TOVID: Trend in COVID-19 Related Academic Literature

Boyang Yu[1*]
[1]Center for Data Science, New York University
by2026@nyu.edu

Abstract

The massive size and the diversity of topics in thousands of COVID-19 related publications makes it an overwhelming task to manually detect the sub-categories, not to mention the change in research trend over time. Statistics-bases topic models are effective tools to analyze the semantic style in large document collections. However, incorporating temporal coherence and maintaining a fair interpretation of the results remain as challenging tasks, especially when the study object is highly real-life based. The aim of this study is to integrate a machine learning point of view and a comprehensive topic evolution framework into the recently flourishing field of SARS-COV-2 studies and propose a retrospective thinking on how the science community could better fight against the disease.

1 Introduction

The year of 2020 witnessed a unprecedented global pandemic which brought up tons of opportunities for novel research objectives in the coronavirus disease 2019 (COVID-19) related literature. Subjects like disease prevention[Cummings *et al.*, 2020], epidemic spread models[Bertozzi *et al.*, 2020], as well as the understanding of viral structures[Shen *et al.*, 2020], serve as magnificent contributions from the science community to assist our general human kind to in the silence battle against the coronavirus pandemic.

It's been almost a year since the emergence of the first outbreak. During this period, the humankind as a whole experienced faced a series of problems that were never seen before, which introduced never-stopping discussions and explorations, including but not limited the effectiveness of lockdown policy and the origin of the novel coronavirus. Answers and interesting discoveries from researchers all over the world flourished the science world in the forms of growing numbers of related publication in either journals or conferences, as well as well formulated pre-print works. Special sessions or workshop on COVID-19 are set up in conferences in the year of 2020, such as Empirical Methods in Natural Language Processing(EMNLP), in the hope of encouraging and assisting the silence battle with the pandemic using modern technology and transforming technical contribution into perceivable social goodness.

Knowledge discovery with the disease itself is exciting and influential. Knowledge discovery with the knowledge discovery is more than exciting and influential. It is the time for the science community to reflect on what is done and what needs to be done in the future given the fact that the chances of a second outbreak are high if effective measures are not taken. The longitudinal view of documents is would largely guide the direction of future investigations. Researchers should think seriously about this issue.

However, given the nature of academic literature as being sometimes obscure and the understanding of which highly time-consuming. Systematically scanning each documentation and making assessments requires reasonable amount of human expertise and laboring. Besides, the objective disagreement would be another detrimental yet unavoidable problem with human assessment. Therefore, a statistically convincing model describing the dynamics in research themes over time would for sure be a decent starting point since this could later be used as on-the-flow tools keeping track of updated publication data collections. Topic models like Latent Dirichlet Allocation(LDA)[Blei *et al.*, 2003] provide a statistical aspect in understanding text data from a probabilistic point of view. As a result, it is powerful when the number of documentation samples goes so high that no human experts can handle.

In this paper, we proposed a topic evolution framework TOVID which captures the temporal dynamics in the COVID-19 centered articles. Outputs of the framework allows researchers track the change in popular research themes over the year of 2020, as well as compare the similarity across broad categories using further dimension reduction techniques. Features such as on-the-flow modeling with updates in source data streaming, as well as interactive visualization tools are made available.

Our efforts make use of probabilistic topic modeling for topic which can be executed automatically with the research literature cohort with temporal labels indicating when the major work is done. The intent of this work is not to provide new algorithmic contribution to the field of natural language processing(NLP), but rather,a sincere advocate for more needs-driven rather than technology-driven effort to decipher real-

*Contact Author

world mysteries using NLP models. We claim that using topic modeling, especially dynamic topic modeling(DTM) with calibration would serve as a starting point for further work to dive in the retrospective assessment in how science contribution is made and develop unique insights for COVID-19.

2 Related Work

Back on March 16, 2020, an open challenge called COVID-19 Open Research Dataset Challenge (CORD-19)[Kaggle, 2020] was launched by White House along with some leading research groups aiming to encourage researchers extract semantic information through over 200,000 scholarly articles related with the global pandemic. Along with the release of the data set, 17 high priority issues, proposed by NASEM's SCIED (National Academies of Sciences, Engineering, and Medicine's Standing Committee on Emerging Infectious Diseases and 21st Century Health Threats) research topics and the World Health Organization's R&D Blueprint for COVID-19, served as the competition tasks. Among the tasks, 9 aspects about COVID-19 are proposed as the topics, where a complete list is attached in the appendix. Extensive AI(Artificial Intelligence) solutions are made available from researchers worldwide following the topic classification definition as mentioned before.

Not only does the successful use of text mining in this challenge reflect the potential to apply computational linguistics to COVID-19 related literature for semantic information retrieval, but also the importance of sub-grouping with the massive research articles, where the essence is similar to subtyping patients with the same disease helps with downstream treatment.

Nevertheless, though health specialists' judgement in deciding the topics succeed in integrating domain-specific knowledge into decision making process, the categories still take the risk of overlapping and ambiguity. For instance, one may find it difficult to differentiate between "transmission, incubation, and environmental stability" and "medical care". Therefore, probabilistic topic modeling deals with this nuance by giving the probability for the document belonging to a specific category, which means a quantitative way to assign the document class is made available.

Though institutional response to COVID-19 text data collection is quick, the selection criteria might be less satisfactory due to little calibration. Researchers[Doanvo *et al.*, 2020] have discovered that some of the papers in the CORD-19 cohort were neither relevant to SARS-CoV-2 nor other coronaviruses. A more curated data set LitCovid[Chen *et al.*, 2020] was designed by a group of researchers from National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM) and National Institutes of Health (NIH). As a daily updated text data hub focusing on published articles on COVID-19 in PubMed (i.e. articles on other coronaviruses, such as SARS or MERS, are not included) in a comprehensive and precise manner. Curation includes document selection and the assignment of categories. Eight broad topics(general information, mechanism, transmission, diagnosis, treatment, prevention, case report, or epidemic forecasting) are determined through a deep learning model using BioBERT embedding and manually crafted features.

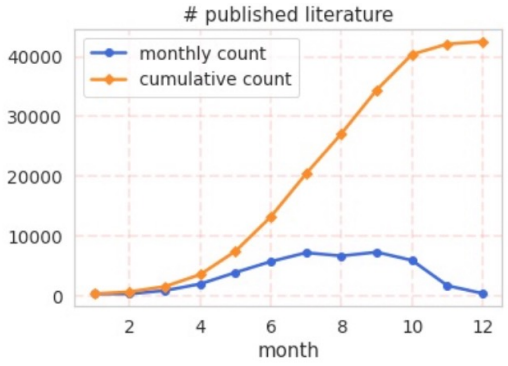

Figure 1: Number of publications in the selected cohort

Even though LitCovid proves success in aligning advanced computational linguistics models like BioBERT to the categorization, it fails to provide an overview of whether research efforts have changed over time. With the development of the pandemic at different stages, the topics are expected to vary. In early outbreak, epidemic models[Holmdahl and Buckee, 2020; Bertozzi *et al.*, 2020] predicting the spread of the disease are popular. Nonetheless, when the spread goes out-of-control, most of the models failed in the prediction and few spread models are proposed by the end of the year. Moreover, the topics are very broad in the sense that words like "treatment", "diagnosis" and "vaccination" are used to represent three categories. Since there are only eight categories in total, sometime one may found the labeling not informative enough to represent the content of the research article. Afterall, the topic assignment in LitCovid serves as a tool to facilitate the searching and filtering in the data base containing thousands of records.

Problems not well addressed by previous work include:

- Documentation-topic proportional investigation along the time axis.
- Topic proportional evolution's connection with real-world disease development

In this work, a framework TOVID integrating previously curated literature collection and probabilistic topic modeling methods is presented to address the temporal change and evolution tendency in novel coronavirus related research publications.

3 Structural design

Our framework proposed in this paper addresses these problems using the LitCovid cohort for its curated inclusion which as a result, provides enhancement in text data quality and the high signal to noise ratio. The selection criteria ensures less effort on down stream document preprocessing and normalization which helps reduce human bias in the final semantic information extraction step.

The design of the proposed framework TOVID, Figure 2, can be divided into three major task, with the first one selecting documents which meets the need for topic modeling,

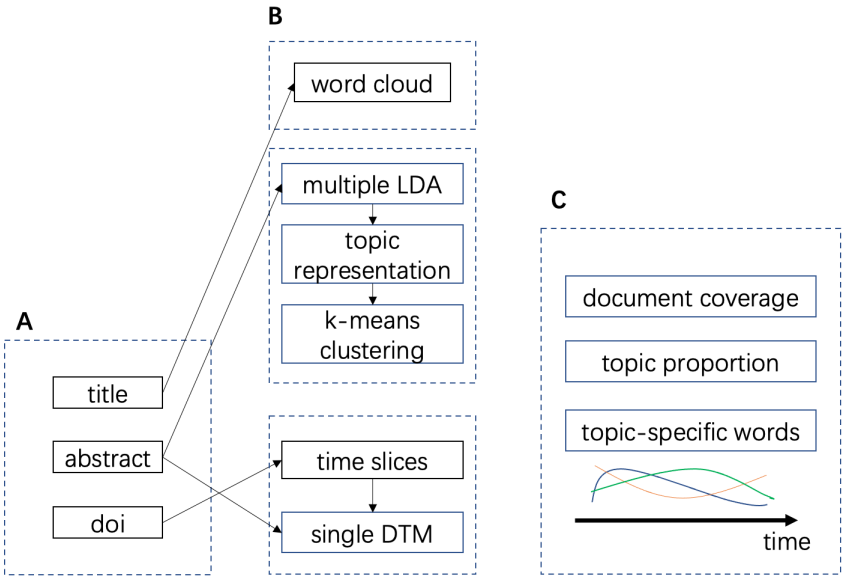

Figure 2: An overview of the workflow(A-C).
(**A**)Collected COVID-19 themed literature are systematically checked, only those with complete information on title, abstract and Digital Object Identifier(doi) are used for the next module. (**B**)The text in title and abstract will be forwarded for various tasks along with the time label extracted from doi number. (**C**)The output of TOVID consists of three major aspects of changes along time involving the coverage of document content, topic proportion and the probability of topic-word matching. **Note**: The abbreviations, LDA and DTM, stand for the method in topic modeling, namely, latent Dirichlet allocation and dynamic topic modeling.

the second to detect the topics using known temporal tag by month level with two coherence alignment strategies, the last to connect evolution phases in the document to real-life development of the spread the disease and its social impact on the human society.

With no doubt, the second module in this proposed framework requires considerable amount of technical calibration with probabilistic modeling in both the topic retrieval task and its integration of time information. It is possible to further divide the two component into sequential processes, extracting major topics for each time group(treating documents belonging to the same time slice separately), followed by applying unsupervised clustering to group the topic representation. Another option, treating the whole document collection as a whole regardless of the time period, puts more stress on the coherence within the same topic along the time indices. Details of this module will be covered in the next two sections.

The first and last module, instead, requires a sharp understanding into the real-life understanding of characteristics of research articles' structure and patterns of how research publications related with real-life events. For instance, in the first module, items in the original LitCovid cohort without explicit doi information are dropped since a valid doi number represents a recognized identity for the article. The use of title in word cloud generation copes with the characteristics of key word visualization using condensed information. While using abstract to develop topics shares the consciences with human scholars as the abstract covers major methodology and discoveries for a particular study. Moreover, although not using the complete article certainly missed part of the information, it saves the time cost by avoiding extensive computation on an extremely large but highly sparse dictionary. In short words, the assumption here is that the use of abstract-only corpus serve as an effective dimension reduction technique in analyzing research article. Besides, the exclusion of empty value in the three parts in module A on Figure 2 ensures that the article is well illustrated with reasonable structures and origins. After the straight-forward filtering, the number of samples in LitCovid(downloaded on November 20, 2020) dropped from 72719 to 42077, which means that 42077 articles will later be used to test the pipeline of our proposed TOVID in topic evolution. An overview of monthly number of publication is illustrated in Figure 1.

Besides data selection, the first module in the proposed framework, TOVID, also includes transforming the text corpus into vectors as the input to the topic models. The dictionary and the implementation of vectorizer. Built-in stopword lists in popular natural language processing tools such as *nltk*[NLTK-Team, 2020] may not be satisfactory since they are not designed for academic writing, not to mention the COVID-19 theme. An obvious problem with the built-in function would be that the synonyms of the novel coronavirus, such as COVID-19 and SAR-COV-2, would never be considered as stop words in these lists. And researchers tend to show their own preferences on the use of the name of the pathogen. Therefore, not using a carefully designed stop words list would lead the model to group the topics with the preference of the usage of the name of the disease, rather the true latent topics aroused around the topic. In this framework,

the calibrated filter includes both COVID-research oriented words and common academic English words, as described in the appendix. For the choice of vectorizer, the unigram count vecterizer is adopted for less sparsity since the dictionary covered is already diverse enough.

4 Topic Retrieval

The concept of topic modeling is to take in a collection of text data, uncover some latent topics characteristic of the input, and denote each article as a mixture of those topics. Each latent topic is in fact characterized as a distribution over the text collection's vocabulary. Generally speaking, a large proportion of the topic models can trace back to a common origin, the Latent Dirichlet Allocation(LDA)[Blei *et al.*, 2003], an unsupervised probabilistic algorithm using a three-level hierarchical Bayesian model to identify the topic-word distribution and document-topic distribution, as described in Figure 3.

To implement LDA, one has to explicitly input an integer indicating the total number of topics. And the optimal choice of that number relies on both quantitative evaluation metrics, like low perplexity or high coherence, and a principled, manual assessment over the resulting topics. A Consistent with previous work[Doanvo *et al.*, 2020] on selecting the number, we found the quantitative approach failed detect optimal value by enumeration the number of topics from 5 to 100, the perplexity score is visualized in the appendix, potentially suggesting a broad yet insubstantial pool of COVID-19 publications. Therefore, a human-in-the-loop judgment was made to select the number of topics by manually checking the top 20 words in each topic, as well as using Principal Component Analysis(PCA) for topic similarity comparison.

The determination comes from experimenting a single LDA model over all the titles in study cohort. In order to keep the number of topics within the scope of human interpretation, the enumeration goes through 5 to 15 and arrive at the number of 6. In the following experiment with time-dependent topic models and monthly LDA, 6 will remain as the total number of topic to be assigned.

4.1 Two-step solution

Using LDA as the building block, the COVID-19 literature cohort is further divided into 12 disjoint subsets by publishing month for separate LDA modeling. In this way, 12 base models and 72 base topics are obtained using the abstract paragraph of the original articles. It is natural to assume some of those topics are highly correlated. However, there's no prior knowledge on the actual topic similarity, which makes manually merging the 72 topics into smaller number of topic-groups unrealistic. In order to accomplish the semantic matching task, a systematic "bundle topic" construction method, Algorithm 1, is applied to the base topics.

The first phase in the major loop utilizes k-means, an unsupervised clustering method to partition some observations into k clusters in which each observation belongs to the cluster with the nearest cluster centroid, is applied to merge similar base topics. To construct the input of k-means, top t words in each topic together with the probability of the word belonging to the topic are used to represent each topic. For each

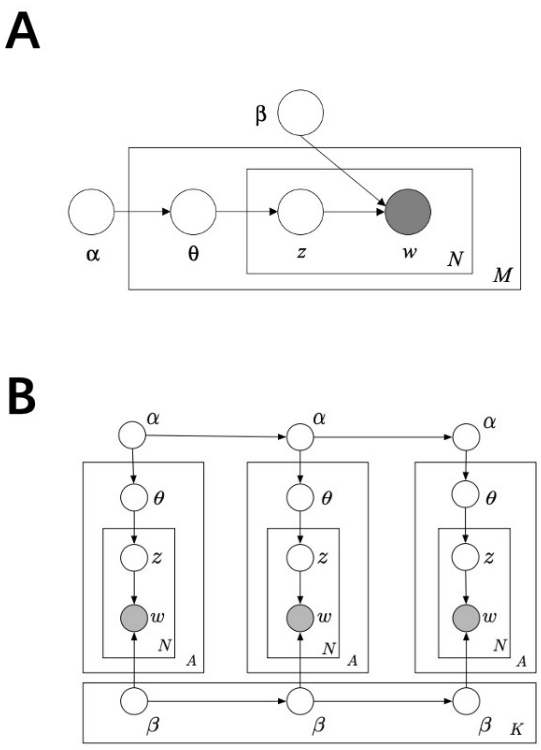

Figure 3: Graphical model representation of LDA and DTM(A-B).
(A) LDA illustration: N and M are the number of words and documents, representing the number of replicates to assign the words in every document. Arrows in the graph indicates the order of the generative process in the sense that α is the parameter in Dirichlet distribution that gives θ, which in turn gives rise to topic z over a multinomial distribution. Finally, the word w is drawn from a multinomial probability conditioning on the topic z and β.
(B) DTM illustration:N is the same as in the LDA model. A and K are the number of documents and time indices, representing the number of replicates to assign the words in every document. The difference with the LDA process is in the outer loop with α and β where they are generated sequentially from a normal distribution with the value in the last time index as the mean, if not the first time slice.

Algorithm 1 Bundle topic algorithm

Input: n base topics, $T = \{t \| t$-word representation$\}$,
K=$\{k \| k$ means cluster, $k < n$ $\}$
Output: k bundle topics ($k \in K$)

for all t,k **do**
  Generate t-representation of all n base topics
  Apply k means clustering with the representation
  Decode k centroids
  if reasonable decoded sequence **then**
    Use k centroids as k bundle topics$k$ bundle topics
  else
    Continue the loop.
  end if
end for
return k bundle topics

value of t, k clusters are generated using the t-representation. The naive cluster representation in k-means algorithm is the geometrical center for each cluster, which fails to map the vocabulary in the text. Therefore, a decoder of the cluster is designed by selecting the first 5 terms in each cluster in descending order of summed probabilistic representation. The optimal choice of t and k come from a grid search with user-defined set of values to choose from.

Bundle topics created using above algorithm would serve as the final latent topics and the articles would be considered as a mixture of those bundle topics. The final document-topic assignment follows an onto map from the base topic in each single LDA model to the bundle topic. The proportional analysis is operated at the bundle level.

The proposed algorithm constituted the second block of sub-figure **B** in Figure 2. On the one hand, treating document at each time slice as independent cohort for LDA modeling allows for deeper exploitation in the vocabulary space for each single smaller cohort. On the other hand, post processing of the base topics require additional calibration and fine-tuning the parameters.

4.2 Single-step solution

Variants of the fundamental LDA model cover a wide range of aspects, including but not limited to temporal structure, hierarchical topics, supervision, temporal structure, discrete covariates. It is natural to directly inherit an enhanced version with adpation in the time domain of the original LDA to the selected LitCovid publications' abstract. Three years after the official release of LDA topic model, the author made temporal structure available to the original idea and proposed Dynamic Topic Models(DTM)[Blei and Lafferty, 2006]. Unlike the static topic model which assumes that the documents are drawn exchangeably from the same set of topics, DTM, illustrated in Figure 3, treats the documents for each time slice with a $K-$component topic model, where the topics associated with slice t evolve from the topics associated with slice $t-1$, reflecting the evolving trend in topics.

The implementation of DTM is similar to that of LDA, requiring a specified number of topics passed on before the fitting. As mentioned before, for better integrity with the two-step solution, six, optimized result in the preliminary LDA using all titles in the LitCovid study cohort, dynamic topics would be generated using DTM. Apart from the number of topics, DTM requires a predefined time slice label indicating specific onto mapping from the document to the time index. Still, the month of the publishing date is used as the time label, that being said, 12 levels of time slices are given to the DTM model.

The advantage of using DTM is obvious since a coherent transition within the same topic is a major contribution. Moreover, less subjective judgement is introduced without merging the topics afterwards. However, temporal coherence comes at a price of sacrificing the diversity of representative vocabulary and a higher running time cost with a more complex optimization.

The comparison between the two proposed approach is indeed seeking the balance between exploration and exploitation, in both time coherence and vocabulary diversity. Using bundle topics built from base topics certainly allows for deeper exploitation in the vocabulary, while a single models that captures the dynamics like DTM pays more attention on exploitation on time coherence. The exploration of time information exists in the form of treating text corpus at each time slice as independent components with the two-step solution. Meanwhile, the exploration in vocabulary lists is maintained within the DTM optimization. Characteristics of the two topic retrieval methods result in a vibrant trending analysis module illustrated in the following section.

5 Literature Trending Assignment

All the word clouds, two-step topic modeling, single-step topic modeling are used for different aspects in assigning the trending. The three aspects will later be compared as both the consistent part and inconsistent part suggest a deeper insight into how researches facilitate the control of the global pandemic.

5.1 Document coverage

Word cloud serve as an exploratory role in unraveling document coverage with a given period of time. A straight forward way to see the words in the corpus with size indicating the significance is the major purpose to directly reflect the coverage of input text data with minimum effort on manual fine-tuning. Example figure can be found in Figure 4. Looking at the subfigures in an up-bottom manner helps one immediately identify the keywords in prevalent research. Words like 'wuhan', 'china' in Figure A reflects that a considerable of articles are related with the reported cases in China since the publications by March are expected to use earlier epidemic statistics studying the spread in China. When it comes to June, with the regional outbreak growing to a global pandemic, a broader range of topics is covered in the literature, especially the viral related research and healthcare experiences. Moreover for the month of November, the word cloud result doesn't convey illustrative information partially due to the complexity of academic topics is too high to be captured by a simple word cloud.

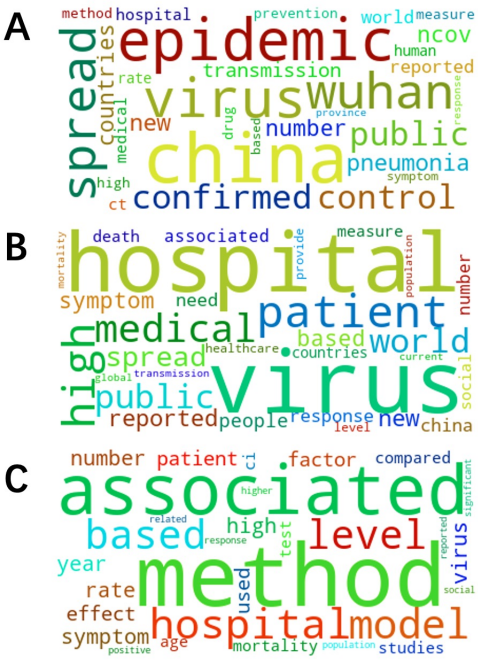

Figure 4: Illustrative document coverage(A-C).
(A) March **(B)** June **(C)** November

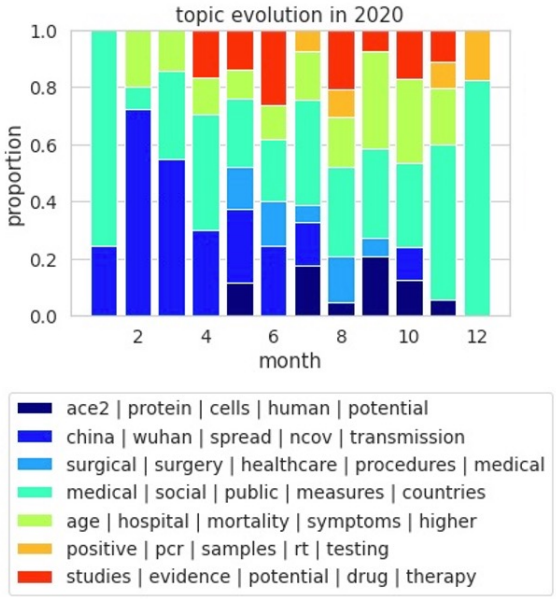

Figure 5: Topic proportion

5.2 Document proportion

The analysis of document proportion in each topic comes from the two-step topic retrieval process described in the previous section where the bundle topic would serve as the categories to assign the fraction. The complete corpus is further divided into 12 subgroup by the month of release. For each element in the smaller cohort, the final category goes to the one with highest probability. Then the proportion is the normalized marginal document distribution over all categories(bundle topics).

In Figure 5, different colors map to different bundle topics discovered using post topic matching algorithm described in 1. The number of bundle topics is selected as 7 and important intermediate result of the parameter tuning can be found in the appendix. At the beginning of the year 2020, only two categories are presented, largely focusing on reporting the cases and generic healthcare measures. And when it comes to February and March, some work indicating mortality and hospitalization is published as more systematic investigation into the pandemic is available. For later studies, the topics grow more and more inclusive as indicated by the colorful bars in the figure. The results for December might be less credible due to the limited number of articles included.

Besides the inter-categorical comparison, the change within the proportion of one specific category is intriguing as well. For instance the orange bar with top 5 terms of 'positive', 'pcr', 'samples', 'rt' and 'testing' representing COVID-19 testing related research emerges in July, indicating the problem that the amount of effort put into this area is not that much in the first half of the year. On the contrary, the light green bar with 'age', 'hospital', 'mortality', 'symptoms' and 'higher' maintains a rather stable appearance in each month, pointing towards an ongoing interest in the medical care and risk factors of the global pandemic.

5.3 Topic-specific words

Dynamics within the same topic, or in other words the distribution of representative key words, also reflect the refined tendency variation at multiple time slices. The proportional analysis could be considered as a macro-level trend discovery while the topic words represent micro-level evolution. The coherence of the DTM topics are suited for this task due to its high coherence within the same topic along the time axis. As mentioned before, 6 topics are retrieved from the 12-level document. The discrete time-dependent distribution over the dictionary in each topic.

There are three major types of representative keywords within one single topic. The first type has a decreasing tendency, such as 'china' and 'wuhan' in the illustrative case in Figure 6. Both the order and the exact probability value shows the dropping passion in mentioning cases in China, which is consistent with the fact that the pandemic is no longer a regional outbreak but a global threat. The second type is exactly the opposite to the first, with an increasing tendency. A typical instance would be with the word 'testing' in Figure 6, which appeared for the first time June and maintained a high probability all the way until December. The third category goes to words that appear to be an ad-hoc topic over several months then gave way to other issues, like 'deaths' on May and June, the violet polyline in sub-figure B in Figure 6.

Besides the three distinguishable patterns, some words have a rather stable contribution to a specific category, such as the word 'transmission' in Figure 6. These rather stable

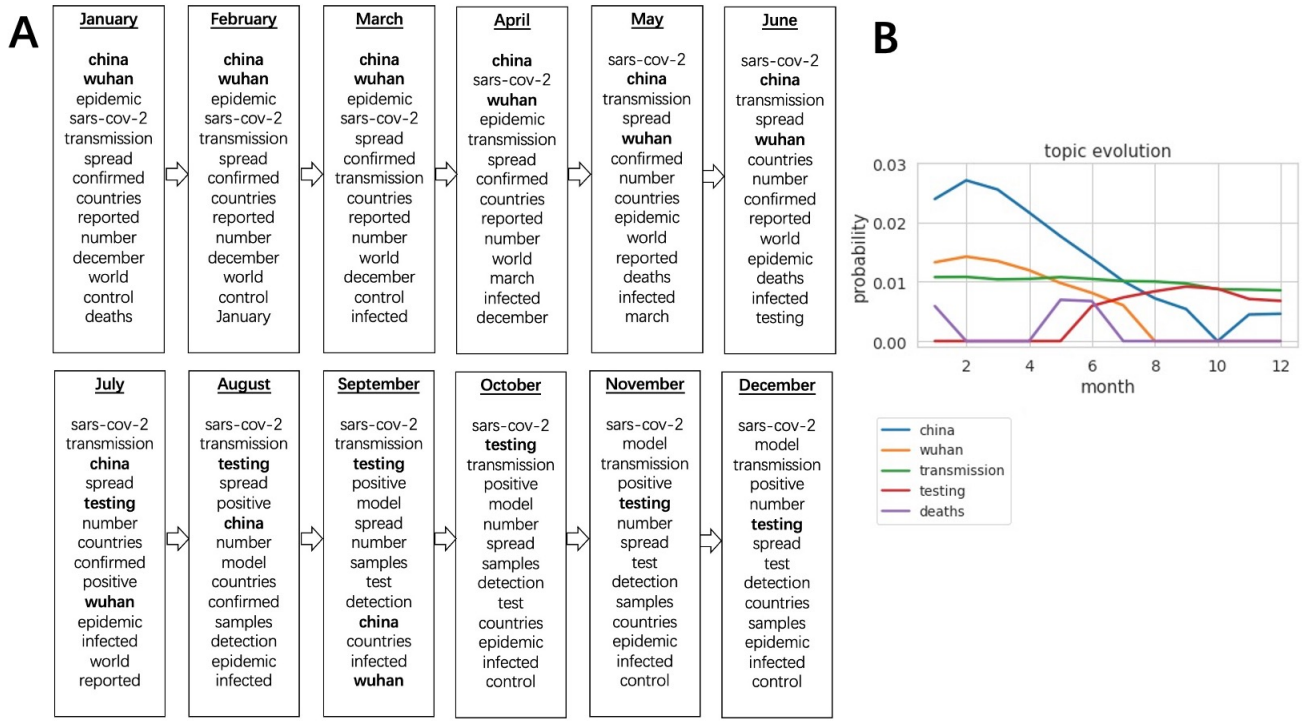

Figure 6: Topic-specific words distribution reflect research trend(A-B).
(A) Topic words ordered by probability in descending order. Bold words indicate informative change of order. **(B)** Probability value fluctuation over 5 terms within the same topic

words are the most important ones to summarize a broader topic or keywords identifying the topic. Althogh they are crusical for semantic analysis, they do not reflect the evolving trend, thus less relevant to this study.

The above alignment is applicable to all the 6 dynamic topics, where the results contented based.

Altogether we presented three ways to identify the change in research sub-domains related with COVID-19, either in a graphical way or quantitative way. The reason to perform such a diversified analysis is to reflect the evolving process as vividly as possible the community involved with the global pandemic. The results are more authentic in the chart or graph, while a few sentences to summarize are less thorough with so many points to talk about in details. Therefore this section is organized by illustrative examples but rather abstract description.

Since the number of research articles are still growing given the fact that we will not stopping fighting with the disease in the near future. The framework can be re-trained with new samples and extended length of time. In other words, the generalization capability of the proposed work has the potential to become an on-the-flow system that can be updated daily with dynamic input.

A sketch of how to generalize the TOVID framework with real-life events include:

- First emergence of a specific topic
- Topic subtyping(vanishing, growing or stable)
- The length of prevalence

6 Discussion

The longitudinal view of documents is largely underemphasized in recent work on natural language processing, except for some exceptions on dynamic topic models. The building blocks of this TOVID framework would be rather intuitive for some experts in topic models. But the results, especially the evolution of topics, either in document proportion or top ranked words' distribution, are meaningful in reminding medical researchers, clinicians, and policy makers of their contribution, possibly guiding the direction of researchers in the near future.

The feasibility of this work implies that a closed loop from designing an original idea to the realization of the idea is already efficient enough in today's scientific research. Improving the productivity of carrying out research projects is not the most urgent task for the community. Instead, it is the time to go back and examine the quality of products of this efficient closed loop.

Limitations with this work exist in both the topic modeling methodology and data collection process. We believe those limitations could further inspire more experienced researchers to conduct in-depth research either from a methodological perspective capturing more confound latent topics in the academic literature or with a more comprehensive multilingual literature cohort.

References

[Bertozzi *et al.*, 2020] Andrea L. Bertozzi, Elisa Franco, George Mohler, Martin B. Short, and Daniel Sledge. The challenges of modeling and forecasting the spread of covid-19. *Proceedings of the National Academy of Sciences*, 117(29):16732–16738, 2020.

[Blei and Lafferty, 2006] David Blei and John Lafferty. Dynamic topic models. volume 2006, pages 113–120, 01 2006.

[Blei *et al.*, 2003] David M. Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[Chen *et al.*, 2020] Q. Chen, A. Allot, and Z. Lu. Keep up with the latest coronavirus research. *Nature*, 579(7798):193, 2020.

[Cummings *et al.*, 2020] Matthew J. Cummings, Matthew R. Baldwin, Darryl Abrams, Samuel D. Jacobson, Benjamin J. Meyer, Elizabeth M. Balough, Justin G. Aaron, Jan Claassen, LeRoy E. Rabbani, Jonathan Hastie, Beth R. Hochman, John Salazar-Schicchi, Natalie H. Yip, Daniel Brodie, and Max R. O'Donnell. Epidemiology, clinical course, and outcomes of critically ill adults with covid-19 in new york city: a prospective cohort study. *medRxiv*, 2020.

[Doanvo *et al.*, 2020] Anhvinh Doanvo, Xiaolu Qian, Divya Ramjee, Helen Piontkivska, Angel Desai, and Maimuna Majumder. Machine learning maps research needs in covid-19 literature. *bioRxiv*, 2020.

[Holmdahl and Buckee, 2020] Inga Holmdahl and Caroline Buckee. Wrong but useful — what covid-19 epidemiologic models can and cannot tell us. *New England Journal of Medicine*, 383(4):303–305, 2020.

[Kaggle, 2020] Kaggle. Covid-19 open research dataset challenge (cord-19). https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/, 2020. Accessed December 14, 2020.

[NLTK-Team, 2020] NLTK-Team. Natural language toolkit documentation. https://www.nltk.org/, 2020. Last updated on Apr 13, 2020.

[Shen *et al.*, 2020] Bo Shen, Xiao Yi, Yaoting Sun, Xiaojie Bi, Juping Du, Chao Zhang, Sheng Quan, Fangfei Zhang, Rui Sun, Liujia Qian, Weigang Ge, Wei Liu, Shuang Liang, Hao Chen, Ying Zhang, Jun Li, Jiaqin Xu, Zebao He, Baofu Chen, Jing Wang, Haixi Yan, Yufen Zheng, Donglian Wang, Jiansheng Zhu, Ziqing Kong, Zhouyang Kang, Xiao Liang, Xuan Ding, Guan Ruan, Nan Xiang, Xue Cai, Huanhuan Gao, Lu Li, Sainan Li, Qi Xiao, Tian Lu, Yi Zhu, Huafen Liu, Haixiao Chen, and Tiannan Guo. Proteomic and metabolomic characterization of covid-19 patient sera. *medRxiv*, 2020.

A CORD-19 Task

- What is known about **transmission, incubation, and environmental stability**?
- What do we know about **COVID-19 risk factors**?
- What do we know about **vaccines and therapeutics**?
- What do we know about **virus genetics, origin, and evolution**?
- What has been published about **medical care**?
- What do we know about **non-pharmaceutical interventions**?
- What has been published about **ethical and social science considerations**?
- What do we know about **diagnostics and surveillance**?
- What has been published about **information sharing and inter-sectoral collaboration**?
- Create summary tables that address relevant factors related to COVID-19
- Create summary tables that address therapeutics, interventions, and clinical studies
- Create summary tables that address risk factors related to COVID-19
- Create summary tables that address diagnostics for COVID-19
- Create summary tables that address material studies related to COVID-19
- Create summary tables that address models and open questions related to COVID-19
- Create summary tables that address population studies related to COVID-19
- Create summary tables that address patient descriptions related to COVID-19

B Coding dependencies

Code is available at https://github.com/fishneck/tovid.

Word cloud visualization is implemented with *wordcloud* version 1.8.1 on *python* version 3.7. The function used to generate the plot is WordCloud(). Parameters adopt the default value except for *max_words*.

Visualization of the topic models is implemented with *pyLDAvis* version 2.1.2 on *python* version 3.7. Default parameters are used to create the web interfaces(add figure).

Implementation of LDA model and DTM model is with *sklearn* version 0.23.2 and *gensim* version 3.8.0 on *python* version 3.7.

C Fine-tuning

Final stop words is the union of *nltk*'s built-in English stop words and the following words = {'19', '2019', 'acute', 'analysis', 'care', 'case', 'clinical', 'coronavirus', 'cov', 'covid', 'disease', 'health', 'impact', 'infection', 'management', 'novel', 'outbreak', 'pandemic', 'patients', 'review', 'risk', 'sars', 'severe', 'study', 'treatment'}

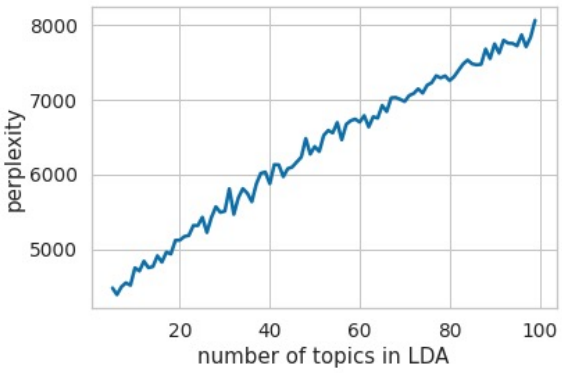

Figure 7: The perplexity score for topics ranging from 5 tp 100

Perplexity score is a negative bijection with log-likelihood, the lower the better. Perplexity might not be the best measure to evaluate topic models because it doesn't consider the context and semantic associations between words. Perplexity curve is illustrated by Figure 7 where no optimal value is reached.

Final selected number of topics is 6. The projection of topic similarity with PCA in a two dimensional space and distribution of word contribution is illustrated by Figure 9-14.

The t-representation are visualized using PCA where $t = 30, 50, 100, 200$. A larger value of t is helpful to differentiate the base topics yet computationally expensive. The choice of 100 is a balance between computation efficiency and shape of PCA projection.

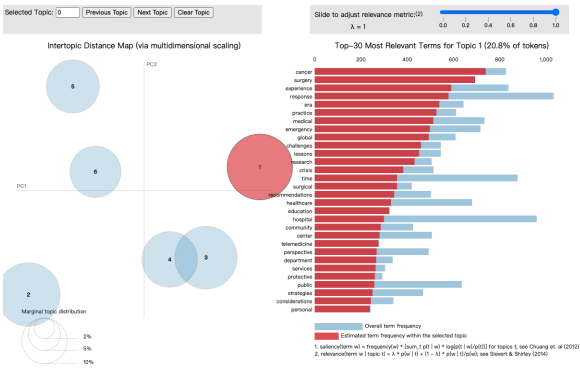

Figure 8: Topic 1

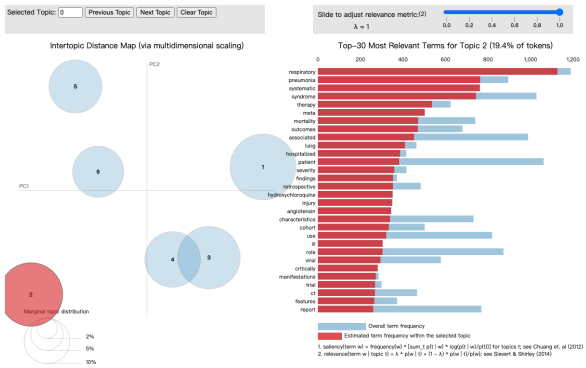

Figure 9: Topic 2

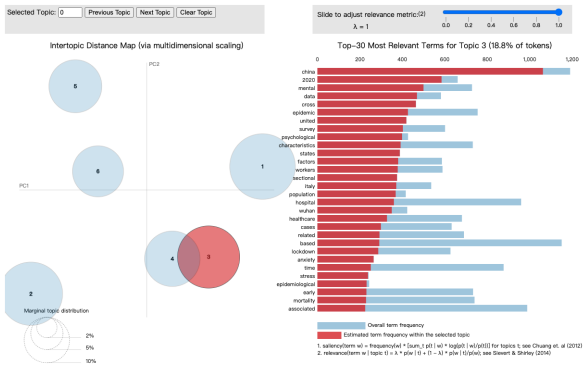

Figure 10: Topic 3

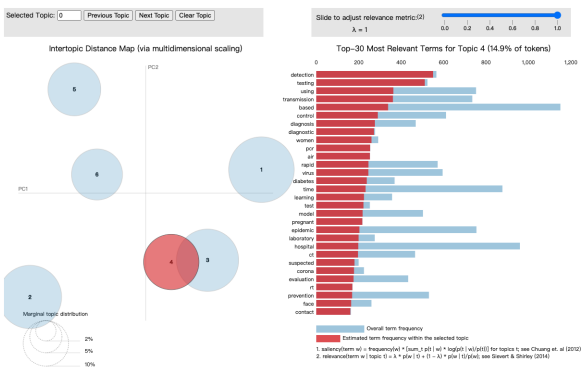

Figure 11: Topic 4

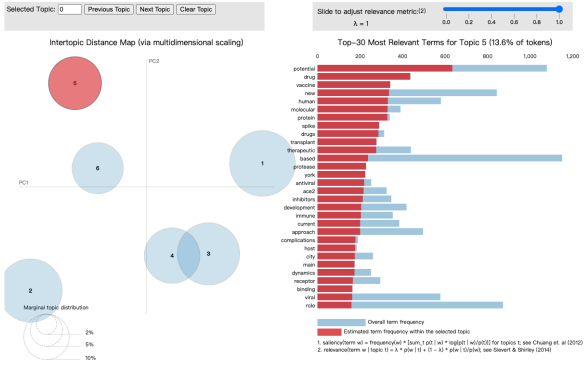

Figure 12: Topic 5

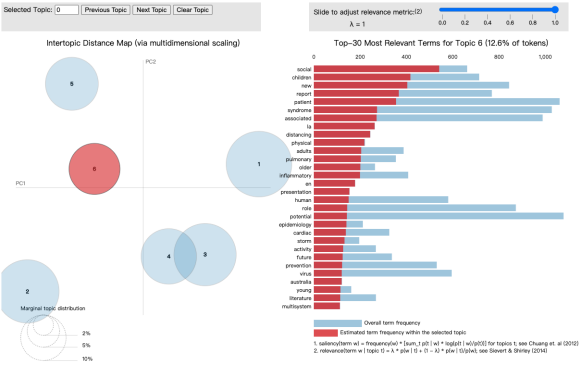

Figure 13: Topic 6

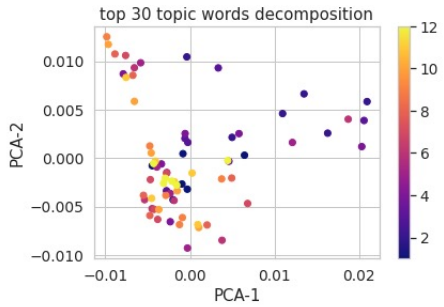

Figure 14: t-representation PCA projection with t=30

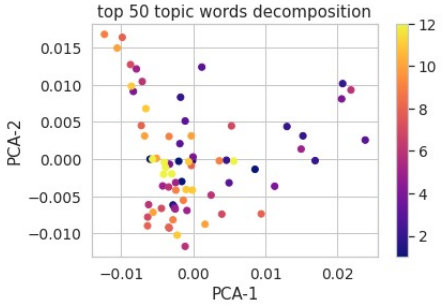

Figure 15: t-representation PCA projection with t=50

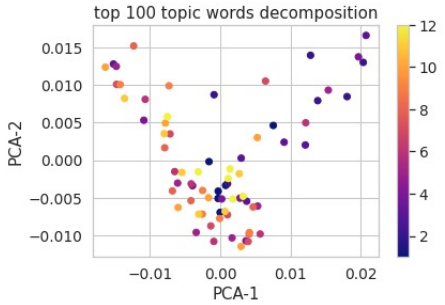

Figure 16: t-representation PCA projection with t=100

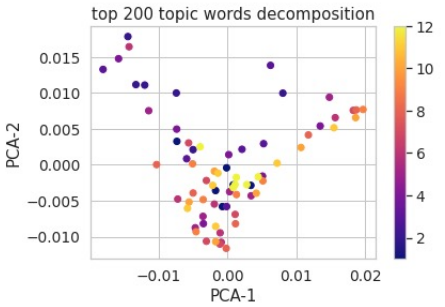

Figure 17: t-representation PCA projection with t=200