic. Naturally, the length of such a sequence might be variable (i.e. no restriction is imposed by the authors on the n-gram size) or fixed. Because of this reason, the provided table also includes data associated with the number of studies utilising fixed-sized unigrams (i.e. one word labels), bigrams, trigrams and 4-grams. For example, Scelsi et al. (2021) proposes both 1-grams and 4-grams thesaurus labels for the generated topics. Because of this reason, the paper is counted as one entry in both sub-categories. Following n-gram labels, the next most popular structures refers to full **Sentence labels** and **Paragraphs** (multi sentence labels). In a sense, these kind of structure extends the basic idea of n-gram identifiers by providing a (set of) complete sentence(s) for a given topic. On a different note, a total of three (Fully Automated) approaches are shown to include **Image labels** (i.e. visual topic identifiers). Interestingly, one can observe that the vast majority of **non n-gram** label structures are found within studies including Fully Automated techniques. The only real exception to this notion is found with regards to Sentence labels, which are also generated (in four cases) using Manual labeling approaches.

	Labeling Approach		
Label	**Manual**	**Partially Automated**	**Fully Automated**
n-gram	66	15	24
1-gram	5	1	4
2-gram	1	0	1
3-gram	1	1	3
4-gram	0	0	1
Sentence	4	0	3
Paragraph	0	0	4
Image	0	0	3

Table 5.10: Encountered label structures

In this regard, the documents associated with the (generally more peculiar) non n-gram labels are extracted and presented with regards to the underlying topic labeling techniques utilised to generate them (such techniques refer to Data Item 8, which is described in detail in Subsection 5.2.4 - *"Encountered labeling techniques"*). The interested documents are highlighted in Table 5.11, which shows the three kinds of non n-gram label structures (Sentences, Images and Paragraphs) together with the studies tied to the 9 different relevant techniques ultimately used to generate these identifiers.

	Label Structure		
Labeling Technique	**Sentence**	**Image**	**Paragraph**
Deep Neural Netowrk		Aletras and Mittal (2017), Sorodoc et al. (2017)	
Graph-based		Aletras et al. (2017)	He et al. (2019), He et al. (2021b)
Long short-term memory model	Lau et al. (2017)		
Manual labeling	Smith et al. (2017), Effrosynidis et al. (2022), Wahid et al. (2022)		
Manual labeling assisted by associated documents	Wang and Hsu (2020)		
Multi-document summarization	Rosati (2022)		Rosati (2022)
Transformer-based	Wang et al. (2021)		He et al. (2021a)

Table 5.11: Techniques associated to non n-gram labels

5.4.2. Label selection and quality evaluation

In order to analyse the last information required to provide a comprehensive answer to Research Question 3, the data associated with the encountered label selection and quality evaluation activities is explored. In this context, it is important to specify that in some instances the label generation process might yield multiple (generally valid) candidates for an individual distribution. Normally, it is considered desirable to tie only a single identifier to a given topic. Therefore, in these cases a selection procedure might be required in order to ultimately determine how such a label might be extracted from the list of generated candidates. Because of this reason, Data Item 15 is used to identify the instances in which authors of the included studies decided to undergo a label selection procedure and serves the purpose of describing how these processes are actually structured. On the other hand, for the purpose of this review it might also be interesting to identify how often the generated labels where evaluated on the basis of their quality and which metrics are generally used to achieve this task. In this regard, Data Item 16 is utilised in order to address this detail and to further clarify the individual organisation of the various evaluation tasks. In both cases, relevant details tied to the human assessor(s) taking part in the label selection or evaluation process (Data Item 17) are also presented in order to better contextualise such activities with regards to the external involvement they generally required.

Starting with **label selection**, it is immediately noticeable how such a process is observed in only a very small minority of the analysed work. In fact, activities that can be tied to a label selection procedure are found in less than 10% of the analysed studies. Given the rather limited collection of documents for which relevant data was found with regards to this item, this paragraph will provide a set of brief description for these procedures will for each of the involved papers. Here, a total of 2 studies (involving some of the same authors) utilise **majority voting** using human assessors in order to make a final selection among the generated candidates. In this regard, Ebadi et al. (2020) and Ebadi et al. (2021) interact with a total of three domain experts in order to carry out a three step process where they: (1) Individually provide a label for the generated distribution; (2) Share their results with the other experts, discussing any incompatibility; (3) Take a formal vote on which candidate should be ultimately assigned to each topic. A less formal approach to candidate selection, which does not involve a formal voting step and that can be generally defined as **"Common Agreement"** is employed by 3 documents. Here, in Ibrahim and Wang (2019) the labeling procedure is conducted by the two authors independently which then compare and discuss their choices in order to reach a final agreement on the individual identifiers. Similarly, two domain experts are consulted

by Chen et al. (2020b). Also in this case, following the labeling procedure, the inconsistent results are *"discussed and unified"*. Finally, Ojo and Rizun (2021) conducts a slightly more involved procedure (also involving two domain experts) which is structured into the following four steps: (1) The two experts independently label each topic; (2) The experts compare, discuss and resolve the differences in the results; (3) Again individually, the labels are refined by inspecting the most prominent 20 documents associated with each distribution; (4) A second round of discussion (and final alignment) is conducted which also includes the relevant documents. Finally, a total of 3 studies present a **joint labeling** procedure which, whilst not strictly including a proper candidate selection step, is tangentially tied to this data item since it is implicitly assumed to include an informal candidate evaluation conducted when the involved annotators discuss possible avenues for labeling a given distribution. In this regard, the interested papers are Nouri et al. (2020); Sakshi and Kukreja (2023) (both involving two authors in the joint labeling procedure) and Symitsi et al. (2021) (where the required input is given by experts in Human Resource Management).

Moving to the encountered **quality evaluation procedure**, it is revealed that a total of 26 distinct documents (around 24% of the analysed work) propose some form of evaluation with regards to the generated identifiers. A tabular summary of the interested studies is presented in Table 5.12. Additionally, the relevant details associated with each approach are provided below.

Nr	Evaluation approach	Papers	Count
1	Human evaluation		15
	Four level scale	Aletras et al. (2017), Aletras and Mittal (2017), Sorodoc et al. (2017), Gourru et al. (2018), He et al. (2019), Popa and Rebedea (2021), Truică and Apostol (2021)	7
	Informal evaluation	Kuhn (2018), Zhang et al. (2018a), Mukherjee et al. (2020)	3
	Multi-choice preference	Smith et al. (2017), Scelsi et al. (2021)	2
	100-level scale	Hosseiny Marani et al. (2022)	1
	Binary evaluation	Allahyari et al. (2017)	1
	Binary evaluation + Fleiss' Kappa	Stamolampros et al. (2019)	1
2	BERTScores	Alokaili et al. (2020), Rosati (2022), Zosa et al. (2022)	3
3	Cosine similarity	Campos et al. (2020), Truică and Apostol (2021), Campos et al. (2022)	3
4	Coverage, Discrimination & Relevance metrics	He et al. (2019), He et al. (2021a), He et al. (2021b)	3
5	Normalised Discounted Cumulative Gain	Aletras and Mittal (2017), Popa and Rebedea (2021)	2
6	(Normalised) Pointwise Mutual Information	Truică and Apostol (2021)	1
7	Rogue-1	Rosati (2022)	1

Table 5.12: Label evaluation approaches

Quite a few papers (13 in total) involve (1) **human ratings** in the quality evaluation procedure.

Of this initial set, 7 documents use a **four level scale** to score a label from totally irrelevant (0) to perfectly fitting (3) with regards to the related term distribution. In this context, Popa and Rebedea (2021) additionally specifies that for each topic, the top 10 labels generated by each of the trained models are submitted for evaluation. Additionally, one extra stopword label is artificially added as a *"distractor"* and use as a metric to discard assessors rating such labels with scores higher than 1 for at least 25% of these labels. In this context, the evaluation is conducted by a total of 35 assessors with varying backgrounds (computer science, medicine, law, and economics). Here, each topic is presented to the assessor with 10 associated terms and two relevant documents. Similarly, the evaluation procedure in He et al. (2019) also provide a collection of relevant documents together with the top 20 terms to the four human annotators involved. In Truică and Apostol (2021), the four level scale defines labels as: Inappropiate, Semantically related, Reasonable and Very Good. Here, the 30 annotators (graduate and undergraduate students of Computer Science) only receive the topic label and the 10 most relevant terms. Gourru et al. (2018) further refines the four level scale by organising the score with value two into two flavors. The the proposed scale, a label can be: (0) Unrelated, Related but not relevant (1), (2a) Related but too broad, (2b) Related but too precise and (3) Perfectly related. Here, thirty terms and seven documents are presented for each label to the six computer scientists acting as the human assessors. Finally, using a similar four level scale, Aletras et al. (2017) performs an assessment of label relevance (with regards to a given document) using at least 10 annotators per topic which are presented (for each document) with four labels. Here, only one of the four labels is truly associated with the given document (instead, the other three are intruders with low marginal probabilities tied to the provided document). In Smith et al. (2017) a **multi-choice preference** task is performed. The assessors are shown the titles of ten documents together with the 5 labels generated for the corresponding topic (4 labels generated by humans on the basis of different visual representation of topics and one automatically generated label). In this context, each of the assessor is asked to choose the best and worst label in the set. Finally, in Scelsi et al. (2021) the 36 assessors were asked to choose, for each topic, the most appropiate of the three presented labels (two of which were generated from the proposed technique and one of which was randomly selected). A document employing a **binary evaluation** procedure is Allahyari et al. (2017), where the three human experts are asked to evaluate each label as either being *"Good"* or *"Unrelated"*. Starting from another binary evaluation procedure marking agreement or disagreement with the assigned labels, Stamolampros et al. (2019) extends this initial evaluation by computing the **Fleiss' kappa** metric representing the agreement between the 8 expert raters taking part in the initial evaluation. In the context of Hosseiny Marani et al. (2022), a subset (60) of the generated topic is evaluated by 300 human assessors recruited via Amazon Mechanical Turk. Here, each evaluation was given on the basis of 15 distinct criteria (e.g. coherence, insightfulness specificity, predictability, etc.), each associated with a continuous visual analog **100-level scale** (representing the level of agreement to each statement). For each evaluation, assessors were presented excepts from five documents associated with each topic (and corresponding label). Finally, for a total of three papers a brief **informal evaluation** is provided with regards to the generated labels. In this regard, Zhang et al. (2018a) and Kuhn (2018) simply state that an inspection of the identifiers reveals that they are successful in achieving their task of more clearly characterising the underlying corpus. Similarly, Mukherjee et al. (2020) examines the labels and finds them to be *"well-aligned"* with therminology found in the existing literature. Three documents (Alokaili et al., 2020; Rosati, 2022; Zosa et al., 2022) employ (2) **BERTScores** (Zhang et al., 2020), a transformer-based metric used to measure the similarity between generated labels and reference text (e.g. gold-standard labels, related documents) by means of contextual embeddings and which is shown to have generally high correlation with human evaluation. Additionally, in this task Rosati (2022) also includes the (7) **Rogue-1** metric (Lin, 2004) as a measure of overlap of the label summaries to the reference text and as an alternative to the semantic interpretation offered by BERTScores. Three studies use the (3) **cosine similarity** metric as a proxy for generating quality estimates for topic labels. In this context, Campos et al. (2020) and Campos et al. (2022) compute such quality metric by considering the distance between generated labels and gold-standard ones. On the other hand, in Truică and Apostol (2021) the embeddings are generated for both the topic label and the associated topic keywords. Then, the cosine similarity measure is computed between the label vector and the average of

the word vectors and the resulting value is used as the relevant quality metric. Additionally, this last paper also provides additional evaluations in the form of (6) **Pointwise Mutual Information (PMI)** (Church and Hanks, 1990) and **Normalised Pointwise Mutual Information (NPMI)**. Here, PMI acts as a measure of relevance for the label with regards to the associated topic (distribution) and is computed by extracting the n-grams from a subset of documents associated with a given topic. In this context, it is stated that: *"By computing the PMI for a label using all the collocations [the sequences of words that co-occur more often than would be expected by chance] that appear in the documents belonging to a topic, it can be determined if a label is indeed representative for a topic"*. Three papers (by some of the same authors) He et al. (2019, 2021a,b) use metrics of (4) **Coverage**, **Discrimination** and **Relevance** in the quality evaluation process. Here, Coverage reflects the idea that labels should not carry incomplete information with regards to the topics they are associated with. On the basis of the notion found in Wan and Wang (2016b), the Coverage metric is computed as the ratio of the top 20 terms found in the corresponding (sentence) label. On the other hand, Discrimination refers to the degree to which a given label uniquely identify a single topic and is clearly discernible (i.e. dissimilar) to the other identifiers. In this context, a lower degree of similarity between labels suggests a better quality of the overall labeling procedure. This level of discrimination is simply measured by computing the (average) cosine similarity value between a given label and all other ones generated in the same labeling phase. As a measure of Relevance of the generated labels with regards to a given topic's content, and following the idea of Peyrard (2019),He et al. (2019) and He et al. (2021a) use the Kullback-Leibler Divergence method to evaluate the relevance of the provided labels as follows:

$$KLD(T,TL) = \sum_{w \in TL \cup T} P_T(w) * \log \frac{P_T(w)}{tf(w,TL)/|TL|} \tag{5.1}$$

Where TL represent the (sentence) label, $tf(w,TL)$ the frequency of term w in TL and $|TL|$ the word count in TL.

On the other hand, in He et al. (2021b) in order to compute the same (Relevance) metric the authors propose the following measure of precision (Equation 5.3) and recall (Equation 5.2) using Doc2Vec embeddings:

$$R_{doc2vec}(T,TL) = \frac{\sum_{w_i \in T} P(w_i) \sum_{w_j \in TL} Sim(w_i,w_j)}{K+l} \tag{5.2}$$

$$P_{doc2vec}(T,TL) = \frac{\sum_{w_j \in TL} tfidf(w_j) \sum_{w_i \in T} Sim(w_i,w_j)}{K+l} \tag{5.3}$$

Where w_i is a word taken from the topic (distribution) and w_j a word taken from the (sentence) label and $Sim(w_i,w_j)$ is the cosine similarity between the two terms. Additionally, K represents the number of generated topics together with the number of generated labels and l is a denominator smoothing parameter. This two measures are then used to compute a final F1 score (Equation 5.4) which is used as the chosen metric for label-topic relevance:

$$F_{doc2vec} = 2\frac{P_{doc2vec} \cdot R_{doc2vec}}{P_{doc2vec} + R_{doc2vec}} \tag{5.4}$$

Finally, Aletras and Mittal (2017) employs (5) **Normalised Discounted Cumulative Gain (NDCG)** (Järvelin and Kekäläinen, 2002) to draw a comparison between the top image labels assigned by the proposed model and the gold-standard rankings generated by human annotators.

5.4.3. Addressing RQ3

In an attempt to address the third Research Question, this subsection summarises the relevant data associated with Data Items 14 to 17. As a reminder, Research Question 3 poses the following issue:

> ***How are the encountered labels generally structured, how are candidates ultimately selected and how is the quality of the final label assignments evaluated?***

Starting from the notion of **label structures** associated with Item 14, it is found that most of the collected research is structured (with regards to the topic labeling activities) around the generation of n-gram labels. The vast majority of papers interested by this structure do not provide any specifics with regards to the length of the generated identifiers. Nonetheless, it is found that for a handful of documents strict constraints associated with the n-gram length are given (such constraints range from 1- up to 4-grams). Building on the general structure provided by n-gram identifiers, 7 documents are found to use full sentences to contextualise the generated distributions. Finally, paragraphs (i.e. multi-sentence labels) are used as the chosen identifiers in four studies and image labels are presented in a total of three documents. Further analysis on the observed structures reveals that most of the non n-gram labels (sentences, images and paragraphs) are proposed within documents containing fully automated techniques, with only 4 cases of sentence labels found in studies employing manual labeling techniques.

Secondly, when considering the notion related to label **candidate selection** (Data Item 15), one can observe that such an activity reveals itself to be for the most part absent within the collected research. In fact, less than 10% of documents are interested by some sort of candidate selection process. Nonetheless, three overarching techniques (all involving human evaluation) can be described for the encountered selection procedure. Firstly, two documents use a majority voting process among assessors to determine the final identifier to be selected from the pool of candidates. Secondly, three studies employ a less formal process of "Common Agreement", were the selection is a result of a discussion among the different assessors. Here, no proper voting activity takes place for determining the final label. Whilst not strictly relating to any specific candidate selection step, another three papers are found to structure their labeling process around a joint labeling activity, involving multiple annotators working together in the generation of the required identifiers.

Finally, the issue of **evaluating the quality** of the proposed identifiers is analysed within the context of Data Item 16. Here, it is revealed that most documents rely on some form of human involvement to generate the required evaluation metrics. In this context, a four level scale is used to rate the labels' relevance with regards to the associated topic (and documents) by six distinct studies. Two documents employ a multi-choice preference approach where the annotators are asked to select the best (and worst) label among a set of identifiers generated for the same distribution. One document is found to use a binary approach (good / unrelated) for the evaluation and another study integrates this approach by also computing the level of agreement (Fleiss' Kappa) between annotators with regards to the given evaluation. Finally, a document uses a 100-level scale presented over 15 different criteria to form a thorough evaluation of the proposed labels. With regards to human-based evaluation, it is also worth mentioning that three studies provide an informal quality assessment, which simply underlines (following manual inspection) the relevance of the generated labels with regards to the domains at hand. Moving to automated techniques for label evaluation, the transformer-based BERTScores metrics is utilised by three documents to provide a measure of similarity between generated and gold-standard labels. Similarly, two studies use the cosine distance to draw a similar comparison. Additionally, one document uses the same metric to compare the label's vector with the average of the topic terms vector. Additionally, an alternative to BERTScores is found in the form of the Rogue-1 overlap metric in one publication. Three papers (by the same authors) evaluate the generated labels using three distinct metrics representing the level of coverage, relevance and discrimination. Finally, (Normalisd) Discounted Cumulative Gain is used in the evaluation of the (ranks associated with the) generated image labels and Pointwise Mutual Information acts (in one study) as a measure of representativness of the label for the corresponding distributions.

5.5. RQ4 - Overview of domains

The contextualisation of the last Research Question proposed within the context of this review is achieved by means of an analysis of the insights found within Data Items 18 to 20. Here, Data Item 18 allows to draw a general overview of how the encountered corpora of analysed documents are structured across the included research. Then, such corpora are further characterised with regards to the characteristics making up the individual documents which serve as the basis for the topic

modeling and labeling activities (Data Item 19). Finally, a brief analysis is provided with regards to the pre-processing steps employed to prepare such documents for the subsequent modeling phase.

5.5.1. Corpus and Documents

The generation of the underlying topics enabling the labeling procedure representing the focus of this review stems, in each of the included studies, from an automated analysis of the set of documents making up a given corpus via the various topic modeling approaches previously described. In this context, it might be interesting to provide a general characterisation of such collections of documents by defining the overarching types of corpora analysed within the boundaries of the selected research. Here, Table 5.14 organises the encountered corpora in 19 distinct categories loosely defined on the basis of the origin and focus of the included documents. Additionally, for each category further details are presented with regards to the individual sources (platform and repositories) from which the documents were originally sourced. For each corpus category and source the number of interested documents is also provided.

In an attempt to further profile the previously described labeling approaches with regards to the information obtained in the context of this Section, Figure 5.6 attempts to relate the identified categories of corpora with each one of the described labeling approaches. Here, the full distribution of categories associated with each approach is presented. As expected, the number of encountered categories grows together with the total number of documents found to contain a given approach, with Partially Automated, Fully Automated and Manual approaches presenting a total of 8, 11 and 16 categories of corpora respectively. Interestingly, the types of corpora characterising the top 50% of documents is different for each of the three labeling approaches. In fact, studies including Manual Labeling techniques are most oftenly found to work on corpora taken from Social Media platforms (26,9%), built from the collection scientific publications (17.9%) or (only in a relatively small percentage of cases) by collecting articles from News outlets (10.4%). This is in strong contrast with publications making use of Fully Automated approaches, where almost half of the interested studies (44.4%) focus on corpora generated from such News outlets.

Another important information, which stems from the analysis of the data associated with Item 19, pertains to the kinds of documents found within the analysed corpora. In this context, Table 5.15 highlights these document types together with the category of corpora to which they are found to belong to. Here, the document type variable provides an intuitive description of the nature of the individual document categories and (in a similar format to what is presented in Table 5.14), each document type is ultimately associated with a counter specifying the number of studies that it characterises.

Starting from these information, and in a similar vein to what is presented in Figure 5.6, the data tied to the individual document types is associated with the encountered labeling approaches. Here, Figure 5.4 visually represents the five most prevalent document categories for each approach. As expected, the documents taking part in studies presenting Fully Automated approaches are found to generally match the corresponding (most prevalent) categories of corpora for the same labeling approach. Here, news articles and paper's contents and abstracts represent the two most prominent document categories and can be tied directly to the "News", "Scientific literature" and "Medical literature" corpora types. Similarly, in Partially Automated approaches one can see that the dominant category of "Food" (and "Tourism") corpora ties closely with the high number of review documents found in the interested studies. Additionally, the "Social Media" and "Medical literature" categories are (as expected) reflected by corresponding documents taking the shape of Tweets and paper's abstracts. Somewhat surprisingly, the most prominent document typology (Review and metadata) observed for Manual approaches does not reflect the most popular type of corpora ("Social media") found in this kind of documents. In fact, in this case Tweets are generally less prominent than review documents. This fact should be attributed to the long tail of corpus types including categories such as "Tourism", "Shopping", "Employment" and "Food" which tend to be strongly represented by review type documents.

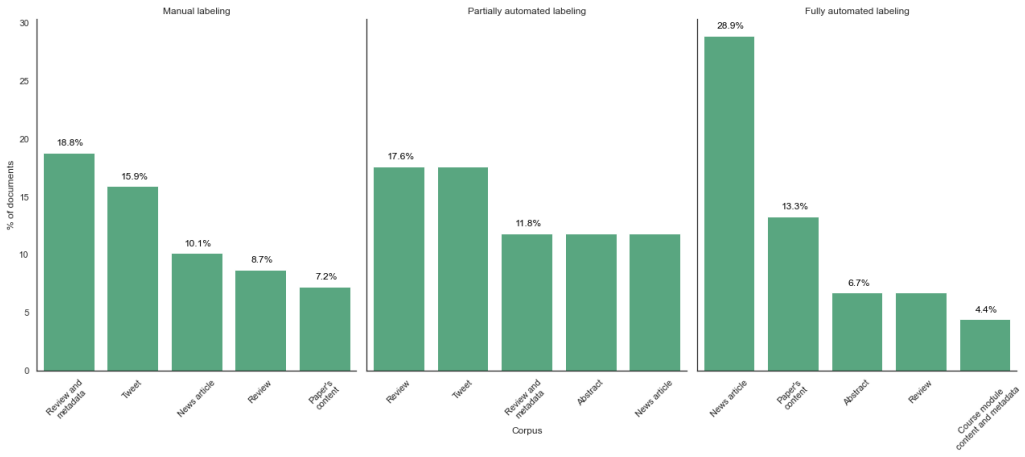


Figure 5.4: Most prevalent document types for each labeling approach

5.5.2. Documents pre-processing techniques

The analysis of the last Data Item pertains to the pre-processing steps applied on the individual documents making up the encountered corpora. In this context, pre-processing activities are a generally necessary procedure for preparing the raw documents collected from the various sources with the goal of converting them into a format that is machine processable and usable by the given topic models. The pre-processing steps recorded in the work taking part in this review are presented, ordered on the basis of their prevalence in the overall collection of documents, in Figure 5.5. Here, it is important to clarify that (1) only activities found to be present in more than one study take part in the presented visualisation and that (2) the individual pre-processing tasks have been recorded on the basis of the corresponding sections mentioning such activities in the related work. In other words, the authors' description related to the employed pre-processing steps are used as a means to populate the information tied to Data Item 20. Any pre-processing step which is not explicitly recorded in the related study is therefore not included here. Among the 108 documents forming the analysed research, a total of **19** pre-processing activities are found occurring in more than one paper. Here, most of the encountered techniques (e.g. stopwords removal, tokenisation, lemmatisation, etc.) are widely common in studying containing a topic modeling step. Somewhat surprisingly, the removal of terms using a previously established word list is encountered fairly frequently (in 15.7% of the documents). Generally, such an approach is useful to remove words that are known to be generally common in a given corpus but that do not add any meaningful contribution with regards to the insightfulness of the generated distributions. Another technique, represented by the the removal of document artifacts, was found to be especially common for Tweets (where it was generally used to remove hashtags) and for documents generated starting from web pages (where the HTML content needed to be cleaned during the pre-processing phase).

Considering the fact that the result of the encountered pre-processing steps creates the input necessary to the various topic models needed to generate the relevant topics, it might make sense to highlight how such steps are distributed among the topic modeling techniques shaping the included research. Therefore, a further characterisation of the highlighted pre-processing scenarios is provided in Table 5.13, where the most frequently observed **topic modeling techniques** are shown together with the pre-processing activities by which they were generally preceded. Here, it is decided to include all those modeling techniques appearing in more than one study (this includes techniques 1 to 7 in Table 5.8. Additionally, papers using traditional LDA and documents using LDA-based techniques are merged into a single category (which is simply named "LDA"). In this context, the prevalence of each pre-processing activity is shown as a percentage value over the total number of documents characterising a given topic modeling technique. Additionally, notice that the modeling technique names are highlighted with the corresponding number of documents in which they

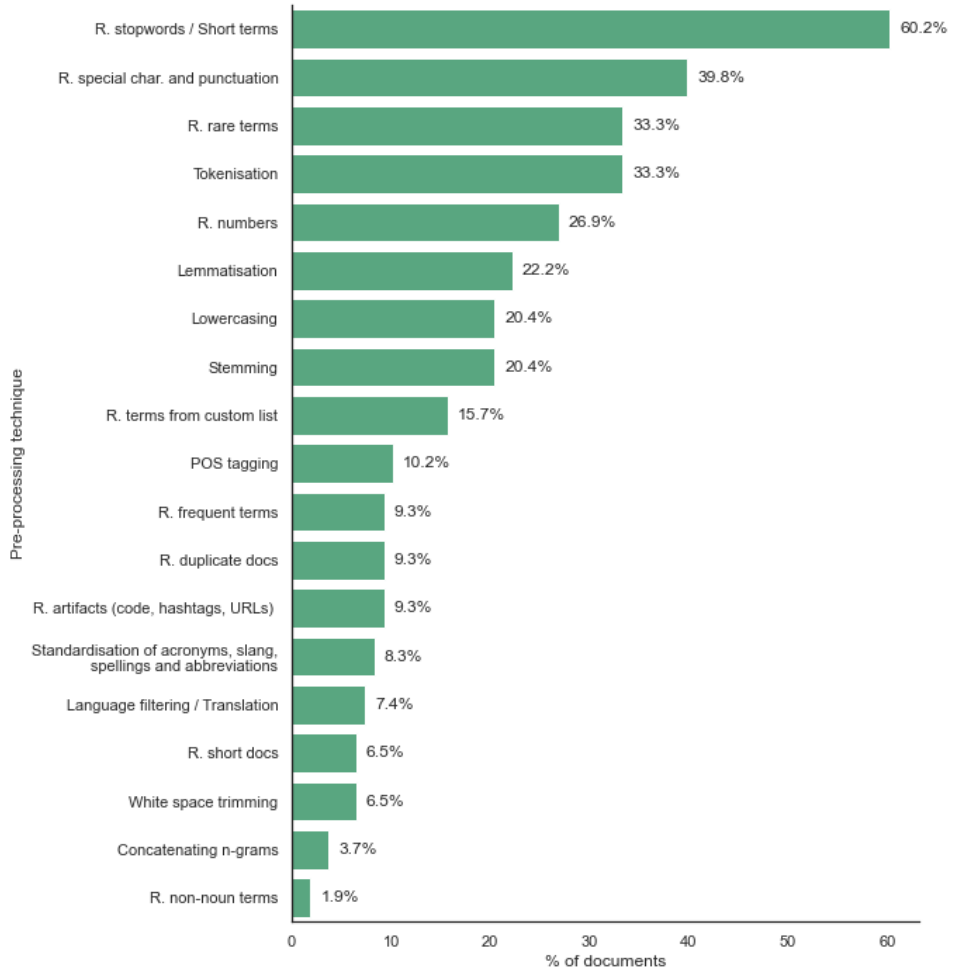


Figure 5.5: Pre-processing techniques encountered in more than one document

are found (here shown in parenthesis). For readability purposes, and in order to better highlight the more relevant information, the upper 50% of pre-processing approaches found in each modeling technique are highlighted in green.

5.5.3. Addressing RQ4

As a last attempt to fully describe the information tied to the collection of primary studies included in the presented review work, the final Data Items (18 to 20) are briefly summarised and their associated data presented as a way to address Research Question 4, which asks:

> ***Which are the prevalent corpora on which topic labeling techniques have been applied to and how are they shaped?***

Firstly, a fairly complete overview of the **encountered corpora** representing the basis for the described topic modeling (and labeling) activities is highlighted with regards to Data Item 18. In this context, a total of 19 different categories of corpora (defined on the basis of the origin and focus of the associated documents) are identified and presented together with the relevant **sources** (from which the underlying document were collected) in Table 5.14. Here, the three most commonly encountered corpora in the included research are made up of: (1) Scientific literature (i.e. papers, present in a total of 24 distinct studies), (2) News (from various outlets and characterising 23 documents and (3)

		Topic modeling techniques						
Nr.	**Pre-processing technique**	**LDA (64)**	**STM (20)**	**NMF (8)**	**K-m. (4)**	**LSA (4)**	**NTM (4)**	**DTM (4)**
1	R. stopwords / Short terms	65,6	80	50	25	75	75	33,3
2	R. special char. and punct.	39,1	70	50	25	25		33,3
3	R. rare terms	25	50	25	100	75	50	33,3
4	Tokenisation	40,6	35	12,5			75	66,7
5	R. numbers	20,3	50		75		25	33,3
6	Lemmatisation	28,1	20	12,5	25		25	
7	Lowercasing	21,9	40	12,5		25	25	
8	Stemming	17,2	35		50	25		33,3
9	R. terms from custom list	10,9	35		25	25		33,3
10	POS tagging	9,4	20	12,5				33,3
11	R. frequent terms	12,5			25	25	25	
12	R. duplicate docs	9,4	10	12,5	25			
13	R. artifacts	14,1		12,5				
14	Stand. of acronyms, slang, spellings and abbreviations	10,9	5	12,5	25			
15	Language filtering / Trans.	6,3	20	12,5				
16	R. short docs	6,3	10	12,5				
17	White space trimming	7,8	10					
18	Concatenating n-grams	3,1	10					
19	R. non-noun terms	1,6		12,5	25			

Table 5.13: Prevalence of pre-processing steps for each modeling technique

Social media content (gathered from Twitter and Reddit and found in 22 publications). Associating the described corpora with the identified **labeling approaches** shows that the most prominent categories are quite different between the three approaches. In fact, almost half of the studies including Fully Automated approaches are based on collection of documents obtained from various news outlets. On the other end, the dominating categories for Manual approaches tend to be more diverse, showing (in addition to news) the presence of corpora stemming from social media platforms and scientific literature. To further describe the 19 highlighted categories, Table 5.15 describes the different shapes that individual **documents** take in the included research. Here, it is possible to observe that is not uncommon for studies to construct their corpora only around "portions" of what would normally be considered a full document (e.g. using only the title and abstract of scientific publications). Additionally, it is also found that often times metadata is added as additional information to the document content. One again, a comparison is drawn with the presented labeling approaches. As expected, this comparison reveals that in most cases the prevalent document types match the previously identified categories of corpora. For example, the most prominent document type for Fully Automated approaches is a news article (which matches the first corpora category represented by "News"). Surprisingly, here one can additionally observe that (despite having its most prominent corpora types suggesting otherwise) Manual approaches tend to include in a substantial amount of cases documents represented by reviews.

Finally, a brief overview is given with regards to the observed document **pre-processing steps** (Data Item 20). In the included research, a total of 19 different techniques are non-uniquely (i.e. in more than one paper) used to prepare the collected corpus for the subsequent modeling phases. In a general sense, the registered pre-processing activities are found to be in-line with the kind of techniques one would expect to find in relation to work including a topic modeling step. Here, the activities of (1) removing stopwords and short terms, (2) removing special characters and punctuation, (3) removing infrequent words and (4) tokenising the documents are the most prominent across the entirety of the analysed collection. As a last effort to further describe the information found within this last Data Item, the encountered pre-processing steps are presented (in Table 5.13) together with the topic modeling approaches that they characterise.

Nr.	Corpus	Sources	Total
1	Scientific literature	WoS (7), Scopus (5), ACM (4), IEEExplore (3), arxiv (2), DBLP (2), ScienceDirect (2), Springer (2), Wiley (2), Aminer (1), British Academic Written English Corpus (1), Elsevier (1), Google Scholar (1), Sc-art (1)	24
2	News	20NewsGroup (5), Associated Press (4), NYT (3), Reuters (2), AgNews (1), Finnish News Agency (1), Huffington Post (1), LexisNexis (1)	23
3	Social media	Twitter (18), Reddit (2)	22
4	Medical literature	PubMed (6), WoS (4), Arxiv (1), BioASQ (1), Public health system (Spain) (1), Scopus (1)	12
5	Miscellaneous	Wikipedia (7), British National Corpus (1)	7
6	Online shopping	Amazon (5), Google Play Store (2), App Store (1)	7
7	Tourism	TripAdvisor (4), AirBnB (3), Expedia (2), Yelp (1)	7
8	Food	TripAdvisor (3), Citysearch New York (1), Yelp (1)	5
9	Finance	Consumer Financial Protection Bureau (1), Dow Jones Industrial Average (DJIA) (1), Refinitiv EIKON (1), Netquest (1), Reuters (1)	4
10	Online teaching	Udemy (3), edX (2), Khan Academy (2), Class central (1), Coursera (1), Crehana (1), Domestika (1), Platzi (1)	4
11	Employment	Glassdoor (3), Turkopticon (1)	4
12	Narrative literature	GoodReads (1), History Museum (1), Internet Archive American Libraries collection(1)	3
13	Q&A	StackExchange (2), StackOverflow (1)	3
14	Entertainment	IMDB (1), Storium (1)	2
15	Policy	TheyWorkForYou (1), WtP website (1)	2
16	Aviation	Aviation Safety Reporting System (1)	1
17	Law	Hathi Trust (1), Internet Archive (1), Project Gutenberg (1), The Text Creation Partnership for Early English Books Online (1)	1
18	Manufacturing	European Patent Office (EPO) (1), United States Patent and Trademark Office (USPTO) (1), World Intellectual Property Organization (WIPO) (1)	1
19	Music	Pitchfork.com (1)	1

Table 5.14: Encountered corpus' domains and associated sources

Nr.	Corpus	Document type	Total
1	Scientific literature	Paper's content (10), Abstract (5), Title and Abstract (2), Title, Abstract and Metadata (2), Abstract and Metadata (1), Abstract, Title and Metadata (1), Paper's content and Abstract (1), Citation statement (1), Title, Abstract and Keywords (1)	24
2	News	News article (23)	23
3	Social media	Tweet (14), Tweet and Metadata (5), Blog post (1), Comment (1), Post content (1)	22
4	Medical literature	Paper's content (4), Title and Abstract (3), Citation statement (1), EHR (1), Title, Abstract and Author (1), Title, Abstract, Author and Metadata (1)	11
5	Miscellaneous	Wikipedia entry's content (6), Extract (1)	7
6	Online shopping	Review (5), Review and Metadata (2)	7
7	Tourism	Review and Metadata (5), Reivew (2)	7
8	Employment	Review (2), Review and Metadata (3)	5
9	Food	Review and Metadata (3), Review (1), Webpage content (1)	5
10	Finance	Analyst report (2), Complaint record and Metadata (1), Earnings call (1), Survey response (1)	4
11	Online teaching	Course module content + Metadata (2), Course module content (1), Review (1)	4
12	Q&A	Question thread (1), Review (1), Single user Q&A (1)	3
13	Narrative literature	Book (1), Diary entry (1), Review (1)	3
14	Entertainment	Review (1), Story (1)	2
15	Policy	Single MP contribution (1), Petition title and Content (1)	2
16	Aviation	Safety report (1)	1
17	Law	Document chunk and Metadata (1)	1
18	Manufacturing	Patent document (1)	1
19	Music	Review (1)	1

Table 5.15: Encountered corpus' domains and associated document types

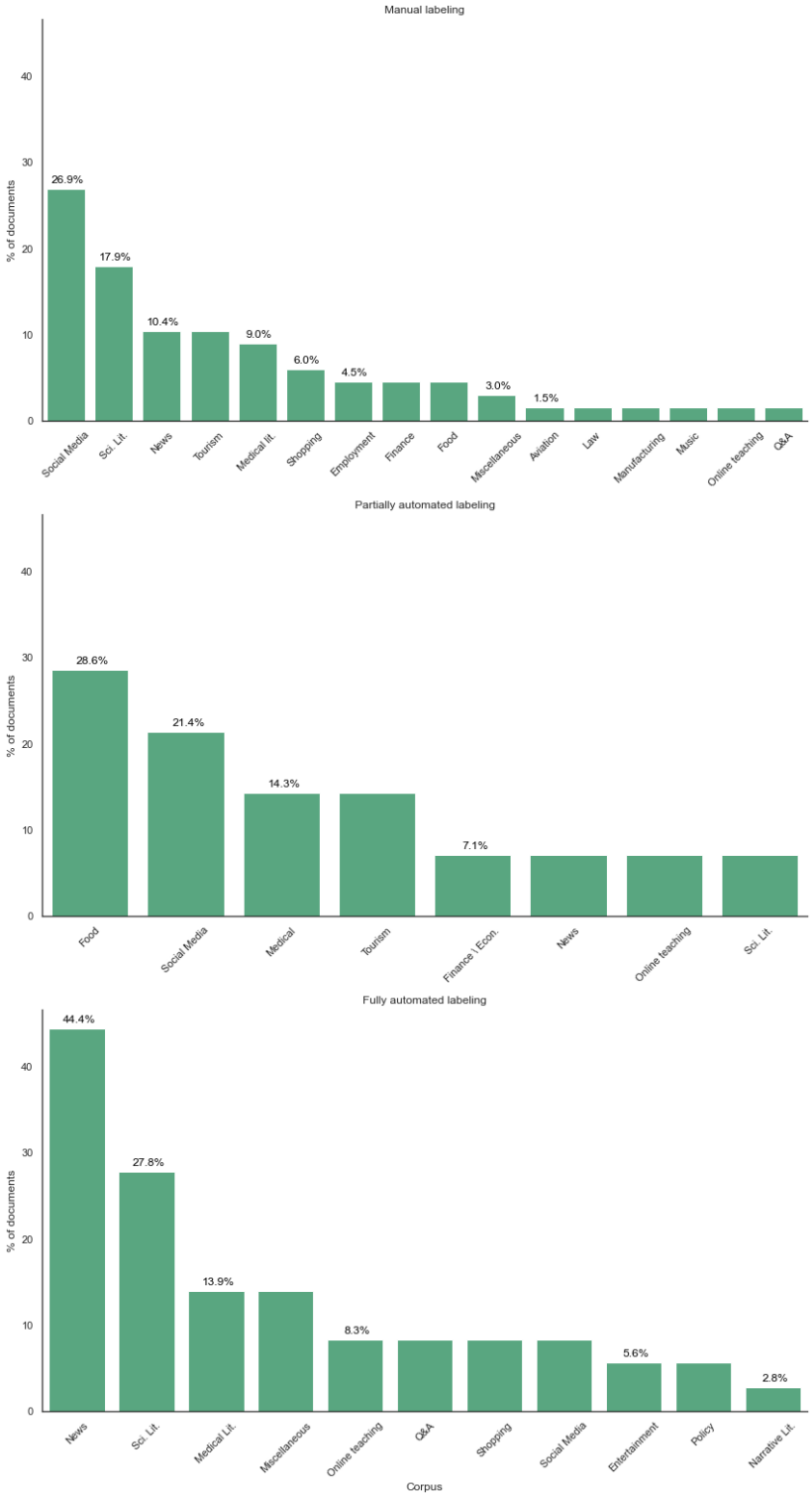


Figure 5.6: Labeling approaches across the identified corpora

Chapter 6

Identified Gaps in the research

6.1. Gap 1 - Lack of forward applications of labeling techniques

In a general sense, meaningful contributions tied to a given study should be expected to carry over to future research, where they can be used as practical tools in the research process, improved and extended in their structure or simply used as baselines methodologies when introducing conceptually comparable approaches. The analysis of the first gap therefore pertains to (1) the degree to which the labeling techniques identified within the collected studies are **applied (in a practical manner) by future researchers** operating in the field, and to (2) the **role that such applications play** in the interested research. Here, the scope of the localisation step tied to such an observation is limited to the set of **29 documents** that were found to present novel fully automated approaches for labeling topics (i.e. the studies represented in Table 5.3). In this regard, it should be underlined how this analysis could conceptually be further extended to all documents presenting novel techniques (even among Partially Automated and Manual approaches).

Verifying the gap

In this context, the activity associated with collecting the data required to understand whether this gap is actually present (i.e. **verification**) can be expressed as a set of **forward searches** performed on the subset of interested papers. For this purpose, the SemanticScholar repository is used to automatically gather the set of citing documents tied to each source study. Then, for each of the newly gathered papers the relevant **citation statements** (i.e. the sentences were a reference is made to the source document) and associated contexts are analysed in order to understand the **role** that the given source document (and corresponding labeling technique) plays in the citing study. Once again, it is important to understand that the main goal for this activity is identifying only the citing work where the identified Fully Automated labeling techniques are implemented in a practical sense. In this context (and as a general example), citations made in the related work section are generally not relevant, because the normally represent only a "superficial" reference to the source work. On the other hand, citations made in "core" sections of a given study should be looked at more carefully, because they are more likely to refer to the implementation of practical approaches. Additionally, it is reasonable to assume that forward implementations of the proposed labeling techniques would (at least in some cases) require the original authors to share some additional information (generally expressed in the form of **code repositories**) in order to allow prospective researchers to easily re-implement the described approaches. Because of this reason, in the context of this analysis the information tied to the presence of external repositories is also recorded for each of the respective source document. Also, one might be interested in verifying whether the recorded forward implementations are performed by (some of) the **same authors** that contributed to originally presenting the corresponding technique in the related source paper. All these information are collected and ultimately presented in Table 6.1. An analysis on the general meaning of the collected data immediately follows.

Paper	Total citations	Forward appl.	Shared repo.	Role	Citing paper	Author overlap
Zhang et al. (2018a)	75	7	✓	Baseline	Lee et al. (2022a)	✓
					Huang et al. (2020)	✓
					Lee et al. (2022b)	✓
				Customisation	Budiarto et al. (2021)	
				Main approach	Mireles et al. (2022)	
					Yang et al. (2019)	✓
					Zhao et al. (2023)	
Lau et al. (2017)	55	7	✓	Baseline	Gao and Ren (2019)	
					Guo et al. (2019)	
					Wang et al. (2017)	
					Gupta et al. (2018)	
					Tang et al. (2019b)	
					Rezaee and Ferraro (2020)	
					Wang et al. (2019)	
Wang et al. (2021)	29	1	✓	Baseline	Li et al. (2022)	
Zhang et al. (2019)	26	1		Main approach	R. and Nag (2023)	
Zhang et al. (2021)	26					
Allahyari et al. (2017)	21					
Aletras and Mittal (2017)	13	1		Baseline	Sorodoc et al. (2017)	✓
Hu et al. (2020)	13					
Alokaili et al. (2020)	12	1	✓	Baseline	Zosa et al. (2022)	
Sorodoc et al. (2017)	12					
Adhitama et al. (2017)	10					
Popa and Rebedea (2021)	10	2	✓	Baseline	Zosa et al. (2022)	
				Main approach	Llerena et al. (2022)	
Gourru et al. (2018)	9	2		Support activity	Popa and Rebedea (2021)	
				Main approach	Velcin et al. (2018)	✓
He et al. (2019)	8	2		Improvement	He et al. (2021b)	✓
				Baseline	He et al. (2021a)	✓
Pergola et al. (2021)	8	1	✓	Main approach	Caton et al. (2021)	✓
Béchara et al. (2021)	7					
Kim and Rhee (2019)	7					
Chin et al. (2017)	4					
Truică and Apostol (2021)	4	1		Main approach	Apostol et al. (2023)	✓
He et al. (2021b)	2					
Campos et al. (2020)	1					
He et al. (2021a)	1					
Zosa et al. (2022)	1		✓			
Hosseiny Marani et al. (2022)						
Rosati (2022)						
Scelsi et al. (2021)			✓			
Yin et al. (2022)						

Table 6.1: Forward applications of fully automated topic labeling techniques

Encountered forward implementations

Starting from the presented data, the first observation that can be made is that among the set of 23 publications cited by at least one document, **only 11 cases** show some practical application of the presented labeling techniques in future research. By itself, this information already suggests a general **lack of practical use** for the presented approaches in subsequent work. Additionally, this notion is further reinforced when comparing the **total number of citations** characterising the individual papers with the associated number of future applications. Here, one can notice that in all cases the citations that were identified to contain forward approaches represent only **small minority** of the total work collected during the forward search process. In this regard, for most of the interested papers (9 of the 11) one can see that only 1 or 2 practical implementations are found among studies having collections of citing papers with sizes ranging from a minimum of 4 to a maximum of 26 documents. Additionally, it is also necessary to underline that the two documents showing a relatively higher number of future applications (Zhang et al., 2018a; Lau et al., 2017) are both studies proposing topic modeling approaches which happen to include a labeling step. In fact, Zhang et al. (2018a) proposes a novel unsupervised hierarchical clustering method (TaxoGen) to generate topic

taxonomies, where the label generation phase is implemented as part of the main taxonomy construction method. On the other hand, Lau et al. (2017) introduces a technique (tdlm) composed by a topic model associated with a language model and capable of generating topics that can be automatically interpreted via sentences generated by the LSTMs structures implemented as part of the existing language model. Given this notion, it is reasonable to assume that in both cases the fact that the techniques introduced by the two studies also offer **broader approaches** (which are capable of generating topic distributions in addition to providing potentially suitable labels) had an effect on the decision of future researchers of incorporating such techniques into their own studies.

As a final observation, one can also notice that the **most common role** played by the encountered techniques in future research is the one of a **baseline approach** (this is found to be the case in roughly 57% of cases). In a general sense, this further characterises the identified gap by suggesting that the encountered techniques are mostly used as comparison mechanisms for novel approaches, rather than as practical tools aiding in the labeling activities found in subsequent studies.

Recurring authors

Looking more in detail at the newly collected documents, it should be also pointed out that in a substantial number of cases (for almost **40% of the retrieved studies**) one can observe some level of **overlap among authors** between some studies proposing the original techniques and the related documents gathered during the forward search. This information is particularly revealing, and in a sense can be interpreted as further reinforcing the notion that forward uses of labeling approaches are found to be generally lackluster with regards to their practical usefulness in future studies. In fact, having a labeling approach re-used by (some of) the same authors can suggest that the technique has not found a wider audience among researchers operating in the field. This can be considered to be especially true when re-implementations of this kind are the only one found for a given source document (which is indeed the case for 4 of the 7 documents interested by author overlap).

Availability of code repositories

At this point, one could hypothesise that a relatively straightforward way to help in addressing this gap would be to make the proposed technique more reproducible to prospective authors that might be interested in implementing them into their own research. In fact, it is reasonable to assume that requiring future authors to re-implement the proposed approaches from scratch is very likely to represent an hindrance in fostering the future use of such approaches. This assumption is further reinforced by Truică and Apostol (2021), which states that:

> *"We could not consider for comparison the majority of the existing methods in the literature [...] as the authors do not provide the raw dataset or the source code, only the extracted topics and labels"*

To see if this hypothesis is also supported by the collected data, the public availability of the labeling techniques making up the original collection of documents is verified by searching for related **code repositories**. Here, **only 8 documents** (roughly 30% of the analysed collection) are recorded to be **supplemented by publicly available code**. Of these 8 studies, **6 are recorded to have forward implementations**. For the other cases, it is possible to verify that such an omission generally signifies a lack of re-implementation of the proposed techniques in future papers. In fact, one can observe that of the **19 documents lacking any form of supplemental technical material, 14 do not present any forward implementation** of the labeling approaches (notice also that 3 of these 14 are also lacking any citation in the total count). Additionally, another three documents (Aletras and Mittal, 2017; He et al., 2019; Truică and Apostol, 2021) belonging to this subset have subsequent implementations found in studies showing some level of author overlap. In this context, one can assume that the original techniques were made accessible directly by the authors that contributed to the original research, and would have not been usable otherwise. Here, the only real exception to this trend is

shown in the cases of Popa and Rebedea (2021) and R. and Nag (2023), where the techniques proposed in Gourru et al. (2018) and Zhang et al. (2019) respectively are re-implemented despite the lack of publicly available sources and no overlap between authors of each study pair.

Addressing the gap

Given the collected information presented throughout this Section, a few considerations can be made to (at least partially) address the shortcomings dictated by this first identified gap and with the hope of aiding prospective authors in avoiding similar limitations in their own research. Firstly, the conducted analysis revealed a distinct lack of forward use for most of the identified (fully automated) approaches. In other words, it is observed that the contributions contained within the analysed novel approaches tend to remain confined to the studies in which they are originally presented. Often times, subsequent implementations are only brought forward by the same authors that contributed to the original work, suggesting that the interested techniques ultimately failed to find more widespread applications. Additionally, when such approaches are indeed re-proposed in subsequent studies, they are often relegated to the role of baseline measures (used as a comparison metric for the main method of the given document) rather than primary tools exploited for the generation of novel findings. In this regard, it has been noted that, when proposing a new approach, making the relevant implementations freely available to the readers substantially helps in boosting subsequent utilisations.

Therefore, it is suggested that prospective authors considering to work in this field should strive to **make the relevant code repositories accessible** to readers and possibly **supplement them with comprehensive documentation** allowing for external users to easily re-implement the described techniques in the context of their future work.

On the other hand, the lack of forward uses for the proposed approaches suggests a general disinterest in utilising already existing labeling techniques as a tool to (partially) automate the topic labeling process. In this context, it is suggested that researchers including such an activity in their own studies **should make themselves aware of the previously introduced techniques** capable of satisfying this requirement. For this purpose, readers of this review can use the content of Table 6.1 and the insights brought forward in Chapter 7 as guidelines to accessing an initial set of (reproducible) methodologies relevant for this task.

6.2. Gap 2 - Shortcomings in the quality evaluation procedures

As one could infer from the information gathered with regards to Data Item 16 and presented in Subsection 5.4 - *"Label selection and quality evaluation"*, the collected research tends to very rarely show implementations of quality evaluation procedures conducted on the generated labels. This creates a situation where the provided identifiers rarely come with a formal guarantee (expressed by means of suitable quality metrics) verifying their ability to address the content-level information tied to the associated distributions. Naturally, one can imagine that the issue caused by this omission might have varying degrees of impact depending on the specifics of the individual studies affected by it. In fact, it is reasonable to assume that the supervision of domain experts when conducting a given topic modeling and subsequent labeling activity might be enough to (informally) ensure that the provided labels are indeed sufficient to address (and describe) the overarching narrative required to address the posed research questions. Nonetheless, having procedures in place to properly demonstrate the qualitative properties of the proposed identifier can be considered as generally desirable trait for a given study to include.

Given this premise, and starting from the relatively straightforward observation bringing this gap to light (represented by the fact that **only roughly 20% of studies include label quality evaluation procedures** of any form), this analysis attempts to further characterise this issue by also exploring **how the few encountered evaluation steps are conducted**, their potential shortcomings and how these might be addressed by prospective authors.

Verifying the gap

As previously stated, the fact that only a relative minority of the included studies (22 distinct documents) go to the lengths required to compute any quality metrics tied to the generated identifiers generally suggests that **such activities should be considered an exception** (rather than a commonly encountered task) for any research including some form of topic labeling step. In a first attempt to further profile this gap, the primary studies interested by these quality evaluations are compared to their respective **focus with regards to topic labeling** (as observed from Data Item 6). Here, the goal is to observe whether the evaluation of labels is conducted only (or primarily) by publications focused on proposing labeling techniques, but otherwise ignored when practical applications of such activities are conducted within more general studies, where the focus is instead shifted to the results produced by the topic modeling and subsequent labeling endeavors. In this context, it is reasonable to assume that studies primarily focusing on topic labeling would be more likely to provide evaluations for the produced results. Indeed, the collected data shows that **17 of the 22 interested document show such a primary focus**. Additionally, it is also revealed that for **3 of the remaining 5 documents** where labeling represents a secondary step only an **informal evaluation** is conducted, where authors simply reaffirm the suitability of the generated labels with regards to the underlying topics (and related corpus).

In other words, the collected data shows that (1) studies where topic labeling is not of primary focus tend not to conduct quality evaluation steps on the generated identifiers and that (2) the few that do generally don't go beyond a very brief and informal evaluation procedure. Unfortunately, this information further aggravates the implications of this gap, and suggests that the provision of suitable quality metrics for topic identifiers is **severely lacking** among the included studies.

Evaluation criteria and dimensions

At this point, given the fact that the existence of the presented gap has been verified and its nature further contextualised, it might be useful to observe how the few instances of the encountered quality evaluation activities are shaped, and which are the potential pitfalls that might characterise them. In this regard, a good starting point for this analysis lies in the observation of the **number of different evaluation criteria** observed among the relevant documents, and the degree to which they **explicitly consider multiple dimensions** in the evaluation of the labeling results. Here, it is reasonable to expect that an evaluation based on multiple dimensions (or factors) might be more effective in assessing the quality of the given labels than a singular goodness-of-fit measure (i.e. good fit / bad fit). Additionally, including more than a single evaluation criteria (or technique) might ultimately be useful to better ensure the soundness of the provided quality evaluation results. Here, notice that utilising more than one criteria does not automatically ensure that multiple dimension are indeed taken into account when assessing the generated labels and that, conversely, it is possible include a single evaluation technique capable of capturing multiple factors in the generated labels.

Looking again at the collected work, it is shown that **7 studies** do include **more than one criteria** for evaluation (in quantities ranging from 2 to 5). In this regard, **four documents show a mix of manual and automated criteria** for producing the relevant quality metrics whilst the remaining three (He et al., 2021a,b; Rosati, 2022) solely rely on approaches not requiring human intervention.

Of these, **6 papers** are set up in such a way as to **evaluate more than one factor** in the generated identifiers. In particular, He et al. (2019, 2021a,b) use the same (previously described) metrics of **relevance, coverage and discrimination** to numerically assess the labels with regards to their ability to (1) accurately convey the meaning of the topic, (2) comprehensively cover the top terms associated with the distribution and (3) uniquely (with minimal overlap to the other labels) describe the given topic. Additionally, Aletras and Mittal (2017); Popa and Rebedea (2021) both evaluate the top label candidates in relation to the provided **gold-standard** ones (using the nDCG metric) and then proceed in a further **manual assessment** (using a four level scale) made with regards to the associated topic terms and documents.

Importantly, a last paper (Hosseiny Marani et al., 2022) titled *"One Rating to Rule Them All? Evidence of Multidimensionality in Human Assessment of Topic Labeling Quality"* goes deeper into the

issue of **multidimensional evaluation** by providing **concrete evidence** showing that (human) evaluation of topic label quality does indeed have multiple underlying dimensions and that exploiting them can ultimately lead to findings that would not be apparent when using single-factor metrics. Here, technically the labels are manually evaluated by means of fourteen distinct items which are ultimately grouped together (by Exploratory Factor Analysis) into **two distinct dimensions** (highlighting to which degree a given label is **Suitable** and **Objectionable**). The observation associated to the potential importance tied to considering multiple dimension in the label quality assessment process is explored in more details in the final part of this Section, where the findings of Hosseiny Marani et al. (2022) are used for supporting the idea that multiple factors should indeed contribute to the quality evaluation procedure, at the very least in those instances where human assessors are involved in such a process.

In summary, whilst acknowledging that a few of the interested studies ultimately included more than one evaluation criteria, one must notice that (1) only a small minority of documents (4 out of the total 22 including a label evaluation procedure) ultimately considered a mixture of human-based and automated techniques and that (2) the included research was similarly lackluster in the task of encapsulating multiple dimensions (or factors) in the evaluation procedure.

Information considered by the evaluation criteria

Another important consideration to be made which is conceptually relevant for both human-driven and automated evaluation procedures relates to the data that is actually considered (or made available) at evaluation time. In fact, it is reasonable to believe that, for instance, a human assessor would be more likely to accurately evaluate the produced identifiers if in addition to the most prevalent terms characterising a given distribution he/she could also be exposed to additional information such as the top documents most strongly associated with the given topic. For a given topic, **4 individual types of information are identified** as basis for such evaluation procedures: (1) the top topic terms, (2) its most strongly associated documents, (3) available gold-standard labels and (4) other generated labels.

By far, the most commonly encountered way used to ground the generation of the provided quality metrics is the use of prominent topic terms. Here, the available data shows that a total of **9 studies** include approaches for evaluation that **only consider such terms** in the generation of the corresponding quality metrics. Naturally, this can be seen as somewhat limiting take on topic quality evaluation, considering the fact that depending on the domain at hand, the terms representing a given topic distribution might not always sufficiently gauge the expressive abilities of a given label. In a more general sense, one can expect that most evaluation criteria using only a single kind of information would be likely to present the same drawback. Therefore, a similar kind of limitation can also apply to those criteria **solely considering the most prevalent documents** tied to a given topic (which are found in 5 separate instances). Conceptually, one potential way to address this issue would be to include both kinds of information when assessing a given label. Unfortunately, among the set of primary study for which an evaluation step is provided, **only for 2 documents** (Gourru et al., 2018; Popa and Rebedea, 2021) it is explicitly stated that for each label **a combination of prevalent terms and documents is taken into account** by the human assessors involved in the process.

As mentioned in the previous Subsection, it is also possible to **exploit as (part of) the quality evaluation procedure the other identifiers** produced during the topic labeling phase. Naturally, the Coverage metric originally proposed in He et al. (2019) based on this kind of information represent a fairly specific kind of evaluation methodology and would hardly be sufficient (on its own) to fully characterise the qualitative properties of the topic labeling output. This is demonstrated by the three papers in which Coverage is used, and where the metric is paired up with additional criteria (namely Relevance and Coverage) for a more complete assessment of the produced results.

Finally, the use of **Gold-standard labels** as a comparison mechanism to the generated ones is found (as a last category of information) to be used for quality evaluation purposes by different approaches in six distinct studies. Here, it is important to mention that whilst representing one of the most valuable sources for evaluating the quality of the generated identifiers, gold-standard labels are still a relatively niche information, which tends to be only accessible in those instances where

characterising the underlying corpus is generally not the primary focus of the given study. In fact, the simple existence of reference labels implies that the existing corpus (and associated topics) have been explored prior to the proposed research.

In conclusion, the collected data shows that (1) most of the encountered evaluation criteria only consider one kind of information (i.e. topic terms, documents, gold-standard labels or other generated labels) for generating a quality assessment value and that (2) in only two instances (involving a human rating procedure) both topic terms and documents are used in the assessment of the individual identifiers.

Considerations on human ratings

As previously specified, having some form human involvement in the quality evaluation procedure represents by far the most common approach for assessing the generated labels. In the documents interested by such evaluation steps, one can notice that criteria characterised by some form of **human rating are found in 16 out of the total 22 documents**. This is not surprising, as similarly to what as been observed for the generation of topic labels, human ratings represents the most intuitive (and easily accessible) form of evaluation available to authors. Despite this fact, researchers should generally take notice of the fact that the involvement of human evaluators is inherently tied to the introduction of **bias** and **subjectivity** into the evaluation process. Depending on the domain of interests and the involved assessor, these factors might have varying degree of influence on the produced evaluations. For instance, one can expect that authors providing an assessment to the labels generated as a result of their proposed methodologies might be naturally biased towards favouring a more positive outlook with regards to their overall quality. In a more general sense, one should consider that such pitfalls might be fairly difficult to address with success when introducing a human factor in the evaluation process. Indeed, when examining research on the matter in the field of natural language processing, Hämäläinen and Alnajjar (2021) recognise these issues and further add that often times a **lackluster definition of problems** that specific tasks (e.g. topic labeling) are trying to address can lead to human assessor not having sufficient context for conducting a proper evaluation procedure. Ultimately, it is stated that:

> *"Human evaluation [...] is certainly not a straight forward problem due to a variety of different reasons, the largest of them being **subjective interpretation** and **limited understanding** the human evaluators have of the evaluation task, questions and the actual output that is to be evaluated".*

Here, the authors take notice of the fact that for natural language processing tasks: *"human evaluation is [often] not conducted in the same rigorous fashion as in other fields dealing with human questionnaires"* and that, consequently *"at the current stage, the **validity** of many human evaluation methods is **questionable**"*.

Now, whilst recognising that at the present time no definitive solution exists to fully address these issues, it is hypothesised that the involvement of **domain experts** in the evaluation procedure might help in partially alleviating this problems. Unfortunately, among the relevant documents only three studies (Allahyari et al., 2017; Gourru et al., 2018; Stamolampros et al., 2019) explicitly state the involvement of such experts in the quality assessment phase.

Addressing the gap

To conclude this analysis, a set of pointers is formulated with the purpose of providing a level of guidance to prospective authors that might want to avoid the shortcomings that emerged from the analysis of the identified gap. Notice additionally that a tabular summary of the data used to describe this gap is highlighted in Table 6.2. To this end, an obvious observation made with regards to the collected work to initially profile this gap is that in most cases (roughly 80% of the included studies) a quality evaluation phase is found to be completely absent with regards to the generated labels. Naturally, such a frequent omission of a proper assessment phase makes it generally difficult

Paper	Sec. Focus	Criteria	Multiple factors	Considered info.	Expert inv.
Aletras et al. (2017)				Docs	
Aletras and Mittal (2017)		2	✓	Gold-standard l., Terms	
Allahyari et al. (2017)				Terms	✓
Smith et al. (2017)				Docs	
Sorodoc et al. (2017)				Terms	
Gourru et al. (2018)				Terms + Docs	✓
Kuhn (2018)	✓			Docs	
Zhang et al. (2018a)	✓			Terms	
He et al. (2019)		5*	✓	Terms + Labels	
				Terms	
Stamolampros et al. (2019)	✓			Terms	✓
Alokaili et al. (2020)				Gold-standard l.	
Campos et al. (2020)	✓			Gold-standard l.	
Mukherjee et al. (2020)	✓			Terms	
He et al. (2021a)		3*	✓	Terms + Labels	
He et al. (2021b)		3*	✓	Terms + Labels	
Popa and Rebedea (2021)		2	✓	Terms + Docs, Gold-standard l.	
Scelsi et al. (2021)				Terms	
Truică and Apostol (2021)		4		Terms	
Campos et al. (2022)				Gold-standard l.	
Hosseiny Marani et al. (2022)			✓	Docs	
Rosati (2022)		2		Docs	
Zosa et al. (2022)				Gold-standard l.	

Table 6.2: Information considered in the characterisation of Gap 2

* Notice that Relevance, Coverage & Discrimination have been counted as separate criteria.

to ensure the credibility of the (labeling) outputs associated with an individual study and reveals itself to be even more problematic when one tries to draw a clearer picture of the current state of the research. Additionally, these issue become even more apparent when topic labeling does not represent a primary focus in the proposed research, where formal evaluation procedures are observed to be virtually absent in such settings. To this end, the general suggestion made to future researchers is to **strive to include at least a basic form of** (preferably non-informal) **quality evaluation activity** in the presented studies. Notice that for further details on such activities one can refer back to the individual descriptions of evaluation techniques provided in Subsection 5.4.2 - *"Label selection and quality evaluation"*.

Following this initial observation, it should additionally be noted that if one indeed decides to provide numerical ratings for the generated labels, the **inclusion of more than one criteria** (i.e. technique) for evaluation is advisable as a way to introduce some level of redundancy which can help in increasing the credibility of the proposed results (by showing that multiple techniques agree on the quality of the identifiers). Additionally, one can consider to **propose a mix of human-based and automated criteria** (as observed in four distinct studies) to further diversify the overall procedure.

In relative close relation to the notion of utilising distinct criteria in the evaluation process, one of the most important observation made within the context of this gap relates to the potential relevance of trying to **employ a multidimensional approach** when assessing the quality of topic labels. In this regard, the findings brought forward by Hosseiny Marani et al. (2022) reveal themselves to be fundamental in providing a way to fully address this gap, especially in relation to human ratings. Here, the authors hypothesise that (H1) human judgment associated with the quality of topic labels might ultimately have multiple underlying dimensions and that (H2) exploiting these addi-

tional dimension can ultimately provide more robust assessment when compared to the use of a single-dimensional metric (i.e. good fit vs bad fit). To prove the first hypothesis (and as previously highlighted) a thorough label evaluation phase is conducted on the basis of 14 distinct items which are ultimately grouped together by means of Exploratory Factor Analysis (i.e. based on their tendency to co-vary amongst assessments) into two overarching groups (or factors). Here, the first factor is named **"Suitable"** and refers to the degree to which a given label suits the related documents (i.e. the label can be considered *"sensible, meaningful, or expected given the text to which it is applied"*). On the other hand, the second factor is named **"Objectionable"** and ties to the notion that the choice of the given identifier might offer a biased viewpoint and be prone to be contested (i.e. the label could be *"biased, offensive, or likely to spark disagreement when applied to the given text"*). In the context of the second hypotheses, the authors notice that the use of the two identified dimensions for evaluation allows to outline observations on the identified labels that would not have been otherwise possible when using single item measures. Here, the concluding statement of this work provides a suitable suggestion (to prospective authors) to address this particular issue:

> *"This work [...] suggests that researchers must account more fully for the complexity involved in topic labeling [...] We must* ***develop means of assessing performance that go beyond asking whether a given data point is assigned the correct label****. [...] Moreover, as suggested by the results present here,* ***quality may be multifaceted****. A given label may seem both well suited yet simultaneously offensive or biased. Some labels or topics may seem initially confusing but, upon further inspection, reveal unanticipated insights."*

When evaluating topic labels, another important aspect relates to the information that are considered to produce the relevant quality metrics. Here, it is noticed that the four individual types of data used to ground such metrics (topic terms, documents, gold-standard labels or other generated labels) are almost always used in isolation by the encountered techniques. To address this issue it is suggested for a **combination of information types** to be used within the same evaluation criteria (e.g. topic terms and associated documents). This suggestion finds particular applicability in the context of manual ratings, where human assessors can incorporate additional information in the evaluation procedure with relative ease.

As a final step in fully addressing the presented gap, a pointer is provided to alleviate the previously highlighted shortcomings of human-based evaluation tied to inherently existing biases and subjectivity concerns. Here, authors should consider that (when possible) **involving domain experts** (separate from the authors of the respective work) should be considered in order to include more informed (and likely less biased) evaluations required to assess the performance of the presented process.

6.3. Gap 3 - Lack of sufficient research on alternative representations

As highlighted in Subsection 5.4.1, the vast majority of analysed work uses some form of n-gram identifier for labeling the generated word distributions. In fact, it is noted that only **13 distinct studies** (out of the total 108) are found to contain labels differing from the prevalent category. Here, outputs taking the form of images, sentences and paragraphs offer alternative takes on topic labeling which might be potentially useful in providing more meaningful and informative descriptions than their n-gram counterparts. Therefore, within this section the collected data is used as a starting point to:

1. Understand to which degree the current state-of-the-art reveals itself to be **lackluster** with regards to non-traditional (i.e. non n-gram) representations for topic labels.
2. Verify whether there exists any evidence that these more elaborate forms of expression are better suited (or **more effective**) in capturing the underlying meaning of the topics they describe.

As a first step, the gap is verified and further described by conducting a closer observation on the interested (already collected) studies with regards to their respective abilities to justify whether the

proposed (non n-gram) representations can be formally shown to yield any improvement on more established approaches to labeling. Then, an attempt is made to extend the scope of the collected research on both sentence-based and image-based approaches by conducting a backward search starting from (some of) the relevant documents. Here, the objective is to find whether other meaningful research which might exists outside the scope of the review can be capable of bridging the aforementioned shortcomings. Finally (and in a similar fashion to the previous gaps), the analysis is concluded with the formulation of some pointers with objective of helping to address the identified limitations.

Verifying the gap

Whilst the verification of (1) relies on the straightforward observation that the presence of **alternative label representations tends to be generally lacking** in the collected research (as stated in the introduction to this Section), it is clear that (2) requires a more thorough investigation in order to be confirmed. In this regard, the **13 studies** interested by such representations are **individually analysed** to asses whether they offer any solid (i.e. formal) comparison with traditional n-gram labels. Additionally, when such a formal comparison step is missing, it is observed whether the interested studies provide at least an n-gram representation to accompany the sentence- or image-based labels and whether any relevant comment is made on the different nature (and effectiveness) of the existing approaches. Notice also that the data collected at this stage is used as a starting point to ground the backward search described in the following Subsection and required to fully characterise this third gap.

Starting from the data associated with this initial analysis, it is possible to observe that only **two** of the relevant studies provide any form of **formal comparison** approach to traditional labeling representations. In this context, Aletras et al. (2017) verifies the impact of different topic representations (term lists, n-grams and images) with regards to (1) their effectiveness in an information retrieval task and (2) their relevance as assessed by human annotators. Here, the results of (1) show that given a certain query, human assessors are generally faster at identifying relevant topics when those are represented using n-gram labels or images rather than word lists. At the same time, it is shown that, among the three representations, term lists (i.e. raw topics) show a slightly higher precision value (given by the ratio of correctly identified topics for a given query) than the other representations. Because of this reason, the authors conclude that:

> *"Overall, textual phrases [i.e. n-grams] and image labels **can be interpreted more quickly** than term lists, **but not as accurately**".*

For (2), the evaluation given by humans using a four level scale reveals that term lists are the most effective way to represent the meaning of a given topic. Interestingly, whilst n-gram labels still show precision values relatively similar to term lists, one can observe that:

> *"**Results for image labels are substantially lower**. This suggests that the image labels are not as clear as are the other two types [of representations], making it difficult for annotators to identify the correct [topic]".*

As a general remark to (2), the authors also underline that: *"The information in term lists was found to be more accurate, which is to be expected since the labels are effectively summaries of the topics"*. On the other hand, in Smith et al. (2017) the generated sentence and n-gram labels undergo (separate) evaluation procedure where the best and worst identifiers among the set of labels associated with a given topic (and originally elicited starting from different topic representations) are selected by human assessors. Here, whilst not drawing a direct (numerical) comparison between the two techniques, the authors note (from the conducted label analysis that:

> *"The labels voted as "best" were **shorter on average** than those voted "worst"".*

In addition to the two aforementioned documents, one can observe that an additional two studies (Wang et al., 2021; Wang and Hsu, 2020) show the generated sentences associated with corresponding n-gram labels and that one document (Sorodoc et al., 2017) does the same with image labels. All three of these papers **lack more formal comparisons** between the different representations. It is also worth mentioning that, when discussing the proposed methodology, the authors in Sorodoc et al. (2017) note that: *"The multi- modal evaluation yields the highest rating across all systems. This suggests that [...]* ***different topics may have different optimal label representations*** *(image or textual)"*. Finally, despite not providing corresponding n-gram labels, He et al. (2019, 2021a,b) make similar remarks with regards to the usefulness of summarisation task in generating topic identifiers. In particular, it is stated that: *"phrases are usually insufficient to interpret the discovered topics owing to their finite length, and images do not fit into most NLP scenarios. [...] Recently, researchers have been paying more attention to the use of automatic summarization technology to label topic models for obtaining more informative and meaningful topic labels"*.

Nonetheless, at this point the analysis carried out on the collected work **fails to provide any substantial evidence** that these alternative representations are indeed more effective at capturing the meaning of the related topics when compared to the more canonical n-gram labels. In fact, the little evidence available currently suggests a general preference of users for shorter label representations and highlights how images might be generally ineffective at accurately representing the meaning of a given distribution.

Backward search of alternative representations

In an attempt to further characterise both major shortcomings associated with the presented gap (the lack of alternative label representations and of sufficient evidence that such representation are beneficial in better describing the associated topics), the **introductory and related work sections** of the 13 relevant studies are analysed with the hope of finding additional existing research (falling outside of the presented review) that might be potentially informative for the current task. Here, any **referenced work** found within these two sections is collected and **manually inspected** in order to establish whether it contains findings that might ultimately help in bridging the gap brought to light by the studies originally included in the presented review. Naturally, the context provided by the original papers with regards to the referenced work is used as an initial hint in deciding which studies should be further inspected. In other words, it is important to notice that here the interest is not in conducting a "full" backward search where all the references tied to the original set of studies are collected and analysed. Instead, a more focused approach is taken to specifically identify those insights that might help in address the identified shortcomings.

Conducting such a reference collection procedure on the initial set of (13) relevant studies leads to the identification of **three additional documents** Aletras and Stevenson (2013); Wan and Wang (2016a); Barawi et al. (2017) containing additional information tied to the current gap. In this context, Wan and Wang (2016a) propose a summarisation algorithm based on submodular optimization capable of generating representative topic summaries. When evaluating the proposed methodology, the results of the summarisation task are compared to baseline approaches using term lists and n-gram labels. Here, **summaries** are found to have **substantially higher metrics for Relevance, Coverage and Discrimination** (which in this case are manually computed by human assessor using a four level scale) compared to the more canonical representations offered by term lists and n-grams. This information can be seen as a first evidence that a more elaborate (multi-sentence) label might indeed help in benefiting topic interpretability. On the other hand, Barawi et al. (2017) propose a sentence label ranking function (based on two metrics of relevance and sentiment co-coverage) used to label sentiment-bearing topics. Once again, a 4-point scale is used to evaluate the sentence topics against the results of some n-gram baselines. The result of this evaluation phase highlight that:

> *"The sentence labels [...] outperform all baselines significantly [. . .] and demonstrates the* ***effectiveness of our sentence labels in facilitating topic understanding*** *and interpretation"*.

So, once again a more advanced representation seems to ultimately be highlighted as beneficial by human assessors. Finally, Aletras and Stevenson (2013) provides an additional study on the generation of image labels (using a web querying technique) that unfortunately fails in providing suitable comparison with text-based representations. At this point, Table 6.3 is provided as an extension of the previously presented Table 5.11 and highlights all the identified documents including non n-grams identifiers. Here, the studies selected from the backward search are marked in green and all studies are additionally complemented by checkmarks indicating whether they contain n-gram labels and formal comparisons between the different types of label representations.

	Label structure		
Labeling tech.	**Sentence**	**Image**	**Paragraph**
Deep Neural Netowrk		Aletras and Mittal (2017), Sorodoc et al. (2017) $*^{1}$	
Graph-based		Aletras et al. (2017) $*^{1,2}$	He et al. (2019), He et al. (2021b)
Long short-term memory model	Lau et al. (2017)		
Manual labeling	Smith et al. (2017) $*^{1,2}$, Effrosynidis et al. (2022), Wahid et al. (2022)		
Manual labeling assisted by associated documents	Wang and Hsu (2020) $*^{1}$		
Multi-document summarization	Rosati (2022)		Wan and Wang (2016a) $*^{1,2}$, Rosati (2022)
Probabilistic approach	Barawi et al. (2017) $*^{1,2}$		
Search engine querying		Aletras and Stevenson (2013)	
Transformer-based	Wang et al. (2021) $*^{1}$		He et al. (2021a)

Table 6.3: Papers using non n-gram labels with additional results from backwards search

$*^{1}$ The paper includes n-gram labels.
$*^{2}$ The paper offers a formal comparison to n-gram labels.

In summary, among the three newly collected documents only two are found to give a better insight on the use of different kinds of topic label representations. Ultimately, one should note that the additional research brought forward by the conducted **backward search**, whilst useful in providing some data supporting the use of sentence and paragraph labels, reveals itself to be **insufficient to fully disprove point (2)** introduced at the beginning of this section. Additionally, some evidence exists (Smith et al., 2017) that in some cases users might prefer shorter labels. Also, one should note that no additional data is found with regards to the evaluation of image labels against textual representations. Here, the only identified comparison (Aletras et al., 2017) actually suggests that images might not be able to fully capture the meaning of the topics they are tied to. Therefore, at the present time the gathered data ultimately supports the identified gap and implies that:

1. The research focusing on non n-gram representation of topic labels is **severely lackluster**.
2. Whilst some evidence exists that certain alternative representations (namely sentences and paragraphs) might be ultimately beneficial in interpreting the associated topics, **more investigation is clearly required** to more confidently support this thesis.

Addressing the gap

Once again, the Section pertaining to the identified gap is closed by proposing a few indications to prospective authors stemming from the observed shortcomings identified in the collected data. In this case, giving clear pointers with regards to the usage of alternative label representations reveals

itself to be a somewhat trickier endeavor due to the distinct lack of application that such approaches are found to have in the current state-of-the-art. In fact, the currently gathered information are arguably not sufficient to confidently recommend any alternative representation over the more traditional approach represented by n-gram labeling. Here, whilst some evidence is found suggesting improvement brought upon by some of these label types (namely sentences and paragraphs), one must also state that counterevidence exists underlining the preference shown by users in certain instances for shorter labels and the limitations found for visual identifiers (i.e. images) in their ability of being informative with regards to the content of the associated topic. Intuitively, one can imagine that under specific conditions (dictated by the type of corpus being analysed and the associated domain) more elaborate representations such as sentences, paragraphs or images might be beneficial in providing more relevant identifiers for the generated topics. Unfortunately, it is currently believed that in practical applications the **domain knowledge** of authors and experts tied to a given task should still play a central role in determining **the most appropriate representation** for a given research objective and that a "catch-all" statement suggesting a best approach applicable to all scenarios cannot be formulated given the available information.

In this context, a relatively "safe" suggestions that can potentially lead to more complete studies would be to simply **produce** (within the same labeling task) **multiple label representations** tied to the same set of underlying distributions. In fact, whilst requiring more work during the topic characterisation phase, one can expect that having more than one identifier tied to each of the generated topics should generally lead to a more easily interpretable set of results. Additionally, when unable to determine an obvious (single) best technique for representing such topics, authors should strive to **include formal comparison metrics** assessing the difference in performance characterising the various representations.

Ultimately, the use of different kinds of topic labels and their performance in distinct application scenarios still represents an underexplored avenue of research in this domain. Further research in this direction is therefore fundamental to fully address the profiled gap.

Chapter 7

Insights from selected studies

Up until this point, the information included in this review with regards to the collected research was mostly focused on describing relevant characteristics common across multiple studies. This holistic approach was fundamentally required in order to successfully answer the posed research questions (Chapter 5) and to ultimately surface a set of gaps (Chapter 6) that could highlight noteworthy shortcomings in the current state-of-the-art. In an attempt to further integrate the information provided so far with a more in-depth look on a selected subset of relevant studies, this chapter provides a set of brief summaries tied to **five distinct approaches** for automatic topic labeling. Each of these methodologies is ultimately extracted from documents conforming to the general guidelines described in Subsection 3.8.2. Fundamentally, this means that prospective readers will be able to **freely access** (1) the documents in which these approaches were originally presented and (2) the relevant code repositories containing the practical implementations tied to such methodologies. In addition to direct links pointing to the related resources, the content of each study is briefly summarised in terms of its problem definition, proposed methodology, utilised dataset and obtained results. Additionally, an example label is shown for each of the highlighted approaches. Ultimately, with this Chapter prospective readers should be able to form a general idea on some (reproducible) techniques for automatically generating topic identifiers that could potentially be exploited in future research. Additionally, notice that the five chosen methodologies ultimately provide fundamentally distinct solutions to the task of topic labeling.

7.1. Thesaurus-based approach

Title: Principled Analysis of Energy Discourse across Domains with Thesaurus-based Automatic Topic Labeling
Authors: Thomas Scelsi, Alfonso Martinez Arranz & Lea Frermann
Venue: Australasian Language Technology Association

Open Access: Gold OA
Resource: https://aclanthology.org/2021.alta-1.11/

Repository: https://github.com/tscelsi/dtm-toolkit
License: Not specified

Problem definition

This work highlights a relatively simple method for the automatic generation of topic labels based on domain knowledge extracted from (political) thesauri. Here, the proposed methodology is tested on a two diachronic corpora generated for this study and covering more than twenty years of energy discourse.

Proposed methodology

The proposed methodology draws as the source of its labels from **existing (domain-specific) thesauri**. In this context, the EuroVoc thesaurus (see "Dataset" Subsection) is utilised, but it is stated that the same approach is conceptually capable of generalising to any thesaurus organising keyphrases in a similary way. Here, each label l belonging to the set of labels L associated with the selected thesaurus is represented as set of keyphrases v each made up of one or more tokens

Renewable Energy: bioenergy, biogas, geothermal energy, etc.

The proposed method is organised in two main steps, where (1) the set of labels L are filtered to retain only the domain-relevant ones and (2) an algorithm to pair a term distribution to one or more labels (and associated keyphrases) is executed.

The initial **label filtering** phase (which might be considered optional depending on the used thesaurus and the domain at hand) is conducted using a few different approaches. Firstly, entries that are too specific (in this case the ones pertaining to EU-specific policies) are manually removed. Then, close-duplicates are removed and labels whose embeddings have high cosine similarities are merged. Finally, log-odds ratios with informative Dirichlet (Monroe et al., 2008; Lucy et al., 2020) prior are used to score each label l with regards to the corpus of interest. Here, the top 40 terms with the highest relevance score are retained.

Following this filtering phase, the remaining task is to assign to a topic distribution the set of most candidate labels gathered from the thesaurus at hand. In this work, two distinct approaches are provided for this issue. The first approach (defined as **Importance-based topic labeling**) assumes that a candidate label should be taken into consideration for a given distribution if (1) it contains (as keyphrases) the topic's most prominent terms and (2) the interested keyphrases are discriminative for the given label (they do not tend to be overly present across the thesaurus). Here, given a topic $\hat{k}$ (which in this context is represented as a unit vector made up of the top-10 most prominent terms w), a term-topic relevance indicated as $\hat{k}[w]$ and a uniqueness score of a keyphrase represented as $TFIDF[w,l]$, the score of a label with regards to $\hat{k}$ is computed as:

$$\sigma^{imp}_{k,l} = \sum_{w \in \hat{k} \cap l} \hat{k}[w] \times TFIDF[w,l] \tag{7.1}$$

This first approach has no need for additional resources (it only requires the filtered thesaurus and the topic distributions), but by virtue of being a string matching method lacks any form of consideration tied to term similarity (e.g. synonyms).

In the second method (named **Embedding-based topic labeling**, the top-word vectors $\hat{k}$ are turned into 50-dimensional embedding vectors (emb_k) by computing a weighted average of the embeddings of the individual terms. As one might expect, the embeddings tied to the labels l (emb_l) are computed in a similar fashion by performing an (unweighted) average over the embeddings of the individual keyphrases. In this case, the label score is computed as:

$$\sigma^{emp}_{k,l} = cos(emb_k, emb_l) \tag{7.2}$$

Each topic is then associated with the top labels maximising this score.

Datasets & Results

As previously mentioned, the implementation of the proposed methodology uses as a label pool the **EuroVoc thesaurus**. This represents a (multilingual) thesaurus created to support the search of EU produced documents. It is made up of 127 labels which are exclusively considered in the English language. The authors recognise that capabilities for future work exists for exploiting the introduced methodology in a multilingual setting.

The described labeling approach is tested on corpora associated to: documents produced by the US Energy Information Administration and scientific literature tied to two energy policy journals.

The specifics of the utilised corpora are not covered in this Section (since they are not strictly relevant to the proposed labeling approach.

In order to **evaluate** the results of the topic labeling efforts, human judgments are obtained by asking assessors to select the most appropriate label for a given topic among three candidates (two generated by the proposed technique and a random one). The procedure is conducted for the top-1 and top-4 label sets. As one could expect, both approaches significantly outperform the random selection baseline. From this results, it is stated that the two approaches are found to produce meaningful labels. Additionally, upon insepcting the generated output it is noted that: *"(a) the labels are varied and cover intuitively relevant aspects of energy-related discussions in government and academia; and (b) that the label distribution differs across corpora in meaningful ways."*

Examples of the top-2 labels generated by the two approaches are shown as follows:

> **Topic:** resource, oil, production, natural gas, tight, gas, shale gas, drilling, estimate, technology
> **TFIDF approach:** oil & gas industry, production
> **Embedding approach:** oil & gas industry, renewable energy

7.2. Multilingual approach (Ontology-based method)

Title: Multilingual Topic Labelling of News Topics Using Ontological Mapping
Authors: Elaine Zosa, Lidia Pivovarova, Michele Boggia & Sardana Ivanova
Venue: ECIR 2022

Open Access: Green OA
DOI: https://doi.org/10.1007/978-3-030-99739-7_29
(Alternative) resource: https://helda.helsinki.fi/handle/10138/342489

Repository: https://github.com/ezosa/topic-labelling
License: MIT

Problem definition

The traditional representation of topics via a ranked list of most prominent terms can be generally difficult for humans to understand and correctly interpret. To more easily capture the semantic content of a given distribution, this selected study proposes an **ontological mapping method** capable of automatically mapping topics to relevant concepts (and associated labels) taken from a language-agnostic (news) ontology. The proposed method is tested on both English and Finnish datasets and is shown to work on topics expressed using **languages unseen during training**.

Proposed methodology

The idea associated with the proposed method is to map topics to ontological concepts extracted from a language-agnostic (news) ontology. Here, the (multi-language) labels associated with each concept are ultimately used as the topic identifiers. In this context, the problem is approached as a classification task where the classifier (represented by two fully connected layers with a ReLU nonlinearity and a softmax activation function) receives as input a term sequence $X = \{x_1, ..., x_n\}$ (encoded using SBERT) and computes the conditional probability $P(c_i|X)$ for each concept c_i in the ontology. Then, suitable labels are extracted form the distribution $P(c_i|X)$ by:

1. Selecting a subset of concept candidates where $P(c_i|X) > t$ (for an arbitrarily set threshold t).
2. Propagating the selected concepts to the top of the ontology.
3. Selecting the most frequently occurring concepts (after propagation) as the source for the top topic labels.

A visual summary highlighting the training and inference procedures for the proposed methodologies are presented in Figure 7.1. Further details on the training procedure are provided in the next Subsection.

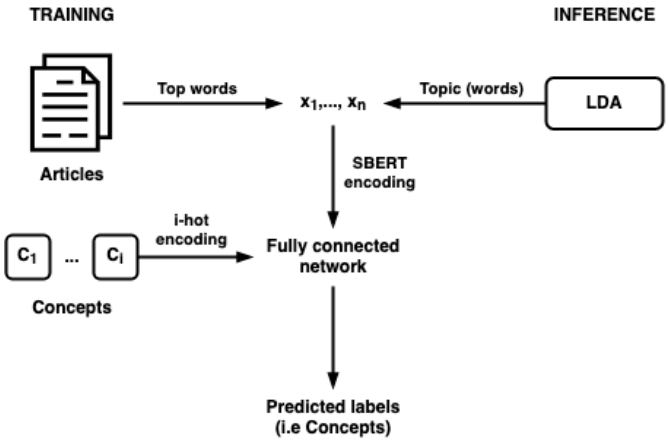


Figure 7.1: Training and inference procedures of the ontological mapping method

Datasets

As the (News) **Ontology** dataset used as the source for the topic labels, the IPCT Subject Codes ontology is selected. Each ontology concept contains a corresponding identifier in multiple languages. The ontology is organised over three levels (17 high-level, 166 mid-level and 1221 fine-grained concepts) and each mid-level concept is set to have one parent and multiple children.

As **training data** for the classifier, two dataset are constructed starting from 2017 news articles of the Finnish News Agency (STT) which are labeled with (multiple) concepts from the IPCT ontology. Here, each entry in the training datasets corresponds to the top n words extracted from the corresponding article and is tagged with the associated concepts. In this context, the top words are extracted by means of tf-idf (first dataset) or by simply selecting the top unique content words in each article (second dataset). Classifiers are built from both training datasets.

On the other hand, the **test data** is made up of 100 topics generated from 2018 Finnish news articles using LDA. Here, gold-standard labels are manually generated for each topic. On average, seven labels are ultimately obtained for each distribution. Additionally, the existing NETL dataset is used for English topics.

Results

The generated models are tested on both datasets against alternative methodologies based on existing research (Alokaili et al., 2020; Popa and Rebedea, 2021) and the generated labels are evaluated with regards to the gold-standard ones by means of BERTScore (Zhang et al., 2020). Here, it is found that: *"All models outperform the baseline [top five terms of each topic] by a large margin which shows that labels to ontology concepts are more aligned with human-preferred labels than the top topic words"*. Additionally, the proposed methodology is shown to be capable of **generating labels in a language unseen during training** and that: *although the ontology models do not outperform the baseline, they are still able to generate English labels that are very close to the gold labels considering that the models have been trained only on Finnish data.*

An example for a topic, the associated Gold-standard interpretation and the generated English label is shown as follows:

Topic: film, movie star, director, hollywood, actor, minute, direct, story, witch
Gold-standard labels: fantasy film, film adaptation, quentin tarantino, a movie, martin

scorsese, film director, film
Generated labels: film festival, cinema industry, human interest, culture (general), media

7.3. Sequence-to-Sequence method

Title: Automatic Generation of Topic Labels
Authors: Areej Alokaili, Nikolaos Aletras & Mark Stevenson
Venue: SIGIR 2020

Open Access: Green OA
DOI: https://doi.org/10.1145/3397271.3401185
(Alternative) resource: https://arxiv.org/abs/2006.00127

Repository: https://github.com/areejokaili/topic_labelling
License: Not specified

Problem definition

To aid in the interpretation of term lists, this papers proposes a **Sequence-to-Sequence model** (trained over large dataset created using a distant supervision approach) which, unlike some previously existing techniques, does not restrict the set of label candidate to a pre-defined pool (e.g. Wikipeida articles titles, ontology concepts, etc.).

Proposed methodology

The proposed Sequence-to-Sequence (S2S) architecture (having a two RNN encoder-decoder structure) is designed to take as input a list of terms (the topic) and generate another sequence of terms to be used as label. In the proposed model, the input terms $X = (x_1, ..., x_t)$ are passed through an embedding layer which encodes the initial terms into 300 dimensional vectors $(e_1, ..., e_t)$ followed by a bidirectional GRU. The forward and backward output of the GRU form the hidden state h_t and pass it to the Decoder which, at a given time step t, uses it to predict the hiddent state s_t as follows:

$$s_t = GRU(y_{t-1}, s_{t-1}, c_t) \tag{7.3}$$

Where y_{t-1} and s_{t-1} are the previous output term and hidden state respectively. Differently from traditional S2S models, where normally the last hidden state of the encoder is used to compute the context vector c_t, in this case such vector is computed as the weighted sum over all encoder hidden states by means of an attention mechanism.

The decoder hidden state s_t is ultimately used to generate probabilities across all vocabulary terms by passing it as input to a dense layer with softmax activation function:

$$P(y_t|y_1, ..., y_{t-1}, X) = Dense(s_t) \tag{7.4}$$

These probabilities are then used to output the most likely label term y_t:

$$y_t = argmax(P(y_t|y_1, ..., y_{t-1}, X)) \tag{7.5}$$

The structure of the proposed S2S at inference is shown in Figure 7.2

Dataset & Results

In order to build a suitable **training dataset**, a distant supervision approach is employed to create two distinct datasets made up of topics and corresponding labels. Here, articles are extracted from

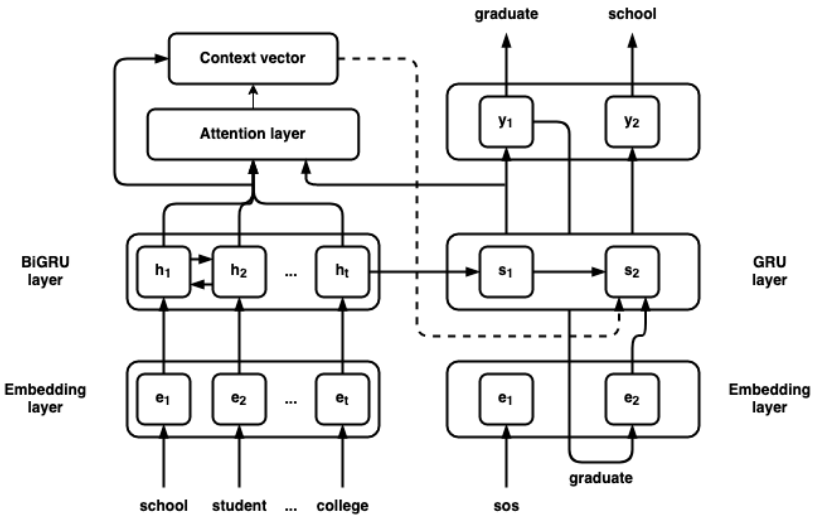


Figure 7.2: Inference using the proposed S2S model

Wikipedia and either the first or the top scoring (in terms of tfidf) 30 terms tied to each document are used to represent the **synthetics topics**. In this context, the article titles are used as the topic labels.

To **test** the generated model in order to compare it to baseline methods (top-two and top-three topic terms), existing datasets containing gold-standard labels (Bhatia et al., 2016) are used. The results obtained by scoring the generated labels using BERTScores reveals that all variations of the trained S2S model have significantly higher scores than the baseline methods and the generated identifiers are observed to be close to their gold-standard counterpart. An example of label generated by the proposed method is shown as follows:

> **Topic:** vote, house, election, poll, bill, republican, party, voter, candidate, senate
> **Gold-standard labels:** election, by-election, general election, primary election, electoral college
> **Generated label:** hall of representatives elections, united states house of representatives elections in Illinois

7.4. Transformer-based approach

Title: BART-TL: Weakly-Supervised Topic Label Generation
Authors: Cristian Popa & Traian Rebedea
Venue: EACL 2021

Open Access: Gold OA
DOI: https://doi.org/10.18653/v1/2021.eacl-main.121
Resource: https://aclanthology.org/2021.eacl-main.121/

Repository: https://github.com/CristianViorelPopa/BART-TL-topic-label-generation
License: MIT

Problem definition

This study proposes a method for automatically generating topic labels by exploiting a pre-trained **transformer model** (BART) fine-tuned using a dataset of labeled distributions generated by exploiting multiple (non-neural) weak labelers. The proposed weakly-supervised BART-TL model is found

to be able to produce novel labels. Additionally, the use of a transformer architecture for the task of topic labeling introduces several advantages tied to the characteristics of the fine-tuning step and the wide availability of existing pre-trained models.

Proposed methodology

The starting point for the proposed methodology is a pre-trained BART model (Lewis et al., 2020). For the purpose of fine-tuning the selected model for the task of topic labeling, a **"weakly supervised" dataset** of labeled topics can be built using the NETL labeler (Bhatia et al., 2016). This labeler (which exploits Wikipedia article titles as candidate identifier), is used to extract an initial set of candidate labels for each topic in the given dataset. Then, the generated candidate identifiers are **enriched** (exploiting other weak labelers) by: (1) Adding (not necessarily consecutive) **n-gram** entries sampled from the most prominent topic terms. (2) Extracting groups of **relevant sentences**, joining them together and adding the resulting paragraphs as targets (Gourru et al., 2018). (3) Including popular **noun phrases** appearing in documents relevant to the topic.

The resulting dataset generated following this steps can then be used to **fine-tune** the pre-trained BART model, creating a transformer (BART-TL) capable of making prediction on topic terms sequences. Notice also that: *"The final models are fine-tuned based on unsupervised labelers and are, thus, weakly-supervised."*

The full process is visually highlighted in Figure 7.3.

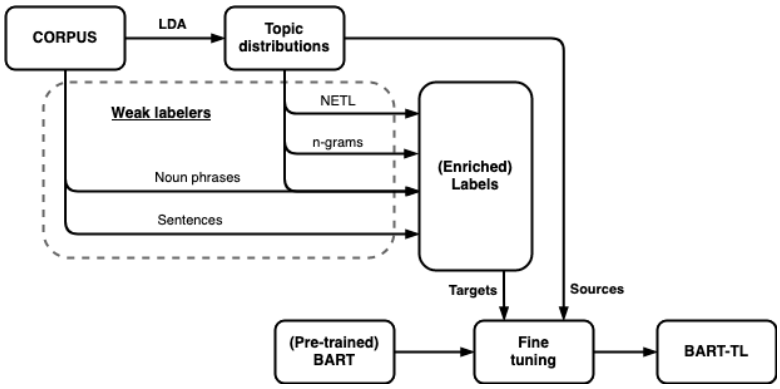


Figure 7.3: Dataset generation and fine-tuning procedure for BART-TL

Dataset & Results

The fine-tuning steps are executed on corpora extracted from StackExchange on five distinct subjects (English, Biology, Economics, Law and Photography). Here, 100 topics are generated for each corpora using LDA and 100 candidate labels are assigned to each topic using the previously described NETL labeler. Additionally, labels tied to each distribution are enriched by adding 5 n-grams, 10 noun phrases and group of sentences with a character limit of 120.

The proposed BART-TL model is executed on previously unseen topics and its results are evaluated by human assessors based on their relevance with regards to the most prominent topic terms and some of the associated documents. In this regard, the labels are found to have comparable results with the existing NETL labeler. Additionally, it is noted that the usage of a transformer model can be advantageous due to (1) its capacity to generate previously unseen identifiers, (2) the ability that authors have to condition the generation of labels with specific properties (by acting on the data used during the fine-tuning step) and (3) the wide availability of models pre-trained in different languages.

An example of a label generated by BART-TL is shown as follows:

Topic: crime, center, institution, chain, prison
Generated label: criminal justice system

7.5. Image-based approach

Title: Labeling Topics with Images Using a Neural Network
Authors: Nikolaos Aletras & Arpit Mittal
Venue: ECIR 2017

Open Access: Green OA
DOI: https://doi.org/10.1007/978-3-319-56608-5_40
(Alternative) resource: https://arxiv.org/abs/1608.00470

Repository: https://paperswithcode.com/paper/very-deep-convolutional-networks-for-large
License: Various

Problem definition

This last selected paper proposes an alternative representation of topic labels which relies on **images** to describe the generated term distributions. Here, the similarity between a given topic and the image identifier (with captions) is computed using a deep neural network. Notice that, whilst not directly sharing the code related to the specific implementation introduced in this work, it is believed that the information describing the network architecture together with the public availability of the model used to generate the required image vectors (VGG-net, the reporisotry of which is linked above) should be sufficient for prospective readers to successfully exploit the described approach.

Proposed methodology

Given a topic T and and image I, the proposed approach computes a value $s \in \mathbb{R}$ representing the suitability of the image identifiers with regards to the given topic. To prepare the input information for the described model (pertaining to one topic and a single image), the topic terms $T = \{t_1, ..., t_10\}$ and the image captions $C = c_1, ..., c_n$ are each transformed into 300-dimensional vectors and used to compute the respective mean vectors x_T and x_C. Similarly, for the given image V, the **VGG-net network** (Simonyan and Zisserman, 2014) is utilised in order to obtain a 1000-dimensional vector representations x_V that, when concatenated with the other numerical representation makes up the (1600-dimensional) input vector:

$$X = [x_T][x_C][x_V] \tag{7.6}$$

In this context, the proposed network is made up of four hidden layers $H_1, ..., H_4$ (with output sizes 256, 128, 64 and 32 respectively) and the output of each layer is computed as:

$$h_i = g(W_i^T h_{i-1}) \tag{7.7}$$

Where g is the ReLU function and $h_0 = X$ (i.e. the input vector). The network's output layer maps the last h to a real value $s \in \mathbb{R}$ representing a score for the given image (and caption) with regards to the topic. In the network, weights are trained to minimise the means absolute errors and a mini-batch gradient descent method is used.

Datasets & Results

For **training and testing** purposes, a dataset of Wikipedia and NYT news articles (Aletras and Stevenson, 2013) made up of topics associated with 20 candidate images with captions (and associated human ratings) is used. The original lack of negative examples in the chosen dataset is ad-

dressed by randomly selecting, for each topic, another set of 20 images (which are assigned a score of 0) from the other topics in the set. In the context of this work, 240 topics are used for training and 60 for testing. The evaluation procedure is based on human evaluation on the top-1 image label assigned by the model and a comparison (via nDCG) of the generated image rankings with the gold-standard ones. For all the established metrics, the proposed methodology is shown to perform better than the other tested state-of-the-art methods. Additionally, it is stated that: *"[the] model is generic and works for any unseen pair of topic and image"*.

An example of "good" and "bad" labels rankings produced by the neural network are shown in Figure 7.4.

Topic #288: surgery, body, medical, medicine, surgical, blood, organ, transplant, health, patient

(a) *3.0* (b) *2.8* (c) *2.9*

Topic #99: wedding, camera, bride, photographer, rachel, lens, sarah, couple, guest, shot

(d) *0.4* (e) *0.8* (f) *0.8*

Figure 7.4: Good and bad examples of image rankings

Chapter 8

Conclusions & Future work

8.1. Conclusions

This work represents a **comprehensive summary** of the relevant state-of-the-art research associated with Topic Labeling in the period 2017-2022 built following established guidelines for conducting Systematic Literature Reviews and further enhanced by introducing a set of **methodological contributions** with the purpose of increasing the validity of overall work. To the best of the authors knowledge, this represents the first secondary study primarily dedicated to exploring research tied to the task of naming topic distributions. Here, the included (novel) contributions (which can be seen as a by-product of the conducted Review process) are believed to be relevant even outside the analysed domain and should allow prospective authors that decide to employ them in their respective work to more meaningfully ground some of the choices made throughout the Review process with the ultimate goal of introducing a qualitative improvement to the final output of the produced secondary study. In this final Section, a concluding statement is provided in order to concisely summarise the results of the conducted activities, together with the main findings extracted from the included work and the observation such findings allowed to formulate with regards to the posed Research Questions and the Research Gaps they ultimately allowed to surface.

As a first major activity in this work, the numerical results stemming from the execution of three distinct searches conducted on five literature repositories were used to justify the selection of an appropriate query to utilise in the **Venue Selection Procedure** which, together with multiple existing (venue-related) quality metrics (SJR(2), CORE & GGS), ultimately allowed for the identification of 10 conferences and 6 journals acting as the starting set of consulted Information Sources. Following this initial task, a proper **Search Strategy** established by further refining the chosen query with the introduction of proximity operators designed to increase the topical relevance of the retrieved publications enabled the authors to conduct a thorough inspection of the identified Sources and allowed to form a selection of 424 relevant documents on which the defined inclusion and exclusion criteria were applied by means of manual inspection. This step led to the identification of a core set of 55 papers representing the starting collection of studies taking part in the proposed review. Starting from this collection, and in an attempt to broaden the original scope of the presented work, the activities of reference and citation extraction were conducted in the context of the **Backward and Forward Snowballing** phases. In this regard, this additionally gathered work underwent the same query- and content-based filtering procedure applied on the core collection of papers. Additionally, in a further attempt to maintain an acceptable level of reliability with regards to the reviewed studies, citation documents collected within the Forward Snowballing activity were also filtered based on established quality metrics tied to their respective venues. In a similar vein, an additional check was conducted to highlight potentially predatory venues among the set of Snowballing papers. Ultimately, this phase resulted in the inclusion (among the two Snowballing tasks) of an additional set of 53 studies, bringing the collection making up the proposed review to a total of 108 documents acting as the ultimate sources of information utilised for exploring existing research on Topic Labeling.

Following the completion of the activities required for building a comprehensive Study Selection, the documents were individually analysed with regards to the established set of (20) Data Items and the **Analysis and Synthesis** of the information tied to the generated Data Collection Forms was structured into a three-tier architecture where the collected data was explored with regards to: (1) The posed Research Questions; (2) Eventual shortcomings emerged from the initial analysis and represented as relevant Research Gaps; (3) Insights from individual studies describing the details of selected (reproducible) approaches.

As a way to address the established **Research Questions**, and after an initial overview of the included work which briefly summarised its year-by-year trends, the collected data was analysed starting with the Items tied to **RQ1**. In this regard, it was ultimately possible to group the inspected topic labeling activities into three distinct macro-categories, represented by Manual, Fully Automated and Partially automated approaches. Here, it was found that studies including Fully Automated techniques were substantially more likely to show a primary focus on the task of Topic Labeling and to ultimately propose novel methodologies for addressing it. Additionally, an exploration of the specific (underlying) techniques guiding the various labeling activities revealed Fully Automated approaches to be the most varied among the three (with a total of 16 distinct techniques characterising the category). Finally, the motivational context provided by the various authors allowed to infer that most studies use topic labeling as a way to either increase the interpretability of the generated topics or to provide a better characterisation of the underlying corpus. On the other hand, automated approaches are often seen as a way to reduce the cognitive load and effort associated with the traditional manual labeling process. For **RQ2**, it was found that most studies exploited LDA (and LDA-based) models to build the underlying topics. Additionally, the total number of generated distributions was observed to be increasingly higher as one moves from the studies utilising Manual Labeling to the ones exploiting Partially and Fully Automated approaches. This information might suggest the existence of a bottleneck imposed by manual labeling procedures imposing an upper bound on the number of presented topics. In order to address **RQ3**, the encountered label structures were analysed across the collected research which ultimately showed an almost absolute prevalence of n-gram labels over alternative representation taking the shape of sentences, paragraphs and images. Furthermore, the label candidate selection and quality evaluation procedure were also explored revealing the current research to be generally lackluster in this regard. Here, the few observed candidate selection procedures were mostly informal and guided by human assessors. Whilst not as rare, quality evaluations were also similarly skewed towards human-centered approaches, even though some efforts were also shown with regards to the proposal of approaches automatic this procedure. Finally, the analysis tied to **RQ4** allowed to recognise the (19) categories of corpora (with related sources and document types) on which the encountered topic modeling and labeling activities were conducted. Comparing these categories with the three previously identified topic labeling approaches showed differences with regards to the most frequently observed categories. For instance, it was noticed that most Fully Automated approaches were used on corpora constructed from data stemming from news outlet, whilst Manual Approaches also focused on data coming from social media and scientific literature. Additionally, the inspection of the conducted document pre-processing steps found them to be generally aligned with the expectations brought forward by the registered topic modeling techniques.

Ultimately, the information utilised for addressing the posed RQs were used as a starting point for the identification of three shortcomings (i.e. **Research Gaps**) characterising the included research. Firstly, the described topic labeling approaches were further characterised with regards to their application in subsequent research. Following the conducted forward search, it was discovered that the utilisation of labeling techniques stemming from existing research is a relatively rare occurrence and that often times such techniques are re-implemented solely by the same authors from which they were originally proposed. As a way to address the gap, it was suggested that prospective authors should strive to share the technical details (i.e. code repositories) associated with the proposed approaches, thereby making them more accessible to future audiences. Secondly, the general lack of encountered quality evaluation procedures led to a deeper analysis on the limitations recognisable in the observed evaluation approaches. In this context, in addition to the straightforward suggestion

given to prospective readers to include a quality assessment step in their work, it was noted that the inclusion of multiple evaluation criteria (potentially characterised by a mixture of human-based and automated metrics), the use of varied information types and the inclusion of domain experts in the evaluation procedure are all factors that can help in increasing the validity of the presented results for a given study. Additionally, it was found that existing research suggests that a multidimensional approach to quality evaluation, characterised by the consideration of more than one factor in the assessment of each label, can lead to findings that would not be discernible when utilising single-item metrics. Finally, the last gap focused on exploring the general lack of alternative label representations and on determining whether such representation might be more effective in describing the topic they are tied to. In this regard, the conducted backward search found little additional formal evidence in support of the use of sentences, paragraphs and images over the traditional n-grams identifiers and it was ultimately determined that, given the currently available research, domain knowledge should still be applied in order to determine the most appropriate type of label for a given collection of topics.

As a concluding Chapter in the presented review, the **Individual Insights** obtained from a subset of studies selected on the basis of their accessibility and the reproducibility of the associated methodologies were described. Here, the five highlighted approaches (each making use of a distinct underlying technique) should help in aiding prospective authors in the process of (fully) automating the conducted topic labeling procedure by exploiting already existing (publicly shared) techniques.

8.2. Future Work

As one might expect, the work conducted within the context of the presented Review has been somewhat influenced by the given availability of resources, expressed in terms of time and manpower, that has ultimately determined the scope of the conducted analysis. Therefore, it is reasonable to believe that under more permissive conditions, additional avenues for research that are currently missing (or only partially explored) within the given document might have been eligible for further consideration. In this Section, some potential directions of Future Work are given as general pointers for suggesting viable extensions to the current secondary study or to other SLRs that might share similarities with the presented work.

In this context, one of the most straightforward suggestions simply pertains to the idea of **broadening the initially established restrictions** associated with the covered time period and the considered repositories (defined in Subsection 3.1.1 and 3.1.3 respectively) as a way to cast a wider net when collecting the relevant primary studies ultimately making up the Literature Review. In fact, although it is believed for the conducted Snowballing activities to have allowed for the collection of most of the relevant research tied to the explored domain, it is also reasonable to assume that the presented analysis does not necessarily offer an entirely complete overview over the chosen time period. In a similar fashion, it would also be conceptually possible to **relax the qualitative requirements** imposed during the Venue Selection Process (Subsection 3.1.3) and the Forward Snowballing phase (Subsection 3.4.2) as a way of allowing a more inclusive outlook over the encountered venues. This general idea could even be extended to the point of allowing for the inclusion of primary studies released outside of traditional publication channels. In this document, the option of building a Multivocal Literature Review taking into account Grey Literature as a viable source of information is briefly touched upon in Section 3.5.

On the other hand, Appendix D offers an initial methodological setup defining how one might go about in **building a literature network** starting from a set of `.bib` files originally constructed during the development of the Review process and offers the required pointers (in terms of suggested libraries and associated code snippets) that should aid prospective readers in automating (as much as possible) the process of constructing suitable Citation Network. In this context, and as suggested in Nooy et al. (2018), the type of analyses that one can conduct on a Literature Network for a given domain of interest can offer additional valuable insights on the shape of the collected research. For example, knowing that: *"Researchers operating within a particular subject area or scientific specialty tend to cite each other and common precursors"* implies that the extraction of k-cores (representing

clusters of **cohesive subgroups**) from a given graph could offer a potentially *"penetrating view"* on the domain of interest, allowing for the identification of specialities among the analysed research that one might fail to recognise without such an in-depth analysis. Additionally, one might even decide to explore the **evolution of research** over time in an attempt to highlight *"how knowledge flows [over the years] through a scientific community"*. Here, conducting a Main Path Analysis on the network at hand could be a viable option for profiling such a flow of knowledge and to identify those publications representing *"crucial links"* in the literature. Ultimately, in the context of this survey such thorough network analyses are left as suitable avenue for Future Work.

Appendices

Appendix A

Study characteristics - Initial selection

Year	Venue	Title	Reference
2017	ECIR	Labeling Topics with Images Using a Neural Network	(Aletras and Mittal, 2017)
2017	SIGIR	TOTEM: Personal Tweets Summarization on Mobile Devices	(Chin et al., 2017)
2017	ACL	Topically Driven Neural Language Model	(Lau et al., 2017)
2017	EACL	Multimodal Topic Labelling	(Sorodoc et al., 2017)
2017	EMNLP	Adapting Topic Models using Lexical Associations with Tree Priors	(Yang et al., 2017)
2017	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)
2018	Knowledge-Based Systems	Identifying topical influencers on twitter based on user behavior and network topology	(Alp and Öğüdücü, 2018)
2018	COLING	Model-Free Context-Aware Word Composition	(An et al., 2018)
2018	ACL	Neural Models for Documents with Metadata	(Card et al., 2018)
2018	ACL	PhraseCTM: Correlated Topic Modeling on Phrases within Markov Random Fields	(Huang, 2018)
2018	Expert Systems With Applications	Document-based topic coherence measures for news media text	(Korenčić et al., 2018)
2018	Expert Systems With Applications	W2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis	(García-Pablos et al., 2018)
2018	Journal of Informetrics	Does deep learning help topic extraction? A kernel k-means clustering method with word embedding	(Zhang et al., 2018b)
2018	KDD	TaxoGen: Unsupervised Topic Taxonomy Construction by Adaptive Term Embedding and Clustering	(Zhang et al., 2018a)

Year	Venue	Title	Reference
2019	Knowledge-Based Systems	Influence Factorization for identifying authorities in Twitter	(Alp and Öğüdücü, 2019)
2019	Expert Systems With Applications	Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints	(Bastani et al., 2019)
2019	Knowledge-Based Systems	Experimental explorations on short text topic mining between LDA and NMF based Schemes	(Chen et al., 2019)
2019	Decision Support Systems	A text analytics approach for online retailing service improvement: Evidence from Twitter	(Ibrahim and Wang, 2019)
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	ACL	Spatial Aggregation Facilitates Discovery of Spatial Topics	(Maiti and Vucetic, 2019)
2019	Knowledge-Based Systems	Learning document representation via topic-enhanced LSTM model	(Zhang et al., 2019)
2020	SIGIR	Automatic Generation of Topic Labels	(Alokaili et al., 2020)
2020	Knowledge-Based Systems	A topic-sensitive trust evaluation approach for users in online communities	(Chen et al., 2020a)
2020	Journal of Informetrics	Application of machine learning techniques to assess the trends and alignment of the funded research output	(Ebadi et al., 2020)
2020	EMNLP	Neural Topic Modeling with Cycle-Consistent Adversarial Training	(Hu et al., 2020)
2020	KDD	CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring	(Huang et al., 2020)
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)
2020	KDD	Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding	(Meng et al., 2020)
2020	SIGIR	Read what you need: Controllable Aspect-based Opinion Summarization of Tourist Reviews	(Mukherjee et al., 2020)
2020	COLING	Mining Crowdsourcing Problems from Discussion Forums of Workers	(Nouri et al., 2020)
2020	Decision Support Systems	How do consumers in the sharing economy value sharing? Evidence from online reviews	(Xu, 2020)

Year	Venue	Title	Reference
2020	Journal of Informetrics	Topic-linked innovation paths in science and technology	(Xu et al., 2020)
2020	EMNLP	Condolence and Empathy in Online Communities	(Zhou and Jurgens, 2020)
2021	NAACL	Topic Model or Topic Twaddle? Re-evaluating Semantic Interpretability Measures	(Doogan and Buntine, 2021)
2021	Expert Systems With Applications	Criteria determination of analytic hierarchy process using a topic model	(Fang and Partovi, 2021)
2021	Expert Systems With Applications	Supporting digital content marketing and messaging through topic modelling and decision trees	(Gregoriades et al., 2021)
2021	NAACL	A Disentangled Adversarial Neural Topic Model for Separating Opinions from Plots in User Reviews	(Pergola et al., 2021)
2021	EACL	BART-TL: Weakly-Supervised Topic Label Generation	(Popa and Rebedea, 2021)
2021	EMNLP	Phrase-BERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration	(Wang et al., 2021)
2021	EACL	Adversarial Learning of Poisson Factorisation Model for Gauging Brand Sentiment in User Reviews	(Zhao et al., 2021)
2022	Expert Systems With Applications	Social media analysis by innovative hybrid algorithms with label propagation	(Altınel, 2022)
2022	Journal of Informetrics	Is it all bafflegab? – Linguistic and meta characteristics of research articles in prestigious economics journals	(Amon and Hornik, 2022)
2022	Expert Systems With Applications	Providing recommendations for communities of learners in MOOCs ecosystems	(Campos et al., 2022)
2022	Expert Systems With Applications	The climate change Twitter dataset	(Effrosynidis et al., 2022)
2022	Decision Support Systems	Sourcing product innovation intelligence from online reviews	(Goldberg and Abrahams, 2022)
2022	Expert Systems With Applications	Large scale analysis of open MOOC reviews to support learners' course selection	(Gomez et al., 2022)
2022	CIKM	One Rating to Rule Them All? Evidence of Multidimensionality in Human Assessment of Topic Labeling Quality	(Hosseiny Marani et al., 2022)
2022	Journal of Informetrics	Developing a topic-driven method for interdisciplinarity analysis	(Kim et al., 2022b)
2022	Journal of Informetrics	Exploring scientific trajectories of a large-scale dataset using topic-integrated path extraction	(Kim et al., 2022a)

Year	Venue	Title	Reference
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	Decision Support Systems	Data-driven decision-making in credit risk management: The information value of analyst reports	(Roeder et al., 2022)
2022	Expert Systems With Applications	Recent trends in mathematical expressions recognition: An LDA-based analysis	(Sakshi and Kukreja, 2023)
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	COLING	Improving Deep Embedded Clustering via Learning Cluster-level Representations	(Yin et al., 2022)
2022	ECIR	Multilingual Topic Labelling of News Topics Using Ontological Mapping	(Zosa et al., 2022)

Appendix B

Study characteristics - Backward Snowballing

Year	Venue	Title	Reference
2017	ICICoS	Topic Labeling Towards News Document Collection Based on Latent Dirichlet Allocation and Ontology	(Adhitama et al., 2017)
2017	Journal of the Association for Information Science and Technology	Evaluating Topic Representations for Exploring Document Collections	(Aletras et al., 2017)
2017	International Journal of Advanced Computer Science and Applications	A Knowledge-based Topic Modeling Approach for Automatic Topic Labeling	(Allahyari et al., 2017)
2017	Management Science	Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach	(Huang et al., 2017)
2017	Social Forces	Managing the Boundaries of Taste: Culture, Valuation, and Computational Social Science	(Light and Odden, 2017)
2017	Journal of Advertising	An Investigation of Brand-Related User-Generated Content on Twitter	(Liu et al., 2017)
2017	TACL	Evaluating Visual Representations for Topic Understanding and Their Effects on Manually Generated Topic Labels	(Smith et al., 2017)
2017	DSAA	Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation	(Syed and Spruit, 2017)
2017	Tourism Management	A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism	(Xiang et al., 2017)

Year	Venue	Title	Reference
2018	Natural Language Processing and Information Systems	United We Stand: Using Multiple Strategies for Topic Labeling	(Gourru et al., 2018)
2018	International Journal of Information Management	Characterizing diabetes, diet, exercise, and obesity comments on Twitter	(Karami et al., 2018)
2018	Transportation Research Part C	Using structural topic modeling to identify latent topics and trends in aviation incident reports	(Kuhn, 2018)
2018	Communication Methods and Measures	Applying LDA topic modeling in communication research: Toward a valid and reliable methodology	(Maier et al., 2018)
2019	Small-scale Forestry	Modelling Research Topic Trends in Community Forestry	(Clare and Hickey, 2019)
2019	Social Network Analysis and Mining	Topic modeling and sentiment analysis of global climate change tweets	(Dahal et al., 2019)
2019	Journal of Comparative Economics	Toward understanding 17th century English culture: A structural topic model of Francis Bacon's ideas	(Grajzl and Murrell, 2019)
2019	Journal of Information Science	Twitter speaks: A case of national disaster situational awareness	(Karami et al., 2019)
2019	Journal of Information Processing Systems	An Ontology-Based Labeling of Influential Topics Using Topic Network Analysis	(Kim and Rhee, 2019)
2019	Media and Communication	Narratives of the Refugee Crisis: A Comparative Study of Mainstream-Media and Twitter	(Nerghes and Lee, 2019)
2020	SBSI	Recommendation System for Knowledge Acquisition in MOOCs Ecosystems	(Campos et al., 2020)
2020	Lecture Notes in Computer Science	What Are MOOCs Learners' Concerns? Text Analysis of Reviews for Computer Science Courses	(Chen et al., 2020c)
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)
2021	Ecological Economics	Free associations of citizens and scientists with economic and green growth: A computational-linguistics analysis	(Savin et al., 2021)

Appendix C

Study characteristics - Forward Snowballing

Year	Venue	Title	Reference
2019	SAICSIT	A Computational Analysis of News Media Bias: A South African Case Study	(Cornelissen et al., 2019)
2019	Government Information Quarterly	Open data visualizations and analytics as tools for policy-making	(Hagen et al., 2019)
2019	IEEE Access	Automatic Labeling of Topic Models Using Graph-Based Ranking	(He et al., 2019)
2019	Tourism Management	Job Satisfaction and Employee Turnover Determinants in High Contact Services: Insights from Employees'Online Reviews	(Stamolampros et al., 2019)
2020	Cognitive Computation	A Structural Topic Modeling-Based Bibliometric Study of Sentiment Analysis Literature	(Chen and Xie, 2020)
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)
2020	International Journal of Hospitality Management	Employing structural topic modelling to explore perceived service quality attributes in Airbnb accommodation	(Ding et al., 2020)
2020	Journal of Travel & Tourism Marketing	Topic modelling for theme park online reviews: analysis of Disneyland	(Luo et al., 2020)
2020	Annals of Tourism Research	Harnessing the "Wisdom of Employees" from Online Reviews	(Stamolampros et al., 2020)

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)
2021	Transportation Research Part D: Transport and Environment	Listen to E-scooter riders: Mining rider satisfaction factors from app store reviews	(Aman et al., 2021)
2021	Scientometrics	Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing	(Ebadi et al., 2021)
2021	IJCNN	Automatic Topic Labeling model with Paired- Attention based on Pre-trained Deep Neural Network	(He et al., 2021a)
2021	Neurocomputing	Automatic topic labeling using graph-based pre-trained neural embedding	(He et al., 2021b)
2021	Scientometrics	The use of citation context to detect the evolution of research topics: a large-scale analysis	(Jebari et al., 2021)
2021	Journal of Medical Internet Research	Topics and Sentiments of Public Concerns Regarding COVID-19 Vaccines: Social Media Trend Analysis	(Monselise et al., 2021)
2021	ICIS	What matters most to patients? On the Core Determinants of Patient Experience from Free Text Feedback	(Ojo and Rizun, 2021)
2021	WIESP	Moving beyond word lists: towards abstractive topic labels for human-like topics of scientific documents	(Rosati, 2022)
2021	ALTA	Principled Analysis of Energy Discourse across Domains with Thesaurus-based Automatic Topic Labeling	(Scelsi et al., 2021)
2021	European Journal of Operational Research	The Informational Value of Employee Online Reviews	(Symitsi et al., 2021)
2021	IEEE Access	TLATR: Automatic Topic Labeling Using Automatic (Domain-Specific) Term Recognition	(Truică and Apostol, 2021)
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	Topic evolution, disruption and resilience in early COVID-19 research	(Zhang 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)

Year	Venue	Title	Reference
2022	Cognitive Computation	A Decade of Sentic Computing: Topic Modeling and Bibliometric Analysis	(Chen et al., 2022)
2022	Electronic Commerce Research and Applications	Understanding music streaming services via text mining of online customer reviews	(Chung et al., 2022)
2022	Scientometrics	Identification of topic evolution: network analytics with piecewise linear representation and word embedding	(Huang et al., 2022)
2022	Journal of Retailing and Consumer Services	Online food delivery companies' performance and consumers expectations during Covid-19: An investigation using machine learning approach	(Meena and Kumar, 2022)
2022	Journal of Big Data	Modeling the public attitude towards organic foods: a big data and text mining approach	(Singh and Glińska-Neweś, 2022)
2022	Food Quality and Preference	Exploring the nexus between food and veg*n lifestyle via text mining-based online community analytics	(Yoo et al., 2022)

Appendix D

Literature network analysis

D.1. Methodology

The graphical representations of the literature network can be 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. Furthermore, 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 D.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 generated 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.

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 Section.

D.2. Tools

In this Section, the relevant libraries and an high level description of the logic that can be employed 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, and using libraries available in the Python programming language, a series of methods to automate the graph generation process is proposed. As a starting point, the functionalities offered by the NetworkX Library can be 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 D.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:

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

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 enough 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 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 can be 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 can be 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 Matplotilib 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)
```

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.

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.

N. Aletras and M. Stevenson. Representing topics using images. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 158–167, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1016.

N. Aletras, T. Baldwin, J. H. Lau, and M. Stevenson. Evaluating topic representations for exploring document collections. *Journal of the Association for Information Science and Technology*, 68(1): 154–167, jul 2017. doi: 10.1002/asi.23574. URL https://doi.org/10.1002%2Fasi.23574.

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.

M. Allahyari, S. Pouriyeh, K. Kochut, and H. R. Arabnia. A knowledge-based topic modeling approach for automatic topic labeling. *International Journal of Advanced Computer Science and Applications*, 8(9), 2017. doi: 10.14569/IJACSA.2017.080947. URL http://dx.doi.org/10.14569/IJACSA.2017.080947.

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. Influence factorization for identifying authorities in twitter. *Knowledge-Based Systems*, 163:944–954, 2019. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2018.10.020. URL https://www.sciencedirect.com/science/article/pii/S0950705118305069.

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.

J. J. Aman, J. Smith-Colin, and W. Zhang. Listen to e-scooter riders: Mining rider satisfaction factors from app store reviews. *Transportation Research Part D: Transport and Environment*, 95:102856, 2021. ISSN 1361-9209. doi: https://doi.org/10.1016/j.trd.2021.102856. URL https://www.sciencedirect.com/science/article/pii/S1361920921001589.

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.

E. S. Apostol, C.-O. Truică, and A. Paschke. Contcommrtd: A distributed content-based misinformation-aware community detection system for real-time disaster reporting. *ArXiv*, abs/2301.12984, 2023.

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.

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.

M. H. Barawi, C. Lin, and A. Siddharthan. Automatically labelling sentiment-bearing topics with descriptive sentence labels. In *Natural Language Processing and Information Systems*, pages 299–312. Springer International Publishing, 2017. doi: 10.1007/978-3-319-59569-6_38. URL https://doi.org/10.1007%2F978-3-319-59569-6_38.

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.

I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer, 2020.

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. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 113–120, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143859. URL https://doi.org/10.1145/1143844.1143859.

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 and J. D. McAuliffe. Supervised topic models, 2010.

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.

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008. doi: 10.1088/1742-5468/2008/10/p10008. URL https://doi.org/10.1088%2F1742-5468%2F2008%2F10%2Fp10008.

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, D. Blei, and X. Zhu. A topic model for word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1024–1033, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/D07-1109.

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.

A. Budiarto, R. Rahutomo, H. N. Putra, T. W. Cenggoro, M. F. Kacamarga, and B. Pardamean. Unsupervised news topic modelling with doc2vec and spherical clustering. *Procedia Computer Science*, 179:40–46, 2021. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2020.12.007. URL https://www.sciencedirect.com/science/article/pii/S1877050920324431. 5th International Conference on Computer Science and Computational Intelligence 2020.

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.

L. Cassi, A. Lahatte, I. Rafols, P. Sautier, and É. de Turckheim. Improving fitness: Mapping research priorities against societal needs on obesity. *Journal of Informetrics*, 11(4):1095–1113, 2017. ISSN 1751-1577. doi: https://doi.org/10.1016/j.joi.2017.09.010. URL https://www.sciencedirect.com/science/article/pii/S1751157717301542.

C. Caton, G. Pergola, T. Novack, and Y. He. Evaluating public consultation in urban planning via neural language models and topic modelling. 2021.

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.

X. Chen, H. Xie, G. Cheng, and Z. Li. A decade of sentic computing: Topic modeling and bibliometric analysis. *Cognitive Computation*, 14(1):24–47, Jan 2022. ISSN 1866-9964. doi: 10.1007/s12559-021-09861-6. URL https://link.springer.com/content/pdf/10.1007/s12559-021-09861-6.pdf.

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.

J. Chuang, C. D. Manning, and J. Heer. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, page 74–77, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450312875. doi: 10.1145/2254556.2254572. URL https://doi.org/10.1145/2254556.2254572.

J. Chung, J. Lee, and J. Yoon. Understanding music streaming services via text mining of online customer reviews. *Electronic Commerce Research and Applications*, 53:101145, 2022. ISSN 1567-4223. doi: https://doi.org/10.1016/j.elerap.2022.101145. URL https://www.sciencedirect.com/science/article/pii/S1567422322000291.

K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990. URL https://aclanthology.org/J90-1003.

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. Collaboration. *Cochrane Reviewers' Handbook*. December 2003.

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.

A. De Lucia, M. Di Penta, R. Oliveto, A. Panichella, and S. Panichella. Using ir methods for labeling source code artifacts: Is it worthwhile? In *2012 20th IEEE International Conference on Program Comprehension (ICPC)*, pages 193–202, 2012. doi: 10.1109/ICPC.2012.6240488.

A. De Lucia, M. Di Penta, R. Oliveto, A. Panichella, and S. Panichella. Labeling source code with information retrieval methods: an empirical study. *Empirical Software Engineering*, 19(5):1383–1420, Oct 2014. ISSN 1573-7616. doi: 10.1007/s10664-013-9285-5. URL https://link.springer.com/content/pdf/10.1007/s10664-013-9285-5.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.

U. Dissemination. Undertaking systematic reviews of research on effectiveness: Crd's guidance for carrying out or commissioning reviews. 03 2001.

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.

G. Erkan and D. R. Radev. LexPageRank: Prestige in multi-document text summarization. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 365–371, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-3247.

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].

C. Gao and J. Ren. A topic-driven language model for learning to generate diverse sentences. *Neurocomputing*, 333:374–380, 2019. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2019.01.002. URL https://www.sciencedirect.com/science/article/pii/S0925231219300128.

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. Guo, B. Chen, R. Lu, and M. Zhou. Recurrent hierarchical topic-guided neural language models. *ArXiv*, abs/1912.10337, 2019.

P. Gupta, Y. Chaudhary, F. Buettner, and H. Schütze. texttovec: Deep contextualized neural autoregressive models of language with distributed compositional prior. *ArXiv*, abs/1810.03947, 2018.

L. Hagen, T. E. Keller, X. Yerden, and L. F. Luna-Reyes. Open data visualizations and analytics as tools for policy-making. *Government Information Quarterly*, 36(4):101387, 2019. ISSN 0740-624X. doi: https://doi.org/10.1016/j.giq.2019.06.004. URL https://www.sciencedirect.com/science/article/pii/S0740624X1830368X.

M. Hämäläinen and K. Alnajjar. The great misalignment problem in human evaluation of nlp methods. 04 2021. URL https://arxiv.org/pdf/2104.05361.pdf.

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.

D. He, Y. Ren, A. M. Khattak, X. Liu, S. Tao, and W. Gao. Automatic topic labeling model with paired-attention based on pre-trained deep neural network. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9, 2021a. doi: 10.1109/IJCNN52387.2021.9534093.

D. He, Y. Ren, A. Mateen Khattak, X. Liu, S. Tao, and W. Gao. Automatic topic labeling using graph-based pre-trained neural embedding. *Neurocomputing*, 463:596–608, 2021b. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.08.078. URL https://www.sciencedirect.com/science/article/pii/S0925231221012686.

R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1036. URL https://aclanthology.org/P17-1036.

N. Health, M. R. C. (Australia), and N. Staff. *How to Review the Evidence: Systematic Identification and Review of the Scientific Literature : Handbook Series on Preparing Clinical Practice Guidelines*. Handbook series on preparing clinical practice guidelines. National Health and Medical Research Council, 2000. ISBN 9781864960327. URL https://books.google.it/books?id=G8JmAAAACAAJ.

A. Hindle, C. Bird, T. Zimmermann, and N. Nagappan. Relating requirements to implementation via topic analysis: Do topics extracted from requirements make sense to managers and developers? In *2012 28th IEEE International Conference on Software Maintenance (ICSM)*. IEEE, sep 2012. doi: 10.1109/icsm.2012.6405278. URL https://doi.org/10.1109%2Ficsm.2012.6405278.

A. Hindle, C. Bird, T. Zimmermann, and N. Nagappan. Do topics make sense to managers and developers? *Empirical Software Engineering*, 20(2):479–515, Apr 2015. ISSN 1573-7616. doi: 10.1007/s10664-014-9312-1. URL https://link.springer.com/content/pdf/10.1007/s10664-014-9312-1.pdf.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997. doi: 10.1162/neco.1997.9.8.1735.

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 multi-dimensionality 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.

L. Huang, X. Chen, Y. Zhang, C. Wang, X. Cao, and J. Liu. Identification of topic evolution: network analytics with piecewise linear representation and word embedding. *Scientometrics*, 127(9):5353–5383, Sep 2022. ISSN 1588-2861. doi: 10.1007/s11192-022-04273-1. URL https://link.springer.com/content/pdf/10.1007/s11192-022-04273-1.pdf.

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].

R. Jacobs. *Developing a Research Problem and Purpose Statement*, pages 125–142. Jossey-Bass, United States, Mar. 2011. ISBN 978-0-470-39335-2.

K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, oct 2002. ISSN 1046-8188. doi: 10.1145/582415.582418. URL https://doi.org/10.1145/582415.582418.

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.eswa.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.

J. Lafferty and D. Blei. Correlated topic models. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005. URL https://proceedings.neurips.cc/paper_files/paper/2005/file/9e82757e9a1c12cb710ad680db11f6f1-Paper.pdf.

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.

D. Lee, J. Shen, S. Kang, S. Yoon, J. Han, and H. Yu. Taxocom: Topic taxonomy completion with hierarchical discovery of novel topic clusters. *Proceedings of the ACM Web Conference 2022*, 2022a.

D. Lee, J. Shen, S. Lee, S. Yoon, H. Yu, and J. Han. Topic taxonomy expansion via hierarchy-aware topic phrase generation. In *Conference on Empirical Methods in Natural Language Processing*, 2022b.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

J. Li, J. Shang, and J. McAuley. Uctopic: Unsupervised contrastive learning for phrase representations and topic mining. *ArXiv*, abs/2202.13469, 2022.

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.

C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

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. Llerena, F. Alva-Manchego, and N. Murrugarra-Llerena. Improving embeddings representations for comparing higher education curricula: A use case in computing. In *Conference on Empirical Methods in Natural Language Processing*, 2022.

S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

L. Lucy, D. Demszky, P. Bromley, and D. Jurafsky. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in texas u.s. history textbooks. *AERA Open*, 6 (3):2332858420940312, 2020. doi: 10.1177/2332858420940312. URL https://doi.org/10.1177/2332858420940312.

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.

X.-L. Mao, Z.-Y. Ming, Z.-J. Zha, T.-S. Chua, H. Yan, and X. Li. Automatic labeling hierarchical topics. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, oct 2012. doi: 10.1145/2396761.2398646. URL https://doi.org/10.1145%2F2396761.2398646.

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).

S. Medini, G. Antoniol, Y.-G. GuÃ©hÃ©neuc, M. Di Penta, and P. Tonella. Scan: An approach to label and relate execution trace segments. In *2012 19th Working Conference on Reverse Engineering*, pages 135–144, 2012. doi: 10.1109/WCRE.2012.23.

P. Meena and G. Kumar. Online food delivery companies' performance and consumers expectations during covid-19: An investigation using machine learning approach. *Journal of Retailing and Consumer Services*, 68:103052, 2022. ISSN 0969-6989. doi: https://doi.org/10.1016/j.jretconser.2022.103052. URL https://www.sciencedirect.com/science/article/pii/S096969892200145X.

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.

D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression, 2012.

V. Mireles, P. Bourgonje, J. Moreno-Schneider, M. Khvalchik, and G. Rehm. Learning ontology classes from text by clustering lexical substitutes derived from language models. 2022.

B. L. Monroe, M. P. Colaresi, and K. M. Quinn. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403, 2008. doi: 10.1093/pan/mpn018.

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.

D. Movshovitz-Attias and W. W. Cohen. KB-LDA: Jointly learning a knowledge base of hierarchy, relations, and facts. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1449–1459, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1140. URL https://aclanthology.org/P15-1140.

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.

C. Müller-Bloch and J. Kranz. A framework for rigorously identifying research gaps in qualitative literature reviews. *ICIS 2015 Proceedings*, 2015. URL https://aisel.aisnet.org/icis2015/proceedings/ResearchMethods/2.

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.

W. Nooy, A. Mrvar, and V. Batagelj. *Genealogies and Citations*, pages 269–312. 07 2018. ISBN 9781108474146. doi: 10.1017/9781108565691.018.

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.

A. Ojo and N. Rizun. What matters most to patients? on the core determinants of patient experience from free text feedback. *ICIS 2021 Proceedings*, 2021. URL https://aisel.aisnet.org/icis2021/is_health/is_health/19.

P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994. doi: https://doi.org/10.1002/env.3170050203. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/env.3170050203.

M. Page, J. McKenzie, P. Bossuyt, I. Boutron, T. Hoffmann, C. Mulrow, L. Shamseer, J. Tetzlaff, E. Akl, S. Brennan, R. Chou, J. Glanville, J. Grimshaw, A. Hróbjartsson, M. Lalu, T. Li, E. Loder, E. Mayo-Wilson, S. McDonald, L. McGuinness, L. Stewart, J. Thomas, A. Tricco, V. Welch, P. Whiting, and D. Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021.

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.

M. Peyrard. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1101. URL https://aclanthology.org/P19-1101.

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.

P. R. and P. K. Nag. Text-based emotion recognition using contextual phrase embedding model. *Multimedia Tools and Applications*, Mar 2023. ISSN 1573-7721. doi: 10.1007/s11042-023-14524-9. URL https://link.springer.com/content/pdf/10.1007/s11042-023-14524-9.pdf.

D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, Singapore, Aug. 2009. Association for Computational Linguistics. URL https://aclanthology.org/D09-1026.

D. Ramage, C. Manning, and S. Dumais. Partially labeled topic models for interpretable text mining. pages 457–465, 08 2011. doi: 10.1145/2020408.2020481.

M. Rezaee and F. Ferraro. A discrete variational recurrent topic model without the reparametrization trick. *ArXiv*, abs/2010.12055, 2020.

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.

M. Roberts, B. Stewart, D. Tingley, and E. Airoldi. The structural topic model and applied social science. *Neural Information Processing Society*, 2013.

M. E. Roberts, B. M. Stewart, and D. Tingley. stm: An r package for structural topic models. *Journal of Statistical Software*, 91(2):1–40, 2019. doi: 10.18637/jss.v091.i02. URL https://www.jstatsoft.org/index.php/jss/article/view/v091i02.

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.

K. A. Robinson, I. J. Saldanha, and N. A. Mckoy. Development of a framework to identify research gaps from systematic reviews. *Journal of Clinical Epidemiology*, 64(12):1325–1330, dec 2011. doi: 10.1016/j.jclinepi.2011.06.009. URL https://doi.org/10.1016%2Fj.jclinepi.2011.06.009.

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.

D. Rosati. Moving beyond word lists: towards abstractive topic labels for human-like topics of scientific documents. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 91–99, Online, Nov. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wiesp-1.11.

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.

T. Scelsi, A. M. Arranz, and L. Frermann. Principled analysis of energy discourse across domains with thesaurus-based automatic topic labeling. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, pages 107–118, Online, Dec. 2021. Australasian Language Technology Association. URL https://aclanthology.org/2021.alta-1.11.

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

C. Sievert and K. Shirley. Ldavis: A method for visualizing and interpreting topics. 06 2014. doi: 10.13140/2.1.1394.3043.

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.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. 09 2014. URL https://arxiv.org/pdf/1409.1556.pdf.

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.

A. Smith, T. Y. Lee, F. Poursabzi-Sangdeh, J. Boyd-Graber, N. Elmqvist, and L. Findlater. Evaluating visual representations for topic understanding and their effects on manually generated topic labels. *Transactions of the Association for Computational Linguistics*, 5:1–16, 2017. doi: 10.1162/tacl_a_00042. URL https://aclanthology.org/Q17-1001.

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

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].

P. Stamolampros, N. Korfiatis, K. Chalvatzis, and D. Buhalis. Job satisfaction and employee turnover determinants in high contact services: Insights from employees'online reviews. *Tourism Management*, 75:130–147, 2019. ISSN 0261-5177. doi: https://doi.org/10.1016/j.tourman.2019.04.030. URL https://www.sciencedirect.com/science/article/pii/S0261517719300925.

P. Stamolampros, N. Korfiatis, K. Chalvatzis, and D. Buhalis. Harnessing the "wisdom of employees" from online reviews. *Annals of Tourism Research*, 80:102694, jan 2020. doi: 10.1016/j.annals.2019.02.012. URL https://doi.org/10.1016%2Fj.annals.2019.02.012.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014. URL http://arxiv.org/abs/1409.3215.

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.

E. Symitsi, P. Stamolampros, G. Daskalakis, and N. Korfiatis. The informational value of employee online reviews. *European Journal of Operational Research*, 288(2):605–619, 2021. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2020.06.001. URL https://www.sciencedirect.com/science/article/pii/S0377221720305269.

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

H. Tang, M. Li, and B. Jin. A topic augmented text generation model: Joint learning of semantics and structural features. In *Conference on Empirical Methods in Natural Language Processing*, 2019b.

A. Tong, K. Flemming, E. McInnes, S. Oliver, and J. Craig. Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. *BMC Medical Research Methodology*, 12(1), nov 2012. doi: 10.1186/1471-2288-12-181. URL https://doi.org/10.1186%2F1471-2288-12-181.

A. C. Tricco, E. Lillie, W. Zarin, K. K. O'Brien, H. Colquhoun, D. Levac, D. Moher, M. D. Peters, T. Horsley, L. Weeks, S. Hempel, E. A. Akl, C. Chang, J. McGowan, L. Stewart, L. Hartling, A. Aldcroft, M. G. Wilson, C. Garritty, S. Lewin, C. M. Godfrey, M. T. Macdonald, E. V. Langlois, K. Soares-Weiser, J. Moriarty, T. Clifford, Ö. Tunçalp, and S. E. Straus. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7):467–473, oct 2018. doi: 10.7326/m18-0850. URL https://doi.org/10.7326%2Fm18-0850.

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

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

J. Velcin, A. Gourru, E. Giry-Fouquet, C. Gravier, M. Roche, and P. Poncelet. Readitopics: Make your topic models readable via labeling and browsing. In *International Joint Conference on Artificial Intelligence*, 2018.

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.

X. Wan and T. Wang. Automatic labeling of topic models using text summaries. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2297–2305, Berlin, Germany, Aug. 2016a. Association for Computational Linguistics. doi: 10.18653/v1/P16-1217. URL https://aclanthology.org/P16-1217.

X. Wan and T. Wang. Automatic labeling of topic models using text summaries. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2297–2305, Berlin, Germany, Aug. 2016b. Association for Computational Linguistics. doi: 10.18653/v1/P16-1217. URL https://aclanthology.org/P16-1217.

B. Wang, S. Liu, K. Ding, Z. Liu, and J. Xu. Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: a case study in lte technology. *Scientometrics*, 101(1):685–704, Oct 2014. ISSN 1588-2861. doi: 10.1007/s11192-014-1342-3. URL https://link.springer.com/content/pdf/10.1007/s11192-014-1342-3.pdf.

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.

W. Wang, Z. Gan, W. Wang, D. Shen, J. Huang, W. Ping, S. Satheesh, and L. Carin. Topic compositional neural language model. In *International Conference on Artificial Intelligence and Statistics*, 2017.

W. Wang, Z. Gan, H. Xu, R. Zhang, G. Wang, D. Shen, C. Chen, and L. Carin. Topic-guided variational auto-encoder for text generation. In *North American Chapter of the Association for Computational Linguistics*, 2019.

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.

G. Wong, T. Greenhalgh, G. Westhorp, J. Buckingham, and R. Pawson. RAMESES publication standards: realist syntheses. *BMC Medicine*, 11(1), jan 2013. doi: 10.1186/1741-7015-11-21. URL https://doi.org/10.1186%2F1741-7015-11-21.

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.

C. Yang, D. Teng, S. Liu, S. Basu, J. Zhang, J. Shen, C. Zhang, J. Shang, L. M. Kaplan, T. Harratty, and J. Han. Cubenet: Multi-facet hierarchical heterogeneous network construction, analysis, and mining. *ArXiv*, abs/1910.01451, 2019.

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.

R. Yoo, S.-Y. Kim, D.-H. Kim, J. Kim, Y. J. Jeon, J. H. Y. Park, K. W. Lee, and H. Yang. Exploring the nexus between food and veg*n lifestyle via text mining-based online community analytics. *Food Quality and Preference*, 104:104714, 2022. ISSN 0950-3293. doi: https://doi.org/10.1016/j.foodqual.2022.104714. URL https://www.sciencedirect.com/science/article/pii/S0950329322001896.

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.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert, 2020.

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.

Y. Zhang, X. Cai, C. V. Fry, M. Wu, and C. S. Wagner. Topic evolution, disruption and resilience in early covid-19 research. *Scientometrics*, 126(5):4225–4253, May 2021. ISSN 1588-2861. doi: 10.1007/s11192-021-03946-7. URL https://link.springer.com/content/pdf/10.1007/s11192-021-03946-7.pdf.

H. Zhao, L. Du, W. Buntine, and G. Liu. Metalda: a topic model that efficiently incorporates meta information. 09 2017. URL https://arxiv.org/pdf/1709.06365.pdf.

Q. Zhao, L. Fan, Y. Zhang, J. Li, Y. Shi, W. Rao, and X. Liu. Dualtaxovec: Web user embedding and taxonomy generation. *Knowledge-Based Systems*, 271:110565, 2023. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2023.110565. URL https://www.sciencedirect.com/science/article/pii/S0950705123003155.

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.