16_11_2022

Selection process (Inclusion/Exclusion criteria) (Further refined during the selection process)

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**:

- The paper does not actively apply any topic labeling techniques
- The labels do not posses descriptive properties with regards to the specifics of the topics content
  - Tang et al., 2019, 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 if it's covered in the subsequent time slice.
  - Figueiredo & Jorge, 2019, use a SVM classifier to assign binary labels to LDA topics in order to classify tweets as either "relevant" or "irrelevant".
- 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)

Selection process

The inclusion/exclusion criteria are applied to the set of 424 papers obtained from the proposed query and proximity operator. The final selection of **65 papers** (divided between journals and conferences) is described in the following section.

Paper selection (Journals)

Among the six analysed journals, a total of 34 papers was selected:

Decision Support Systems

- A Text Analytics Approach for Online Retailing Service Improvement: Evidence from Twitter
- Data-driven decision-making in credit risk management: The information value of analyst reports

- How do consumers in the sharing economy value sharing? Evidence from online reviews
- Sourcing product innovation intelligence from online reviews

Expert Systems with Applications

- Criteria determination of analytic hierarchy process using a topic model
- Document-based topic coherence measures for news media text
- Large scale analysis of open MOOC reviews to support learners' course selection
- Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints
- Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews
- Preliminary exploration of topic modelling representations for Electronic Health Records coding according to the International Classification of Diseases in Spanish
- Providing recommendations for communities of learners in MOOCs ecosystems
- Recent trends in mathematical expressions recognition: An LDA-based analysis
- Social media analysis by innovative hybrid algorithms with label propagation
- Supporting digital content marketing and messaging through topic modelling and decision trees
- The climate change Twitter dataset
- Topic2Labels: A framework to annotate and classify the social media data through LDA topics and deep learning models for crisis response
- W2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis
- Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis

Information Sciences

- A weakly-supervised graph-based joint sentiment topic model for multi-topic sentiment analysis
- Author topic model for co-occurring normal documents and short texts to explore individual user preferences
- Identifying impact of intrinsic factors on topic preferences in online social media: A nonparametric hierarchical Bayesian approach

Journal of Infometrics

- Application of machine learning techniques to assess the trends and alignment of the funded research output
- Developing a topic-driven method for interdisciplinarity analysis
- Does deep learning help topic extraction? A kernel k-means clustering method with word embedding

- Exploring scientific trajectories of a large-scale dataset using topic-integrated path extraction
- Improving fitness: Mapping research priorities against societal needs on obesity
- Is it all bafflegab? – Linguistic and meta characteristics of research articles in prestigious economics journals
- Topic-linked innovation paths in science and technology

Knowledge-Based Systems

- A topic-sensitive trust evaluation approach for users in online communities (Knowledge-based systems)
- Detecting and predicting the topic change of Knowledge-based Systems: A topic-based bibliometric analysis from 1991 to 2016
- Experimental explorations on short text topic mining between LDA and NMF based Schemes
- Identifying topical influencers on twitter based on user behavior and network topology
- Influence Factorization for identifying authorities in Twitter
- Learning document representation via topic-enhanced LSTM model

Paper selection (Conferences)

Among the ten analysed conferences, a total of 31 papers was selected:

CIKM

- ConCET: Entity-Aware Topic Classification for Open-Domain Conversational Agents
- One Rating to Rule Them All? Evidence of Multidimensionality in Human Assessment of Topic Labeling Quality

COLING

- Community Topic: Topic model inference by consecutive word community discovery
- Improving Deep Embedded Clustering via Learning Cluster-level Representations
- Mining Crowdsourcing Problems from Discussion Forums of Workers
- Model-Free Context-Aware Word Composition

EACL

- Adversarial Learning of Poisson Factorisation Model for Gauging Brand Sentiment in User Reviews
- BART-TL: Weakly-Supervised Topic Label Generation
- Multimodal Topic Labelling

ECIR

- Labeling Topics with Images Using a Neural Network
- Multilingual Topic Labelling of News Topics Using Ontological Mapping

ECML PKDD

- A Semi-discriminative Approach for Sub-sentence Level Topic Classification on a Small Dataset
- Survival Factorization on Diffusion Networks

KDD

- Automatic Phenotyping by a Seed-guided Topic Model
- CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring
- Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding
- TaxoGen: Unsupervised Topic Taxonomy Construction by Adaptive Term Embedding and Clustering

EMNLP

- Condolence and Empathy in Online Communities
- Adapting Topic Models using Lexical Associations with Tree Priors
- Neural Topic Modeling with Cycle-Consistent Adversarial Training
- Phrase-BERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration

ACL

- Neural Mixed Counting Models for Dispersed Topic Discovery
- Neural Models for Documents with Metadata
- PhraseCTM: Correlated Topic Modeling on Phrases within Markov Random Fields
- Spatial Aggregation Facilitates Discovery of Spatial Topics
- Topically Driven Neural Language Model

NAACL

- A Disentangled Adversarial Neural Topic Model for Separating Opinions from Plots in User Reviews
- Topic Model or Topic Twaddle? Re-evaluating Semantic Interpretability Measures

SIGIR

- Automatic Generation of Topic Labels
- Read what you need: Controllable Aspect-based Opinion Summarization of Tourist

Reviews
- TOTEM: Personal Tweets Summarization on Mobile Devices

A note on supervised topic modeling

Some of the selected research makes use of supervised topic labeling techniques.
In this context, the set of available labels is generally contained in the dataset used to build the model. The topic model learns to generate the topics together with a set of probability distributions relating each label to the set of topics.
In this context, the selection of papers making use of supervised topic modelling techniques is the result of an evaluation process performed on a **case-by-case basis** where the expressiveness of the training labels for a given corpus has been taken into account in relation to the guidelines imposed by the inclusion and exclusion criteria.

A brief overview on the impact of the 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 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 this review.
The reasoning behind this choice is 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.
Adding a proximity constraint generally allows to filter out those 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 are provided below:

- "…, **topics** can be more readily interpretable when they are assigned semantically meaningful **labels** ." - 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 & Buntine, 2021

The choice of using 20 terms as a (somewhat broader) constraint can be justified by the information found in Griffies et al., 2020 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.

Another potentially interesting insight can be provided by observing the influence of

stricter proximity constraints over the set of selected papers.

Papers discarded by imposing stricter constraints

20 → 5

- KDD, "CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring"
- EMNLP, "Adapting Topic Models using Lexical Associations with Tree Priors"

5 → 3

- Expert Systems with Applications, "Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis"
- SIGIR, "Read what you need: Controllable Aspect-based Opinion Summarization of Tourist Reviews"

In a general sense, this shows that choosing a stricter criteria with regards to the imposed proximity constraints would only tangentially affects the set of gathered publications and would ultimately lead to a similar set of results.

Conferences with pending dates for 2022

EMNLP: **December 7-11, 2022**
emnlp 2022

- Repeat the research for EMNLP'22 after December 11?

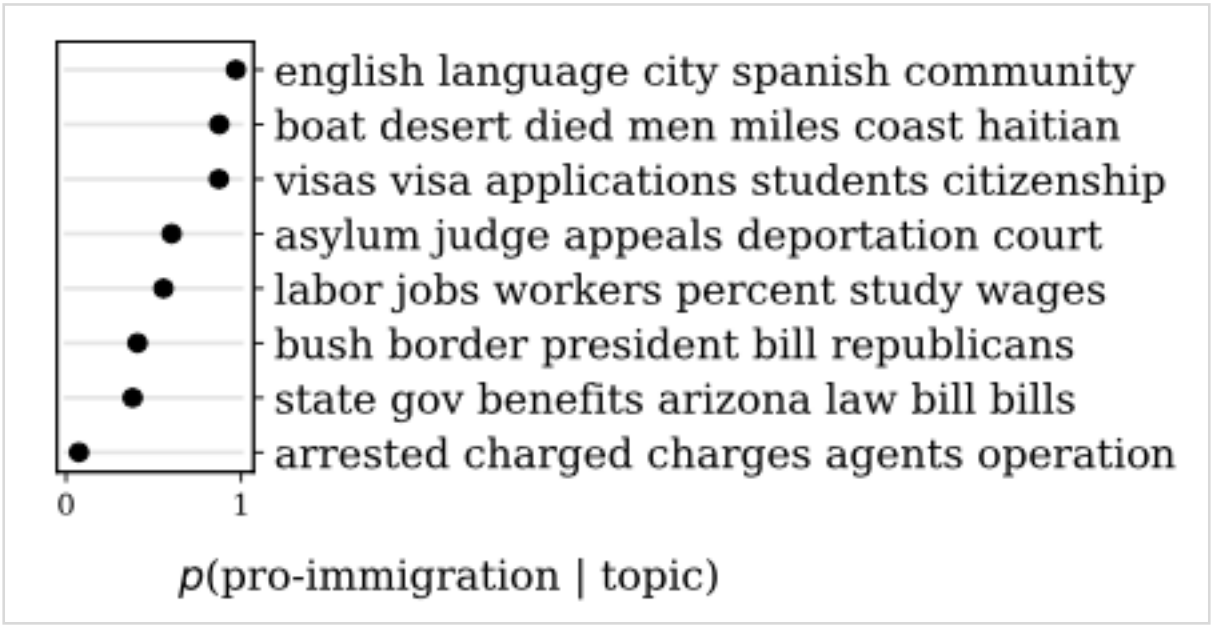
Next steps

- Extract references from the selected papers and use the results to:
  - Perform snowballing
  - Build a graph using Pajek
- Depending on the size of the final selection (after snowballing)
  - Increase/Decrease the time-frame / selected venues
- Start to rewrite the notes into the "Methods" section (i.e. transcribe the work done so far on Overleaf)

Personal notes

SLDA note (lda - sLDA. How much values response variable may have? - Stack Overflow)

sLDA has a response variable that is a label, but that really has nothing to do *directly* with the topics. The topics are still inferred exactly as they are with regular LDA, using probability calculations to build up N topics. Each document ends up with a vector of length N indicating how strongly it "contains" each topic. In sLDA it goes one step further - where it also in the model internally correlates the response label with the topics, to be able to predict what the response label should be for a never before seen document based upon its topic vector.



#Thesis/Temporary notes#