Tennessee TECH

Identifying Transmission and Risk Factors for COVID-19

Brittany Harbison, Logan Capes, Dustin Lee, Eric Cabarlo

Introduction

The COVID-19 pandemic has triggered major interest in research related to the virus. The Allen Institute for AI recently partnered with several major research groups to provide an open source dataset, providing researchers with a foundation for Natural Language Processing of current research applicable to the virus, with the intent of enabling new insights into the disease. With this in mind, we have utilized this data to convert the current corpus of research into a more accessible format, using various techniques to generate visualizations and summaries, working towards an end goal of identifying factors indicated by the research as relating to COVID-19. Because of this, our methodology emphasizes identification of effective methods for uncovering text features relating to a specific topic in large volume textual data.

Methodology

◆ **Data Wrangling**

- After the Cord-19 dataset was downloaded, pre-processing was performed to convert the data from a JSON format to a dataframe. Some initial data cleaning was also performed during this process by replacing null values obtained from the JSON-to-dataframe conversation with NA values.
- Additional pre-processing was performed by using regexes to filter out all papers that did not contain an explicit reference to coronavirus or COVID-19 and remove irrelevant words such as those referring to copyright, preprint, attribution, journal references, etc. During this stage, stray characters such as slashes, were also removed.
- While a dataset comprising all four license categories of documents was created, it proved too large to use for testing purposes (1GB of text data). Due to this, the smaller subset dataset of only pre-published works from the MEDRXIV and BIORXIV journals was used.

◆ **Text Summarization**

- LSA (Latent Semantic Analysis): The first summarization method attempted involved use of a library implementing LSA, a technique for identifying the higher-order structure of a text by decomposing the feature-matrix into a reduced vector space using SVD.
- Topic Modeling: A second method was also attempted using the textmineR library. Term co-occurence and document term matrices were created and a LDA (Latent Dirichlet Allocation) model was created from it. Summaries for each paper were generated using this model.

◆ **Frequently used terms**

Word frequency was measured after removing stopwords (components of language that are irrelevant to the topic or use case). A generic list of stopwords from the English dictionary was used.

◆ **Filter papers based on keywords**

Three subsets were created by filtering out papers that did not contain a reference to a keyword in a topic group. These sets included,

- An antiviral drugs keyword set
- A set composed of selected diseases, to establish risk factors
- A set composed of weather-related environmental effects.

◆ **Visualization**

Frequently used terms were gathered from the body text, from both filtered and non-filtered sets. Basic word clouds were generated from the non-filtered set. Some terms were then gathered from the summaries, basic word clouds were generated and compared with those created from the body text. Commonality and comparison clouds were created from the filtered-set only, to demonstrate changes in word use based on the topic.

Results

◆ **Summarization**

- LSA was simple to use and processed documents quickly; however, LSA was unable to process some papers. While the LSA summaries that were generated were *mostly* readable and usually provided a vague gist of the paper, too much information was missing from the summaries for more than that. Additionally, the summaries for some papers were completely incomprehensible.
- The output from topic modeling provided much better output. For most papers, the summaries from textmineR were easily readable and provided an accurate summary of the paper; however, implementation was difficult, and summary generation for the 11MB pre-published dataset took several hours.

◆ **Visualization**

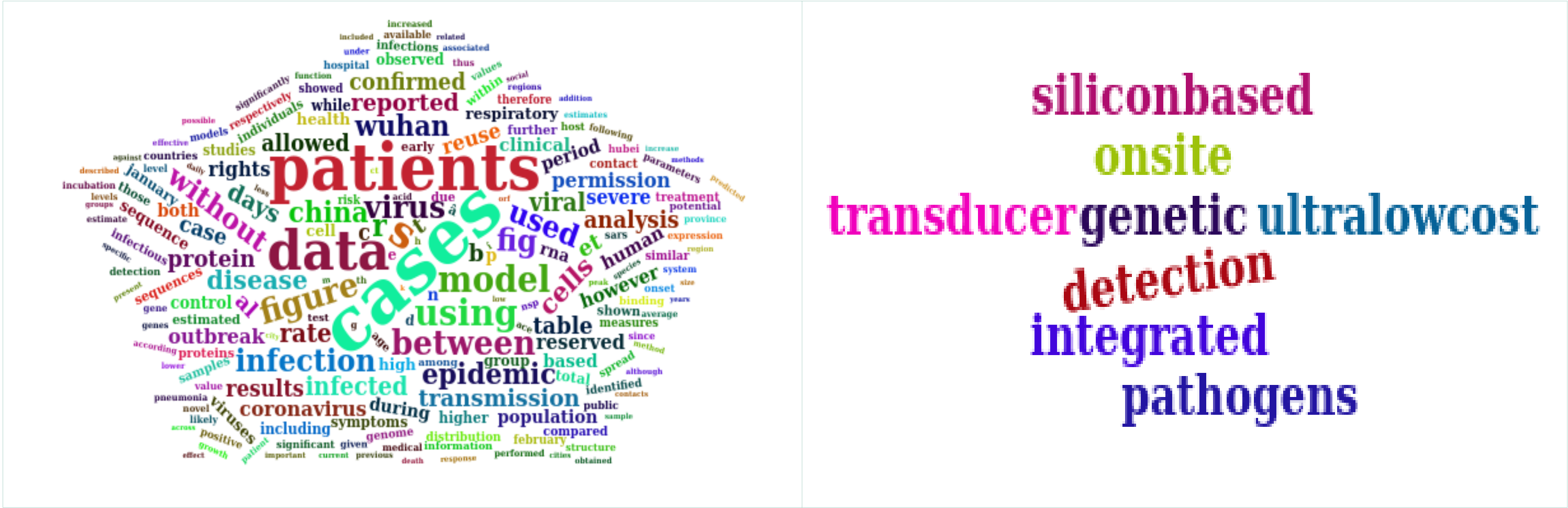

Figure 1.1. Side by side comparison of a word cloud from all body texts with a word cloud from all summaries.

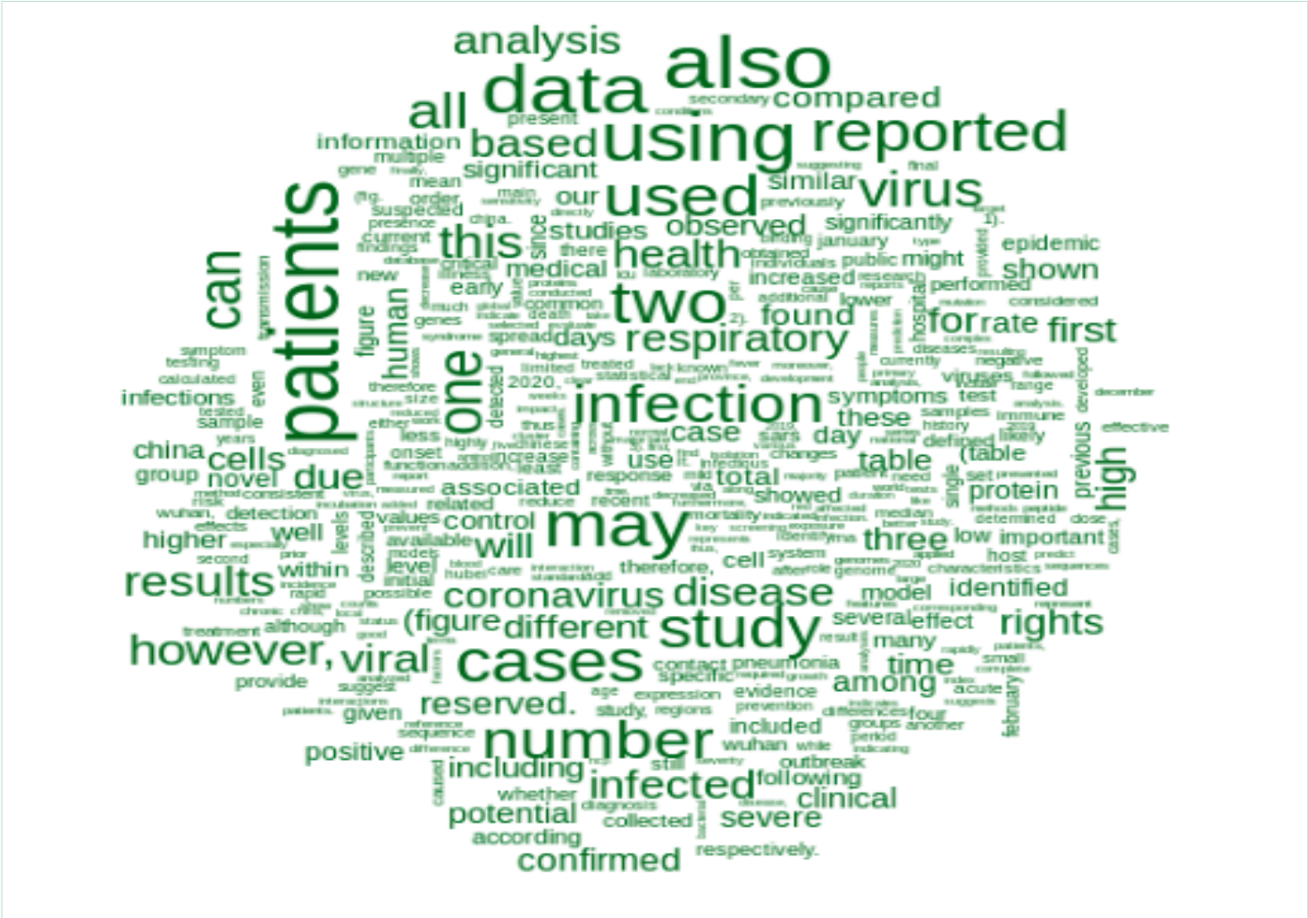

Figure 1.2. Commonality Cloud displaying words contained in all of the three subsets, (antiviral drug, risk factor, and environment subsets.)

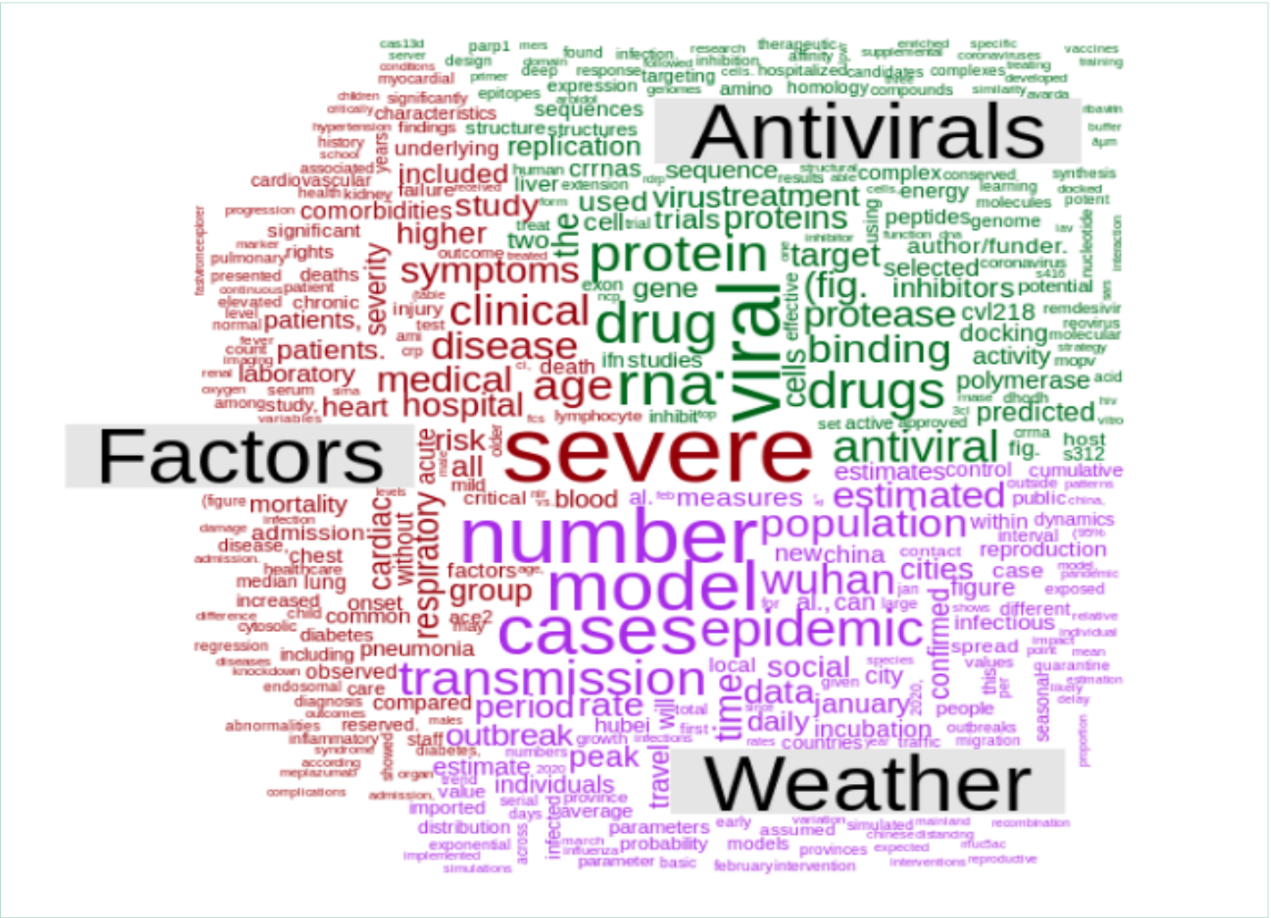

Figure 1.3. Comparison cloud showing the words distinct to each subset (same subsets as in Figure 1.2.)

References

COVID-19 Open Research Dataset (CORD-19). 2020. Version 2020-03-27. Retrieved from https://pages.semanticscholar.org/coronavirus-research. Accessed 2020-03-30. doi:10.5281/zenodo.3715505

Discussion

As can be seen from Figure 1.1, word frequency and summarization methods do not appear to complement each other, as was previously theorized. It is unclear why this is, but it is possible it is related to the smaller number and variety of words.

Also from Figure 1.1, the word cloud generated from the body texts of all documents shows a high frequency of references to patients, data, cases, epidemic, and infection. This is somewhat unsurprising considering the topic of coronavirus. In comparison with Figures 1.2 and 1.3, while the commonality cloud appears fairly similar, the comparison cloud reveals that certain topic keywords may have a hidden relationship with other words within the documents. For example, the cloud shows that for weather-related topic keywords, transmission, Wuhan, population, and January all appear in these papers but not in papers in the other two subsets.

Conclusions

TextmineR topic modeling may be better suited for generating summaries on small sets of textual data and wholly unsuited for large scale, serial text analysis. LSA may be useful for generating initial insights into large volumes of text data, though its accuracy should not be relied upon.

Though it initially seemed promising, generating a summary first and then building a word cloud off of it rather than the full data may lead to a word cloud that is *too* specific.

Subsetting documents by topic and generating a comparison cloud off those subsets may be useful in uncovering hidden traits of that topic and specific keywords target for further EDA.

Future work

As differences in word use were revealed by the comparison cloud visualization, a method utilizing word frequency may be a good approach to uncovering specifics about a topic from this dataset. Uncovering and targeting unique and rare words from the body text of each topic subset and analyzing the widening differences between the subsets may narrow scope enough to show specific factors in each topic.

It is possible that the failure of summarization to sufficiently narrow word frequencies to uncover specifics was simply due to the small amount of data. Further analysis using the complete dataset, techniques to parallelize textmineR, and/or greater computing resources may be fruitful.

Acknowledgments: Thanks to Dr. Eberle, our Advisor, for enabling and assisting this research project, and to Tennessee Tech for hosting this event and encouraging student research.