04_04_2023

Placeholder thesis title

A novel approach to Systematic Literature Reviews: A case study on topic labeling research in the period 2017-2022

Writing progress

Theoretical Foundations

- Primary, secondary & tertiary studies
- Systematic Literature Review
- Systematic Mapping Study
- Topic modeling
- Topic labeling

Study selection

- Note on predatory venues

Review Protocol (Methods)

- Analysis and synthesis methods (intro)

Reading progress

- 65 papers out of 124

Upcoming activities

- Start populating the analysis and synthesis (Results) with the data gathered so far
- Continue the theoretical foundations section
  - Next up: description of common techniques (LDA, LSI, STM, etc…)
- Literature networks
  - Tools & Methodology
  - Analysis on main components of selected papers network

Reminder: Data items

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

Table 3.3: Data extraction form

Analysis and synthesis methods

“Data synthesis involves collating and summarising the results of the included primary studies. Synthesis can be descriptive (non-quantitative). However, it is sometimes possible to complement a descriptive synthesis with a quantitative summary. Using statistical techniques to obtain a quantitative synthesis is referred to as meta-analysis. ” -Kitchenham (2004)

“Report the methods used for synthesis of primary study outcomes.
Describe the process used to decide which studies were eligible for each synthesis.
Describe any methods required to prepare the data for presentation or synthesis, such as handling missing summary statistics, or data conversions.
Describe any methods used to tabulate or visually display results of individual studies and synthesis.
Describe any methods used to synthesize results and provide a rationale for the choice(s) [8, Sec. 3.16]. Required for all types of review except mapping studies. Qualitative studies should, where necessary, identify constructs analyzed, explain how findings from different studies were compared, and specify how synthesized findings were validated.”
-Kitchenham (2022)

Descriptive (/ qualitative) synthesis

"Extracted information about the studies (i.e. intervention, population, context, sample sizes, outcomes, study quality) should be tabulated in a manner consistent with the review question.

Tables should be structured to highlight similarities and difference between study outcomes.

It is important to identify whether results from studies are **consistent** one with another (i.e. homogeneous) or **inconsistent** (e.g. heterogeneous).

Results may be tabulated to display the impact of potential sources of heterogeneity, e.g. study type, study quality, and sample size." -Kitchenham (2004)

"A qualitative synthesis is a narrative, textual approach to summarizing, analyzing, and assessing the body of evidence included in your review. It provides a general summary of the characteristics and findings of the included studies and Analyses the relationships between studies, exploring patterns and investigating heterogeneity" - Ohio State University

Quantitative Synthesis (/ meta-analysis)

"A quantitative synthesis, or meta-analysis, uses statistical techniques to combine and analyze the results of multiple studies. The feasibility and sensibility of including a meta-analysis as part of your systematic review will depend on the data available." - Ohio State University

"To synthesis quantitative results from different studies, study outcomes must be presented in a comparable way."

"Quantitative data should also be presented in tabular form."

-Kitchenham (2004)

Structural organisation of the analysis and synthesis task

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 analysis and synthesis tasks would be structured accordingly.

In this context, the data extracted from the items associated with each question is analysed within the context of three distinct point of views.

These different frames of reference 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:

- **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 are 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.
- **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 (for each research question and data item) what are those desirable methodologies and insights that are usually missing (i.e. rarely or never found) in the collected work. Once again (and in a similar fashion to the first point), this step requires comparing the content of multiple studies.
- **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 SLR 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.
    - [Open Science]

As previously mentioned, an analysis based on these three macro-categories is conducted within the context of each of the presented Research Questions. Specifically, the individual data items found in the data extraction form 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.

RQ0. Basic information about the collected studies (Overview)

Data items: Year, Author(s), Title, Venue

In a similar way to what is seen in Silva et al., 2021, this Section is used to offer a brief overview / recap of the analysed research in terms of:

- Year range
- Total nr of publications
  - By category (initial selection, FS, BS)
  - By venue-type (conferences, journals)
  - By individual venue

Additionally, a table is included highlighting the number of papers by venue and year.

RQ1. What are the different approaches for topic labeling, how are they used and in which context?

Data items: Topic labeling, Focus, Type of contribution, Underlying technique, Topic labeling parameters, Label generation, motivation

Synthesis of the primary studies

1. Topic labeling
   - Manual / Partially automated / Fully automated approaches (figures and trends).
3. Focus
   - Primary (Abstract / Introduction) / Secondary focus
5. Type of contribution
   - Established / Novel approach
   - In the case of established approaches, what are the most commonly used pre-existing techniques for topic labeling?
7. Underlying technique
   - Focus on non-manual approaches:
     - Which are the prevalent models / techniques used to fully automate the labeling process?
     - How is additional metadata (see `Document` data item) used to guide partially automated approaches?
9. Topic labeling parameters
   - Overview of the commonly used parameters (and associated values)
11. Label generation
    - Describe label generation steps common to multiple studies:
      - Useful to quickly summarise manual labeling approaches and approaches deriving from supervised topic modeling.
      - More advanced approaches will be described in the `Insights from individual studies` section
12. Motivation
    - What is the prevalent reasons given to justify the assignment of topic labels?

Identified gaps in the research

1. Topic labeling
2. Focus
    - Papers entirely dedicated to proposing a topic labeling approach are still the vast minority. Is this trend changing over time?
4. Type of contribution
5. Underlying technique
    - Manual labeling: Who are the annotators? Does the choice of having labels assigned by authors rather than domain experts affect the quality of the final results?
7. Topic labeling parameters
8. Label generation
9. Motivation
    - Is the decision to use labeling techniques sufficiently justified?
    - In those studies that provide no explicit motivation for labeling their topics, can such reason be inferred from the presented results?

RQ2. Which underlying topic modelling techniques are used and how do they affect the choice of a topic labeling approach?

Data items: Topic modeling, Topic modeling par., Nr. of topics

Synthesis of the primary studies

1. Topic modeling
    - Summary of encountered techniques (also showing trends over time)
    - Focus on most used approach (LDA)
        - Description / Contextualisation of LDA-based techniques
        - Motivation: Is it possible to identify a reason for such a preference (also in relation to the domain, corpus and subsequent labelling step)?
    - Associating modeling and labeling: Is there some correlation between the chosen approaches?
    - Overview of supervised topic modeling (and how it affects labeling)
2. Topic modeling parameters
    - Overview of the commonly used parameters (and associated values)
3. Nr. of topics
    - How do we related this value to:
        1. the chosen corpus,
        2. the chosen modeling technique and (more importantly)
        3. the chosen labeling approach (i.e. do manual labeling approaches favor

fewer topics)?

Identified gaps in the research

1. Topic modeling
2. Topic modeling parameters
    - Does the chosen domain affect how detailed the hyperparameter tuning approach is? (i.e. Do papers from less "technical" domains execute more succinct model tuning processes?)
    - Looking at outliers: What are the parameters considered only in a minority of papers (for a given technique)? Why? Do they have any apparent effect on the produced topics (and related labels)?
4. Nr. of topics:
    - Are the advantages of partially / fully automated labeling techniques being harnessed? i.e. Do we see an higher number of generated (and therefore labeled) topics when such techniques are proposed?

RQ3. How are candidate labels ultimately selected and how is the quality of the final label assignments evaluated?

Data items: Label, Label selection, Label quality evaluation, Assessors

Synthesis of the primary studies

1. Label
    - Overview and prevalence of structure (single / multi word labels, sentence labels, image labels, etc…)
    - … and of category (ad hoc generated, pre-defined categories, extracted from manuals, etc…)
2. Label selection
    - Overview of candidate sets shape and label selection techniques.
4. Label quality evaluation
    - Overview of evaluation techniques.
    - Comparison of (semi-)automated vs manual approaches for evaluation
5. Assessors
    - When human assessors are involved, how are they selected? How does the evaluation process usually look like?

Identified gaps in the research

1. Label
    - Is the use of a predefined set of labels (e.g. taken from a manual) generally beneficial to the few papers the implement this approach?

- Does the use of more elaborate label structures (e.g. sentences, images) help in improving topic interpretability?

3. Label selection
    - Why are candidate sets rarely provided in favor of a single label per topic approaches? (i.e. Why is the label selection step usually skipped?)
4. Label quality evaluation
    - Why is the quality of labels rarely evaluated?
    - What does the lack of proper evaluation tell us about the proposed methods?
5. Assessors
    - Linking back to the lack of quality evaluation, how does the general lack of domain experts involvement affect the evaluation procedure?

RQ4. Which are the prevalent domains on which topic labeling techniques have been applied to and how are they shaped?

Data items: Domain, Problem statement, Corpus, Document, Pre-processing

Synthesis of the primary studies

1. Domain
    - What are the macro-areas and sub-domains in the collected work?
    - Does the domain of reference affect the likelihood of utilising novel / established techniques?
3. ~~Problem statement~~
4. Corpus
    - What is the size and origin of the encountered collections?
    - Is there any correlation between such features and the chosen modeling & labeling approaches?
4. Document
    - How are individual documents generally shaped?
    - What does the additional metadata associated with each document generally looks like (when present)? How does it help in the topic modeling and labeling activities?
5. Pre-processing
    - What are the more commonly used pre-processing approaches?
    - How does the used topic modeling approach affect the extent of the pre-processing phase?

Identified gaps in the research

1. Domain
2. ~~Problem statement~~

3. Corpus
4. Document
5. Pre-processing

#Thesis/Weekly notes#