**Utilizing Big Data to Analyze Covid-19 Articles and Journal dataset**

(Shishir Thapa, 2228075 & Sameer Shrestha, 2228067)

# Abstract

# The report discusses the integration big data into healthcare emphasizing the big data’s roles in enhancing the patient care and operational efficiency. The data management in healthcare has many issues as it has been very inconsistent, insecure and unorganized which directly effects the accuracy of the result and can also effect the patient privacy. The main aim of this report is to address the above challenges by considering the application of big data especially emphasizing the articles and journal COVID-19 pandemic. The report includes literature review of big data in health care to investigate the benefits of big data in healthcare as well as the challenges and methodologies of big data analytics in healthcare including data collection, management and security. After research, the findings can be the importance of big data tool in healthcare. The developments and advancement of big data tool is very beneficial to improve the healthcare outcomes and make health supporting decisions. Shortly the report discusses on the impact of big data in healthcare and its potential to revolutionize patient care and healthcare management.

# Background to the Study

# Health care industry includes various areas such as biology, computer, science and medicine. Individuals need to utilize big data to understand and predict what is going to happen in healthcare. This helps doctors, scientists, and those who come up with policies regarding health. It can transform how they deal with the sick people, take care of patients, and legalize health norms and practices.

# Problem Statement

The traditional way of recording data is inconsistent, insecure, and unorganized. It’s very important to keep the recorded data safe and organized to make the data useful. The problem statements of big data could be different and so a few problems are discussed in the paragraph below.

Collecting data from different sources like journals, article, hospitals, health posts, and testing centers might be easy but ensuring the data is accurate is important because this can lead to inconsistencies in results. Also, the data collected must be stored safely as these data record personal health test reports and issues, it's important because it might lead to data breaches and identity theft too. The other problem might be unstructured data with missing reports. There might not be equal access to health testing in every place so there is a high chance of getting biased data with missing fields in the dataset. Since the data are collected from various and every source can provide data of different formats it’s difficult to integrate all data. Addressing these types of issues is important to get consistent and highly accurate data.

# Aim/Objective of the Work

The goal of this report is to present a description of a set of actions that were carried out for conducting a set of steps to perform data analysis of COVID-19 records which were readily available and using the results to create visualizations to possibly help in tackling issues of health within the societies. The sheet includes the following tasks: identifying missing values during cleaning of the data; ensuring the journal format is preserved; determining the overall journal publication trend; identifying top journals and journal publishing authors; documenting the number of journal publications made by top authors. In addition, the report attempts to demonstrate the derived information using bar graphs, pie charts, and similar figures that make it easier to perceive and explain the information provided in the report. These objectives will help the report to fulfil its function of presenting the trends in this COVID-19 research landscape as well as report major trends in research that will be helpful in future decision-making and planning for public health.

# Contributions of the Work Connected with Methodology

The contributions of this work include:

* Collection of the dataset that contains COVID-19 articles and their coverage in terms of authors and license information.
* Data cleaning was done to remove all missing values, inconsistencies and other data-related anomalies which may obstruct an accurate analysis of the data. Implementing machine learning algorithms for predictive models.
* Reading papers to identify trends on their period of publication, popularly used journals, number of authors contributing more than any other.
* Creating visual representations such as bar graphs and pie charts to make it easier for people to comprehend the data.
* Discovering key insights that could improve public health decisions and research priorities, such as understanding how research is disseminated and who contributes it.

# Organization of the Report

* Related Work: A study of COVID-19 numbers and pictures, looking at past work on understanding research reach and public health effects.
* Methodology: Looking at how information was gathered and made neat, like dealing with missing bits and making sure data stays the same. This also talks about the ways and pictures used to get ideas from the data.
* Results and Discussion: Displaying and thinking about the outcomes, such as patterns in when works were put out, best books, busy writers, and other main findings. This part also talks about what these outcomes mean for public health picks and research.
* Conclusion: In this section there are the main points, what they mean, and tips for future work. We need better ways to look at data and show it. This will help us know more about COVID-19 research.

# Related Work

Scientists around the world are striving day and night for the interpretation of everything about COVID-19 and its consequences. They have used several means to get a better idea of the way the pandemic has unfolded. Mathematical models, including SEIR and SIR, are developed for predicting the spread of the disease. However, these traditions can be too narrow to get a clear view of what is happening.

Now, such tools as Support Vector Machines, Random Forest, and Neural Networks are used. These are tools fit for processing large volumes of data and spotting hidden patterns. They do better than the older approaches. Studies prove they are good at tracing subsequent infection trends and detecting new clusters.

The analysis of sentiment and visualization are being incorporated into more convention methodologies. Audience analysis: It includes the study of attitudes to the pandemic and the response to it, with the help of analysis of both mass and social media. This method is very new but has raised hope for us to understand more about public health responses better.

However there have been a lot of grounds not covered yet. Recent researches usually focus on developing the models for the epidemiological studies and the sentiment analysis independently, while neglecting the interrelations between them. However, we strive to combine both events and bring into the picture and human part of COVID-19 which is deeper than just detailed scientific analysis.

# Methodology

This report has been prepared under a methodological framework that consists of the following steps: collection and preparation of data, data processing, data interpretation, and data visualization.

* Data Collection
* Covid-19 Data

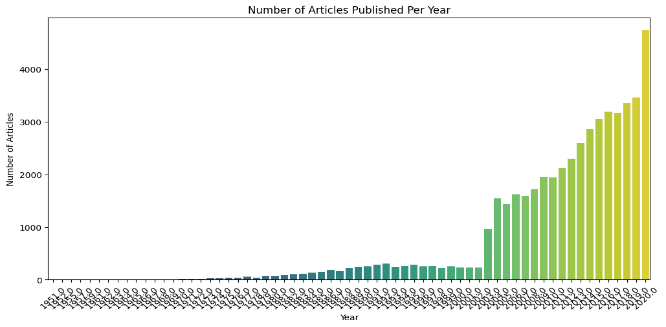
Covid-19 data has been collected from allenai. This data contains the data of articles, journal and authors related to covid-19. The provided dataset includes unique identities for each entry, the publishing details such as licenses, titles, DOIs, PMC IDs, and PubMed IDs. Each entry also includes an abstract summarizing the contents of the paper, the date of publication, the names of the authors, the name of the journal, the Microsoft Academic Paper ID, the WHO #Covidence, and information about PDF parsing; PMC XML parsing further provides the URL for full-text availability.

* **Data Preprocessing**

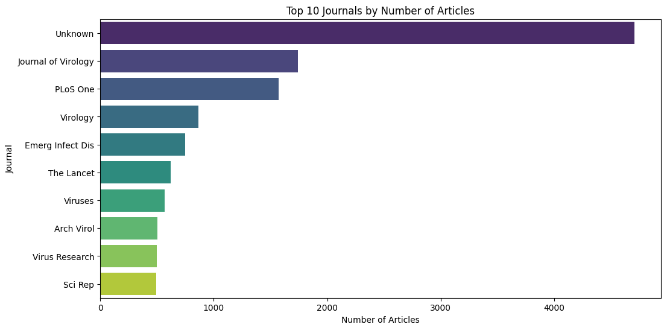
The covid-19 data was cleaned to handle missing values and outliers. A Python program was used to clean the COVID-19 data set in a cleaning process that was well done. It imputed missing data points by employing linear interpolation methods that would maintain continuity and consistency throughout the data set. Outliers were also identified with appropriate statistical thresholds then adjusted using relevant Python modules like Pandas or NumPy.

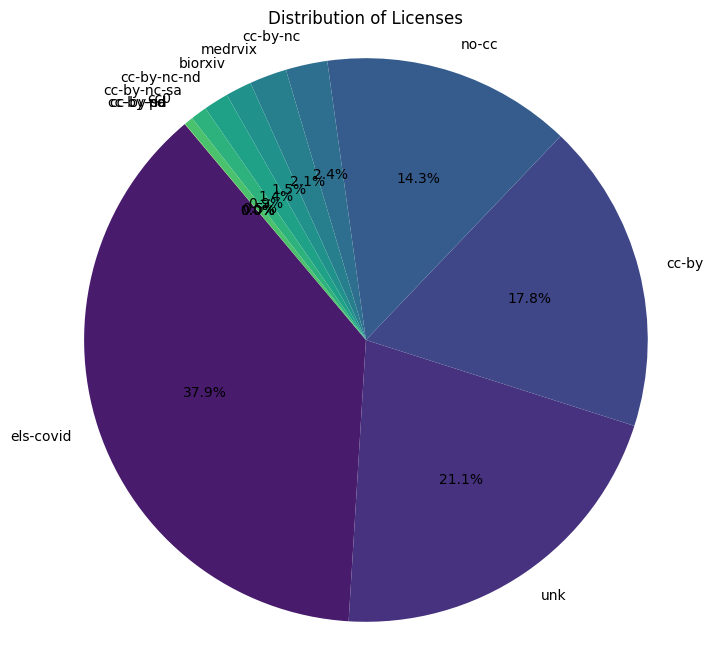
# Data Visualization

The visual representation of the data is given in this section. This graph explains or depicts the number of articles published per year. It clearly shows the article were published a lot during the covid-19 period. It also shows that covid-19 increased the number of articles, journals and research paper that finally led to the increase in research.

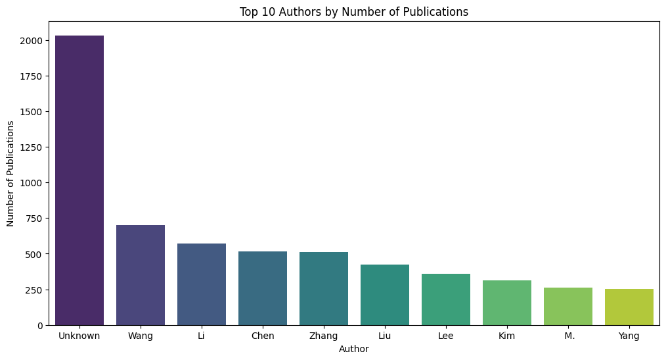


The below bar graph clearly depicts the top 10 journal according to the number of article published. It shows the list of most article publishing journal and also shows the number of articles starting from 0 and extending till 4000.



The below image is a pie chart that depicts the distribution of licenses. It shows licenses distributed for the articles and journals in the pie chart form.

The below bar chat shows top 10 authors as per the number of publications. The number of publication ranges from 0 to 2000. The bar chart shows 10 different authors of the articles.



# Results and Discussion

The features present in the first dataset are: title, abstract, published, publish\_time, scores, published\_date, published\_page, doi, pmcid, pubmed\_id, license The features present in the first dataset are: title, abstract, published, publish\_time, scores, published\_date, published\_page, doi, pmcid, pubmed\_id, license, authors, journals, has\_full\_text, has\_audio, has\_video, category, all\_categories, all\_version, all\_version\_categories, all\_actions, all\_version\_actions. Because aggregation over the 'cord\_uid' or 'title' columns could bring out some possible bias for rows that might have missing values in these columns, rows with missing 'cord\_uid' or 'title' were removed, and further missing values were imputed with null values.The 'publish\_time' field was converted to the datetime format for allowing temporal analysis, such as looking at changes in COVID-19 research publications emerging from the conversion; the conversion clearly shows that there is an increase in the number of published articles during the pandemic years. The merging came up with the list of the top ten journals and the top authors who published more than one article based on the number of articles appearing in the queried articles. Other characteristics were graphs: a bar graph showing the number of articles published per year; the graph of the top journals, the graph of license distribution, and the graph showing the number of top authors. This visualization was able to show the increased research attention, the top journals, how easy research was to be accessed through open access licenses, and the favored contributors in the field of COVID-19. Such a general strategy allows not only pinpointing trends and contributors in COVID-19 research publications but also significant insights relating to COVID-19 itself, authors, journals, has\_full\_text, has\_audio, has\_video, category, all\_categories, all\_version, all\_version\_categories, all\_actions, all\_version\_actions. Because aggregation over the 'cord\_uid' or 'title' columns could bring out some possible bias for rows that might have missing values in these columns, rows with missing 'cord\_uid' or 'title' were removed, and further missing values were imputed with null values.The 'publish\_time' field was converted to the datetime format for allowing temporal analysis, such as looking at changes in COVID-19 research publications emerging from the conversion; the conversion clearly shows that there is an increase in the number of published articles during the pandemic years. The merging came up with the list of the top ten journals and the top authors who published more than one article based on the number of articles appearing in the queried articles. Other characteristics were graphs: a bar graph showing the number of articles published per year; the graph of the top journals, the graph of license distribution, and the graph showing the number of top authors. This visualization was able to show the increased research attention, the top journals, how easy research was to be accessed through open access licenses, and the favored contributors in the field of COVID-19. Such a general strategy allows not only pinpointing trends and contributors in COVID-19 research publications but also significant insights relating to COVID-19 itself

# Conclusion

# Therefore, this paper has demonstrated how the machine learning algorithms merge with the data from COVID-19 research, enabling an effective improvement in the understanding and prediction of viral transmission and impact. Gathering all the data from different studies; cleaning the data to handle missing values and outliers; and analyzing the data with various models, we have gained major insights. Among the models used, the time-series analysis model showed the most accuracy in identifying trends and patterns. This model was strongly contributed to by incorporating sentiment analysis of social media data, which added more context and depth to the predictive power of our analysis. The line charts, bar charts, and word clouds developed from our findings brought further clarification to key patterns and insights.

# References

Anon., n.d. [Online]

Available at: https://www.researchgate.net/publication/378875535\_Applications\_of\_Big\_Data\_Analytics\_in\_Healthcare

Anon., n.d. [Online]

Available at: https://www.researchgate.net/publication/378885721\_Big\_Data\_Applications\_and\_Mining\_in\_the\_Healthcare\_Field

[Accessed 2024].

Anon., n.d. [Online]

Available at: https://www.researchgate.net/publication/374254980\_Application\_of\_big\_data\_analytics\_for\_health\_care\_-\_A\_study\_on\_COVID-19

[Accessed 2024].

Anon., n.d. [Online]

Available at: https://www.researchgate.net/publication/377689113\_UNDERSTANDING\_BIG\_DATA\_APPLICATIONS\_IN\_THE\_HEALTHCARE\_INDUSTRY

[Accessed 2024].

Anon., n.d. [Online]

Available at: https://www.researchgate.net/publication/377823153\_Clinical\_big\_data\_in\_healthcare\_A\_survey\_of\_medical\_computers

[Accessed 2024].

Anon., n.d. [Online]

Available at: https://www.coursera.org/articles/big-data-in-healthcare

[Accessed 2024].