MSR76

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]

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

BACKGROUND

The process of conducting a **systematic literature review (SLR)** involves the retrieval of a large set of citations from multiple bibliographic databases which are then screened to identify those that address our research question, according to pre-specified criteria based on **titles and abstracts (TAs) and full texts (FTs)**. SLRs require **a considerable effort** in terms of knowledge, time, cost and resources. This is expected to only increase further, with the **continuously growing** volumes of scientific research published every year.

The **need for automation** in conducting evidence review is becoming substantial and is a critical element to the way we identify and use evidence in the future.

Several **Natural Language Processing (NLP) and Machine Learning (ML)** techniques have been proposed to assist the screening process.

OBJECTIVE

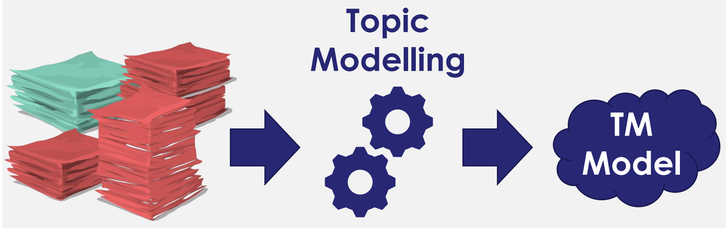
SLRs often require updating to meet research or policy requirements. In this study, we explore if **Topic Modelling (TM)** can enable more efficient identification of relevant publications in the screening process of an updated dataset based on the same study selection criteria, aiming **to exclude a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

METHODS

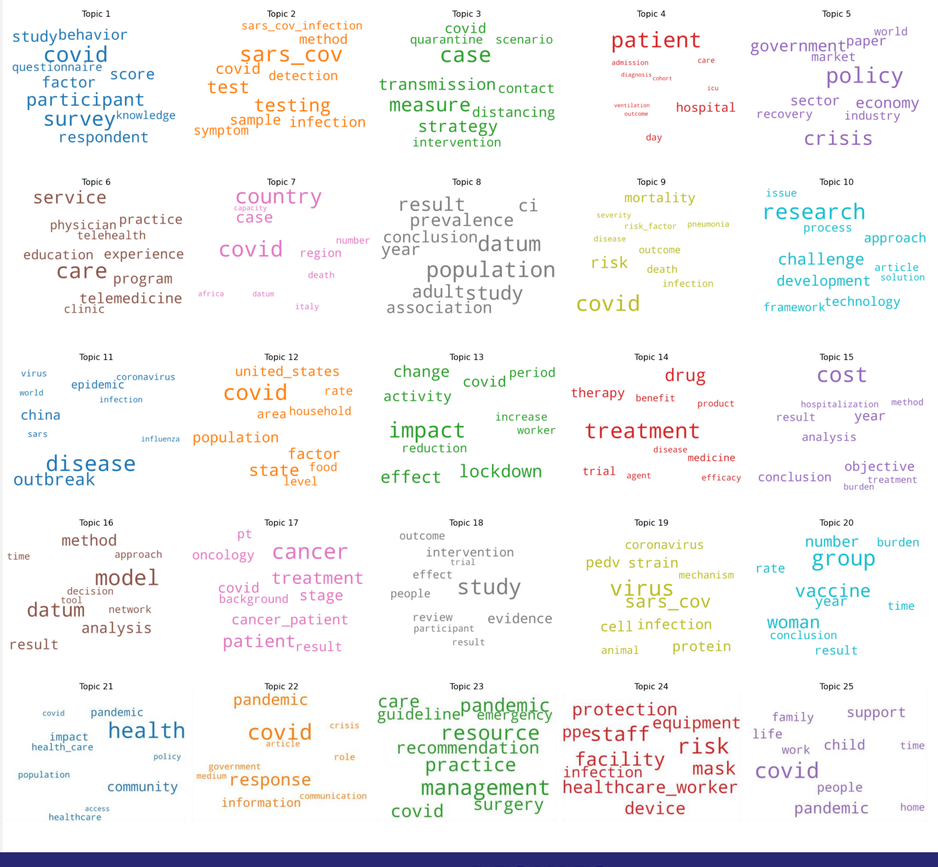
TM is a **ML** technique used **to extract hidden topics** from large volumes of documents, which is often the case in SLRs. Here, we applied **Latent Dirichlet Allocation (LDA)**, an unsupervised TM method, to analyse the previous SLR.

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

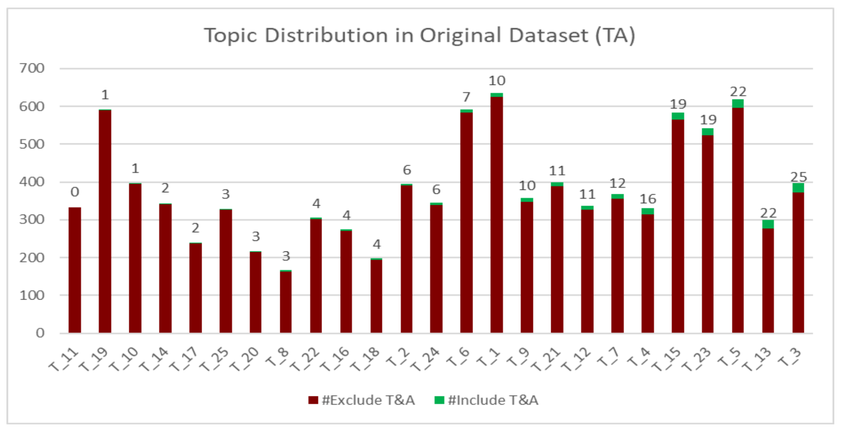
Based on the observed proportions in the Original Dataset, we will **evaluate the performance of removing the topics with less relevant citations**.

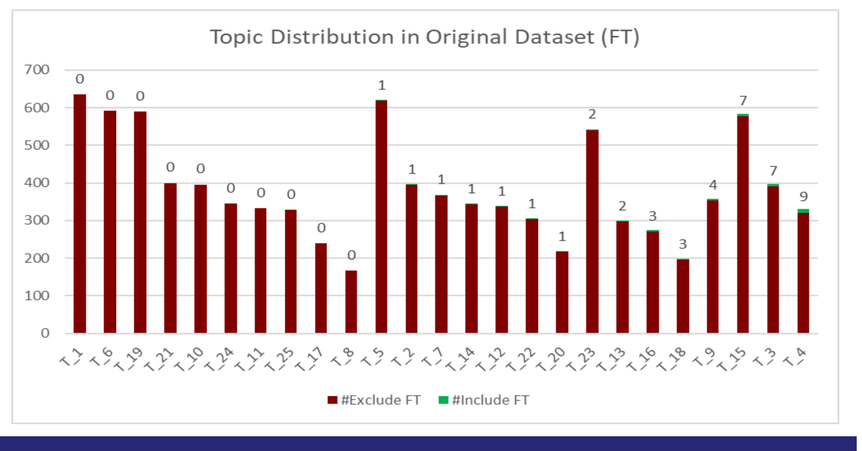


The best number of topics was **25** based on the **Coherence** score:



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

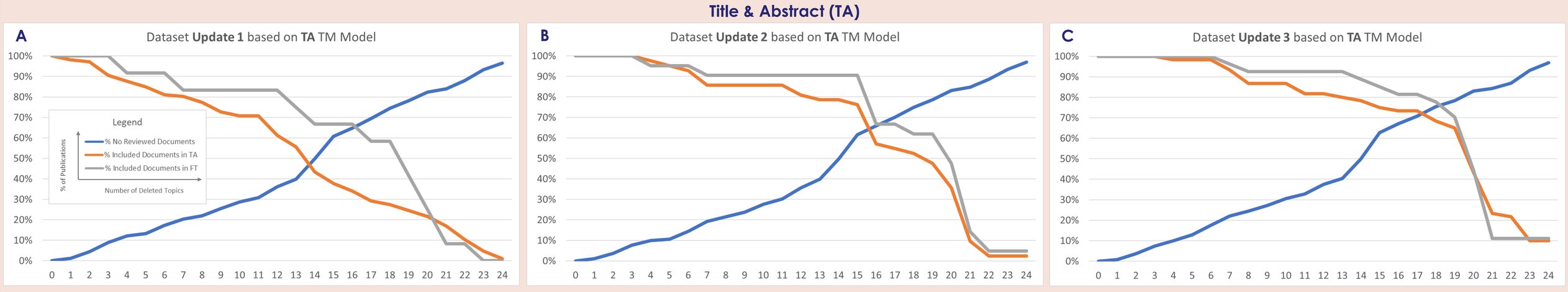




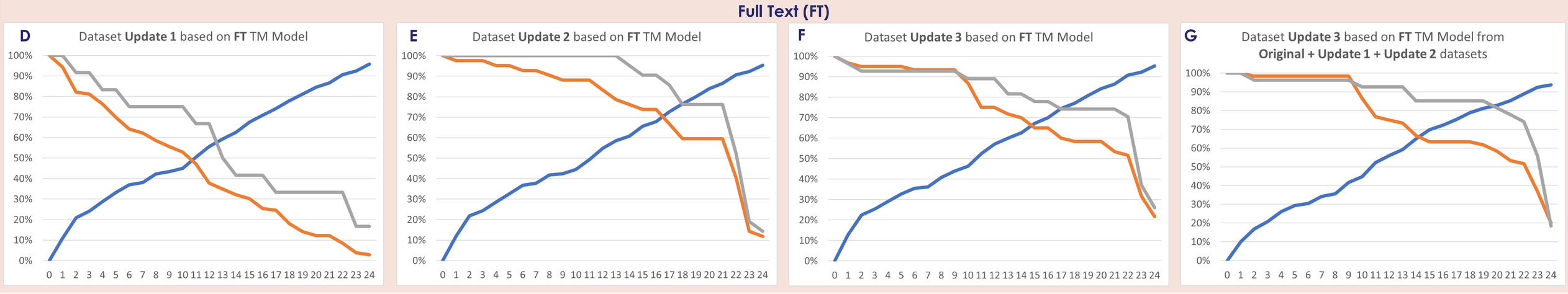
RESULTS

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).

Title & Abstract (TA)



Full Text (FT)



CONCLUSIONS

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.

AMARIS · CONSULTING ·