January 2021

Text mining of biomedical literature: discovering new knowledge

Saikat Goswami
Department of Mass Education Extension and Library Services, Government of West Bengal- 734001, India, goswami0408saikat@gmail.com

Sourav Mazumder
University of North Bengal, smazumderlis91@gmail.com

Sumana Chakrabarty
University of Calcutta, jayee2212@gmail.com

Text mining of biomedical literature: discovering new knowledge

Saikat Goswami[1], Sourav Mazumder[2], Sumana Chakrabarty[3]

[1]Assistant District Library Officer, Department of Mass Education Extension and Library Services, Government of West Bengal- 734001, India
[2]Department of Library and Information Science, University of North Bengal, West Bengal- 734014, India
[3]Department of Library and Information Science, University of Calcutta, 700073, India

Abstract:
Biomedical literature is increasing day by day. The present scenario shows that the volume of literature regarding *"*coronavirus*"* has expanded at a high rate. In this study, text mining technique has been employed to discover something new from the published literature. The main objectives of this study are to show the growth of literature (Jan-Jun, 2020), extract document section, identify latent topics, find the most frequent word, represent the bag of words, and the hierarchical clustering. We have collected 16500 documents from PubMed. This study finds most number of documents (11499) belong to May and June. We explore: *"*betacoronavirus*"* as the leading document section (3837); *"*covid*"* (29890) as the most frequent word in the abstracts; and positive-negative weights of topics. Further, we measure the term frequency (TF) of a document title in the bag of words model. Then we compute a hierarchical clustering of document titles. It reveals that the lowest distance the selected cluster (C133) is 0.30. We also have made a discussion over future prospects and mentioned that this paper can be useful to researchers and library professionals for knowledge management.

Keywords: Data extraction, Text preprocessing, Latent semantic analysis, Bag of words, Hierarchical clustering.

1. Introduction

Text mining is a method which analyzes extensive data and represents it in an exceedingly new form. It deals with raw data extraction, data mining, and knowledge revelation. The leading function of text mining is to depict the structured information from unstructured large datasets. An enormous collection of data can be found in web pages, news articles, social media (e.g. Facebook, Twitter, and YouTube etc.), bibliographic and statistical databases (Zhai & Massung, 2017). To determine an active information retrieval system, text mining is one among the most effective practices that we use to work in several research purposes. Text mining ascertains the text analysis activities through Natural Language Processing (NLP) which is also used in automatic text classification (Larsson et al., 2017). Text mining can be applied in many sectors such as educational (Ferreira-Mello et al., 2019), corporate, medicines, security, research (Hearst, 2003) as an analytical tool to deal messy and large data. Text mining also generates structured data that emphatically consolidates into databases for modeling and organizing. Furthermore, Supervised Learning (SL) and Unsupervised Learning (UL) are the two algorithms used in text mining methods. SL algorithms are for predicting a target variable. On the other hand, the UL algorithms use a set of predictors to reveal hidden structures in the data (Hassani et al., 2020). Text mining can be applied in text clustering, concept extraction, sentiment analysis, summarization, and Knowledge Discovery in Database (KDD) of digital libraries (Dhiman, 2011; Talabis et al., 2015).

The application of text mining in biomedical literature is unprecedented and can be exemplified by PubTator, an online application of the National Library of Medicine (NLM, USA) for automatic literature curation. It offers users to access annotations through a RESTful application program interface for interoperability (NCBI, 2020). As we can see the rise of scholarly publication of medical and biosciences in several online databases viz. PubMed, DOAJ, and EMBASE, etc. (University of Illinois at Chicago, 2020) in the recent past. A regular search provides retrieval of information by querying author, title, keyword, date, and subject term, etc. but it would not be adequate to uncover the thematic structure of the large collection. Hence, text mining technique helps to explore

the hidden information by using NLP-based software packages (Tao et al., 2020). Also, the use of text mining enables us to extract abstract and meaningful information from a big amount of scientific literature. The purpose of this study is to form and visualize meaningful data from large set of data which includes text of titles and abstracts of biomedical literature. The purpose of this study is to form and visualize meaningful information from a large set of data which includes the text of titles and abstracts of biomedical literature.

This paper is designed as follows: Related works on text mining in the biomedical literature are briefly described in Section 2. Section 3 presents the objectives of this study. It shows the different applications made in this study. Section 4 discusses the methodology of this study. It shows the retrieval of raw data and other methods. Section 5 presents a comparative study on the growth of literature regarding *"*coronavirus*"* in PubMed. It also presents a mathematical equation for predicting future trends in publication. Section 6 describes the development of the workflow. It includes the whole applications of text mining for this study from preprocessing text to clustering the text data that are presented in Section 7. The rest of this paper deals with limitations and the perspectives for future research.

2. Related work

Many scholarly studies on text mining are done throughout the past decades. This section combines some related studies. Tan et al. (2000) presented a general framework for text mining consisting of text refining and knowledge distillation. They surveyed the state-of-the-art text mining applications on the text refining and knowledge distillation functions. They highlighted the challenges of text mining and therefore the opportunities it offers. Natrajan (2005) gave a definitional analysis of text mining and acknowledged some valuable reasons for text mining. The reasons are adequate retrieval of useful information from global databases; identification of authors, journals, and organizations; generation of technical taxonomies; a roadmap for tracking any research impact. The author stated some limitations of text mining. In conclusion, it was manifested that text mining methods would be needed to enable greater acceptance

within the biomedical community. To grasp human diseases and treatment better, Ye et al. (2016) asserted that text mining could help to get more knowledge from large dataset. They designed and developed a text mining framework called SparkText on a big Data environment. That framework was made by Apache Spark data streaming, machine learning (ML) methods, and the Cassandra NoSQL database. They extracted data from PubMed and built prediction models. On text mining and visualization techniques, Yang et al.(2008) identified many tools within the marketplace for uncovering the hidden data. They presented a comparative overview of some key text mining and visualization tools for chemical, biological, and patent information. They also discussed the integration of sophisticated tools on text mining, full-text searching, and data visualization with more sophisticated software packages. Zhou et al.(2010) surveyed traditional Chinese medical information sources for data mining. The authors expressed *"*Traditional Chinese Medicines*"* (TCM) provides distinct methodologies for treating human diseases. Observing the development of knowledge discovery approaches in TCM, they tried to contribute their paper with an introduction to TCM and modern biomedicine. They showed innovative and advanced text mining techniques. Lastly, they discussed research issues and prospects of TCM text mining. Keeping eyes on the massive growth of biomedical literature, Shatkay and Craven(2012) introduced significant ideas in biomedical text mining in their book *"*Mining the Biomedical Literature*"*. They covered up such areas: text-analysis methods, NLP, information extraction, information retrieval system, and text mining systems. They described the applications of recognizing the entities in text, their linkages with other entities, further prediction, and discovery. Gong(2018) expressed the huge growth of biomedical literature and the importance of text mining. The author explored adequate *"*static biomedical information recognition*"* and *"*dynamic biomedical information extraction*"* using text mining techniques like NLP and machine learning. Simon et al.(2019) researched the classification of medical literature using *"*BioReader*"* (Biomedical Research Article Distiller). It enables users to classify articles and abstracts based on two different categories (positive and negative text mining training corpora). They retrieved the abstracts from PubMed. Next, they preprocessed and classified the retrieved text and showed how *"*BioReader*"* worked for mining the database.

3. Objectives of the study

In the previous section, one of the significant outcomes reveals that most of the authors found the rapid increase of biomedical literature. They also narrated how text mining became a definite mechanism to take on the large, unstructured, and unsupervised data. From the biomedical perspective, text mining finds applications in many diverse areas namely drug discovery, predictive toxicology, protein interaction, competitive intelligence, identification of new product, and many more (Natarajan, 2005). So, this paper tries to show:

- The current trends of publications relating to the *"*coronavirus*"*;
- Data retrieval and construction of the text preprocessing;
- Section of each document;
- Topic Modeling of document titles using latent semantic indexing;
- Concordance of the text(titles);
- Top frequent words appeared in abstracts;
- Statistics of the unique titles and abstracts;
- Bag of words in titles; and
- Hierarchical clustering of titles.

4. **Methodology**

There are many proprietary and open-source tools available for text mining. For this paper, Orange data mining software was used to accomplish the tasks like fetching, preprocessing, extraction, and visualizing data. Orange is an open-source machine learning and interactive data visualization tool(Demšar et al., 2013). Biomedical literature is exclusive due to its exponentially increasing volume and interdisciplinary nature(Xie et al., 2017) and it can be seen in the PubMed. We selected the PubMed database as the data source. PubMed updates tons of literature (new and revised) every day (PubMed Help, 2020). We chose *"coronavirus"* as the search-query for information retrieval. First, we aimed to fetch data from 2000 to 2019 and from January to June 2020. Second, we retrieved only the data of January-June (6 months) based on our desirable text features (title and abstract). Section 5 shows a comparative study of

two different periods. The rest of this paper gives an overview of various text mining techniques that we applied to get meaningful information.

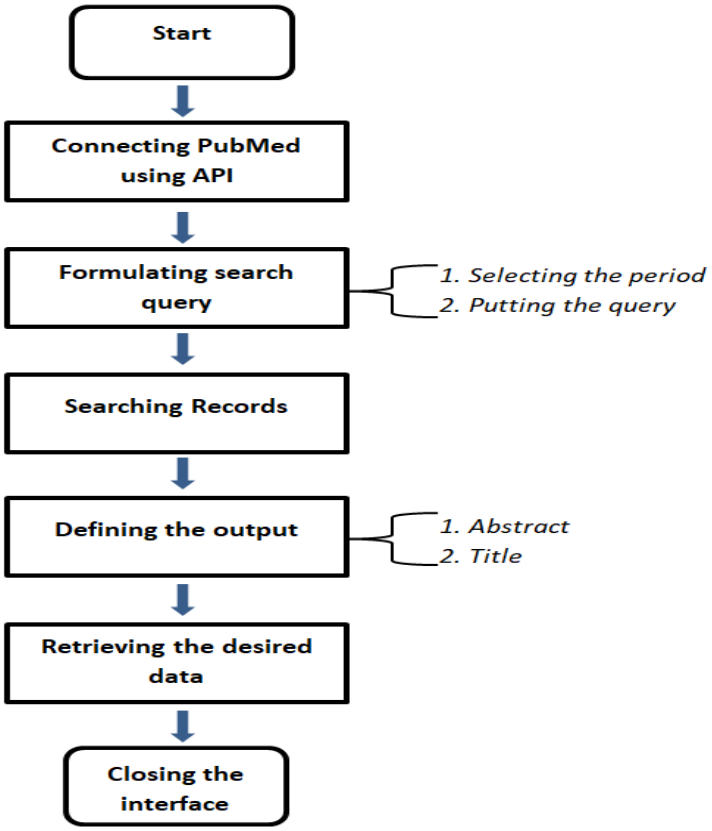

Fig.1: Simple flowchart of data fetching from PubMed

Data fetching is a common and primary task for the retrieval of data from any server or database in the digital era. Most of the online databases require API (Application Program Interface) keys for syntax-based (Lioma & Ounis, 2008; Nai-Lung Tsao et al., 2009) information retrieval operations. PubMed is functionalized by the Entrez Programming Utilities (E-utilities) that are a set of nine server-side programs for e-search. These programs are structured with the Entrez system which receives requests or queries made by users for getting data from NCBI databases**(**Sayers, 2017**).** For this study, a valid Email ID (pubmedcentral@ncbi.nlm.nih.gov) as an API key of PubMed was applied to retrieve the data. Fig.1 indicates the flowchart of the data fetching process. We formulated a regular search based on the query "coronavirus" and the period. After finding all the records available in the PubMed, we selected *"*abstract*"* and

*"*title*"* of documents as the desirable data (includes the word "coronavirus") for this study. Since all the setting was done, we retrieved 16500 records from all retrievable records that were initially prompted after searching records. This was the pre-task that we performed before starting the final work.

5. Comparison of published (available) literature between 2000-2019 and January-June, 2020: based on the query *"Coronavirus"*

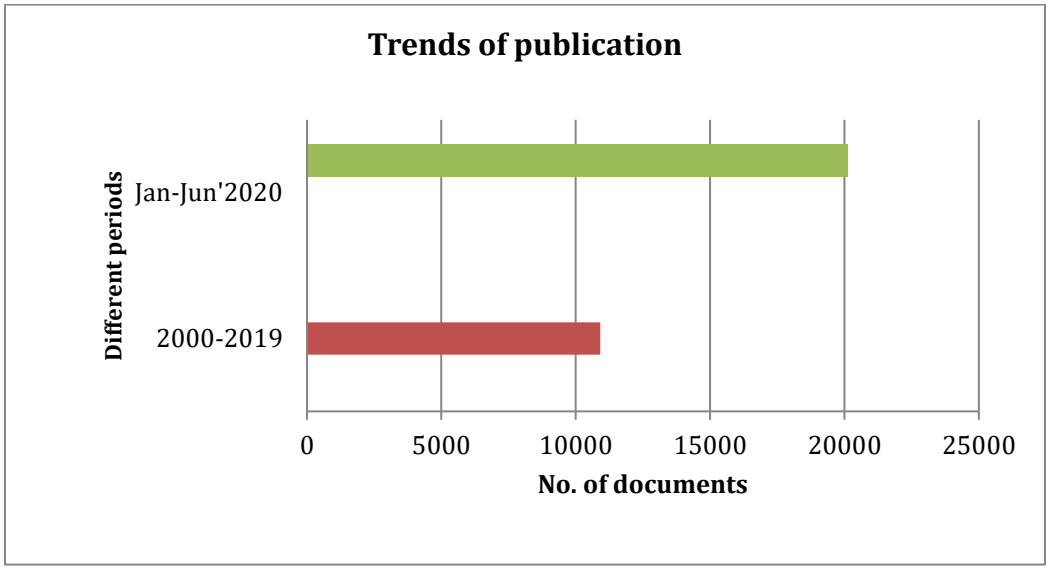

Fig.2: Comparison of two different periods (As the data was retrieved on June 30, 2020 and the data is based on the online availability)

A pre-survey was conducted to measure the growth of literature on coronavirus in the PubMed Database. Fig.2 shows a comparative study of two different periods. The first period implies the time between 2000 and 2019. A total of 10912 documents were found during the period. On the other hand, in the second period from January 2020 to June 2020, a total of 20132 documents were found. The comparison indicates how the growth of literature has been increased since January 2020. It also shows that the last six months' publications are significantly doubling the published literature of 2000-2019. The rate of growth is 0.84(84%).The following equation (Eq.1) displays the trends in the publications of literature related to *"coronavirus"* starting from the months of January 2020.

Lagrange Interpolation Polynomial (LIP) is used to find out a polynomial equation that matches with the trends in the research outputs related to *"coronavirus"*.

Table-1: Number of documents: January-June, 2020

Months	January	February	March	April	May	June
No of documents	1247	659	1835	4892	6102	5397

To find the quadratic equation, we have taken the bin size that comprises the data of two consecutive months in succession. Also, we have taken on 1st January the number of publications to be zero.

Table-2: Data of two consecutive months in succession

Months	No of Days	Documents
Jan-Feb	59(x_0)	1906(f_0)
Mar-April	120(x_1)	8633(f_1)
May-June	181(x_2)	20,132(f_2)

The second order LIP equation that passes through all the three data points is Eq.1

$$f(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}\, f_0 + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}\, f_1 + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}\, f_2$$

After putting the values, we got the polynomial equation:

$$f(x) = 0.642x^2 - 4.501x - 60.566$$

This equation will help us in predicting publications in future also. To acquire a more accurate polynomial, we have to increase the order of the polynomial (Balagurusamy, 1999). Fig.3 shows the trends in published literature during the months from January to June, 2020.

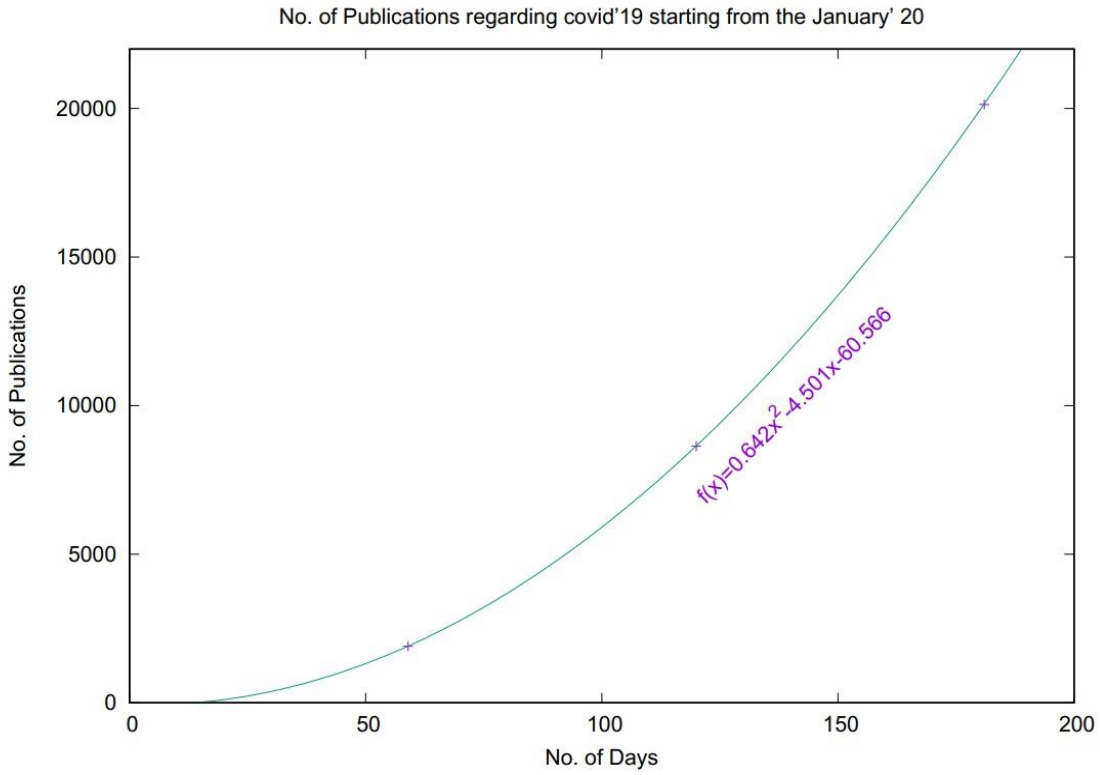

Fig 3: Publication of Jan-June'2020

6. Development of the workflow

The workflow was deployed in the Orange. Each part of the workflow gives interactive information regarding the mining of the documents. Fig.4 visualizes the components involved in developing the entire workflow for mining the documents.

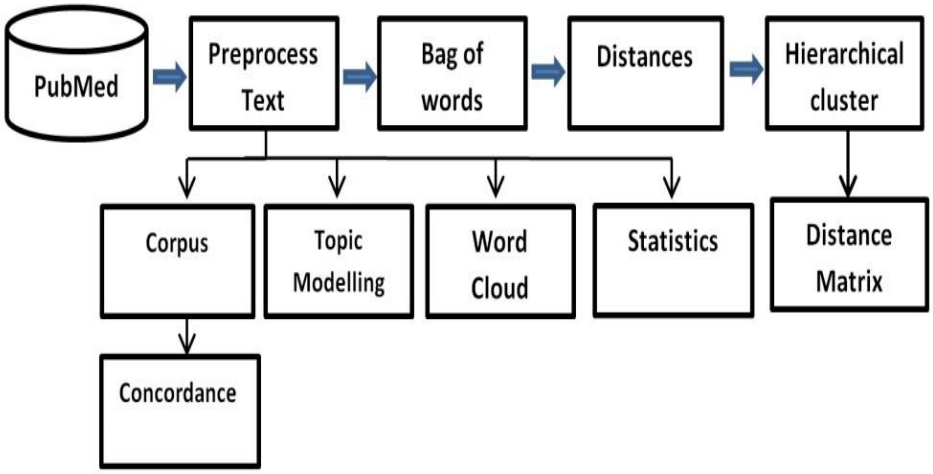

Fig.4: Workflow diagram of the present study

Fig.1 already showed how the data was fetched from the PubMed. The above diagram shows the plenary activities of this present study. Here, preprocess text, bag of words, distances, and hierarchical cluster are the core components. The preprocess text supported to represent: the corpus viewer for documents' section extraction and concordance; topic modeling; word cloud and text statistics. Hierarchical clustering is one of the powerful techniques to analyze the connection between two or more than two clusters in the text. It also helps to detect the distances between two clusters. So, it can be said that text mining is more than just a data extraction process.

6.1 Text preprocessing

Text Preprocessing or preprocess of the text is the preliminary task in text mining. It involves the transformation of raw data into an understandable format. Raw data is not sufficient to identify any object properly in the whole text. It always creates errors during data analysis and standardizing research output. The preprocessing of text helps to split the raw data into smaller tokens (words or units) and understand the data. In this study, we followed 4 different methods to preprocess the retrieved text.

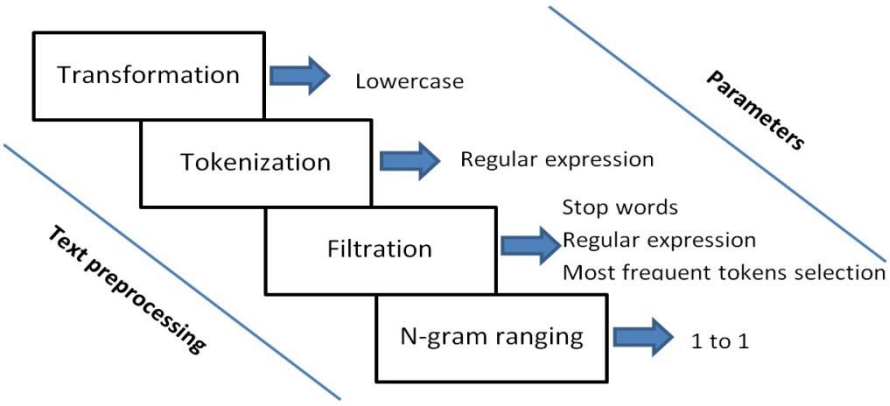

Fig.5: Construction of text preprocessing

The application of the four methods (Transformation, Tokenization, Filtration, and N-gram ranging) converted 16500 documents' text into 123528 tokens. First, the transformation method kept all the text in lowercase. No other inputs like removal of accents, removal of URL, and parsing HTML(Read the Docs, 2020) were executed. Second, the tokenization method divided the entire text into smaller elements (words, sentences, phrases, parts of speech, and n-grams, etc.). Then **"**regular expression*"* (Regex) technique (Thompson, 1968) was employed to justify and remove the punctuations. The expression pattern was *"*\w+*"* which omitted the punctuations and matched the exact word within the text. Third, "stop words" input removed articles, prepositions, and conjunctions, etc. within the filtration method. On recall values of text categorization, it is necessary to omit the stop words for the improvement of text data (Silva & Ribeiro, 2003). In addition, a text (*.txt) file containing a list of numerals and words *("19", "2019", "april", and "three",* etc.) was loaded in the software to filter more. Again a regex input **"**\.|,|:|;|!|\?|\(|\)|\||\+|'|"|‘|’|“|”|'|\’|…|\-|–|—|\$|&|*|>|<|\/|\[|\]" was induced for further filtration. The last input in the filtration method was the selection of the most frequent tokens (1000 tokens). Fourth, the *n*-gram range (1 to 1) specified the all unigrams (*n*=1) of the text. Section 7.1 and 7.6.1 reveals more about the application of *n*-gram.

7. Results

7.1 Document section extraction

The section or the class of each document is an important feature to analyze the context of the document. In the corpus, all the 16500 documents was filterable (RegExp) by section, title, abstract, and date. We extracted all the sections for a statistical representation.

Table 3: Top tweleve document sections (number of documents that appeared at least 100(>=) times appeared under document sections)

Sl. No	Document Section	No. of documents

1	betacoronavirus	3837
2	covid-19	3160
3	*anonymous*	*2405*
4	coronavirus	830
5	adult	671
6	aged	390
7	adolescent	292
8	acute	220
9	animals	214
10	antiviral	132
11	antibodies	109
12	ace2	100

In this paper, we listed out those document sections that consolidate a minimum of 100 documents. We found that most of the documents belong to the 'betacoronavirus' section. Betacoronavirus is one of four genres (Alpha, Beta, Gamma, and Delta) of coronaviruses ('Betacoronavirus', 2020). A total of 3837 documents are under this section followed by *"*covid-19*"*(3160), *"*coronavirus*"* (830), *"*adult*"* (671), *"*aged*"* (390), and rest in the list. The remaining 4140 documents belong to different sections. We also recognized that 2405 documents belonged to *"*anonymous*"* sections (underlined in the Table 3). These were showed with question mark (?) symbols in corpus.

7.2 Latent Semantic Analysis (LSA)

Latent Semantic Analysis or Latent Semantic Indexing is a method for analyzing relationships between a set of documents and the terms related to the documents ('Latent Semantic Analysis', 2020). It comes under *"*Topic Modeling*"* which is an important text mining technique. It discovers the co-occurring keywords and reduces

large textual data. Also, it helps to identify latent topics in a particular record or document. Further analysis is discussed below:

Topic	Topic keywords
1	19, covid, pandemic, patients, coronavirus, disease, 2019, 2, care, health
2	2, sars, cov, coronavirus, infection, disease, covid, 19, 2019, patients
3	coronavirus, disease, 2019, cov, sars, 2, covid, 19, novel, patients
4	pandemic, patients, health, 19, care, covid, clinical, infection, severe, disease
5	patients, pandemic, clinical, coronavirus, 19, covid, study, cancer, management, health
6	acute, respiratory, syndrome, severe, disease, coronavirus, 2019, infection, cov, sars
7	health, care, mental, public, pandemic, outbreak, disease, workers, patients, treatment
8	infection, clinical, disease, acute, analysis, review, novel, study, respiratory, syndrome
9	infection, review, analysis, clinical, disease, systematic, care, meta, coronavirus, novel
10	review, care, coronavirus, disease, novel, analysis, systematic, clinical, meta, acute

Fig.6: Topic modeling (10 topics in the text)

Fig.6 shows the keywords associated with each document. LSA method was executed to find out the positive and negative topic weights. Both weights reveal two different meanings. A positive weight (green colored word) signifies that the word is more representative regarding a topic, while a negative weight (red-colored word) is less representative. We can see both positive and negative weights occurred in the above figure (Fig.6). The group of words is the keywords of documents. It constructed a topic. We can find multiple contexts and themes of a document. For instance, a string of keywords (in green color) "*19", "covid", "pandemic", "coronavirus", "patients", "disease", "2019", "2", "care",* and "*health"* in the first row(1) represents a topic *'Coronavirus'*. These keywords not only describe a single topic but also describe more than one topic. The topics can be *"Health care"* or *"COVID-19"*. In contrast, the second row shows both positive and negative weights. It uncovers *"covid"* and *"19"* as unrepresentative weights that occurred less in the document. Other techniques involved in Topic Modeling are Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP)(Mulunda et al., 2018). Though LSA is easy to presume and perform. Furthermore, the topics can be measured by marginal probability (Albright, 2020).

7.3 Concordance

Concordance is one of the common methods in text mining. It is a very useful aspect of the information retrieval process. Luhn(1960) introduced the Keyword-In-Context (KWIC) index. It is relevant to concordances. KWIC is a machine-based activity. It helps to retrieve all documents related to a given query and joins them together. We examined how many documents are associated with a query COVID. The tool we used for text mining only shows exact matches to the query.

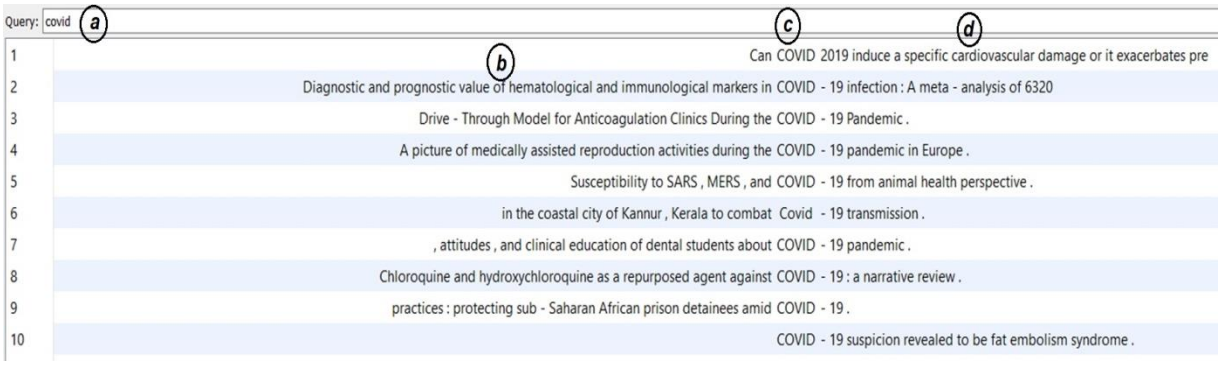

Fig 7: The above snapshot consists of: a) query, b) the preceding context, c) the queried word, and d) the following context

We got three types of information in the corpus. The information includes the total number of tokens, unique tokens, and matching documents. Out of 16500 documents, the queried word retrieved a total of 11756 matching documents. It has 283274 tokens and 18247 unique tokens. Fig.7 shows 10 examples of concordances in the titles (documents). COVID is associated with different prospects of the text and it enables us to know the frequencies of it.

7.4 Word Cloud

Word cloud is a core technique of information visualization. It is applied in many different context of text mining (Heimerl et al., 2014). It is very effective for visualizing the highly frequent words of a collection of documents and the average bag of words count. It displays the most essential words (tokens) of a given text. Also, it assists to

understand the relevancy of text data. Consequently, a word cloud summarizes text and explores various topics that are available in the whole dataset. Summarizing text depicts a statistical significance of the weights of the words. We applied an interactive mechanism to generate a word cloud in the software. It could have been carried out without preprocessing the text. To get more adequate results, text preprocessing was required. Also, stop words and unnecessary numerals were removed by adding a text file (contains numerals like 19, 2019, and 14, etc.) to the filtering preprocessor of the software (see Section 6.1). The word cloud is based on the 1000 most frequent tokens of retrieved abstracts.

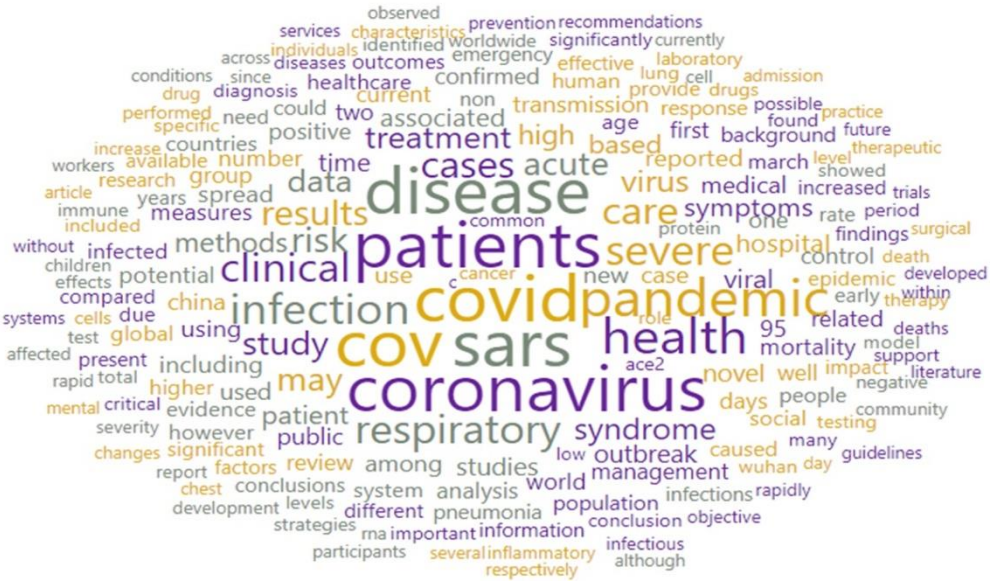

Fig.8: The World Cloud: the words in the figure above are generated by following the unigram model (*n*=1).

Table 4: Top ten frequent words (> = 6000 times occurred) used in abstracts

Words	**Weight($_w$)**
covid	29890
patients	18194

sars	11464
cov	11256
coronavirus	10687
disease	10563
pandemic	8313
health	7510
infection	6312
severe	6090

Table-4 shows the most frequent words used by different authors in the abstracts. We extracted the top 10 words and their weights. Covid ($_w$29890), patients ($_w$18194), sars ($_w$11464), cov ($_w$11256), coronavirus ($_w$10687), and disease ($_w$10563), pandemic ($_w$8313), health ($_w$7510), infection ($_w$6312), and severe ($_w$6090) were found as the most used words.

7.5 Statistics of the text of titles and abstracts

This section describes general statistics of the text of titles and abstracts. A set of 16420(title available) documents containing unique titles and 10429(abstract available) documents containing unique abstracts were extracted from 16500 documents in corpus. Fig.9 visualizes 3 statistical features namely word count, specified token (coronavirus) count, and n-gram count. Six different plots have been presented in a single frame. Plots (a, b, and c) in the first row present the statistics of the unique titles. On the other hand, Plots (d, e, and f) in the second row present the statistics of unique abstracts.

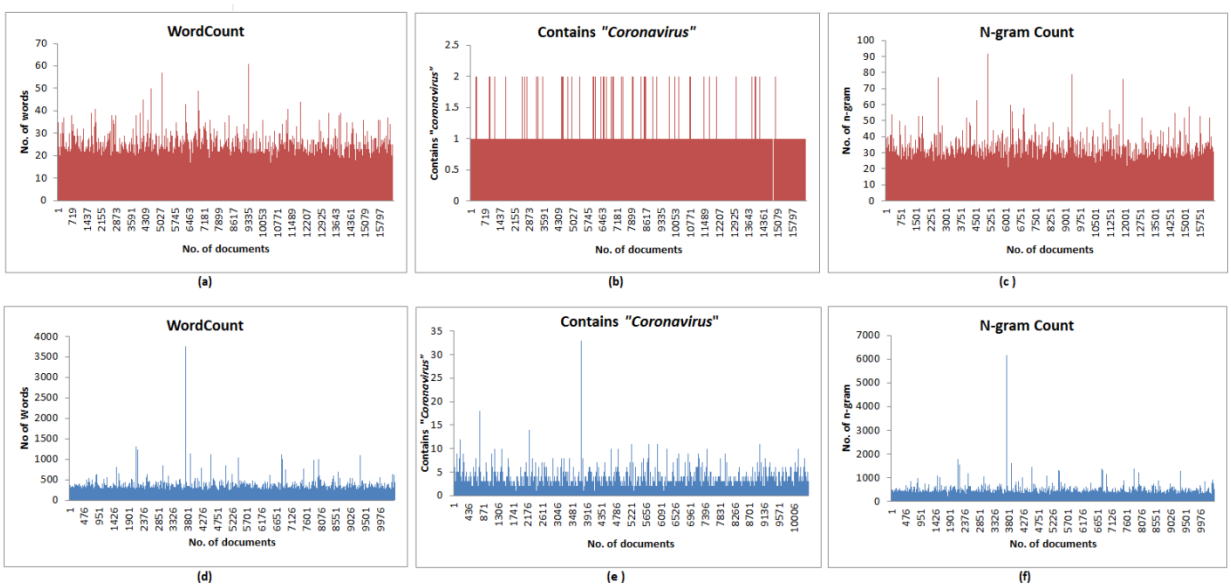

Fig.9: Statistics of the text of unique *Titles* (Plot: a, b, and c) and *Abstracts* (Plot: d, e, and f).The titles and abstracts are equivalent to documents.

It is found that a total of 109 titles contains 30(>=) words, 1443 titles contain words in between 20-29, and the rest of the titles contain less than (<) 20words. The specified token (contains or “coronavirus”) was found 2 times in 15 titles, 1 time in 3143 titles, and 0 times which is most in the rest of the titles. Plot-f shows the statistics of *n*-grams (where *n*=1(size): as default in the software, we made no changes on it). It reports that 50 (>=) unigrams are found in 22 titles followed by 40-49 unigrams in 80 titles, 30-39 unigrams in 556 titles, and 1-20 unigrams in the rest of the titles.

At least 1000 (>=) words are found in 9 abstracts (Plot-a) followed by 500 to 999 words in 67 abstracts and 1-499 words in the rest of the abstracts. Plot-b indicates that the specified token has occurred more than (>) 10 times in 15 abstracts followed by 5-9 times in 187 abstracts, and 1-4 times in the rest of the abstracts. Finally, Plot-c indicates that 27 abstracts contain 1000 (>=) unigrams, 285 abstracts contain 500-999 unigrams, and the rest of the abstract contains 1-499 n-grams. We can see there is an unusual document (plot: d, e, and f) which has more than (>): 3500 words in abstract, specified token occurred 30 times and 6000 unigrams.

7.6 Bag of words

The bag of words is a model used to facilitate the word counts of documents. This model shows the multi sets of words. The parameter used to generate a bag of words is:

- Term Frequency (TF) or counts (generates number of occurrences of a word in document)

Eq.2

$$tf(t,d) = f_{t,d} \div (number\ of\ t\ in\ d)$$

- **Terminologies used:**
  - ***t*** stands for Term or Word
  - ***d*** stands for Document(s) or Title(s).
  - ***f*** stands for Frequency('Tf–Idf', 2020).

Table 5: Example of Bag of words

Title	**Vector space representation(without stop words**	**Vector space representation(with stop words**
"The holistic way of tackling the COVID-19 pandemic: the one health approach."	way=0.111111 tackling=0.111111 one=0.111111 holistic=0.111111 health=0.111111 approach=0.111111 pandemic=0.111111 covid=0.111111 19=0.111111	way=0.0769231 tackling=0.0769231 one=0.0769231 holistic=0.0769231 health=0.0769231 approach=0.0769231 pandemic=0.0769231 covid=0.0769231 19=0.0769231 the=0.230769 of=0.0769231

This data table presents three different columns: title of a document, TF without stop words, and TF with stop words. This information retrieval process is based on vector space model(VSM) and it is used for automatic indexing (Salton et al., 1975). Here, the title represents common vectors (1,1,1,1,1,1,1,1,1) of the nine words(second column) in a common Vector Space Model (VSM). We calculated (Eq.2) to normalize (by regularizing the sum of elements) the vector representation. As per example, the TF of *'way'* is 0.111. The word *'way'* occurred once (one time) in the title and the title contains

9 words without stop words (e.g. "*the" and "of").* The result comes with 1/9=0.111 and it reveals the significance of the each word. In the very next section, we explore the *n*-grams of a bag of words with an example.

7.6.1 N-grams representation

We took the title from Table-5 to demonstrate the *n*-grams. An n-gram is a contiguous sequence of '*n*' items from a given sample of text or speech('N-Gram', 2020). The N-grams model is comprehensively used in text mining and an example is given below:

Title as a sentence:

"The holistic way of tackling the COVID-19 pandemic: the one health approach".

Unigram (n=1):

"Way", "tackling", "one", "holistic", "health", "approach", "pandemic", "covid", "19".

Bigram (n=2):

"way tackling", "tackling covid", "pandemic one", "one health", "holistic way", "health approach", "19 pandemic", "covid 19".

Unigram represents only one word, whereas bigram represents a pair of words. The above example doesn't show the stop words. Usually, *n*-grams are used in supervising data models or language models (Kuznetsov et al., 2016). It is not only useful for showing up a bag of words but also it is effective in generating word clouds and text summarization.

7.7 Hierarchical clustering

This section reveals hierarchical cluster analysis with a representation of dendrogram. A hierarchical clustering method works by grouping data objects into a tree of clusters (Han et al., 2011). It helps to identify the groups in which each cluster is associated. One of the most graceful advantages of clustering is getting meaningful taxonomies. There are two hierarchical clustering approaches. The first one is Agglomerative Clustering and the second one is Divisive Clustering(Teknomo, 2009). For the present

study, we followed the agglomerative clustering method. It is a bottom-up approach to clustering. It gathers small clusters and merges them to represent a single cluster. Another important phenomenon in hierarchical clustering is the distance metrics (or distance measurement). Cluster analysis is not possible without measuring the distances. Euclidean and Cosine are the common distance metrics(Gu et al., 2017). Fig.10 shows clusters of titles of each document based on the Euclidean metric. It is the straight-line distance between two points. Raw data is unstructured by its nature. Working with such unstructured data is extremely critical and sophisticated for any investigation. To overcome such issues, clustering method also makes data more meaningful. It helps to understand each cluster in a dendrogram. We computed the distance between data instances (title rows) to normalize the fetched data and drifted the result into a hierarchical clustering visualization. For the standardization of the clusters, a couple of extensive experiments were performed. The experiments include linkage selection, cluster pruning (selected as *none)*, height ratio selection (11.6%), and analysis of the cutoff value (line) within the dendrogram.

We got a complete of 15592 clusters at point (scale) zero. The amount of clusters decreases as they move towards the last scaling point. But dendrograms do not provide the exact number of clusters. It is an interactive approach. We took a snapshot to show the appearance of the clusters.

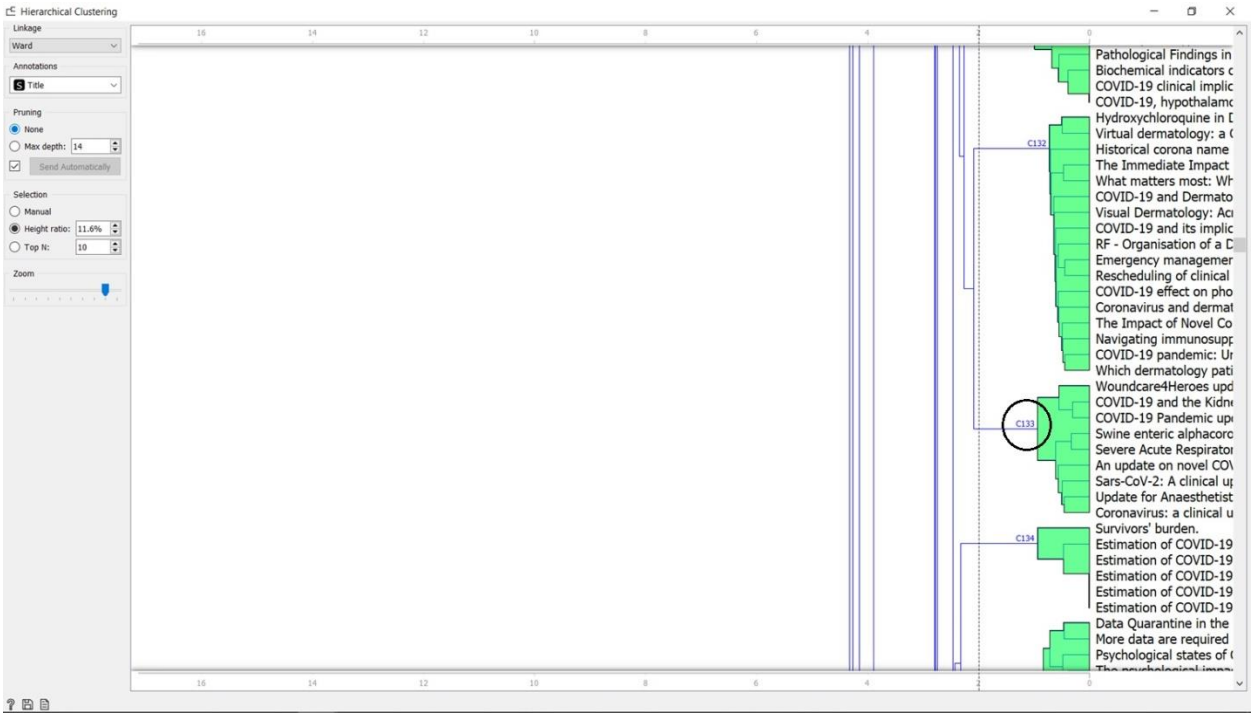

Fig.10: An example of interactive dendrogram (*C133* is specified for analysis)

Fig.10 shows an example of the hierarchical clustering. The cutoff point is 1(scaling). It classifies the titles into a total of 1162 clusters. Ward's linkage method (Ward, 1963) was applied to calculate the similarity of two clusters. The selected (circled) cluster has 6 clades and 7 leaves. The leaves are the titles that developed each clade.

Table 6: Distance matrix of the selected cluster (*C133*)

Sl.no	Full Title	From T1 to T9 is rendered as Title 1 to Title 9								
		T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**
T1	COVID-19 and the Kidneys: An Update.		0.76	0.52	0.68	0.66	0.30	0.74	0.52	0.78
T2	Sars-CoV-2: A clinical update - II.	0.76		0.62	0.55	0.55	0.63	0.49	0.48	0.47
T3	Woundcare4Heroes update: our response to COVID-19 and more.	0.52	0.62		0.57	0.56	0.46	0.60	0.50	0.61

T4	An update on novel COVID-19 pandemic: a battle between humans and virus.	0.68	0.55	0.57		0.52	0.52	0.52	0.52	0.50
T5	Swine enteric alphacoronavirus (swine acute diarrhea syndrome coronavirus): An update three years after its discovery.	0.66	0.55	0.56	0.52		0.55	0.52	0.47	0.32
T6	COVID-19 Pandemic update.	**0.30**	0.63	0.46	0.52	0.55		0.61	0.44	0.64
T7	Update for Anaesthetists on Clinical Features of COVID-19 Patients and Relevant Management.	0.74	0.49	0.60	0.52	0.52	0.61		0.45	0.48
T8	Coronavirus: a clinical update of Covid-19.	0.52	0.48	0.50	0.52	0.47	0.44	0.45		0.52
T9	Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2): An Update.	0.78	0.47	0.61	0.50	0.32	0.64	0.48	0.52	

Table-6 shows the distance matrix of the selected cluster. It also reveals that the distance between the two titles is symmetric. The distance between T1 and T2 is the same as T2 and T1. A distance of 0(zero) is denoted by the black part of the data matrix table. Based on the lower triangular matrix, we computed the closest or lowest distance by the following equation Teknomo (2009).

Eq.3: Number of elements = $\frac{1}{2}m(m-1)$

The titles are the *'m'* objects. Since there are 9 objects, we got 36 elements (value) in the lower triangular matrix. Distinctively, the lowest distance of 36 elements is 0.30 (red-colored). The lowest distance is occurred between T1 and T6. It is more visible in the dendrogram.

8. Discussion

PubMed contains more than 30 million biomedical citation data (as on 10th Oct, 2020). We have discussed about the rapid growth of biomedical literature in the preliminary sections. Hence, text mining tools can help in discovering something new from the raw data. We have tried to apply mining techniques to extract useful information from 16500 documents. The following are the key findings:

- The "*Document Section"* of each document determines the sectional division of the literature. We have extracted the top 12 sections. As the section "betacoronavirus" became the highest ranked section and it is also listed in the MeSH (find in MeSH browser: https://meshb.nlm.nih.gov/search) as controlled vocabulary.
- *"Topic Modeling"* is a popular research area in NLP(Zou et al., 2019). As we looked for extracting the latent topics of the document titles. We have discovered 10 topics (Fig.6) using LSA which produces a classification of text in terms of topics. This shows the representativeness (positive and negative) of a particular topic.
- Context of word or *"concordance"* has been displayed in the corpus. We have found the preceding and following contexts of the queried word "COVID". It shows the number of documents and unique tokens that matched to the queried word.
- In Fig.8, we have generated a *"word cloud"* of the most frequent words that occurred in each abstract. As the stop words were removed for making it more understandable. The result shows that the word "covid" has appeared 29890 times. It reveals its engagement and gives a statistical summarization of weighted texts. There is a drawback of the word cloud. It provides statistical summary of isolated words without taking linguistic knowledge (Heimerl et al., 2014). Though the word cloud is a powerful for text mining.
- We have presented the *"statistics"* (Fig.9) of the whole text retrieved from PubMed. Statistics simplify a general text analysis. It highlights the three different statistical approaches (word count, contains "coronavirus", and *n*-gram count) to the titles and abstract.

- Then we have used the *"Bag of words"* model to classify the documents with TF. An example has been given to demonstrate the classification (Table 5). An attempt has been made to normalize the classification. Then we have measured unigram and bigram of the sentence that we exemplified (Section 7.6.1) with a sentence.
- Clustering of text is very essential in text mining. We have applied the *"hierarchical clustering"* (Fig.10) method to calculate similarity and the distance of 9 titles (C133) from one to another. We executed Ward's method in the software to show a distance matrix (Table 6) of 9 titles and compute the closest distance between two clusters.

8.1 Limitations and future prospects

Text mining consists of many specific features like linguistic analysis, morphological analysis, and domain-specific knowledge filtration, and knowledge management (Chen, 2001). There were some limitations in this study. We tried to discover something new by applying basic text mining techniques. We couldn't show an in-depth analysis of every section. Processing, transforming, and analyzing large dataset is not an easy task due to its heterogeneity. We fetched more than 20 thousand documents but we retrieved only 16500 (82%) documents. We found missing data in our dataset. Somewhere some sections (showed as *"anonymous"*) were missing and again where few abstracts were missing. We already stated about the unique titles and abstracts in Section 7.5.

The application of text mining is diverse. It has not only been applied in biomedical literature also in researching on twitter literature, news media, and other social networks (Karami et al., 2020). Text mining supports to beat the unstructured data that generally is produced within an organisation or outside of the organisation. Therefore, we can assume that most organisations use databases for storing their own data. For instance, libraries use digital library software (e.g. DSpace, https://duraspace.org/dspace/). This study was limited within selected areas but it can lead researchers, knowledge curators, and library professionals towards more in-depth research in future. Our study emphasizes a few prospects that may come out with effective results: a) predictive modeling can be computed based on bag of words, b) to identify and visualize the level

of representativeness of any word to a particular topic, multidimensional scaling can be used while performing topic modeling, and c) the statistics (e.g. word count or word contains) of text can be evaluated using logistic regression. We hope these might help in further research on any context of text mining.

9. Conclusion

Text mining became an essential technique for exploring ontologies in biomedical literature. It not only supports to make decisions over unstructured and raw data but also helps to identify the exact information. We are witnessing exponential growth of biomedical literature; this paper demonstrates several approaches to text mining of literature regarding coronavirus. It shows a general analysis of text data. A comparative overview of two different periods has been shown to present the trends of publications. This paper ascertains: sections of each retrieved document, the word concordances, highly weighted words in abstracts, modeling bag of words along with *n*-grams, topic modeling with positive and negative approaches of keywords, and hierarchical clustering of titles. The role of text mining in useful data extraction and knowledge discovery is going to be indispensable in the coming days.

10. References

Albright, E. (2020). *Probability: Joint, Marginal and Conditional Probabilities – ENV710 Statistics Review Website*. Retrieved from https://sites.nicholas.duke.edu/statsreview/jmc/ . Accessed September 16, 2020

Balagurusamy, E. (1999). *Numerical Methods*. McGraw-Hill Education (India) Pvt Limited.

Betacoronavirus. (2020). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Betacoronavirus&oldid=977635393 . Accessed September 13, 2020

Chen, H. (2001). *Knowledge Management Systems: A Text Mining Perspective*. Knowledge Computing Corporation. Retrieved from http://hdl.handle.net/10150/106481 . Accessed June 12, 2020

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., & Zupan, B. (2013). Orange: Data Mining Toolbox in Python. *J. Mach. Learn. Res.*, *14*(1), 2349–2353.

Dhiman, A. K. (2011). Knowledge Discovery in Databases and Libraries. *DESIDOC Journal of Library & Information Technology*, *31*(6), Article 6. https://doi.org/10.14429/djlit.31.6.1319

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., & Romero, C. (2019). *Text mining in education*. WIREs Data Mining and Knowledge Discovery; Wiley Online Library. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1332 . Accessed September 15, 2020

Gong, L. (2018). *Application of Biomedical Text Mining*. Retrieved from https://www.intechopen.com/books/artificial-intelligence-emerging-trends-and-applications/application-of-biomedical-text-mining . Accessed June 18, 2020

Gu, X., Angelov, P. P., Kangin, D., & Principe, J. C. (2017). A new type of distance metric and its use for clustering. *Evolving Systems*, *8*(3), 167–177. https://doi.org/10.1007/s12530-017-9195-7

Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text Mining in Big Data Analytics. *Big Data and Cognitive Computing*, *4*(1), 1. https://doi.org/10.3390/bdcc4010001

Hearst, M. (2003). *What is text mining*. SIMS. Retrieved from https://www.jaist.ac.jp/~bao/MOT-Ishikawa/FurtherReadingNo1.pdf . Accessed September 15, 2020

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). *Word Cloud Explorer: Text Analytics Based on Word Clouds*. Retrieved from https://www.computer.org/csdl/proceedings-article/hicss/2014/2504b833/12OmNqNG3jl . Accessed September 14, 2020

Karami, A., Lundy, M., Webb, F., & Dwivedi, Y. K. (2020). Twitter and Research: A Systematic Literature Review Through Text Mining. *IEEE Access*, *8*, 67698–67717. https://doi.org/10.1109/ACCESS.2020.2983656

Kuznetsov, V., Liao, H., Mohri, M., Riley, M., & Roark, B. (2016). Learning N-gram Language Models from Uncertain Data. *Interspeech*.

Larsson, K., Baker, S., Silins, I., Guo, Y., Stenius, U., Korhonen, A., & Berglund, M. (2017). Text mining for improved exposure assessment. *PLOS ONE*, *12*(3), e0173132. https://doi.org/10.1371/journal.pone.0173132

Latent semantic analysis. (2020). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Latent_semantic_analysis&oldid=976515199 . Accessed September 14, 2020

Lioma, C., & Ounis, I. (2008). A syntactically-based query reformulation technique for information retrieval. *Information Processing & Management*, *44*(1), 143–162. https://doi.org/10.1016/j.ipm.2006.12.005

Luhn, H. P. (1960). Key word-in-context index for technical literature (kwic index). *American Documentation*, *11*(4), 288–295. https://doi.org/10.1002/asi.5090110403

Mulunda, C., Wagacha, P., & Lawrence, M. (2018). Review of Trends in Topic Modeling Techniques, Tools, Inference Algorithms and Applications. *5th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, 28–37. https://doi.org/10.1109/ISCMI.2018.87032314

Nai-Lung Tsao, Chin-Hwa Kuo, David Wible, & Tsung-Fu Hung. (2009). Designing a Syntax-Based Retrieval System for Supporting Language Learning. *Journal of Educational Technology & Society*, *12*(1), 73–81. JSTOR. http://www.jstor.org/stable/jeductechsoci.12.1.73

Natarajan, M. (2005). Role of Text Mining in Information Extraction and Information Management. *DESIDOC Journal of Library & Information Technology*, *25*(4), Article 4. https://doi.org/10.14429/djlit.25.4.3663

NCBI. (2020, March 26). *Text Mining Tools—NCBI - NLM*. Retrieved from https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/ . Accessed July 8, 2020

N-gram. (2020). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=N-gram&oldid=975867256 . Accessed September 13, 2020

PubMed Help. (2020). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK3827/ . Accessed September 13, 2020

Read the Docs. (2020). Retrieved from https://readthedocs.org/projects/orange3-text/builds/ . Accessed September 16, 2020

Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Commun. ACM*, *18*(11), 613–620. https://doi.org/10.1145/361219.361220

Sayers, E. (2017). Sample Applications of the E-utilities. In *Entrez Programming Utilities Help [Internet]*. National Center for Biotechnology Information (US). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK25498/ . Accessed June 26, 2020

Shatkay, H., & Craven, M. (2012). *Mining the Biomedical Literature*. Retrieved from https://mitpress.mit.edu/books/mining-biomedical-literature . Accessed July 18, 2020

Silva, C., & Ribeiro, B. (2003). The importance of stop word removal on recall values in text categorization. *Proceedings of the International Joint Conference on Neural Networks, 2003.*, *3*, 1661–1666 vol.3. https://doi.org/10.1109/IJCNN.2003.1223656

Simon, C., Davidsen, K., Hansen, C., Seymour, E., Barnkob, M. B., & Olsen, L. R. (2019). BioReader: A text mining tool for performing classification of biomedical literature. *BMC Bioinformatics*, *19*(13), 57. https://doi.org/10.1186/s12859-019-2607-x

Talabis, M. R. M., McPherson, R., Miyamoto, I., Martin, J. L., & Kaye, D. (2015). Chapter 1—Analytics Defined. In M. R. M. Talabis, R. McPherson, I. Miyamoto, J. L. Martin, & D. Kaye (Eds.), *Information Security Analytics* (pp. 1–12). Syngress. https://doi.org/10.1016/B978-0-12-800207-0.00001-0

Tan, A., Mui, H., & Terrace, K. (2000). *Text Mining: The state of the art and the challenges*. Undefined. /paper/Text-Mining%3A-The-state-of-the-art-and-the-Tan-Mui/9a80ec16880ae43dc20c792ea3734862d85ba4d7

Tao, D., Yang, P., & Feng, H. (2020). Utilization of text mining as a big data analysis tool for food science and nutrition. *Comprehensive Reviews in Food Science and Food Safety*, *19*(2), 875–894. https://doi.org/10.1111/1541-4337.12540

Teknomo, K. (2009). *Hierarchical Clustering Tutorial: Distance matrix*. Retrieved from https://people.revoledu.com/kardi/tutorial/Clustering/Distance%20Matrix.htm . Accessed September 12, 2020

Tf–idf. (2020). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Tf%E2%80%93idf&oldid=976859641 . Accessed September 15, 2020

Thompson, K. (1968). Programming Techniques: Regular expression search algorithm. *Communications of the ACM*, *11*(6), 419–422. https://doi.org/10.1145/363347.363387

University of Illinois at Chicago. (2020, March 26). *Health Sciences Databases A - Z - Health Sciences Gateway—Subject & Course Guides at University of Illinois at Chicago*. Retrieved from https://researchguides.uic.edu/c.php?g=252180&p=1682634 . Accessed July 8, 2020

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Xie, B., Ding, Q., & Wu, D. (2017). *Text mining on big and complex biomedical literature*. ResearchGate. https://www.researchgate.net/publication/320046547_Text_mining_on_big_and_complex_biomedical_literature

Yang, Y., Akers, L., Klose, T., & Barcelon Yang, C. (2008). Text mining and visualization tools – Impressions of emerging capabilities. *World Patent Information*, *30*(4), 280–293. https://doi.org/10.1016/j.wpi.2008.01.007

Ye, Z., Tafti, A. P., He, K. Y., Wang, K., & He, M. M. (2016). SparkText: Biomedical Text Mining on Big Data Framework. *PLOS ONE*, *11*(9), e0162721. https://doi.org/10.1371/journal.pone.0162721

Zhai, C., & Massung, S. (2017). Text Data Understanding. In *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Association for Computing Machinery and Morgan & Claypool. https://doi.org/10.1145/2915031.2915035

Zhou, X., Peng, Y., & Liu, B. (2010). Text mining for traditional Chinese medical knowledge discovery: A survey. *Journal of Biomedical Informatics*, *43*(4), 650–660. https://doi.org/10.1016/j.jbi.2010.01.002

Zou, X., Zhu, Y., Feng, J., Lu, J., & Li, X. (2019). A Novel Hierarchical Topic Model for Horizontal Topic Expansion With Observed Label Information. *IEEE Access*, *7*, 184242–184253. https://doi.org/10.1109/ACCESS.2019.2960468