6.864 PROJECT PROPOSAL: INFORMATION EXTRACTION AND UNSUPERVISED METHODS FOR STREAMLINING EVIDENCE SYNTHESIS IN INTERNATIONAL DEVELOPMENT GREY LITERATURE

KRISTEN EDWARDS, JACK GAMMACK, LUIS KUMANDURI, DYLAN LEWIS

1. Motivation

Our goal is to use state of the art natural language processing (NLP) methods to perform information extraction and document understanding techniques for unstructured documents in the field of international development. Much of the research produced in the field of international development is unstructured text or "grey literature" - information produced outside of traditional publishing and distribution channels, such as reports, policy literature, newsletters, government documents, and so on [1]. Within this field, having an understanding of the entire current corpus of research in an area is particularly important to researchers and decision-makers [2]. A full understanding of the current state of research allows for proper funding in areas with the most need, and also informs future research initiatives with past results. However, the manual assessment required to gain this full understanding takes extensive time and effort.

The term evidence synthesis refers to the process of compiling information and knowledge from many sources and disciplines to inform decisions [2]. Evidence synthesis allows for the interpretation of an individual study within the context of the global knowledge about a topic [3]. Thus, evidence synthesis is an incredibly valuable tool for decision-makers in the fields of policy and research funding. However, the nature of grey literature can hinder evidence synthesis. Manually understanding the plethora of research in the field of international development requires manually extracting information - like relevant country, study type, action employed, and population observed - for every document. Researchers must also manually cluster and classify documents in order to sort them.

We propose using NLP to automatically extract information from international development documents. We plan to also perform unsupervised clustering and classification in order to automatically sort documents. Our goal is to use NLP to expedite currently manual document understanding processes and to uncover relationships and broad classes among documents in this field.

2. Related Work

2.1. **Information Extraction from Domain-Specific Free Text.** Machine learning (ML) has produced state of the art results for many Information Extraction (IE) subtasks, e.g. Named Entity Recogition (NER), Relation Extraction, Temporal IE, etc. and produced models like BERT that are capable of performing multiple subtasks at once [4, 5, 6, 7]. Researchers have successfully produced IE models that are specific to certain domains. For example, LexNLP was developed to perform NLP based IE on legal and regulatory texts [8]. LexNLP's key functionality is to take unstructured legal text and a) segment documents, b) identify key text such as titles and section headings, c) extract over eighteen types of structured information like distances and dates, d) extract named entities such as companies and geopolitical entities, e) transform text into features for model training, and f) build unsupervised and supervised models such as word embedding or tagging models.

In addition to the domain-specific system mentioned above, there exists automated systematic literature review system prototypes, such as ExaCT, RobotReviewer, and NaCTeM which extract relevant data such as sample sizes, population, intervention, and outcomes from free-texts [9].

NLP systems such as LexNLP and RobotReviewer utilize tokenizers that are pre-trained on domain-specific text, such as legal documents or scientific research papers, so that the models are able to produce embeddings for concepts that are not common in general literature, such as "LLC" or "hydrophilic". Word embeddings and topic models are created using state of the art methods that rely on word co-occurrence,

such as TF-IDF, or more sophisticated machine learning methods such as transformers and BERT. These word- and document-level embeddings are used to train standard ML models on supervised classification tasks such as document labelling.

2.2. **Uncovering Semantic Similarity between Documents for Literature Review.**

2.2.1. *Topic Modeling for Exploratory Literature Review.* Latent Dirichlet Allocation (LDA) uncovers a fixed number of abstract topics, K, represented in the corpus. Each topic has a distribution over the words in the corpus vocabulary, in which the words with the highest probability collectively convey a theme which can be used to describe the topic. LDA provides a distribution over topics for each document in the corpus, representing a document as a mixture of topics [10]. After determining the optimal number of topics in terms of perplexity, researchers in [11] applied LDA on a corpus of 650 papers at K = 20 topics placing each document into the topic which had the highest probability for that document. The procedure was evaluated qualitatively by determining if the papers grouped together in a topic on the basis of paper titles and the 10 most probable words for that topic made semantic sense grouped together.

2.2.2. *Clustering on Scientific Article Metadata.* To assist in the search phase of systematic literature review (SLR), the authors in [12] iteratively applied K-means clustering algorithm using article metadata including the title, keywords, and the abstract to form distinct topical clusters. A TF-IDF normalized term-document matrix is constructed using the words in the text corpus and documents represented by the words counts of their respective article metadata. Further, Latent Semantic Analysis is applied to get a lower-dimensional representation of the words in the corpus from the term-document matrix. After applying K-means, the clusters are then semantically defined by their most relevant words determined by cluster's centroid. This process is done iteratively so as to refine the search for specific and relevant corpora as part of SLR using the relevant keywords for a cluster of interest. Then the average TF-IDF score among the top 5 words (by TF-IDF score) from the cluster's centroid is used as a proxy for the cluster's informative value [12].

Our work bridges some of the NLP techniques discussed in the related work with other NLP techniques we discuss below by applying them in the domain of international development. Namely, we apply NER, Pretrained Word2Vec Embeddings, and LDA to assist in the process of gaining information from a corpus of international development grey literature.

3. Implementation

3.1. **Dataset Description.** We have access to a dataset of titles, abstracts, descriptions of interventions and outcomes, extracted information, and labeled classes from approximately 250 documents in the international development field. The extracted information includes countries mentioned. The labeled classes come from experts who identified the broad intervention type and outcome type of each document.

3.2. **Data Preparation Pipeline.** To assist in preparation of the unstructured text data, we use NumPy, pandas, Natural Language Toolkit (NLTK), & spaCy packages in Python. Our data preparation pipeline will consist of extracting text fields and labels from csv files using pandas and separating documents into sentences with NLTK if necessary. SpaCy's NER library may be used to perform NER.

3.2.1. *Word & Document Embeddings.* Text fields will be tokenized using pretrained tokenizers in PyTorch or Huggingface, or tokenizers will be pretrained by ourselves if it is determined that there are many out-of-vocabulary words in the corpus. Word embeddings can be created from the tokenized documents using statistical word co-occurence methods, like TF-IDF, or through transformers by either applying a pre-trained BERT model or fine-tuning BERT on the corpus.

Document embeddings may be created through statistical methods like LDA, which also performs topic modeling, or through combinations of the word embeddings present in the document that are trained on semantic similarity tasks.

3.3. **Models.** We will make use of existing ML packages including scikit-learn and Gensim (unsupervised method such as LDA and K-means and evaluation metrics), and PyTorch for utilizing deep learning architectures and extracting word and document embeddings. Scikit-learn's clustering methods may be used for unsupervised grouping of documents for LSA. PyTorch may further be used for the supervised downstream classification tasks of labelling documents.

3.4. **Plotting.** To project our embeddings from a higher dimensional space to 2 dimensions for plotting, we can utilize methods such as SVD or UMAP. We will use seaborn & Matplotlib for plotting results from clustering.

3.5. **Runtime & Computing Resources.** Based on previous experience of building and training NLP models for the class and research projects, we have determined that the GPU runtime provided by Google Colab Pro, which members of our group have accesss to, should be sufficient for required training, inference, and evaluation of the unsupervised and supervised models we create.

4. Evaluation

Our project will have two components, information extraction and uncovering semantic similarity using unsupervised methods, each of which will require different metrics for evaluation.

4.1. **Information Extraction Validation.** We have a dataset of documents with ground truth labels that we will use to train and evaluate our model. We will shuffle the data set and fine tune large pretrained language models on one section of the data. To test our model, we will then evaluate on a test subset of the data that the model was not trained on, and evaluate the extracted information both quantitatively and qualitatively. Quantitatively, we can score the accuracy of our model by comparing the extracted information and label classes to our ground truth. Beyond the quantitative scoring, it is important that the extracted labels are intelligible and conveying the correct information, and we will analyze the output data to see whether or not this is the case. Evaluation here will be inherently qualitative, and we will be trying to answer questions like what kind of information our model is extracting, as well as try to interpret what our model is doing.

4.2. **Analysis of Unsupervised Methods.** Our labeled data naturally clusters in many different ways, for instance by intervention framework or outcomes. One metric we will apply to evaluate the performance of our unsupervised methods is comparing the clusters and similarity that our model learns to our existing knowledge of clustering in the data. We will also evaluate our unsupervised methods qualitatively, testing whether learned topics and clusters make sense, and interpreting what relationships between documents are uncovered by our unsupervised methods.

References

[1] S. F. U. Libraries. (2021) Grey literature: What it is & how to find it. [Online]. Available: https://www.lib.sfu.ca/help/research-assistance/format-type/grey-literature

[2] C. A. Donnelly, I. Boyd, P. Campbell, C. Craig, P. Vallance, M. Walport, C. J. M. Whitty, E. Woods, and C. Wormald, "Four principles to make evidence synthesis more useful for policy," *Nature*, vol. 558, pp. 361–364, 2018.

[3] E. S. International, "What is evidence synthesis?" Available at https://evidencesynthesis.org/what-is-evidence-synthesis/ (2021/08/04).

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2020.

[7] D. Wimalasuriya and D. Dou, "Ontology-based information extraction: An introduction and a survey of current approaches," *J. Information Science*, vol. 36, pp. 306–323, 05 2010.
[8] M. J. B. I. au2, D. M. Katz, and E. M. Detterman, "Lexnlp: Natural language processing and information extraction for legal and regulatory texts," 2018.
[9] I. J. Marshall and B. C. Wallace, "Toward systematic review automation: a practical guide to using machine learning tools in research synthesis," *Systematic Reviews*, vol. 8, 2019.
[10] "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003. [Online]. Available: http://portal.acm.org/citation.cfm?id=944937
[11] C. Boye Asmussen and C. Møller, "Smart literature review: a practical topic modelling approach to exploratory literature review," *Journal of Big Data*, vol. 6, 10 2019.
[12] T. Weißer, T. Saßmannshausen, D. Ohrndorf, P. Burggräf, and J. Wagner, "A clustering approach for topic filtering within systematic literature reviews," *MethodsX*, vol. 7, p. 100831, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2215016120300510