CMSC773 Project Report: Evaluating Metadata-Driven Approaches to Topic Modelling

Abhishek Kumar, Andrew Hian-Cheong, Anjali Mittu, Calvin Bao

May 17, 2020

1 Introduction

This group is made up of Abhishek Kumar, Andrew Hian-Cheong, Anjali Mittu and Calvin Bao. Code for the project can be found at https://github.com/anjmittu/covid-risk-factors. Project roles:

1. Planning/exploration of data: Everyone
2. Scholar
   (a) Epoch: Andrew
   (b) Institution: Abhishek
3. MetaLDA
   (a) Epoch: Calvin
   (b) Institution: Anjali
4. Evaluation: Everyone
5. Write up: Everyone

The COVID-19 pandemic has caused an increase in scholarly articles published about the coronavirus. It has been difficult for the medical community to keep up with the rate that articles are being published. This has created a demand for data mining tools and text analysis techniques to sort through the published findings. One way people have been able to help the medical community is by grouping together articles using topic modeling, or producing salient topics and words with which a researcher could focus on. Identifying the various topics within the dataset can help researchers group documents together and more quickly identify relevant articles.

One theme throughout the semester has been how to unify both data and 'knowledge' in NLP tasks. Several recent papers in the past few years have proposed different ways of incorporating known metadata attributes into topic models that previously only accounted for the word tokens in the documents themselves. In this project, we explored two such models, SCHOLAR (Card et al., 2018) and MetaLDA (Zhao et al., 2017). Incorporating metadata in a topic model is a natural way to include additional information since humans also use metadata when processing text. We chose two metadata attributes about the documents: time period of publication and author's institution. Using these, we explored the effects of both metadata attributes across both models and compared this with the topics generated from similar models without additional metadata. Our initial assumption here is that text with similar metadata would have similar topics, and thus, using the metadata within the model would lead to more accurate models. We compared the model's outputs by calculating the average pairwise cosine distance of the top ten terms in each topic. We tried this using 'off-the-shelf' word embeddings, as well as embeddings trained on the cord-19 dataset. This is a naive variant of what is proposed in the paper "Evaluating Topic Coherence Using Distributional

Semantics” (Aletras and Stevenson, 2013). In addition, we calculated the intrinsic UMass measure for the top ten terms as discussed in the paper “Optimizing Semantic Coherence in Topic Models” (Mimno et al., 2011). Our overall goal was to identify topics relevant to risk factors associated with coronavirus diseases (one of the ’tasks’ from the original Kaggle challenge) and to identify a model which was better at this task. We manually surveyed the output topics to assess if we had been successful.

2 Data and Methods

Our data comes from the COVID-19 Open Research Dataset (CORD-19) (Wang et al., 2020). This is a collection of over 52,000 scholarly articles related to coronavirus from the PubMed’s PMC open access corpus, a corpus of research articles maintained by WHO and bioRxiv and medRxiv pre-prints. Along with the articles, there is a metadata file containing information about every article in the dataset. We focused on the year of publication of the article as well as the top author’s institution as the metadata attributes. As both metadata attributes could potentially affect the different topic priors it is worth investigating what impacts they have on the generative process of topic modeling.

2.1 Data Processing

In the interest of time, we performed quick and dirty filtering of the dataset. As we wanted to explore the impact of different metadata fields and different topic models, we needed to ensure that any document subset had the metadata available. Unfortunately, some of the documents parsed from the XML on PubMed did not have the author’s institutions and some were missing the date of publication. While we could conduct our baseline models against the full subset, it would then introduce inconsistencies so we focused only on the documents parsed out of the PDF documents. In addition, while the vast majority of the documents were published in English, there were a few outliers that had sections published in different languages, including French, German, and Chinese. Some early baseline models highlighted and caught entire topics in these languages. Given time constraints we used the world’s dirtiest heuristic to try and filter these out: filtering out all documents without the word ’the’. We also tried a more time-consuming tactic, using python’s langdetect library (Nakatani, 2010) to detect English vs. not. This yielded similar results as filtering on the word “the”, so we decided to use the quicker and similar heuristic. Once we did this, none of our top models were identifying foreign terms which was a poor, yet sufficient, validation for this early exploration. As a result of this filtering we ended up running our models on the full text of 45225/52000 documents.

In order to use the data with Mallet and MetaLDA, we had to prepare the data using Mallet’s import tool. This tool converts our set of documents into Mallet’s internal data representation. Mallet represents data as lists of “instances”. Each instance contains a name, any labels, and the data itself. When importing the data, Mallet has options to convert word features to lowercase, tokenize the data and remove stop words. We leveraged all of these options. As a requirement of MetaLDA, we also used an option to preserve the document as a sequence of word features, instead of as a vector of word counts.

2.2 Covariates

2.2.1 Disease Epoch

The coronavirus literature dataset contains literature not just from the current global pandemic, which is from the strain SARS-CoV-2, but from other very similar viruses too, that have caused similar afflictions albeit less widely spread. Two other coronaviruses did however have well-documented breakouts especially in parts of South-East Asia over the 2000s and some of the research into various risk factors or traits about those breakout could potentially be useful in research of the 2020 COVID-19 pandemic. We wanted to explore some measure of temporal metadata associated with the literature dataset, yet using year of publication alone did not seem sufficient as the interactions of years and topics would have been needlessly granular. Instead, we chose to divide the document set into 4 ‘epochs’ related to coronavirus events. The first epoch is all literature prior to the outbreak of the SARS virus in November of 2002. The second epoch is all literature between the SARS outbreak and the MERS outbreak which took place in June 2012. The third epoch is everything between MERS and the outbreak of COVID-19 in December 2019 and the 4th epoch is everything since COVID-19 has been a recognized disease. We include an additional empty epoch for documents that did not appear to have a date associated with them. Each document was labeled with one of these epochs and we explored the impact that this metadata had on the topic models for both MetaLDA and Scholar.

2.2.2 Top Author's Institution

In addition to creating a feature representing the time period in which a document was published, we also wanted to capture some feature representing who was writing the publication or some other proxy measure for different funding levels, research standards and reputation. We landed on looking at the institutions of the publication's authors since the set of authors itself would have been too sparse to make a good feature. Some of the models we explored can only handle a single value for a specific metadata feature instead of handling a list or continuous variable. As a consequence we took the 'plurality' institution of all the paper's authors (if a paper had 3 authors from institution X and 2 from institution Y, we'd label the document with X.). Our motivation for why this feature could impact different topics is that different institutions have different standards for quality research as well as different levels of access to funding. These factors could determine what kinds of results or papers are being produced and including this feature could help us identify some of these interactions. We also believe there could be a connection between institutions and topics because researchers from the same institution, or institutions that often collaborate, would tend to have similar research topics. As mentioned above, we did not include the XML documents missing author's information. Even with this exclusion, we had 15793 papers without an author's institution; these papers were labeled with 'Unknown'.

2.3 Models

Scholar (Card et al., 2018) is a Neural Topic model that leverages Variational Autoencoders (VAE) and a logistic normal prior instead of a Dirichlet prior. It seeks to identify the latent representation of a document across a topic distribution. We used the python implementation of the model from the model's author [1] and converted the CORD-19 data into the appropriate format for the CLI tools it provides. One limitation of the model is that it is slow to train with a large vocabulary using even default parameters. Thus, the authors recommend using a fixed vocabulary size and select terms based on frequency metrics. Due to computation and time constraints, we fixed our vocabulary at the top 2000 words after stop words were removed. This is admittedly a pretty small vocabulary however larger vocabulary sizes were unable to complete in a reasonable timeframe for our experiments.

MetaLDA (Zhao et al., 2017) is an implementation of LDA guided by metadata. It is built to be compatible with standard inputs from Mallet, a popular tool used in NLP and uses its LabeledLDA model. Because of this, we used Mallet without any labels as our baseline for this model. The metadata labels of a document are incorporated into the prior of per-document topic distributions, and this helps drive the idea that topic distributions should be generated with similar Dirichlet priors. Unlike regular LDA, MetaLDA uses different Dirichlet priors for the per-document topic distribution and the per-topic word distribution. In contrast to Scholar, no such vocabulary restriction was recommended or posed issues, so we did not restrict our vocabulary for this model. We keep this in mind as a caveat when we attempt direct comparisons between the models. MetaLDA does not even require all words in the training and testing documents to have embeddings, so prior to computing the average cosine similarity for topics generated by this model, we ensure that a word has an embedding representation.

For both Scholar and MetaLDA, we trained the models with a topics size of 30. Since our primary interests were to compare the models, and we had limited time, we only considered one topic size for our evaluation. We picked a value of 30 by testing different values in our mallet baseline model, as well as looking at what others on Kaggle had used with this dataset.

[1] https://github.com/dallascard/scholar

While our overall goal was to try and identify risk factors associated with corona-viruses, we choose not to use some of the 'seeded' LDA or other kinds of topic model that allows for some preset weights or distributions based on external knowledge. We chose not to do this because while we could impart a prior on a term like 'risks', there are plenty of papers that unexpectedly discover correlations between the disease and underlying effects even if they don't structure their terminology as 'risk factors'. In addition, the motivation is to discover the terms that would be in a 'risk factor' topic and thus we cannot properly 'seed' that topic either.

3 Evaluation

While the gold standard for topic evaluation is human assessment, there has been a lot of recent work on various metrics for automatic topic evaluation several of which have been shown to correlate well with human evaluations. We take a simplified approach inspired by the work of Aletras and Stevenson (Aletras and Stevenson, 2013). We collect a set of 300-dimensional word embeddings (GloVe embeddings) trained on a large sample of Wikipedia articles and the Gigaword corpus. Using this, we look at the pairwise cosine similarity of the top 10 words of each of our topics and find the average similarity. Obviously this requires all words in our dataset to also be in the GloVe vocabulary, however, given that the embeddings leveraged Wikipedia, we found few out-of-vocabulary terms. One notable and potentially problematic exception were terms heavily specific to the current Covid-19 outbreak. As a consequence we also did a quick training of GloVe embeddings on top of the CORD-19 dataset itself using only the top 50,000 terms (for computational limitation reasons) and conducted the same cosine similarity metric, using 300 dimensions. This gave us both a measure of 'external' quality (using Wikipedia trained vectors) and a measure of the 'internal' topic quality (trained on covid dataset). These results are shown in Table 1 and 2.

	Scholar	MetaLDA/Mallet
No covariate baseline	0.684	0.553
Disease epoch as covariate	0.686	0.557
Institution as covariate	0.685	0.552
Institution, Disease as covariates	0.664	0.561

Table 1: External avg. Cosine Similarity, k=30, top 10 words (Wikipedia Glove embeddings)

	Scholar	MetaLDA/Mallet
No covariate baseline	0.633	0.545
Disease epoch as covariate	0.629	0.530
Institution as covariate	0.631	0.546
Institution, Disease as covariates	0.598	0.527

Table 2: Internal Avg. Cosine Similarity, k=30, top 10 words (Glove on CORD-19 embeddings)

In addition to cosine similarity, we considered the UMass measure (Mimno et al., 2011). These results are shown in Table 3. This measure is a form of topic coherence, which is a singular score that measures the semantic similarity between topic words. Most forms of topic coherence are based on a sliding window and involve calculating the pointwise mutual information and/or cosine similarity between words in the topic and corpus. We choose to use the UMass measure because it is not based on a sliding window, and thus is quicker to compute with a corpus of our size. The UMass metric is also better at measuring what the model was able to learn with the given data. This is because other measures often use an external dataset such as Wikipedia to determine the PMI value. On the other hand, the UMass metric computes the probabilities against the original corpus, making the score more intrinsic. The work in Stevens et al. (2012) (Stevens et al., 2012) found sliding window-based topic coherence to agree with the UMass measure. The UMass measure is computed as

$$score(v_i, v_j) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m, v_l) + 1}{D(v_l)} \tag{1}$$

where D(x, y) is the count of documents containing words x and y and D(x) is the count of documents containing x. The words are iterated in order of probable words in the topic. This means that v_l will always be more probable than v_m. This measure is similar to PMI, but instead of calculating the difference in joint probability between words, it calculates the conditional probability of a word given a higher-ranked word in a topic.

	Scholar	MetaLDA/Mallet
No covariate baseline	-1.381	-11.503
Disease epoch as covariate	-1.324	-11.057
Institution as covariate	-1.348	-11.146

Table 3: Intrinsic UMass measure, k=30, top 10 words

Given that we are comparing various outputs of different kinds of topic models, as well as the impact of different metadata attributes on the outputs, we wanted to get a sense of how similar the actual topic outputs are from the different models. Our coherence metrics can proxy how good the quality of some of the topics are, but we wanted to know if the actual topics are the same or highly similar. We use a very basic approach and define a ‘topic’ being the same if it shares at least 7 of its top 10 words. We limited ourselves to the top 10 words since that is what is commonly used to evaluate topics and it aligns with our coherence scores. Obviously this metric does not take into account the rankings of words within each topic but errors on the conservative side and requires more than two-thirds of the exact terms to overlap to be called ’similar topics’. We compute a pairwise comparison of all our topic variants and how many topics between them meet our 7/10 similarity threshold. These results are shown in Table 4.

4 Discussion

One major challenge we discovered in working with this dataset is the difficulty in working with data that is highly domain-specific, encompassing many scientific terms. Many of the terms alone were specific to the medical field or even specialized medical subfields. This can be seen by looking at the results of the most probable topic generated by a Scholar model shown in Table 5. It is difficult to determine whether these terms make up a useful topic unless you have knowledge of what these terms mean.

4.1 Scholar vs MetaLDA

According to our cosine metric the scholar model produced ’better’ topics in terms of their embedding coherence. This occurred both when using the Wikipedia based GloVe embeddings as well as the internal embeddings trained on the covid dataset itself. While there are not clear statistical tests for determining significance, the differences (0.55 for metaLDA vs 0.68 for Scholar) are non trivial especially given that this was the average across pairwise comparisons. While Scholar yielded higher cosine scores on average, the vocabulary limitation (2000 vs more than one million) had the potential to bias the model heavily to frequent terms which are likely to co-occur together thus increasing their embedding similarity. Both models

Model	Scholar Baseline	Scholar Disease Epoch	Scholar Authors Institution	Scholar Both	MetaLDA Baseline	MetaLDA Disease Epoch	MetaLDA Authors Institution	MetaLDA Both
Scholar Baseline								
Scholar Disease Epoch	0.1							
Scholar Authors Institution	0.1	0.067						
Scholar Both	0.233	0.2	0.233					
MetaLDA Baseline	0	0	0	0				
MetaLDA Disease Epoch	0	0	0	0	0.7			
MetaLDA Authors Institution	0	0	0	0	0.833	0.633		
MetaLDA Both	0	0	0	0	0.633	0.867	0.667	

Table 4: Coverage of overlapping topics (an overlapping topic meets the 7/10 similarity threshold) over 30 topics. A cell represents how many topics overlap normalized by n=30 topics. Scholar model is shown in blue and the MetaLDA model is in green. The comparison between different model types is shown in yellow.

used word embeddings (either pretrained, or embeddings learned during processing), and therefore neither model necessarily has a strong advantage on the feature front.

The UMass metric also agrees with the cosine metric. We believe one reason the UMass score might be higher for Scholar is again because of the vocabulary limitation. The UMass metric measures how well common words can predict less common words. Since the Scholar model had a limited vocabulary, there were fewer words for it to test against.

If you look at topics created by Scholar in Table 5, you see that they are much more specific and are composed of more highly domain specific words. In contrast, the top terms from MetaLDA shown in that table are more generic. It would make sense that the scholar model had a higher cosine similarity value, since it was mainly composed of highly domain specific words which would have very similar embeddings. This does not necessarily mean that it produces better topics. Since we only had 30 topics and a corpus of thousands of documents, we would expect to have more general topics in order for the topics to cover all of the documents accurately. With this in mind, we could say MetaLDA produced better topics in general.

Looking at our matrix comparison of similar topics in Table 4, we see that none of the scholar models create any similar topics with MetaLDA. Some top terms did cross over, but no single topics met our similarity thresholds across the models. This is again unfortunately most likely due to the vocabulary limitation we had to impose on Scholar, however, it's worth considering that the scholar models themselves had few overlapping topics. The scholar model does appear to be more sensitive to the metadata attributes, relatively few similar topics seem to have appeared in the difference scholar models despite sharing the same vocabulary and parameters. In contrast, the MetaLDA models consistently shared between 19 and 26 similar

Model	Top 10 words
Scholar Baseline	loop, base, pro, stem, energy, nucleotides, pair, industry, strand, residue
Scholar Disease Epoch	compounds, inhibitory, mixture, potent, synthesized, inhibitors, synthetic, protease, drug, plasmid
Scholar Author's Institution	activate, clearance, innate, receptors, delivery, adaptive, cytokines, trigger, induces, signaling
Mallet Baseline	virus, viral, viruses, infection, host, replication, influenza, human, cell, hiv
MetaLDA Disease Epoch	health, care, public, disease, information, medical, data, surveillance, control, diseases
MetaLDA Author's Institution	cells, fig, cell, protein, min, expression, figure, incubated, control, assay

Table 5: The top 10 words for the most probable topic given all models

topics between the various model variations suggesting either the model was more stable or the impact of the covariates was less. It's unclear without further investigation as to why this may be, however, we surmise that the 'neural-ly' nature of Scholar may contribute to less stability of the outputs across runs.

4.2 External vs Internal evaluation

For both the scholar and metaLDA models, the overall cosine scores were lower when we used CORD-19 word embeddings compared to the pre-trained external Wikipedia based embeddings. However, the differences in the scores were smaller than the differences between the models. GloVe embeddings are computed based on the co-occurrence of terms within a context window; within a large generic corpus like Wikipedia, biomedical terms are likely to co-occur with each other frequently compared to how often they co-occur with other terms in the very broad corpus. This makes their embedding representations 'closer' together compared to representations in the covid corpus. Within the covid literature, 2 terms that are highly specific to different medical sub fields may not occur as often as they are more distinct to the medical field. This would lead those 2 words to have representations that are further apart in the embedding space trained only on that biomedical literature. This opens a conundrum for the use of some of these metrics in automatically evaluating topic models. The 'external' scores suggest the model did better than the internal ones, however, there is an argument that since it might be difficult for non-medical experts to understand the topics (or underlying literature), the internal scores better represent how a human medical expert would evaluate the topics and hence is the more appropriate measure to use. Exploring this dichotomy will be left to future work.

4.3 Metadata/Covariates

Neither of the metadata labels/covariates we explored seemed to have substantial impacts on either of our models. The scores when considering either disease epoch or top author's institution seemed to vary significantly from the respective baseline models regardless of which embeddings we used for evaluation. Furthermore, when we tested the models with both features as metadata attributes the scores overall worsened. As previously mentioned, the scholar models varied substantially more when the different metadata attributes were added. The set of topics seemed to shift more. The metaLDA topics did shift between different runs however even at worse over half of the topics remained 'stable'. Investigating the specific nature of these shifts and trying to capture what those changes are is reserved for future work.

Model	Topic Probability	Top 15 words
Mallet Topic 15	0.049	patients, blood, treatment, disease, clinical, therapy, infection, hiv, risk, patient, infections, cases, recipients, hepatitis, liver
MetaLDA Insitutions Topic 4	0.043	patients, study, group, studies, treatment, results, data, significant, asthma, age, compared, clinical, risk, analysis, higher
MetaLDA Disease Epoch Topic 7	0.043	study, group, data, studies, results, analysis, significant, age, groups, compared, higher, risk, included, table, significantly

Table 6: Topics related to risk factors

4.4 Risk Factors

The Scholar models consistently failed to find a topic specifically related to risk factors for coronavirus diseases. While many of the relevant terms that we would expect to be related to risk, such as 'risk', 'patient', 'asthma', 'hiv' etc, are found in the filtered vocabulary, they never made it into any top 10. Even when expended the topic size to 15 words, we were unable to reasonably label any of the Scholar topics as 'risk factors'. Three of the four MetaLDA variants did uncover topics that we were able to somewhat reasonable label as covid risk factors (albeit with some googling help to interpret some of the medical terms) and had similar probability. This became even more apparent when we increased the topic size to the top 15 words. These topics can be seen in Table 6. While these topics look promising, when we looked at the documents most correlated with these "risk topics", we found that the papers were only loosely related to coronavirus. For example, some of the top document titles were "Nasal decongestants in monotherapy for the common cold", "Temporal Trends of In-Hospital Mortality in Patients Treated with Intra-Aortic Balloon Pumping: A Nationwide Population Study in Taiwan, 1998-2008", "Immunology of Multiple Sclerosis" and "Modulation of Multiple Sclerosis and Its Animal Model Experimental Autoimmune Encephalomyelitis by Food and Gut Microbiota". This could mean that while tops words in the topic seems to be related to coronavirus risk factors, that is not what the model is actually detecting.

5 Conclusion

We performed topic modeling on the COVID-19 Open Research Dataset (CORD-19) dataset using two different approaches: Scholar and MetaLDA. This was an interesting task that had its own unique challenges like niche domain-specific scientific terms. We found that the MetaLDA models created topics with more generic words, while the Scholar model produced topics with domain-specific words. We believe these differences came from the vocabulary limitation with the Scholar model. We performed multiple experiments using disease epoch and author's institution as covariates for the topic model. However, these on their own did not have a substantial impact on the topics generated. We also performed topic-modeling using their combination. We used cosine similarity of top-10 words and UMass measure for 30 topics to evaluate the output of Scholar and MetaLDA. We found that for both Scholar and MetaLDA, the pre-trained external Wikipedia embeddings produced better outputs (*higher cosine similarity scores*) than the CORD-19 word embeddings. This might be because the pre-trained external Wikipedia embeddings captured generalizations better.

With more computational resources and time, we would like to look at how the number of topics affected the results. To make the comparisons easier we all used the same number of topics, but it would have been interesting to see if the best value for this parameter changed based on the model. Both the Scholar and MetaLDA took considerable time to run which made it hard for us to do much hyperparameter tuning. MetaLDA took 4 hours per run when running with 16 threads on a desktop computer. The Scholar model took 8 hours per run even with the vocabulary limited to 2000 words (without stopwords) on a high performance computing cluster. MetaLDA has additional options to tune how the parameters of Dirichlet priors, alpha and beta, are sampled. We would have liked to experiment with these values further, but found that changing these values could more than double the execution time. Thus, we concentrated on only the fastest option.

Given more time, we also would have performed more pre-processing on the data. By looking at other submissions on Kaggle, we saw that people had good results when filtering out the most common words that appear in more than 95% of the documents and the least common words in less than 5% of the documents. This may have helped us with the problem we had with domain-specific words. We also would have used a better method for detecting language.

Our biggest limitation in the analysis generally was the lack of domain expertise of our research team in the medical field. Collecting human assessments of the topics would have helped clarify which model was actually generating the more useful topic outputs.

References

N. Aletras and M. Stevenson. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 13–22, 2013.

D. Card, C. Tan, and N. A. Smith. Neural models for documents with metadata. In *Proceedings of ACL*, 2018.

D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 262–272. Association for Computational Linguistics, 2011.

S. Nakatani. Language detection library for java, 2010. URL `https://github.com/shuyo/language-detection`.

K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961. Association for Computational Linguistics, 2012.

L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, et al. Cord-19: The covid-19 open research dataset. *arXiv preprint arXiv:2004.10706*, 2020.

H. Zhao, L. Du, W. Buntine, and G. Liu. Metalda: A topic model that efficiently incorporates meta information. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 635–644. IEEE, 2017.