12_10_2022

Eligibility criteria

Selecting venues (Journals/Publications restrictions)

Selecting a set of candidate venues (Journals and Conferences) that will establish the basis of the research process represents the first step in defining the eligibility criteria of the covered literature.

To perform this selection, an initial **exploratory search** was carried out on a set of scholarly literature repositories (Google Scholar, DBLP and Semantic Scholar) in order to obtain a list of recurring venues related to topic labeling (and topic modeling).
In a similar fashion, recurring references to journals and conferences were observed on 4 different surveys on topic modeling (Chauhan et al. 2022, Silva et al. 2021, Churchill et al. 2021, Xia et al. 2019).
In both cases, venues belonging to papers published from 2017 onwards were considered during this initial research process.
Additionally, these recurring venues were then organised into their corresponding **domains**.

Based on the frequency and relevance of the obtained results, the venues used in this review were then selected starting from the following domains: Computational linguistics, Information retrieval, Software engineering and Machine learning.

It is also worth mentioning that the venues selected from the **software engineering** domain are a subset of the ones contained in the systematic review proposed by Silva et al. (2021). In particular, this subset represents the venues containing the papers that implemented topic labeling techniques.

During this selection, the ratings of the considered venues was also taken into account. In particular, the Computer Science and Information Systems research in the Excellence Research for Australia (**CORE**) ranking was considered for conferences and the SCImago Journal Rank (**SJR**) indicator was considered for journals.

The primary set of selected venues (6 Journals and 7 Conferences), organised into four different domains and presented with their full name, acronym and (CORE / SJR) ranking looks as follows:

- **Computational linguistics**
  - Association of Computational Linguistics (ACL) (A*)

    - European Association of Computational Linguistics (EACL) (A)
    - International Conference on Computational Linguistics (COLING) (A)
    - North American Association for Computational Linguistics (NAACL) (A)
    - Transactions of the Association for Computational Linguistics (TACL) (Q1 - 2.372)
- **Information retrieval**
    - ACM International Conference on Research and Development in Information Retrieval (SIGIR) (A*)
    - European Conference on Information Retrieval (ECIR) (A)
- **Software engineering**
    - Empirical Software Engineering (ESE) (Q1 - 1.890)
    - Information and Software Technology (IST) (Q1 - 1.446)
    - Journal of Systems and Software (JSS) (Q1 - 1.418)
    - IEEE International Working Conference on Mining Software Repositories (MSR) (A)
    - International Symposium on Empirical Software Engineering and Measurement (ESEM) (A)
- **Machine learning**
    - Neurocomputing (Q1 - 1.660)
    - Machine Learning (Springer) (Q1 - 1.640)

The following is a secondary list of venues (and related domains) which also appeared multiple times during the initial exploratory search but that were ultimately not included in the main list:

- Software Engineering
    - European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering (ESEC/FSE) (A*)
    - International Conference on Software Engineering (ICSE) (A*)
- Data Mining
    - IEEE International Conference on Data Mining (ICDM) (A*)
    - ACM International Conference on Knowledge Discovery and Data Mining (KDD) (A*)
- Information and Knowledge Management
    - ACM International Conference on Information and Knowledge Management (CIKM) (A)
- Machine learning
    - IEEE International Joint Conference on Neural Networks (IJCNN) (B)
- Semantic web
    - Semantic Web (Elsevier) (Q1)
- Data Engineering
    - International Conference on Data Engineering (ICDE) (A*)

- System Sciences
  - Hawaii International Conference on System Sciences (HICSS) (A)

Information sources

Based on the main list of selected venues, the repositories on which the literature search process is performed are:

- ACL Anthology (ACL, EACL, COLING, TACL)
- ACM Digital Library (SIGIR, ESEM)
- Springer (ECIR, ESE, Machine Learning)
- ScienceDirect (IST, JSS, Neurocomputing)
- IEEE Xplore (MSR)

Time period

In order to direct the focus of the proposed review on the latest state-of-the-art research while, at the same time, obtaining a set of resulting publications that could be reasonably processed given the available resources, it has been decided to select work published in the time span corresponding to the **five years** period prior to this review process (2017-2022).

Language limitations

This review will only consider papers written in the **English language**.
Similarly, in order to ensure that the described research (and corresponding results) are fully understandable to the reviewers, this work will only analyse literature where the executed topic modeling and labeling techniques have been applied on a corpus containing documents written in the English language.

Search strategy (Search keywords)

During the search process, the strings “topic label[s]” and “topic label[l]ing” are used to query the selected repositories/venues. Additionally, the search keywords “topic model[s]” and “topic model[l]ing” proposed by Silva et al. (2021) are also used in order to capture additional papers containing potentially relevant research.

The proposed research keywords are issued with regards to the paper’s title, abstract, body and tags.

An example of a search query (in this case issued on IEEE Xplore) looks as follows:

(“topic label” OR “topic labels” OR “topic labeling” OR “topic labelling” OR “topic model” OR “topic models” OR “topic modeling” OR “topic modelling”)

AND

("Publication Title":"Association of Computational Linguistics" OR "Publication Title": "…" OR …)

Selection process (Inclusion/Exclusion criteria)

In order to be selected for this review, a paper should 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**:

- The paper does not actively apply any topic labeling techniques
- All the described labeling approaches are taken from existing work and re-proposed as-is (on the same corpus and set of topics)
- The paper and/or the analysed corpus do not match the imposed language restrictions
- The paper is a systematic review (secondary/tertiary study)

#Thesis/Temporary notes#