Topic Modeling of COVID-19 Papers using Latent Dirichlet Allocation (LDA)

Sawyer Jacobson

8/22/2020

Contents

Executive Summary

In 2020, COVID-19 has escalated from a seemingly negligible threat to a global pandemic that has altered nearly ever facet of our lives. Since the initial outbreak in January 2020 to present day, research around the world has been focused on this virus trying to determine information such as what symptoms are most likely, what demographic will be most affected, how to minimize the spread of the virus, and if a potential cure or vaccine can be created to name a few. This research has manifested itself into thousands of published papers and abstracts detailing their results. With this massive amount of information, machine learning techniques such *Latent Dirichlet Allocation* (LDA), a form of topic modeling, can be useful to determine how a collection of documents are related to one another through underlying similarities known as topics. The research question addressed in this paper is what are the topics and related keywords associated with 2020 COVID-19 research? We used LDA to find a predetermined 34 topics, number of topics obtained using the `ldatuning` package in R. The resulting topics can be used to filter abstracts based on which topic is most prevalent in it.

Data Source

The metadata used for this analysis was obtained from the semanticscholar.org/Cord19 (Wang et.al, 2020) website. The data source on this website is updated every day. In this paper, the data from August 8, 2020 were used. The initial data contained information on over 200,000 abstracts. Many of these observations were publications of the same paper in different journals as well as a number published in foreign languages. Additionally, not all articles in these data are based around COVID-19 so publications that did not contain “Coronavirus” or “COVID” in one of the title or abstract were removed. A large number of publications were published in 2019 or earlier, before the initial Coronavirus outbreak, and were filtered out as well. For the purpose of this analysis, the duplicate and foreign publications were removed leaving us with 28,303 English language abstracts from 2020. Further filtering based on abstract length removed another 50 abstracts leaving 28,253 for the full analysis. This dataset contained variables such as publication journal and date, authors, url, etc., but the variables used in this analysis were the unique article ID, the abstract itself, the derived number of words in the abstract, and the month published (if available).

Exploratory Data Analysis

The data source used in this paper comes from a massive collection of papers concerning COVID-19, SARS, etc., and the method used to collect the data by Wang et. al was meant to be all inclusive, therefore exploratory data analysis is necessary to determine the initial quality of these data.

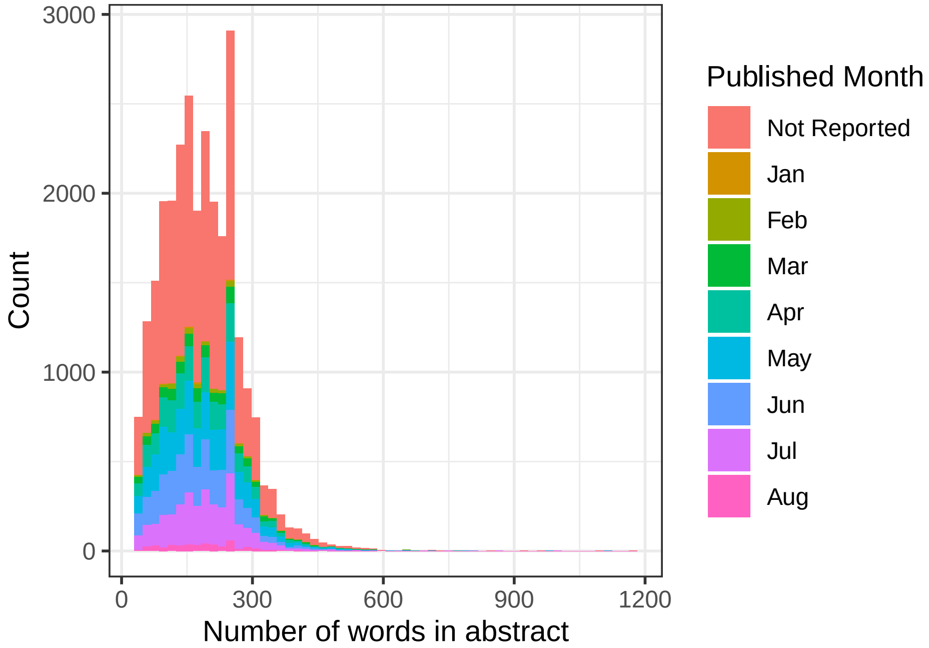
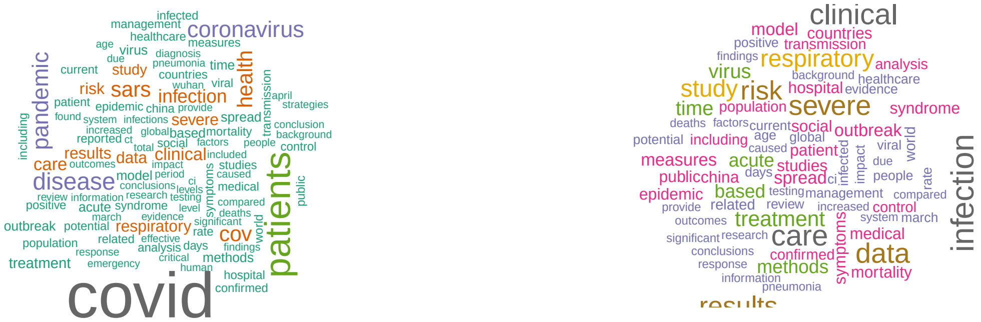


Figure 1: Histogram of number of words in each abstract colored by month published

While abstracts are usually around 100-500 words in length, we see in Figure 1 that, while few, there are abstracts of up to 1200 words in length. To keep more with standards and to remove right skew from the length of these articles, abstracts that had more than 600 words were removed. It is also apparent in this figure that there are significantly more articles missing a publication month as well as there being relatively few January published articles. The January articles will also be removed from the rest of this analysis.



(a) All words included

(b) Eight most common words removed

Figure 2: Wordclouds of most commons words in the corpus.

Text normalization is necessary for our analysis. Using the `tidytext` package, the abstracts were separated into individual words with the stopwords removed (words such as “and”, “a”, “the”, etc. that add no real meaning to the sentence). Figure 2 contains wordclouds of the most commonly occurring words left in the corpus. Larger text indicates more word occurrences. In (a), we can see that there are around 8 words that occur much more often than others such as “coronavirus”, “covid”, and “disease” that contain little meaning since they are essentially the topic we are researching. These words were removed and a wordcloud of the remaining words can be seen in (b). With those words removed, a more even distribution of word occurrences can be seen.

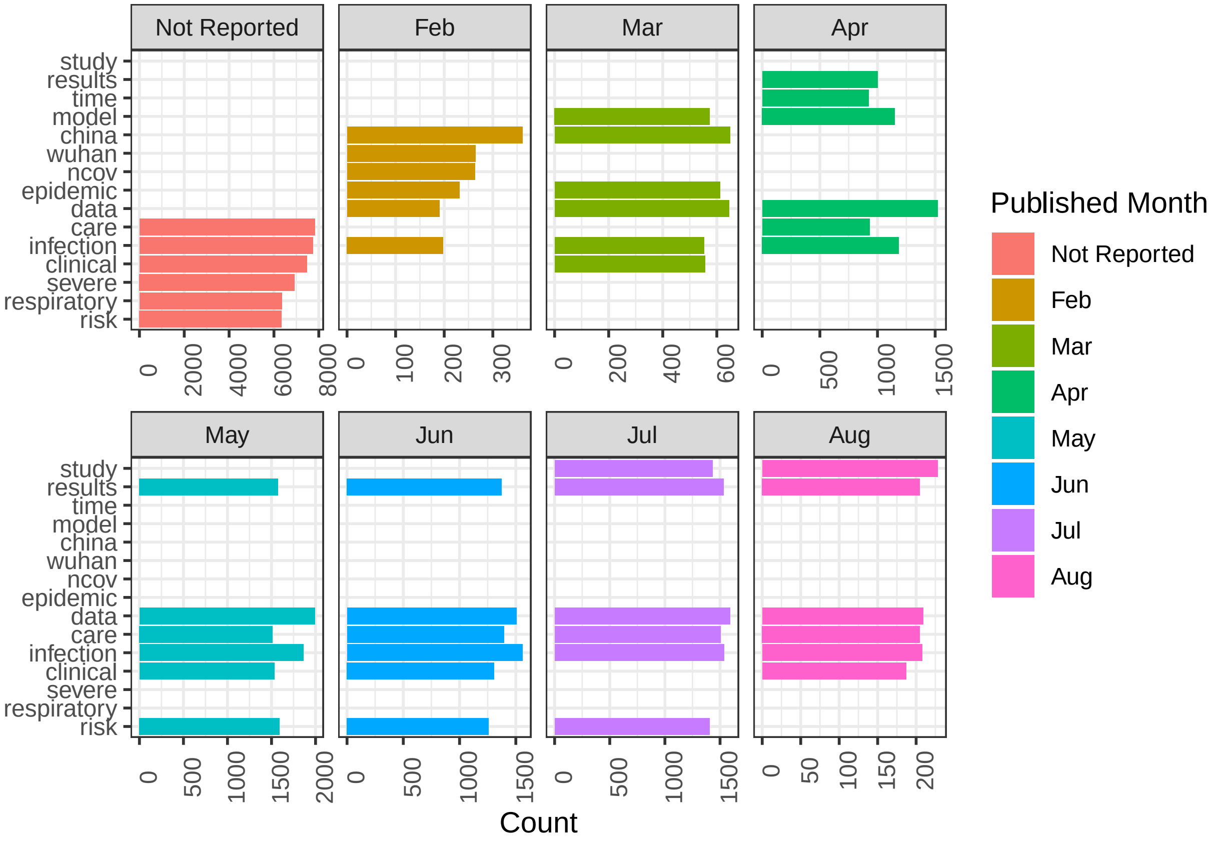


Figure 3: Histograms of 6 highest occurring words by month in the corpus.

In Figure 3, a plot of the top 6 highest occurring words per month can be seen. Only May and June have the same top 6 occurring words, July and August differ by one word, but all other months have different words giving evidence that the point of interest in COVID-19 has changed over time, and that a variety of different foci have been used

Method

The method used to analysis this large corpus of COVID-19 research is *Latent Dirichlet Allocation* (LDA), a generative probabilistic model of a text corpus. LDA is used to discover the underlying (*latent*) structures in texts with the topics themselves representing a *Dirichlet* distribution and the words being *allocated* into the different topics.The assumptions of LDA are that each document can be represented by a mixture of topics and each topic is composed of a collection of words (Blei et. al). The model also acts under the bag-of-words assumption: word order in a text does not matter, only the word frequency. If the number of topics to be fit is not determinable by industry knowledge, the `ldatuning` package can be used to find the ideal number of topics by optimizing metrics that go beyond the scope of this paper. The `quanteda` package can be used to create a document-feature matrix from a document-word count dataset which then can be used with the `stm` package to fit the LDA model with a specified number of topics. One of the downfalls of LDA is that it is not possible to compare the correlation between topics.

Results

An LDA model with 34 topics was fit to the document-feature matrix generated from a document-word count dataset of our corpus. In Figure 4, a histogram of the top 20 topics can be seen based on their prevalence in the corpus. The γ on the x-axis represents the average proportion of documents represented by that topic. Most of the γ values are low indicating that no one produced topic is prevalent in all documents. The words on the right of each bar are the top 7 highest probability words for that topic. We can see the top topic has words dealing with the spread and outbreak of the virus around the world. The next topic has words that deal with modeling the data over time in trying to understand the spread of the pandemic. One very interesting topic is Topic 13 that has words dealing with mental and physical health, stress, and the impact of peoples lives. These are issues that are talked about daily on the news and social media on how take care of oneself while quarantining and following social distance practices. Mental health is an aspect of health that should not be overlooked in these difficult times.

From the LDA model, we also receive β, the probability that a word belongs to that particular topic. In Figure 5, a selection of 6 topics can be seen along with the top 10 highest probability

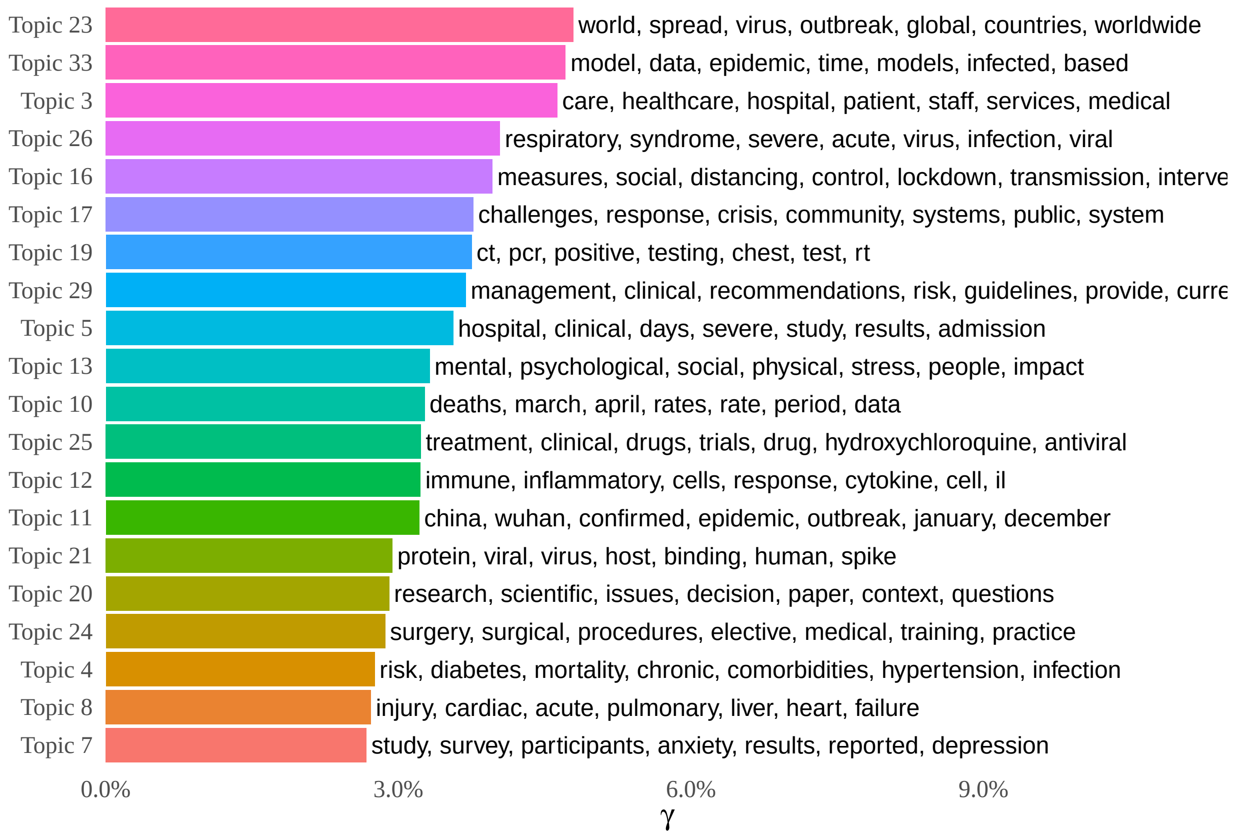


Figure 4: Top 20 topics by prevalence in the COVID-19 corpus with the top words that contribute to each topic.

(β) words belonging to that topic. Note that most of these words have a fairly low probability of belonging to any one topic. There is a probability that each word can belong to any topic. Most of these topics can be see above in Figure 4, but it is interesting to see the probabilities for each word belonging to their respective topics. For example in Topic 11, there is a fairly high probability the word China belongs to this topic as well as other words relating to the initial outbreak of COVID-19.

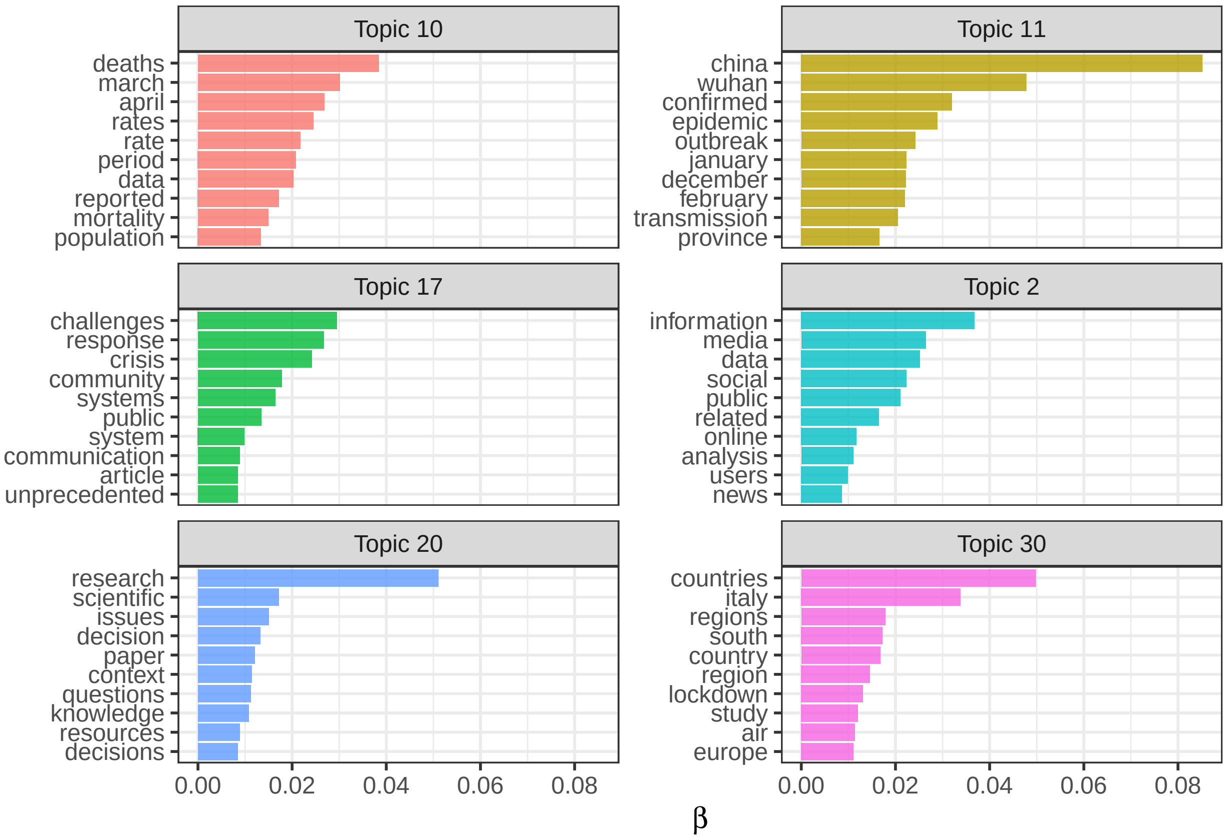


Figure 5: Highest word probabilities for each topic

Conclusion

Overall this model did a good job of creating interpretable topics from a large corpus of publication abstracts relating to COVID-19 research in 2020. The results of this model could have many applications such as information retrieval or classifying abstracts. Individual abstracts could be pulled and analyzed to determine which topics they are primarily composed of. Further work using this method would involve using the full paper for LDA analysis and determining if there is a large difference in topics derived from the the abstracts or the full publication.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, null (3/1/2003), 993–1022.

Lu Wang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wilhelm, C., Xie, B., … Kohlmeier, S. (2020). CORD-19: The Covid-19 Open Research Dataset. ArXiv, arXiv:2004.10706v2.