**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Titles:** | *Advanced Data Analytics*  *Big Data Storage and Processing* |
| **Assessment Title:** | MSC\_DA\_CA2v4 |
| **Lecturers Name:** | David McQuaid  Muhammad Iqbal |
| **Student Full Name:** | Mustafa Gunes Cetin |
| **Student Number:** | 2022456 |
| **Assessment Due Date:** | 26/05/2023 |
| **Date of Submission:** | 26/05/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Table of Contents**

[**1.** **Introduction** 1](#_Toc135957887)

[**2.** **Literature Review** 2](#_Toc135957888)

[**3.** **Data and Methodology** 2](#_Toc135957889)

[**3.1. Data** 2](#_Toc135957890)

[**3.2 Methodology** 3](#_Toc135957891)

[**3.2.1 Hadoop Framework (General)** 3](#_Toc135957892)

[**3.2.2 Database Storage System** 4](#_Toc135957893)

[**3.2.3 Apache Pyspark** 5](#_Toc135957894)

[**3.2.4 Modelling** 6](#_Toc135957895)

[**3.2.5 Visualisation Dashboard** 6](#_Toc135957896)

[**4.** **Results and Discussion** 7](#_Toc135957897)

[**4.1 Results** 7](#_Toc135957898)

[**4.2 Discussion** 10](#_Toc135957899)

[**5.** **Conclusion** 10](#_Toc135957900)

[**6.** **References** 11](#_Toc135957901)

Sentiment Analysis for Pfizer Vaccine using MapReduce and NoSQL databases

# **Introduction**

All around the world, including the industrialized economies of the United States of America, United Kingdom, Japan, Korea, and Russia, Novel Corona Viruses, also known as COVID-19 viruses, have spread. Due to widespread lockdown in many places, the pandemic scenario has an impact on all sectors, but particularly the economy. There are effects of the shutdown on all activity by humans. Many scientists have created vaccines to address this issue.

Due to its adaptability, interactivity, and affordability, social media became the primary means of expression during that difficult time [1]. Twitter and other social networking sites are increasingly used for sharing ideas and forming personal opinions, in addition to for advertising [2]. By examining how individuals express their thoughts on various social media subjects, sentiment analysis may be performed on the public's perception of social media. Sentiment analysis is employed since its customary to consider public opinion while making decisions. As a result, this study tries to determine whether the sentiment in tweets about the Pfizzer Vaccine is good or negative. Any stakeholder in this topic, including vaccine manufacturers, medical facilities, and decision-makers, benefits from the findings.

The COVID-19 situation is quickly changing thanks to vaccination initiatives. Most nations made enormous efforts to obtain and disperse millions of doses throughout the world [3]. Vaccines have been extensively reported, discussed, and argued by news as well as social media since their debut in November 2020, resulting in a flood of information that people have consumed[4, 5]. In the end, the internet and newspapers both serve as important information sources[6, 7]; they support knowledge and particular viewpoints about actual events like vaccinations. Reconstructing major concepts that are read and influenced by large audiences requires examining the linguistic and emotional characteristics of knowledge conveyed by such venues[4]. It is imperative to reconstruct the way news outlets have been structuring knowledge around vaccinations for COVID-19, especially considering the massive audiences they reach. Monitoring emotions like anger or trust in shared knowledge is crucial, but understanding the main concepts that elicit these emotions in widely read content remains a challenge. Therefore, identifying specific emotional and conceptual associations promoted by news media is crucial for combating misinformation and social manipulation respectfully.

This paper focuses on usage of two database of MongoDB and MySQL system under the Hadoop MapReduce environment to understand the sentiment of Pfizer vaccine using the online Twitter data.

# **Literature Review**

The use of language is fundamental for conveying ideas and exchanging knowledge. Consequently, analyzing linguistic data is essential when mapping perceptions and attitudes accurately. Emerging artificial intelligence technologies have been leveraged in computer science to create novel approaches like deep neural networks or sentiment analysis that enable precise predictions of stances within texts [8, 9].

Previous research has used tweets to analyse public opinion of vaccinations other than COVID-19 vaccines. There are currently few studies that have examined global public opinion regarding tweets directly related to the COVID-19 vaccine. For instance, Dubey examined the emotions expressed in tweets about the COVID-19 vaccine that were only posted in India and between January 14, 2021, and the 18th of January 2021 [10]. In order to explain behavioural intention towards COVID-19 vaccinations from a social psychology perspective, Liu et al. assembled a limited set of data using the COM-B model [11]. In order to automatically identify tweets containing behavioural intents in relation to the COVID-19 vaccination, several researchers have also developed transfer learning models [11]. Deep learning models were created by Du et al. for tweets relating to the HPV vaccine, and they discovered that the mood shifted in early 2017 and differed significantly between regions in the U.S. [12]. Pregnant women's participation in research, vaccine promotion, and institutional confidence and openness were the three primary issues with a global concern that Martin et al. effectively discovered after analysing maternal vaccination posters from various countries [13].

# **Data and Methodology**

## **3.1. Data**

The data has been retrieved from the Kaggle dataset containing tweets about the Pfizer Vaccine. The dataset probably contains a variety of information about each tweet, including the tweet text, time stamp, user details, location, and sentiment ratings. It might also include other metadata, such as user demographics, favourites, and retweet counts.

The dataset's goal is to examine online debates, sentiment, and attitudes regarding the Pfizer vaccination. The dataset can be used for exploratory analytics such as sentiment analysis, topic modelling, sentiment trends across time, etc [14].

## **3.2 Methodology**

**A picture containing text, screenshot, circle, font

Description automatically generated**

Figure : Overall Framework

### **3.2.1 Hadoop Framework (General)**

A 2004 Google research paper on the MapReduce computing model and the Google File System (GFS) served as the inspiration for the Hadoop framework. At Yahoo, Doug Cutting and Mike Cafarella began work on the open-source Hadoop implementation in 2006 after being inspired by this study. The project is named after a toy elephant owned by Doug Cutting's son. Hadoop's first iteration was built around two main building blocks: the The MapReduce framework for distributed data processing and the Hadoop Distributed File System (HDFS) for distributed storage serve as the foundation for Hadoops initial implementation. It was later given the name Apache hadoop and submitted to the Apache software Foundation in 2008. Then, Apache Hadoop has become very well-known and is frequently used as a framework for efficiently managing and analysing enormous counts of data. [15].

The Hadoop architecture, which consists of a number of crucial components, enables the distributed processing and storage of masive amounts of data. The key important of Hadoop are:

-Hadoop's main storage system is the Hadoop Distributed File System HDFS. It is a distributed file system made to handle and store huge datasets across a cluster of computers. HDFS offers scalability, high throughput, and fault tolerance.

-The resource management structure in Hadoop is called Yet Another Resource Negotiator (YARN). It controls and distributes CPU and memory resources among the cluster's operating applications. MapReduce, Spark, and Hive are just a few of the processing frameworks that may coexist and effectively share cluster resources thanks to YARN.

-MapReduce is a computing methodology for handling and analysing massive datasets in a distributed, parallel fashion. It separates data into smaller components and processes them concurrently across a number of cluster nodes. Map and Reduce are the two fundamental components of the MapReduce algorithm. Map processes and modifies the data, and Reduce collects the outcomes [16]

The Hadoop MapReduce framework is made up of a number of essential parts that cooperate to handle massive amounts of data in a distributed fashion. The JobTracker, which accepts job requests through users or apps, is the framework's central component. It divides the job into smaller assignments, distributes each one to a cluster node, and tracks each one's development. Every node in the cluster has a TaskTracker that is in charge of carrying out the given tasks. The map tasks are executed by the TaskTracker, and they handle the incoming data processing and produce intermediate key-value pairs.

After that, the result of the map jobs is divided and sorted. The TaskTracker also executes the reduction tasks, which take the intermediate key-value pairs and aggregate and summarise them to provide the final result. The input splits are determined by the InputFormat component, which also specifies how each split is handled by the map jobs. The formatting and storing of the finished output are handled by the OutputFormat component. There are other supporting elements like the DistributedCache, which enables file and archive sharing between nodes, and the Combiner, which aggregates intermediate data locally before sending it to the reduce processes. Together, these elements help a distributed Hadoop cluster analyse massive datasets efficiently and in parallel [17]

-The common tools and libraries utilized by various other Hadoop components are provided by Hadoop Common. It contains the Java libraries, file systems, utilities, and configurations required to run Hadoop.

-Tools and utilities for the Hadoop environment are also provided by the software. These tools and utilities make it easier to manage and oversee the Hadoop ecosystem. These include technologies like Hadoop Oozie for workflow scheduling and coordination, Hadoop Streaming for creating MapReduce programmes in languages other than Java, and Hadoop Distributed File System (HDFS) command-line interface (CLI) [18]

### **3.2.2 Database Storage System**

**3.2.2.1 MySQL**

Instead of placing all of the data in one huge warehouse, a relational database keeps the data in individual tables. The physical files used to organise the database structure are designed for speed. The logical data model provides an adaptable programming environment with objects including tables of data, views, rows, and columns. The relationships between various data fields, such as one-to-one, one-to-many, unique, necessary, or optional, and "pointers" between other tables, are governed by the rules you establish.

The acronym "SQL" in "MySQL" represents "Structured Query Language." SQL is widely recognize as the most popular standardized language for accessing databases. In your programming environment, you have several options for working with SQL. You can directly input SQL statements, allowing you to generate reports and perform database operations.

After the reduction() function of Hadoop, MySQL is the appropriate database for storing the sentiment analysis data from Pfizer vaccines. First off, MySQL is a trusted and well-liked relational database management system (RDBMS) renowned for its robustness and scalability. Large amounts of sentiment data from analysis produced by the reduce() function can be handled effectively. Second, MySQL offers strong indexing and querying features that make it possible to quickly retrieve and analyse sentiment data [19].Due to its support for SQL, complicated queries and aggregations can be used to glean valuable information from the output of sentiment analysis. Thirdly, when working with delicate sentiment analysis data, MySQL maintains transactional integrity, guaranteeing data consistency and reliability. A seamless transfer of information and connection among the MapReduce process and the MySQL database is made possible by MySQL's easy interface with Hadoop through connectors or APIs. In summary, MySQL is the best option for storing and analysing the decreased sentiment analysis information in the setting of Pfizer vaccine sentiment analysis because to its speed, capacity, SQL abilities, reliability of data, and Hadoop connection.

**3.2.2.2 MongoDB**

A well-liked open-source NoSQL database called MongoDB provides a scalable and adaptable method for organizing and storing data. Data is stored in flexible, schema-free documents in a JSON-like syntax using a document-oriented data paradigm. Documents, collection, indexes, recombination and sharding are the main elements of MongoDB. Within the same collection, documents can have different structures and serve as unique records. Similar to how tables in relational databases group relevant documents together, collections do the same. Indexes facilitate quick data retrieval, which enhances query performance. By replicating data across numerous servers, replication provides maximum availability and fault tolerance. By dividing data among several servers, sharding enables horizontal scaling. MongoDB is significant because it offers an adaptable data format that supports agile development and adapts to changing data requirements. It provides seamless scalability, enabling programmes to manage huge and expanding datasets. MongoDB is an effective option for processing varied and high-volume data because to its document-oriented approach, sophisticated query language, and automated sharding, making it suited for modern applications that demand versatility, scalability, and performance.

### **3.2.3 Apache Pyspark**

On top of Apache Spark, PySpark is a freely available distributed computing framework made specifically for processing and analysing huge data volumes. With PySpark, programmers can use Python to create data-intensive applications that take advantage of Spark's powerful distributed processing capabilities.

The Spark Core, which delivers the fundamental functionality and distributed computing APIs, and the PySpark SQL module, that offers a high-level user interface for interacting with organised and semi-structured data, are the two main parts of PySpark. Additionally, PySpark has modules for stream processing and machine learning (PySpark Streaming and PySpark MLlib). PySpark is crucial for a number of reasons. It is appropriate for big data scenarios because it provides effective analysis and processing of data on dispersed clusters. Processing speeds are increased by utilising in-memory computing and optimised execution strategies [20]. The second benefit of PySpark is that it offers a familiar programming language interface, enabling programmers to use their Python expertise and already-existing Python modules for manipulating information and analysis. Python developers may utilise it easily and conveniently because of this. Finally, PySpark is a useful tool for creating from beginning to end data processing pipelines and sophisticated analytics applications because of its connection to other Spark components, like Spark SQL and MLlib, which offer an extensive ecosystem for data machine learning and analytics activities [21].

### **3.2.4 Modelling**

**3.2.4.1 Linear Regression**

A common supervised machine learning approach called linear regression is used to forecast a target variable that is continuous based on one or more input data.

The fluctuation in the values of a continuous parameter over time can be modelled and predicted using linear regression in time series analysis. The goal variable and the input features (time) are presummably linearly related. It is possible to determine the trend or direction of change in the time series by fitting a linear regression model to past information.

In time series analysis, linear regression treats the length of time component as a distinct variable. The general trend, seasonality, and other linear trends in the data are all captured. The slope and intercept of the model used for regression are represented by the coefficients, which shed light on the time series' baseline value and rate of change.

**3.2.4.2 Gradient Boosting Algorithms**

Gradient Boosting is a technique for ensemble learning that combines a number of weak prediction models (usually decision trees) to produce a strong model for prediction. It operates by gradually introducing new models that fix the errors created by the earlier ones. The approach calculates the loss function gradients regarding the predictions in each iteration, and the ensuing models undergo training to minimise these gradients.

Gradient boosting can identify nonlinear correlations, intricate interactions, and temporal dependence in the data in this time-series analytic setting. Gradient boosting in analysis of time series can learn to remedy the mistakes produced by earlier models by training models that are weak progressively. By taking into consideration lag variables or designed features, it may identify trends, seasonality, and additional patterns in the time series.

In time series data, gradient boosting techniques excel in identifying nonlinear correlations and producing precise forecasts. They are capable of coping with noisy and nonstationary time series, modifying to shifting patterns, and dealing with missing data. To prevent overfitting, however, precise parameter tuning and regularisation strategies are crucial.

### **3.2.5 Visualisation Dashboard**

A visualization dashboard is a user-friendly interface that visually presents data, allowing users to explore information, uncover patterns, and make well-informed decisions. It typically consists of diverse graphical elements such as tables, maps, charts, and graphs that dynamically update as new data is selected or received. The primary purpose of a visualization dashboard is to provide an interactive and intuitive way to analyze data, enabling users to gain fresh insights and understand complex relationships more easily.

A visualisation dashboard can be quite helpful in understanding and evaluating the sentiment patterns and trends connected to the Pfizer vaccination in the context of the Pfizer sentiment research. For Pfizer's sentiment analysis, a visualization dashboard is crucial since it offers a visual and dynamic depiction of sentiment patterns and trends. Stakeholders can use it to compare sentiment changes in real-time across several dimensions and engage in interactive data exploration. This makes it possible to make decisions quickly and intelligently, assists in spotting new trends, and offers a thorough understanding of how the public feels about the Pfizer vaccination.

Python Dash has been utilised in this project because it enables programmers to create interactive web-based dashboards and data visualisation apps. It offers a simple framework with many adjustable elements that makes it easier to create aesthetically beautiful and dynamic dashboards for decision-making based on data.

# **Results and Discussion**

## **4.1 Results**

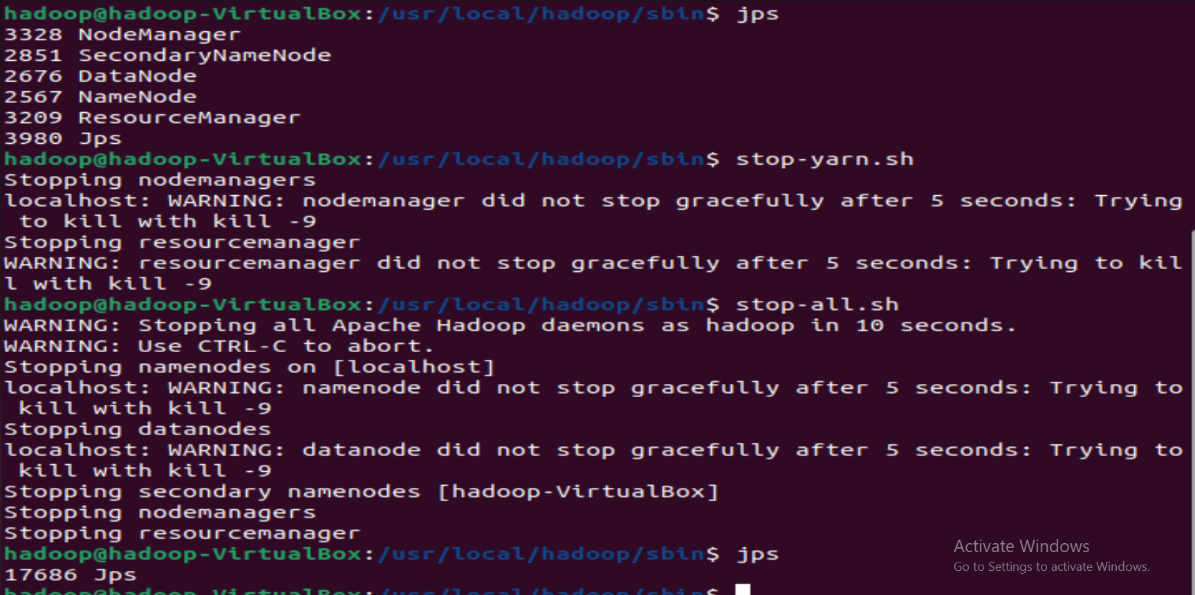
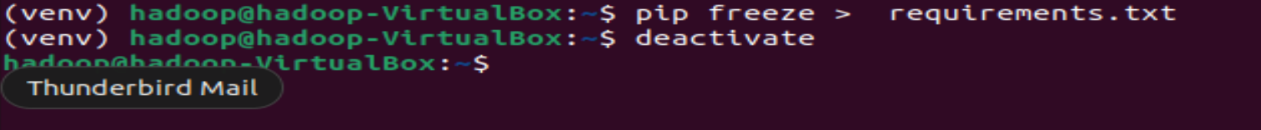
****

Figure : Hadoop Installation

****

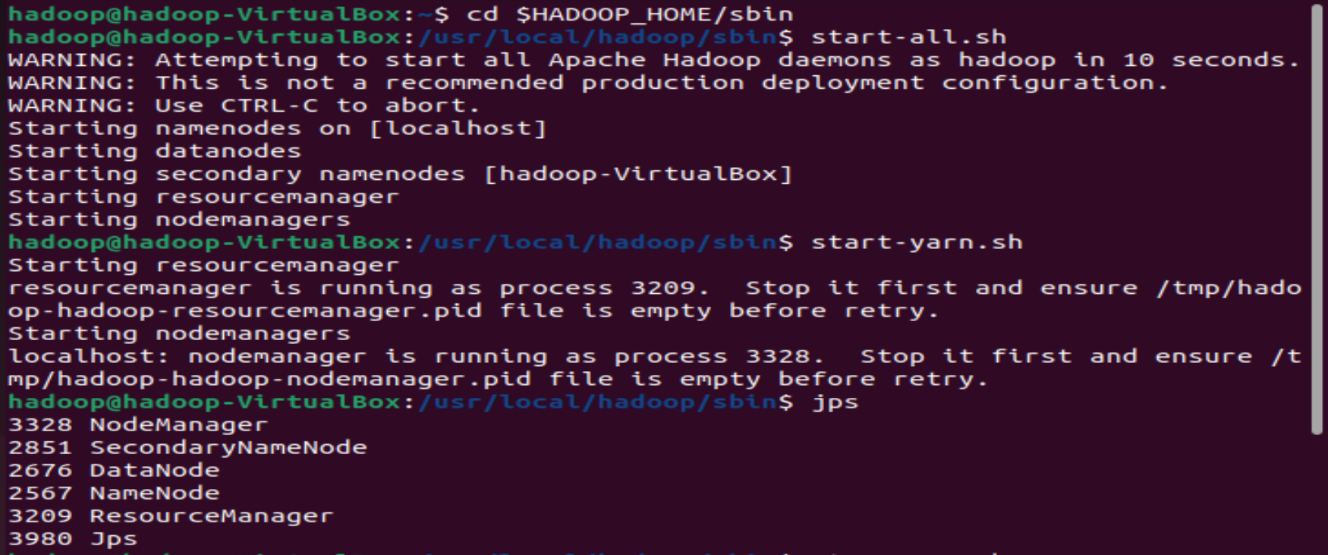
****

Figure : Hadoop Installation: 2



Figure : MongoDB

A screen shot of a computer

Description automatically generated with medium confidence

Figure : Pyspark PreProcessing

A picture containing text, font, screenshot, line

Description automatically generated

Figure : Window Size for Dashboard

A graph with red and blue lines

Description automatically generated with low confidence

Figure : 7 days forecast results

A picture containing text, font, line, plot

Description automatically generated

Figure : One month forecast

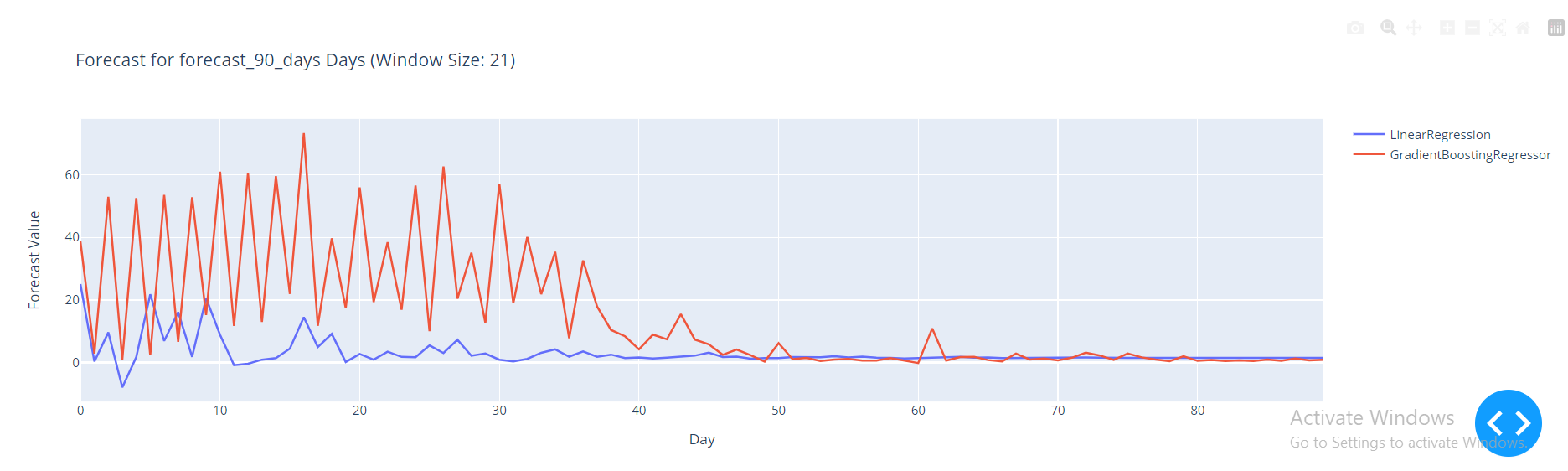


Figure : 3 month forecast

If you see in comparison to 7, 30, and 90 days, there is greater volume of fluctuation for 3 months forecast then it seamlessly smoothens out.

## **4.2 Discussion**

The paper utilizes the mapper reducer function within the Hadoop architecture allowing the data to be basically processed via the reducer and stored into the MySQL db. Post which the data is loaded to the PySpark or Apache spark environment where the generic preprocessing like stemming, stop words removal, hashword removal and lemmatization takes place and finally the data is written over to the MongoDB. This allows the paper to utilize semi-SQL response benefits and fast bulk insertion benefits of MongoDB with respect data sampling. This data is then used for forecasting 7days 1 month and 1 quarter based on window size of 21 with linear regression and gradient boosting architectures.

# **Conclusion**

Finally, the sentiment analysis of the Pfizer vaccine employing a Hadoop MapReduce cluster, MySQL data storage, Apache PySpark for processing and storing information in MongoDB, and integrating linear regression as well as gradient boosting models over time series analysis offers a thorough framework for analysing and forecasting sentiment trends associated to the Pfizer vaccine.

Sentiment analysis is made possible by the Hadoop MapReduce cluster's capacity to analyse massive amounts of data in an evenly distributed and parallel fashion. The data can be stored and retrieved more easily by using an organised and dependable database management system like MySQL. Strong data processing capabilities provided by Apache PySpark enable seamless Hadoop integration as well as quick data reading as well as storage in MongoDB. As a result, data management is adaptable and scalable.

A statistical method to comprehend sentiment trends over time is to include linear regression plus gradient boosting algorithms for time series analysis. Future sentiment trends can be predicted using these models since they capture the connection between sentiment and independent variables.

Overall, by offering insights into public opinion and assisting in the prediction of sentiment patterns, this integrated approach makes it easier to analyze Pfizer vaccine sentiment. To improve sentiment analysis in the setting of the Pfizer vaccine, it combines the scalability of Hadoop, the dependability of MySQL, the adaptability of MongoDB, and the ability to predict of regression and gradient boosting models.

# **References**

1.     Ainin, S., Parveen, F., Moghavvemi, S., Jaafar, N.I. and Shuib, N.L.M., 2014. Determinants of user behaviour and recommendation in social networks. *Industrial Management & Data Systems*, *114*(9), pp.1477-1498.

2.     Garcia, K. and Berton, L., 2021. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Applied soft computing*, *101*, p.107057.

3.     Rolland, Y., Cesari, M., Morley, J.E., Merchant, R. and Vellas, B., 2021. COVID19 vaccination in frail people. Lots of hope and some questions. *The journal of nutrition, health & aging*, *25*, pp.146-147.

4.     Puri, N., Coomes, E.A., Haghbayan, H. and Gunaratne, K., 2020. Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. *Human vaccines & immunotherapeutics*, *16*(11), pp.2586-2593.

5.     Stella, M., Vitevitch, M.S. and Botta, F., 2021. Cognitive networks identify the content of English and Italian popular posts about COVID-19 vaccines: Anticipation, logistics, conspiracy and loss of trust. *arXiv preprint arXiv:2103.15909*.

6.     Castioni, P., Andrighetto, G., Gallotti, R., Polizzi, E. and De Domenico, M., 2022. The voice of few, the opinions of many: evidence of social biases in Twitter COVID-19 fake news sharing. *Royal Society Open Science*, *9*(10), p.220716.

7.     Vilella, S., Semeraro, A., Paolotti, D. and Ruffo, G., 2022. Measuring user engagement with low credibility media sources in a controversial online debate. *EPJ data science*, *11*(1), p.29.

8.     Zarrella, G. and Marsh, A., 2016. Mitre at semeval-2016 task 6: Transfer learning for stance detection. *arXiv preprint arXiv:1606.03784*.

9.     Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š. and Sedlmair, M., 2018. More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, *12*(2-3), pp.140-157.

10.  Dubey, A.D., 2021. Public sentiment analysis of covid-19 vaccination drive in india. *Available at SSRN 3772401*.

11.  Liu, S. and Liu, J., 2021. Understanding behavioral intentions toward COVID-19 vaccines: theory-based content analysis of tweets. *Journal of Medical Internet Research*, *23*(5), p.e28118.

12.  Du, J., Luo, C., Shegog, R., Bian, J., Cunningham, R.M., Boom, J.A., Poland, G.A., Chen, Y. and Tao, C., 2020. Use of deep learning to analyze social media discussions about the human papillomavirus vaccine. *JAMA network open*, *3*(11), pp.e2022025-e2022025.

13.  file:////Users/sheebamoghal/Downloads/mustafa work/pfizer/S. Martin, E. Kilich, S. Dada, P.E. Kummervold, C. Denny, P. Paterson, et al.

14.  Preda, G. (2021). Pfizer Vaccine Tweets [Data file]. Available at: <https://www.kaggle.com/datasets/gpreda/pfizer-vaccine-tweets> (Accessed: 25 May 2023).

15.  Torlone, R., 2020. Hadoop & Map-Reduce.

16.  Apache Software Foundation. (Year). Hadoop Documentation. Available at: <https://hadoop.apache.org/documentation> (Accessed: 25 May 2023).

17.  Hortonworks. (Year). Hortonworks Documentation. Available at: <https://docs.cloudera.com/documentation/enterprise/latest.html> (Accessed: 25 May 2023).

18.  Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. OSDI'04: Proceedings of the 6th Symposium on Operating Systems Design and Implementation, 6, 137-150. Available at: <https://static.googleusercontent.com/media/research.google.com/en/archive/mapreduce-osdi04.pdf> (Accessed: 25 May 2023).

19.  Talend. (n.d). What is MySQL? Available at: <https://www.talend.com/resources/what-is-mysql/> (Accessed: 25 May 2023).

20.  Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauly, M., Franklin, M.J., Shenker, S. and Stoica, I., 2012. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Presented as part of the 9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12)* (pp. 15-28).

21.  Apache Software Foundation. (Year). Apache Spark Documentation. Available at: <https://spark.apache.org/documentation.html> (Accessed: 25 May 2023).