**Mani Sundar Pathuri**

**Srinivas Rao PrithviNag Kolla**

**INFO 5731 Fall 2020**

**TERM PROJECT – FINAL REPORT**

Project name**: SEMI-SUPERVISED LEARNING FOR LEGAL DATA ARGUMENTATION**

**Abstract**

Now a day’s digital data is overflowing, Information is available everywhere and in vast amounts to people that it becomes difficult for an individual to understand in a short period of time.

We took up this Legal Research project as this fall into the above category. We see that Law professionals go through Large volumes of cases to have better grasp on past cases to take up future cases.

This helps them have a whole grip on the relevant previous cases to take better decisions during court sessions.

This is time consuming, so in order to fast track the process, NLP professionals in text classification parse huge documents of cases using machines to support. This will be a boost to law professionals in Legal Domain, which is generally seen down.

*Keywords: LEGAL dataset, Natural language processing, Semi-Supervised Learning*

**Introduction**

The legal domain is an under-explored area where machines have the potential to play an important role in alleviating the arduous tasks committed by law professionals by a daily basis. Previously work had been done on European Court of Human Rights data as legal corpora to classify argumentative and non-argumentative sentences in their argumentation structure(s). They were able to receive an 80% accuracy score in their binary classification task. Some Indian researchers have applied these methods on Indian Law. They used an annotator agreement system called Inter-Annotator Agreement (IAA), to judge the quality of the annotations, along with acknowledging the subjectivity existent in the labels. After the realization that hand-crafted features do not produce high results, the paper utilizes BiLSTM models, which achieves between 80-90% accuracy, for 7-class classification. It is possible to achieve similar results for U.S. law, using a similar process.

* To train the classifiers we took Labeled data set of published court reports with 6 component classes as: Invalid, Facts, Analysis, Conclusion, Issue, Rule/Law/Holding.
* 11 Train Data Sets, 1 Test Data Set, 1 Unlabeled Data Set.
* Columns & Type:

Text (String)

Target (String)

we aim to answer the following research questions:

**RQ1: How to handle the unbalanced data?**

**RQ2: Which type of algorithm is better, semi supervised or supervised?**

**RQ3: If supervised is better or semi supervised, then which classifier will be better?**

1. **Related works**

Compared to the numerous domains ML and NLP knowledge is applied to, the legal domain is an under-explored area where machines have the potential to play an important role in alleviating the arduous tasks committed by law professionals by a daily basis.

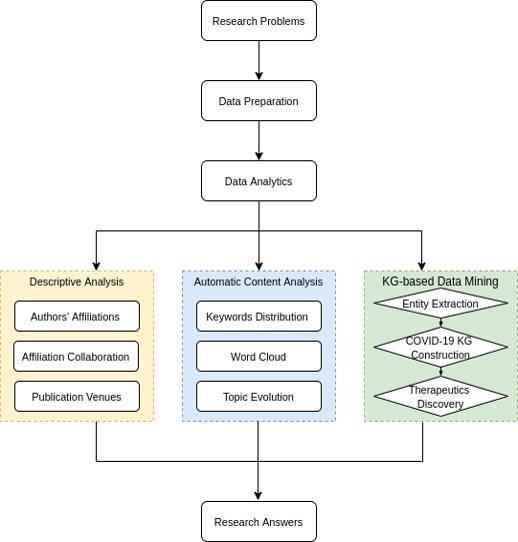
As argumentation is a crucial part of the overall project, explains what argumentation is and how to identify it in legal corpora. They use the European Court of Human Rights (ECHR) data as legal corpora to classify argumentative and non-argumentative sentences in their argumentation structure(s). They were able to receive an 80% accuracy score in their binary classification task.

Corpus annotation is a crucial step for the annotation task, as a principle in machine learning states “garbage in, garbage out”. understands to create an automated annotation system, manual annotation performed by 2 or more people is required, to ensure the quality of data is reliable for training. An important diagram in the paper states that for the highest of quality annotations, an annotator would require domain and linguistics knowledge and need to have a solid theory behind the annotation labels, to reduce ambiguity among annotators.

Potentially the most difficult phases to this research project are phases 3, 4, and 5, as they require a deeper study in the semantics and pragmatics of legal linguistics. A similar goal of implementation is thoroughly documented in, as their main focus is centered around “a ground truth for testing predictions about outcomes in new cases with new evidence; patterns for successful and unsuccessful argumentation; and guidance in retrieving, extracting, and organizing evidence for new arguments and new situations”. A creative approach that is taken for the construction of evidence-based arguments is a DeepQA architecture. This was the same architecture used to build IBM Watson for the game *Jeopardy*, except it is being used here to extract relevant arguments from past information using a relation extraction method.

In another part of the world, the same research project my team and I are researching is being conducted over in India by a group of Indian researchers. However, is applying similar principles and goals to Indian law, instead. They used an annotator agreement system called Inter-Annotator Agreement (IAA), to judge the quality of the annotations, along with acknowledging the subjectivity existent in the labels. After the realization that hand-crafted features do not produce high results, the paper utilizes BiLSTM models, which achieves between 80-90% accuracy, for 7-class classification. It is possible to achieve similar results for U.S. law, using a similar process implemented by this paper.

1. **Methodology**

To address the above research questions, we applied various methods of natural language processing shown in Figure 1 to mine the CORD19 dataset. However, our work primarily focused on three major data analysis parts: **descriptive analysis and automatic content analysis and knowlege-graph data mining (**Figure 1). For the first analysis, we built an affiliation collaboration network and presented journal distributions over the years to answer the research question 1 and 2. In the second part, we applied unsupervised learning methods to automatic extract keywords and topics from the whole dataset and on certain time sliced to figure out how keywords and topics revolute over the time of the pandemic.

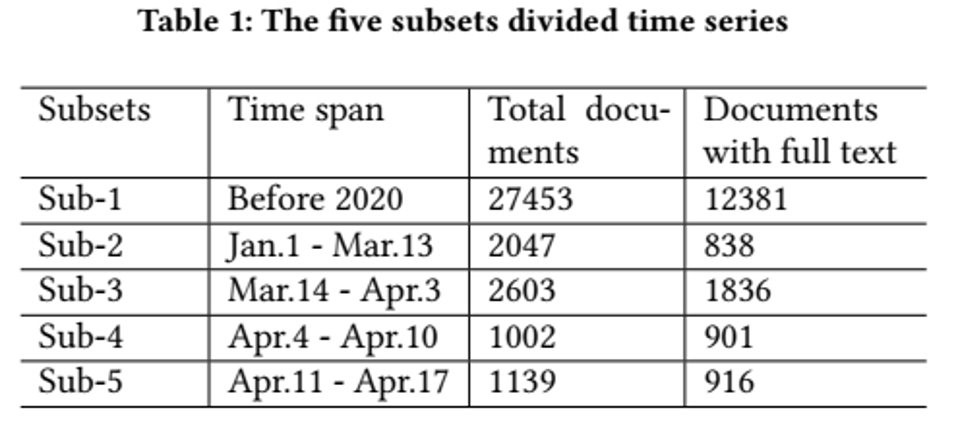
**Figure 1: Workflow**

For the last analysis, we used various method to construct a knowledge graph to discover COVID-19 therapeutics that might have been reported from the dataset to address the research question 4. Text filtering method was used to filter out COVID-19-irrelavent documents to build a knowledge-graph-based corpus. Next, we used a several models for extracting name entity and relation recognition to build the knowledge graph.

**3.1. Data preparation**

In the CORD-19 dataset, data had been reformatted into CSV files for metadata and JSON for full-text papers. Some significant metadata such as affiliations was not included in the CSV metadata files, but in JSON. Based on our data analysis goals, we extract contents from JSON full-text files and combined with metadata CSV formatted files to build our own corpus.

Additionally, we also divided the dataset into time sliced subsets for mining topic and keyword evolutions. We took 2020 as a milestone to divide the dataset into two-time spans: before and after the pandemic started. For the subset after 2020, we based on time releasing data weekly to divide it into time-sliced subsets. All overlapped documents are filtered out to keep only no duplicate among them. Table 1 shows 5 time-slice data subsets.



**Table 1: 5 time slice subsets in relation to data release time**

Data preprocessing was proved to affect performance of many text mining task such as text classification (Vijayarani, S., Ilamathi, M. J., & Nithya, M., 2015) (Uysal, A. K., & Gunal, S., 2014), sentiment analysis (Krouska, A., Troussas, C., & Virvou, M., 2016, July), especially for medical data, (Almuhaideb, S., & Menai, M. E. B. (2016). There is no single standard for data preprocessing but basically including: tokenization, lowercase, non-alphabet removal, stop word removal, stemming, lemmatization, etc. Datasets, domains, language and analysis tasks should be considered to apply the above data normalization. Keep this in mind, we lowercased, removed non-alphabet and stop words from abstracts and full texts before most data analysis tasks. However, stemming and lemmatization were only applied for topic modeling and keyword extraction.

**3.2. Descriptive Analysis**

**3.2.1. Affiliation Analysis**

In this part, we did analysis on affiliations results in the distribution of institutions and their collaborations in the research network. For affiliation distribution, we counted their frequencies and then ranked their frequency on the whole result, which then was selected the top list for visualization.

The analysis on affiliation collaboration could be useful to find which organizations are the center of this network. We counted all collaborations (no duplicates) if there were two or more different affiliations within the same papers. However, the network dimensions were very large, and it would be impossible to display all dimensions and nodes of the network. Therefore, we filtered to only remain the institutions which the number of collaborations with others were more than 90 collaborations in our network. Also, we used color and node size to identify each of institution in the network. The node sizes represent the number of collaborations while colors illustrate countries and region of those institutions. Finally, the edges were used to present the relations among institutions while the node sizes illustrated the number of collaborations. It should be noted that we increased these values to 10 times to get a more good-looking graph. NetworkX [[1]](#footnote-1)– a package for creating complex networks in Python – was used to plot the graph.

**3.2.2. Publication Analysis**

Publication is a crucial attribute in this dataset. It enables us to find journal distributions and see how these publications changes over the years prior to the pandemic if applicable. Journals which papers published was simply counted and then ranked to find their distributions. We selected the top 20 for presenting in visualization. Also, we counted the number of papers over publication years and plotted them by seaborn – a library built on top of Matplotlib in Python for visualization.

**3.3. Automatic content analysis**

**3.3.1. Keyword Distribution**

Automatic keyword extraction is a preferred approach in automatic text summarization. Many methods have been applied and evaluated. YAKE – a completely unsupervised learning model was proved to outperform compared to the previously best models (Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A., 2020). YAKE achieved a better performance compared to RAKE-NLTK (Rose et al., 2010), TF-IDF, and many other state-of-the-art graph-based methods such as TextRank , TopicRank, etc. (Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A., 2020). YAKE does not need any training, and it is independent from domain, language, dataset size. For the above reasons, we decided to apply YAKE to automatically extract keywords from our dataset.

Feature extraction of YAKE considered five features: (1) **casing** that capitalized words their acronyms would be counted by maximum occurrence of them, (2) **word position** which considered the position of the sentence the word occurs, (3) **word frequency** that scores the words’ frequencies are over mean, (4) **word –related to context** which examine the word occurring within the window of size of the candidate words, (5) **Word DifSentence** that was to see frequency of candidate words occurring in different sentences. Importantly, YAKE handles how to generate the candidate keyword list. Sliding window size could be set to create a sequence of candidate keywords; we set **windowSize = 2.** Levenshtein distance – presenting similarity between two terms – was measured; we set a threshold by default. We set max\_ngram\_size = 2 to extract bigram keywords rather than n-gram or trigram.(Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018, March).

Again, we aim to extract keywords on the whole dataset, and on 5 separate time sliced subsets to find how keyword revaluated before and after the pandemic began. We extract 2000 keywords from the whole dataset, and 2000 for each time sliced dataset.

The output was then plotted with word cloud, a package for generating word cloud in Python for keywords from the whole dataset. We also use NEviewer tool (L. Ma et al., 2010), a dynamic tool for displaying evolution of objects over time series for visualize results from 5-time sliced subsets.

**3.3.2. Topic Modeling**

Topic modeling was an unsupervised learning model for generating topic from many documents. This method enables us to rapidly get a comprehensive insight into the content of documents in our huge dataset. We aimed to figure out dominant topics from the whole dataset and to find significant variation of topics over dataset releases.

Topic modeling based on the assumption that every document is a mixture of k latent topics and every topic describes the probability of each word occurring in topics over the documents (Bisgin, H., Liu, Z., Fang, H. et al, 2011) *(* Alghamdi, R., & Alfalqi, K., 2015) . Among various unsupervised topic models, we investigated Latent Semantic Analysis (LSA) and Latent Dirichlet allocation (LDA) because of their popularity in semantic analysis. LSA is basically based on the document representative term matrix: term frequency (TF) or term frequency-inverse document frequency (TFIDF). However, TF or TF-IDF creates sparse and redundant matrices. The key idea of LSA model is truncated Singular Value Decomposition (SVD) – the method for dimensionality reduction on matrices generated by TF or TF-IDF to generate topics. Therefore, LSA topic modeling hardly can control the number of topics. LDA model, on the other hand, gives each of d documents k latent topics, and each of k topics is described by a multinomial distribution over w words (Alghamdi, R., & Alfalqi, K. (2015). Therefore, LDA model allows to decide k topics and control the dense of topics for given documents (called α) and the distribution of words in topics (called β). The higher α is set, the more mixed-up topics are. The higher β results in more words in topics.

For the above distinguished features of LDA, we decided to apply this method in the topic modeling and topic evolution. We used Gensim LDA package in Python to achieve this task [[2]](#footnote-2). It should be noted that selecting k topics has a big impact on results. To find proper k topics, we first ran a baseline model and then observed the topic coherence score to adjust hyperparameters. Topic coherence is a novel algorithm to measure the hanging and fitting of words in a word set (Röder, M., Both, A., & Hinneburg, A., 2015). Gensim LDA has a function to easily compute this score. We set up our baseline model on 10 documents with the following parameters: *num\_topics = 10, chunksize = 100, passes = 10, alpha = auto, eta = auto*, we got around 0.6078 on coherence score. We changed some hyperparameters to achieve a better coherence score. Finally, we set up these parameters to get a slightly better coherence score: *num\_topics = 30, chunksize = 100, passes = 60, alpha = auto, eta = auto.*

**3.3.3. Knowledge Graph-Based Data Mining**

As mentioned previously, we aimed to dive deep into CORD-19 to figure out which treatment therapeutics conducted for treating COVID-19 or closely related to COVID-19. We applied the Knowledge Graph-Based approach to mine the dataset and provide answer for this research question. Several methods were used to construct this knowledge graph.

**Text Filtering**

We assumed that the dataset is likely to contain papers which were not much related to COVID-19. Therefore, we filtered out the papers which were not about COVID-19 before extracting therapeutic-related entities to build the knowledge graph. Either COVID-19 is the major research problems or just mentioned in literature, filtering by using relative names of this disease would be a proper method. We used information retrieval method for this task.

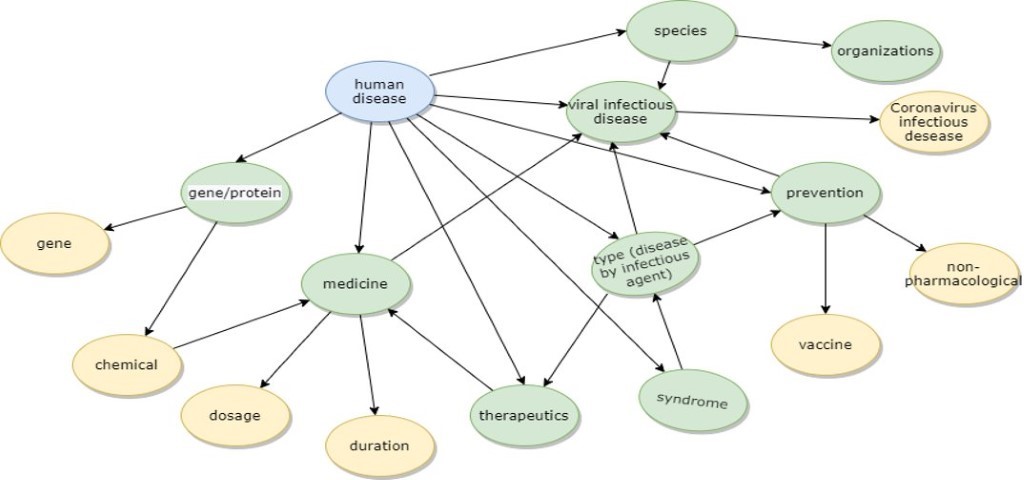
Rank-bm25, a package in Python, was used for this mining task. It was implemented various search engine ranking BM-25 algorithms such as Okapi BM25, BM25L, BM25+ (Trotman, A., Puurula, A., & Burgess, B. (2014, November). First, a query must be designed; the query essentially has an impact on output. Our query consists of possible names and name formats of COVID-19 such as *COVID-19, (COVID-19), COVID19 Coronavirus, Corona, SARS-CoV-2, 2019-nCoV (2019-nCoV) (SARS-CoV-2*) as defined by WHO and CDC. Then we tested this query and the model on one of first released subset, i.e. bioRxiv/medRxiv with 803 documents. The output gave a rank of the documents regarding relatedness score. We first assumed that the documents with 0 in this score were most likely to be irrelevant to the query. We classified them into two clusters: zero- scored (358 documents), and non-zero-scored clusters (445 documents). Then, we manually checked every document in zero- scored clusters and checked randomly one-third of another cluster. We concluded that our query and the Rank-bm25 model gained very high accuracy score (up to 95%).

After testing, we applied the same query and model to filter all dataset releases. The task results **31,572 documents** kept out of 47,000 full-text documents.

The clusters with the documents which were not relevant to COVID-19 will be discarded while documents in the rest cluster would be split into single sentences to build a corpus for constructing the knowledge graph. Finally, we got **6,193,625 sentences** to build a corpus for knowledge graph construction.

**Entities and Relation Extraction**

We pre-defined ontologies and their relations regarding therapeutics, based on MeSH (Medical Subject Headings) of USA medicine library **[[3]](#footnote-3)** and from other resources.

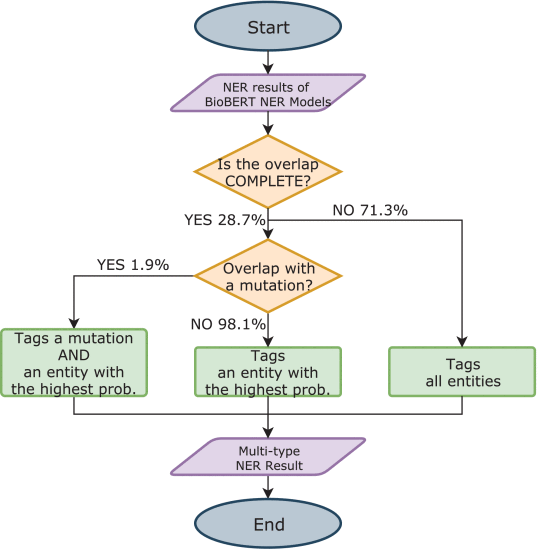


**Figure 2: Pre-defined ontologies and relations**

Figure 2 shows that human disease is a root of our knowledge graph, below is parent class with ontologies such as: gene/protein, medicine, therapeutics, viral infectious disease, species, prevention, types (disease by infectious agent, syndrome and organization. Yellow nodes are child class of the above class, including gene, chemical, dosage, duration, vaccine, non-pharmacological and coronavirus infectious disease.

To extract entities of the above predefined ontologies, we combined multiple Named Entity Recognition (NER) models. NER models were implemented in sentence-level to be able to address the relations among entities within each sentence.

**Neural biomedical named entity recognition and multi-type normalization (BERN) tool**

BERN tool was used to extract entities regarding genes/proteins, diseases, drugs/chemicals, and species. The raw data are first fed in BioBert NER model (Lee et al., 2019) to tag genes/proteins, diseases, drugs/chemicals, and species. BioBert NER model had been evaluated to obtain the highest F1-score in the above entity recognition. However, overlapping among tagged entities was really a challenge for any NER models. BERN can fix this gap by performing decision rules to deal with overlapped entities. Figure 2 presents how BERN deals with overlapped entities (Kim et al., 2019).

**Figure 3: Decision tree to eliminate overlapped entity types (Kim et al., 2019)**

We used RESTful API Web service offered by BERN to accomplish this task. Originally BERN outputted JSON formatted results, including entity ID, ontologies, and indexes of entities in documents. We then computed indexes to extract only ontologies and entities and added them into our knowledge graph-based corpus.

**MED7 model**

MED7 is a self-supervised model for NER in clinical NLP, specified in electronic health records. It was trained to recognize 7 entities in health records: drug names, route, frequency, dosage, strength, form, duration (Kormilitzin et al., 2020). These entities are useful for the COVID-19 therapeutic knowledge graph we supposed to build with rich information about successful therapeutics of this disease.

MED7 was trained in 2 million free-text patients’ records from Medical Information Mart for Intensive Care-III corpora and then fine-tuned by self-learning and annotators. The model achieved 0.957 on micro F1-score (Kormilitzin et al., 2020). The outputs were turned into our corpus for constructing knowledge graph.

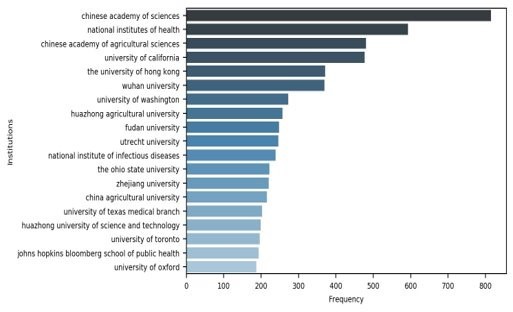
**Knowledge Graph Construction**

Due to the time constrain, we have not extracted enough entities and relation to build the above suppose knowledge graph. We would leave this task for future work.

1. **Results and analysis**

**Author’s affiliations**

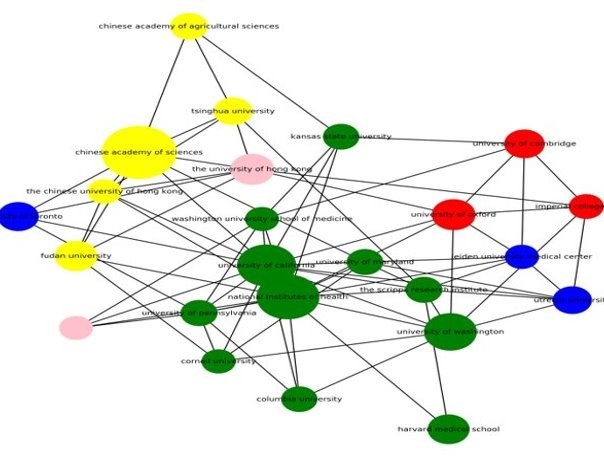
We found that there were at least 1000 affiliations contribute in more than 10 papers. From the top 20 institutions involved in COVID-19 research (Figure 4), Chinese academy of science and national institutes of health are the most active in this area (over 800 and 600 papers repectively). We also found that from the top 100 institutions, there were 37 from United States,27 from China, and others from the United Kingdom, Canada, Japan, Netherlands.



**Figure 4: Top 20 institutions most involved in COVID-19 research**

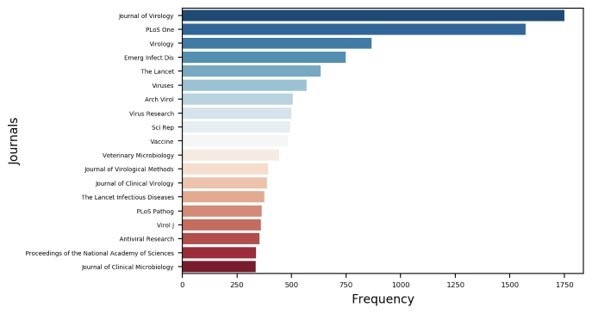
**Affiliation collaboration**

The results on institutions showed that Chinese academy of sciences, national institutes of health, university of California are three institutions with the most collaborations, 459, 324 and 283 collaborations respectively (Figure 5). Additionally, the result indicated that international collaborations could be comparative with domestic collaborations.



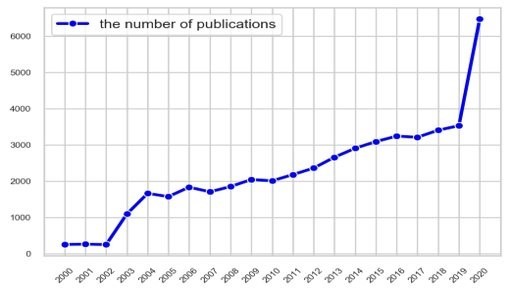
**Figure 5: Affiliation collaboration network. Filtered to display the ones with over 90 collaboration. Colors indicates country/region: yellow-China, green-USA, red-UK,**   
**blue-Canada, pink-Hong Kong**

**Publications**

Our result showed that 6320 venues participating in COVID-19 research from the dataset. Most of them were in medical and biology field, especially virus; the minority came from other fields such as bioinformatics, information science... Journal of Virology had the most papers published from the dataset (1750 papers), followed by PLoS One and Virology (1572 and 868 papers respectively) (Figure 6)

***Figure 6: Top 20 venues contributing in COVID-19 research***

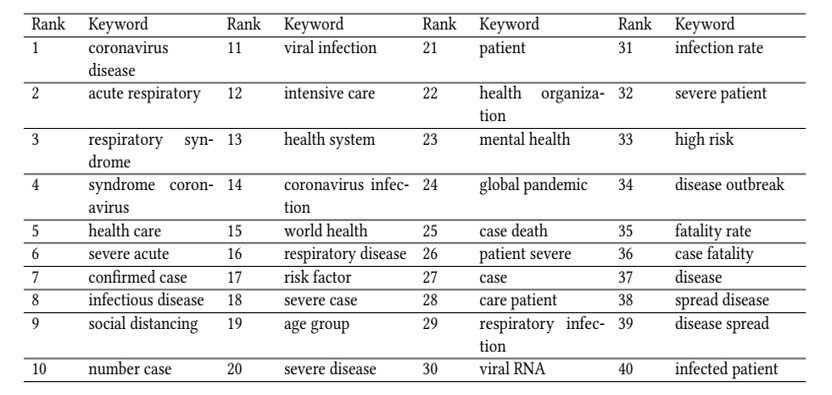
The analysis on publications over the years showed that the number of papers published before 2002 were very limited. The figures after 2000 were presented in Figure 7. From the chart, the number of papers started to increase significantly after 2002. Until 2019, the figure shot up incredibly. From 2019 to 2020, this number increased by the total of increasing over the last 30 years, over 3000 papers.



**Figure 7: The number of papers over years**

***Keyword analysis***

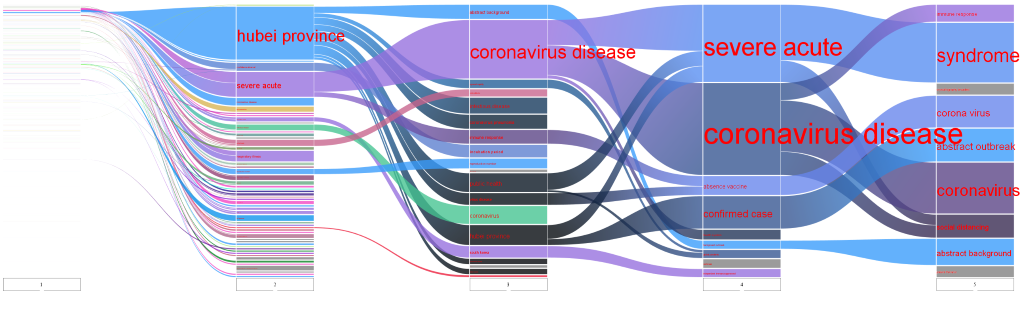
40 most important keywords were extracted from the whole dataset and ranked by their importance (table 2). From the result, coronavirus disease, acute respiratory, respiratory syndrome, health care, confirmed case... were the most significant keywords. Besides, social distancing, age group... were contributing to the list. This indicated how the pandemic has been directing of research.



**Table 2: Top 40 most important keywords extracted from the whole dataset**



**Figure 8: Word cloud of top 40 keywords extracted from the whole dataset**

The analysis of keyword evolution over the 5 time slice subsets also gives meaningful information. From Figure 9 and 10, there is no clear keyword community in the first slice subset; however, the latter ones show more converged. Hubei province is the largest community in the second time slice aligned with COVID-19 outbreak in China at that time. In the next slice, coronavirus disease is the largest one which originated from a small community: severe acute previously. This community then keeps developing as two most dominant ones in next time slices**.** 



**Figure 9 &10: Keyword evolution over 5 and 4 time slice subsets, respectively**

**Topic modeling**

From 30 topics extracted from the whole dataset, we found that Spike glycoprotein of SARS-CoV-2, Antiviral drug, Antimicrobial use, COVID-19 transmission, information and data in the pandemic and COVID-19 testing are 6 most dominant topics (Table 3).

|  |  |  |
| --- | --- | --- |
| **Topic ID** | **Word list** | **Interpreted topic** |
| Topic 1 | SARS-CoV-2, protein, ACE2, bind, cell, structure, human, receptor, spike, covid19, express, antibody | Spike glycoprotein of SARS-CoV-2 |
| Topic 2 | drug, antivirus, active, inhibitor, compound, inhibit, proteas, effect,treatment,approval, develop, target | Antiviral drug |
| Topic 3 | animal, york, microbiology, crystal, mNG, colon, microorganism, feed, parasite, autoimmune, antimicrobial, discuss | Antimicrobial use |
| Topic 4 | case, infect, transmission, contact, symptom, period, asymptomatic, isolate, day, incubation, onset, among | COVID-19 transmission |
| Topic 5 | disease, health, public, social, popular, risk, pandemic, system, information, data, community, global | Data and information system in pandemic |
| Topic 6 | test. detect, use, sample, positive, SARSCoV-2, method, result, assay, sensitive, diagnostic, acid | COVID-19 testing |

**Table 4: Top 6 popular topics extracted from the whole dataset and interpreted**

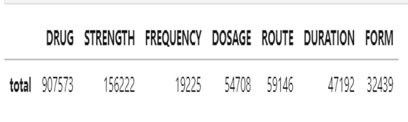
Mining topics from the two-time slices before 2020 and after 2020 also bring about meaningful findings. Before 2020, topics are almost overlapped with one another, especially the top 10 most dominant topics while after 2020, distances among topics are significantly clear. The table 5 presents 30 topics from each time slice interpreted. Among them, virus infects immune response in cell, studies on patient groups, Modeling disease out-break using data, Analysis of gene from different region, bat: virus host are 5 most popular topics from the time slice before 2020. Meanwhile, Model estimating the number of COVID 19 infection rate, clinical studies of severe cases associated with COVID-19 disease in hospital, Report confirms COVID-19 outbreak in Wuhan, China, Public healthcare risk due to COVID-19 pandemic, Severe respiratory syndromes of coronavirus disease are the top 5 in the data subset after 2020.

|  |  |
| --- | --- |
| **Before-2020 subset** | **After-2020 subset** |
| Public health response to the emerged global outbreak | A political narrative to eradicate the border travel |
| Replicated virus genome | sample test detects sars-cov2 shows positive results |
| Recent study developed a system by using biological approach | Viral drug antivir inhibits virus replication |
| Virus infect children respiratory tract | The transit of poultry industry within state during the outbreak |
| Virus infect cell protein via membrane fusion | Improved approaches to research provide efficient system development |
| Binding to protein domain | Sarscov2 virus binds to host cell protein upon interaction |
| Unknown | Mask protects face by filtering droplets |
| Gastroenteritis Noroviruses | Unknown |
| Unknown | Delevoping vaccine by using epitope on antigen |
| Gene expression activates transcription pathway | Model estimating the number of COVID 19 infection rate |
| virus infects immune reponse in cell | PEDV and PDCoV- members of the coronavirus family |
| human immunodeficiency virus (HIV-1) | Influenza virus infect respiratory tract in children |
| Swine virus causes diarrhea | The analysis of genome sequence isolated from virus in Spain and Italy |
| studying on patient groups | The injection of oseltamivir |
| Acute respiratory syndromes of coronavirus | trials of clinical studies using drug as treatment for sarcov2 |
| Effect of active drug antivirus | Specific virus strain in animals cause infection |
| Hospital studies associated risks in patients | Unknown |
| Antibody response to antigen | Information analysis network uses model method to predict disease |
| wash hands and masks : Hygiene behaviors | chromatin transcription effects expression of genes |
| Modeling disease out-break using data | clinical studies of severe cases associated with covid 19 disease in hospital |
| H1N1 and H5N1: Influenza virus infect human | Infected lung cells express ace2 |
| Using sampling and assay methods | Pregnant women transmiss covid2019 to newborn |
| Online database electronic materials contain authors and users | scan image of chest and lung show pneumonia |
| symptoms caused by rotavirus | Report confirms covid 19 outbreak in wuhan, china |
| unknown | Public healthcare risk due to covid 19 pandemic |
| Virus infects host | Prevention and treatment of covid19 infected patients |
| Analysis of gene from different region | Unknown |
| Bat: virus host? | Unknown |
| Patient with severe lung diseases | Onset symptoms after contacting covid 19 patients |
| Invasive cancer | Severe respiratory syndroms of coronavirus disease |

**Table 5: 30 topics from two time slice data subsets interpreted**

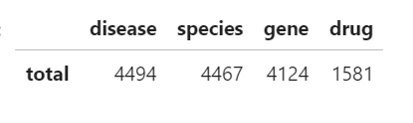
**Knowledge graph**

As mentioned previously, our knowledge graph and the last research question have not been done yet. However, we have got some certain output preparing for building the knowledge graph. So far we have extracted all possible entities recognized by Med7 such as: drug names, route, frequency, dosage, strength, form, duration (Table 6).



**Table 6: The number of entities extracted by using Med7**

For entity extraction by Bern, we have just finished a part of them. The table displays the number of entities we got so far with Bern.



**Table 7: The number of entities extracted by using Bern**

1. **Conclusion**

The outbreak of the coronavirus (COVID-19) has created a global health crisis and had changed the way we perceive our world. In this paper, we conducted a comprehensive analysis of the COVID-19 Open Research Dataset to understand the existing research efforts that may help to combat this pandemic. By conducting a descriptive analysis on the distribution of the leading organizations and venues, we found significant contributions of affiliations and their global collaboration among them. Our analysis also indicated distributions of venues and research fields in the COVID-19 research. The majority comes from medical, biology and related fields. By conducting an automatic content analysis on the titles and abstracts of the publications, including keywords/word cloud analysis and topic modeling, we clarified how COVID-19 research has been covering aspects of this pandemic and reversely how it was directed by the pandemic. Due to the time constraints, we were not able to fully address COVID-19 therapeutics and building a knowledge graph as we were supposed to do. However, we were able to pre-identify the ontologies, relations, and extracted a dramatic number of entities. Thus, it will remain to be a future work. The results of our research using COVID-19 publications show that the global research map is in the right direction. Moreover, it also shows that the research fields are heavily influenced by the pandemic as it changes. For the future, we would like to dive deeper into the data set, hoping to contribute our effort into fighting the battle against COVID-19.

**Our project GitHub link:** [**https://github.com/HuyenNguyenHelen/CORD-19\_Group6**](https://github.com/HuyenNguyenHelen/CORD-19_Group6)

**References**

Aguado-Cort'es, C., & Castaño, V.M. (2020). Translational Knowledge Map of COVID-19. ArXiv, abs/2003.10434.

Aguado-Cort'es, C., & Castaño, V.M. (2020). Translational Knowledge Map of COVID-19. ArXiv, abs/2003.10434.

Chen, B., Shi, M., Ni, X., Ruan, L., Jiang, H., Yao, H., Wang, M., Song, Z., Zhou, Q., & Ge, T. (2020). Visual Data Analysis and Simulation Prediction for COVID-19.

Chen, B., Shi, M., Ni, X., Ruan, L., Jiang, H., Yao, H., Wang, M., Song, Z., Zhou, Q., & Ge, T. (2020). Visual Data Analysis and Simulation Prediction for COVID-19.

Chen, M., Qu, J., Xu, Y., & Chen, J. (2018). Smart and Connected Health: What Can We Learn from Funded Projects?, Data and Information Management, 2(3), 141-152. doi: https://doi.org/10.2478/dim-2018-0015

Gysi, D.M., Valle, '.D., Zitnik, M., Ameli, A., Gan, X., Varol, O., Sanchez, H.N., Baron, R.M., Ghiassian, D., Loscalzo, J., & Barabási, A. (2020). Network Medicine Framework for Identifying Drug Repurposing Opportunities for COVID-19.

Gysi, D.M., Valle, '.D., Zitnik, M., Ameli, A., Gan, X., Varol, O., Sanchez, H.N., Baron, R.M., Ghiassian, D., Loscalzo, J., & Barabási, A. (2020). Network Medicine Framework for Identifying Drug Repurposing Opportunities for COVID-19.

Koubaa, A. (2020). Understanding the COVID19 Outbreak: A Comparative Data Analytics and Study. ArXiv, abs/2003.14150.

Koubaa, A. (2020). Understanding the COVID19 Outbreak: A Comparative Data Analytics and Study. ArXiv, abs/2003.14150.

Li, L., Zhang, Q., Wang, X., Zhang, J.J., Wang, T., Gao, T., Duan, W., Tsoi, K.K., & Wang, F. (2020). Characterizing the Propagation of Situational Information in Social Media During COVID-19 Epidemic: A Case Study on Weibo. IEEE Transactions on Computational Social Systems, 7, 556-562.

Li, L., Zhang, Q., Wang, X., Zhang, J.J., Wang, T., Gao, T., Duan, W., Tsoi, K.K., & Wang, F. (2020). Characterizing the Propagation of Situational Information in Social Media During COVID-19 Epidemic: A Case Study on Weibo. IEEE Transactions on Computational Social Systems, 7, 556-562.

Lopez, C.E., Vasu, M., & Gallemore, C. (2020). Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset. ArXiv, abs/2003.10359.

Lopez, C.E., Vasu, M., & Gallemore, C. (2020). Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset. ArXiv, abs/2003.10359.

Ma, L. et al., "NetViewer: A Universal Visualization Tool for Wireless Sensor Networks," 2010 IEEE Global Telecommunications Conference GLOBECOM 2010, Miami, FL, 2010, pp. 1-5, doi: 10.1109/GLOCOM.2010.5683876.

Nguyen, T. (2020). Artificial Intelligence in the Battle against Coronavirus (COVID-19): A Survey and Future Research Directions. 10.13140/RG.2.2.36491.23846.

Nguyen, T. (2020). Artificial Intelligence in the Battle against Coronavirus (COVID-19): A Survey and Future Research Directions. 10.13140/RG.2.2.36491.23846.

Vattay, G. (2020). Predicting the ultimate outcome of the COVID-19 outbreak in Italy.

Vattay, G. (2020). Predicting the ultimate outcome of the COVID-19 outbreak in Italy.

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks, 5(1), 7-16.

Ye, Y., Hou, S., Fan, Y., Qian, Y., Zhang, Y., Sun, S., Peng, Q., & Laparo, K. (2020). α-Satellite: An AI-driven System and Benchmark Datasets for Hierarchical Community-level Risk Assessment to Help Combat COVID-19. ArXiv, abs/2003.12232.

Ye, Y., Hou, S., Fan, Y., Qian, Y., Zhang, Y., Sun, S., Peng, Q., & Laparo, K. (2020). α-Satellite: An AI-driven System and Benchmark Datasets for Hierarchical Community-level Risk Assessment to Help Combat COVID-19. ArXiv, abs/2003.12232.

|  |
| --- |
| WHO. Naming the coronavirus disease (COVID-19) and the virus that causes it. Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it |

1. <https://networkx.github.io/> [↑](#footnote-ref-1)
2. <https://radimrehurek.com/gensim/auto_examples/tutorials/run_lda.html> [↑](#footnote-ref-2)
3. 6 <https://meshb.nlm.nih.gov/search> [↑](#footnote-ref-3)