Evaluation of the use of topic modelling (TM) to support systematic literature review (SLR) updates for screening truncation of non-relevant citations.

Bravo À.¹, Kodjamanova P.², Atanasov P.¹

¹Amaris Consulting, Barcelona, Spain, ²Amaris Consulting, Sofia, Bulgaria

BACKGROUND

which are then screened to identify those that address our research question, in the future. according to pre-specified criteria based on titles and abstracts (TAs) and full and resources. This is expected to only increase further, with the continuously techniques have been proposed to assist the screening process. **growing** volumes of scientific research published every year.

The process of conducting a systematic literature review (SLR) involves the need for automation in conducting evidence review is becoming retrieval of a large set of citations from multiple bibliographic databases substantial and is a critical element to the way we identify and use evidence

texts (FTs). SLRs require a considerable effort in terms of knowledge, time, cost Several Natural Language Processing (NLP) and Machine Learning (ML)

OBJECTIVE

SLRs often require updating to meet research or policy requirements. In this study, we explore if **Topic Modelling (TM)** can enable more efficient identification of relevant publications in the screening process of an updated dataset based on the same study selection criteria, aiming to exclude a safe proportion of non-relevant publications to save working time.

METHODS

topics was **25**

based on the

work child

pandemic home

TM is a ML technique used to extract hidden topics from large volumes of documents, which is often the case in SLRs. Here, we applied Latent Dirichlet Allocation (LDA), an unsupervised TM method, to analyse the previous SLR.

We apply TM to the original dataset and will analyse the results. Then, the TM model will be applied to 3 updates.

Based on the observed proportions in the Original Dataset, we will evaluate the performance of removing the topics with less relevant citations.

Topic Modelling

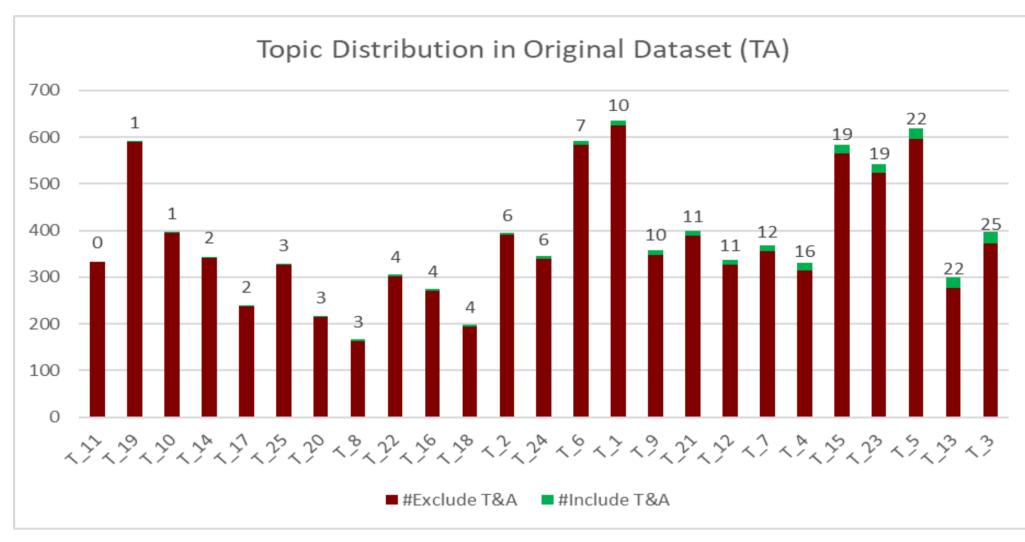
Coherence score: sars_cov_infection covid
quarantine scenario governmentpaper studybehavior covid questionnaire factor score sars_cov case covid detection policy transmission_{contact} sector economy recovery industry participant measure_{distancing} testing surveyknowledge strategy

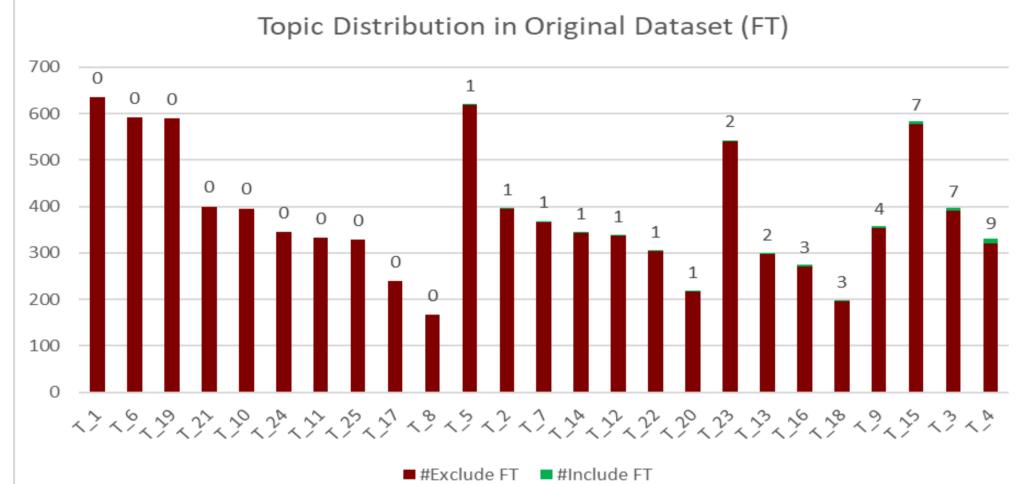
The best number of

sample infection symptom respondent crisis intervention service country mortality result research physician practice telehealth prevalence risk_factor pneumonia process conclusiondatum education experience challenge article risk population care program adultstudy frameworktechnology covid Topic 15 covid ^{period} united_states cost covid activity area household impact treatment reduction factor disease state food level effect Topic 16

number burden coronavirus oncology cancer intervention trial group pedv strain mode] virus sars_cov background stage cancer_patient cellinfection analysis conclusion patient_{result} result result Topic 23 pandemic covid

proportion of relevant and non-relevant citations (in TA and FT) for each topic

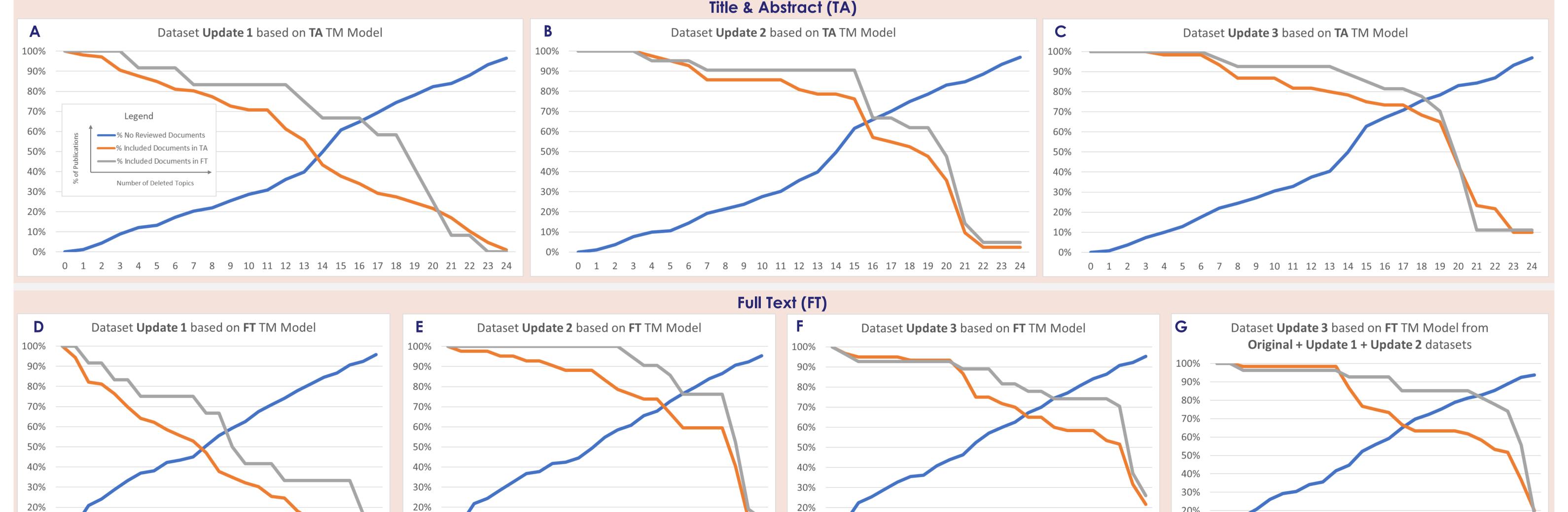




RESULTS

management healthcare_worker

The following figures represent the result to apply the TA TM model (top row) and the FT TM model (bottom row) to the three updates. The x-axis represents the number of removed topics, the y-axis represents the percentage of publications. The blue line indicates the percentage of publications (y-axis) discarded to screen depending the number of topics removed (x-axis). The orange and grey lines represent the percentage of Included citations available in the TA phase and FT phase, respectively. For example, Figure E shows that removing the 13 least relevant topics (59% of the dataset) out of 25 topics would results in no loss of relevant citations in the FT (and 21% of relevant citations in TA). On average in the three updates, by removing the 2 least relevant topics (22% of the citations), we lose 5% of relevant citations in FT (and 12% in TA). In addition, we can combine the original dataset with the following updates to improve the performance of the TM. Figure G shows the results combining datasets from the original SLR and first two updates, by removing the 9 least relevant topics (>41% of the dataset) out of 25, we lose 3.7% of relevant citations in FT (and 1.7% in TA).



10%

CONCLUSIONS

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

10%

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

10%

We show that TM and ML applied to SLRs can be an effective method that can assist the SLR process by accelerating the identification of relevant citations.

The Spanish Ministry of Science and Innovation granted a TORRES QUEVEDO Project to Amaris Consulting with reference PTQ 2019-010389, that is co-financed by the European Social Fund.

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

20%



0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24