Lagged Dependant Variable for Time Series Sentiment Analysis:

Opinions of COVID-19 Vaccines in Tweeter Data

Raúl Martín Sánchez   
Data Analytics  
CCT CollegeWord Count: **2746**  
[rmsryu@gmail.com](mailto:rmsryu@gmail.com)  
<https://github.com/sbs22021/ca2-twitter>

*Abstract*

This paper presents a study of public tweet data sentiment on COVID-19 vaccines from May 2020 to May 2021. In the study, the use of time series forecasting and sentiment analysis leverages the public data source available from the Twitter stream at [archive.org](https://archive.org/details/twitterstream?sort=-publicdate). For the forecast of sentiment, three distinct machine learning models: Linear Regression, SARIMA, and the Prophet (Samal *et al.*, 2019) (a model developed by Facebook), were considered. Furthermore, the labelling process of sentiment analysis utilised RoBERTa (Liu *et al.*, 2019), a replication study of BERT pretraining, which outperforms the original BERT model, for sentiment scoring.

The study found that using features such as Lagged Dependant variables (LDV) with aggregate sentiment scores and the daily retweet rate, the Linear Regression model produced the most accurate predictions, with an RMSE of 0.088. Then, using hyperparameter tuning and considering key vaccine-related events, the Prophet model achieved an RMSE of 0.114. Finally, the SARIMA model, despite having the highest RMSE of 0.144, provided valuable insights due to its capability to handle complex data patterns incorporating trends and seasonality.

The results highlight the power and potential of machine learning in analysing public sentiment regarding critical public health issues such as the COVID-19 vaccine roll-out. Furthermore, these findings have important implications for public health policy, contributing to a better understanding of public sentiment information in tweet data and potentially aiding in formulating strategies to improve vaccine acceptance and uptake.

Keywords: Sentiment Analysis, Time Series, COVID-19, Vaccines, Linear Regression, SARIMA, Prophet Model, RoBERTa

# Introduction

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, was declared a global health emergency by (WHO (World Health Organization), 2020) in January 2020. The pandemic created implications affecting nearly every aspect of life worldwide. It has not only threatened global health, causing millions of infections and deaths, but also disrupted countries’ economies, completely rewriting human social interactions at all levels of education and social structures to an unprecedented scale.

A key measure in the fight against this pandemic has been the development and distribution of vaccines. Throughout 2020, the global scientific community rallied to discover effective vaccines at a pace never seen before in the history of medicine. However, several key events marked these vaccines’ development, trials, and roll-out, each impacting public sentiment.

On March 27, 2020, as the virus surged globally (The Guardian, 2020b) Ireland went into lockdown, a measure shared by many countries at the time. This move, necessitated by the threat of the virus, set the stage for an urgent global need for a vaccine.

The first hope in this “war” against the virus was on May 18, 2020 (The Guardian, 2020a), when Moderna announced the start of its vaccine trials, eliciting positive sentiment towards the potential vaccine.

However, not all developments were positive. On July 21, 2020, (Bastian, 2020) concerns about potential vaccine side effects were raised, influencing public sentiment negatively. Similarly, on September 8, 2020, AstraZeneca (Beasley, 2020) temporarily suspended its vaccine trials due to a participant’s adverse reaction, raising doubts about vaccine safety and efficacy.

As 2020 progressed, more hopeful news arrived. On October 13, 2020, (Pfizer Inc, 2020) Pfizer announced positive preliminary results from their vaccine trials, a significant milestone in the fight against the pandemic. This trend continued when, on December 3, 2020, the UK (Ledford, Cyranoski and Van Noorden, 2020) became the first country to approve the Pfizer-BioNTech vaccine for emergency use, marking a pivotal moment in the global effort to curb the pandemic.

These key events in the COVID-19 vaccine timeline profoundly impacted public sentiment towards vaccines, highlighting the importance of understanding and analysing these sentiments, especially in guiding public health strategies.

# Big Data

This study used the extensive streamed dataset available from archive.org for the period included. In addition, over 300 compressed files in tar and zip formats were collected. Each file contained daily tweet data in compressed bz2 format, which provided individual tweet data in JSON format.

The significant size of the data posed considerable challenges. In addition, the limited bandwidth for downloading from the archive.org site necessitated using parallel processing techniques to simultaneously download multiple files, followed by extraction, processing, publishing, and removal of data files.

Innovative solutions for storage requirements were required for the ingestion phase of the project. For example, on average, a month’s compressed data requires about 70 Gb of space. Which gracefully exceeded the capacity of the available developer machine used in the study (TODO APENDIX). A virtual machine was deployed on Microsoft Cloud Azure to facilitate the required data ingestion and parallelisation of the tasks. Furthermore, three additional HDD disks of 500 Gb were mounted to this VM, for the download, process, and extract folders. These independent disks significantly improved the performance of the read and write operations required for over two weeks period that the process lasted.

The extracted JSON data were pre-filtered to retain tweets pertinent to the study. Keywords and hashtags were used to sift through the enormous dataset and extract information relevant to the investigation with a filter applied for tweets in English. This preselection of JSON files was uploaded onto a Hadoop Cluster for subsequent processing and accessibility. The initial DataLake, with over 3.2 million tweets, underscored the scale and complexity of the data ingestion process for this research.

The application of big data processing techniques, including parallel processing, cloud computing, and distributed storage, thus played a critical role in this study, making it possible to harness the power of millions of tweets for analysis.

## Database

After preparing the Data Lake, the subsequent step involved establishing a database suitable for analysis. Three databases, Redis, MongoDB, and Cassandra DB, were evaluated for their compatibility with the study’s existing dataset.

Redis, an in-memory Key-value store database, provides access to rapid read and write operations. In addition, its schemaless nature is a good fit for the data structure employed in our study. However, one downside to the open-source version is that data is not persistent, (EDUCB, 2023) meaning that service restart would result in data loss. Fortunately, data persistence is a feature offered in the enterprise version of Redis.

MongoDB, on the other hand, is a distributed No-SQL database tailored for document storage using BSON format.

Lastly, Cassandra is another distributed No-SQL database but is column-oriented. Given Cassandra’s data structure, a transformation of the JSON files would be necessary.

To evaluate performance across these databases, YCSB benchmark was conducted on each, focusing on the primary workloads a, b and e (the other workloads were also generated for completeness). The table below presents the results of those three most relevant workloads:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Redis avg. ms** | **Mongo Db avg in ms** | **Cassandra Db Avg. in ms** |
| ***Workload a (50% read, 50 % update)*** | 234.023 | 327.267 | 705.014 |
| ***Workload b (95% reads)*** | 205.1990 | 399.558 | 485.089 |
| ***Workload e (Range Queries)*** | 263.5132 | 486.337 | 468.665 |

**Table 1:** (\*) Average of 10k, 100k and 1Million operations

|  |
| --- |
| **Figure 1:** Workflow a |

# Benchmarking

The objective is to benchmark (Cooper *et al.*, 2010) Redis, MongoDB, and Cassandra for the following workloads.

|  |  |
| --- | --- |
| **Workload** | **Description** |
| A | Update heavy workload: 50% read, 50% update operations. |
| B | Read mostly workload: 95% read, 5% update operations. |
| C | Read only: 100% read operations. |
| D | Read latest workload: New records are read more frequently. |
| E | Short ranges: Range queries over the latest records. |
| F | Read-modify-write: The client read a record, modifies it, and writes back the changes. |

***\*Workload E*** *uses the Zipfian distribution to choose the first key in the range and the Uniform distribution to choose the number of records to scan* (Cooper *et al.*, 2010)

|  |
| --- |
| **Figure 2:** Seonality anlysis of sentiment |

The testing strategy will be applied to the different workload loads on these databases with operation counts of 10,000, 100,000, and 1,000,000 per iteration.

## Reproducibility and test conditions:

The test environment will be a docker container to ensure the isolation and reproducibility of the test environment.

The host computer is a Dell XPS 13, Intel(R) Core(TM) i7-8565U CPU @ 1.80GHz (4 cores, 8 logical processors), 16 Gb RAM @2133MHT and 500 Gb SSD.

Docker images:

|  |  |  |  |
| --- | --- | --- | --- |
| **Container Name** | **Docker Image** | **Base Image** | **Repo / Documentation** |
| YCBS | ycbs-loadgen:latest | maven:3.6.1-jdk-8-slim | [BlueMedoraPublic/docker-ycsb: YCSB in docker (github.com)](https://github.com/BlueMedoraPublic/docker-ycsb) |
| Redis | redis/redis-stack-server:6.2.6-v7 |  | [Run Redis Stack on Docker | Redis](https://redis.io/docs/stack/get-started/install/docker/) |
| Mongo | mongodb/mongodb-community-server:6.0.4-ubuntu2204 |  | [Docker & MongoDB | Containers & Compatibility | MongoDB](https://www.mongodb.com/compatibility/docker) |
| Cassandra | cassandra:4.1.1 |  | [cassandra - Official Image | Docker Hub](https://hub.docker.com/_/cassandra) |

The test starts with setting up the database and ensuring it runs optimally for each database. Then each workflow load into the database using YCSB’s load functionality is executed with the results recorded as per Appendix V.

From these results, it is clear that Redis demonstrates the fastest execution times across all workloads, but it has the caveat of non-persistence in its open-source variant. Meanwhile, MongoDB also performed reasonably well, outperforming Cassandra in workloads A and B. Moreover, MongoDB’s native support for BSON formatted document storage it is an excellent fit for JSON-like document storage, eliminating the need for transformations required by Cassandra. MongoDB characteristics, performance results, and architecture fit this study perfectly.

# Data Preparation

Over 3.4 million tweets were processed and stored in a Mongo DB database. This data constitutes the initial collection of tweets obtained from the data lake after applying specific filters based on relevant hashtags and keywords (See [Appendix I](#_Appendix_I_–)). In addition, data preparation at this stage involved an additional filter to isolate tweets containing variations of the term “vaccine.” This resulted in approximately 332k tweets for other sentiment and time-series analysis.

Retweets represent an essential feature within this dataset. As indicated in (Metaxas, 2017), retweets are commonly viewed as endorsements or expressions of agreement with the original message. Given the study’s focus on vaccine sentiment analysis, capturing this information is relevant. In addition, it enables the amplification or attenuation of sentiments based on the popularity or frequency of reposting original messages.

The labelling process for the dataset utilised RoBERTa (Liu *et al.*, 2019), a pre-trained model with an extensive 160GB text. The robustness of RoBERTa stems from a unique masking process during its development. By exposing the model to a random subset of tokens in each input sentence, RoBERTa (Nikolopoulos, 2023) remains resilient against input data changes. In addition, a distinct “No-Mask-Left-Behind” technique was applied, ensuring all tokens were masked at least once during training for better sentence representation and more accurate sentiment analysis.

Before sentiment labelling, certain elements, such as usernames and URLs were removed from the tweets keeping the integrity of the remaining text content. The final sentiment score value generated ranges between -1 and 2 with a range [-1,0) for negative, (0,1] neutral and (1-2] positive.

## Trends and key events

Previously highlighted key events were analysed to identify patterns and trends that could reveal the seasonality and existing trends in the dataset’s sentiment. In Figure 2 above, the sentiment time series is presented against these key events, generating trends between them to comprehend better the impact these events had on public sentiment.

These events, categorised as positive or negative, often align with trend shifts, signifying the prevailing trend’s end and suggesting they substantially influence public sentiment, directly affecting the direction of sentiment trends.

## Seasonality

Given the insights provided during this analysis, it is visible as per **Figure 2**, season periods of 1 and 2 months, from left to right, this is 1-2-2-1-2-2. With this in mind, features of 28 and 56 days lagging periods representing this seasonality would be added for a simple linear regression model. Although linear regression models do not directly handle seasonality, including these calculated features, which are found cyclical in the dataset, can allow the linear regression model to account for these seasonal effects.

For the preparation of the SARIMA model (Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors model), although it is not strictly required, autocorrelation plots and stationary tests were conducted on the dataset, with no further transformation required as both conditions were satisfied.

|  |
| --- |
| **Figure 3:** Sentiment score autocorrelation |

# Machine Learning Models

## Linear Regression

The Linear Regression model obtained the best performance of the three models evaluated in this study, registering the lowest Root Mean Square Error (RMSE) (Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman, 2021) of 0.088 for the forecast of the test dataset. This model also surpassed others in the one-week and one-month period, as depicted in ***Table 2.***

The Linear Regression model incorporated eight independent variables. The retweet rate and seven LVDs with sentiment scores on different lagging periods ranging from 1 to 84 days.

The author of (Wilkins, 2018) suggests that regression with the proper lag structure often gives the best estimates of the independent variable’s effect. However, they also indicate paying close attention to autocorrelation in the error term, which may necessitate additional lags of the dependent variable for accurate coefficient estimation and confirming stationary of the dataset (**See Figure 3**).

Incorporating these features, reflecting rolling sentiment scores at diverse lagged timeframes, along with the retweet\_rate for each day, contributed to the model’s superior performance. The lower RMSE of the Linear Regression model indicates its robustness in accurately predicting the sentiment score based on historical data.

By order of importance, the top 3 features’ coefficients values

1. **sentiment\_score\_lag\_1**: 0.344
2. **retweet\_rate**: 0.15
3. **sentiment\_score\_lag\_28**: 0.096

All else being equal, a one-unit change in 1-day lag variableleads to approximately a 0.344 change in the dependent variable, followed by the retweet rate and the 28-days lag period.

This highlights the efficacy of relatively simple models in generating accurate forecasts. Despite their simplicity, these models can produce results that are on par with, or in this case, superior to, their more complex counterparts.

## SARIMA

The SARIMA model (M. Pirani *et al.*, 2022), despite being a powerful tool for time series analysis, has obtained the highest RMSE of the three models (0.144). This model was fine-tuned using the autoarima function from the pmdarima python library, which helped identify the best parameters for the model (Order: p = 1, d = 0, q = 1; Seasonal Order P = 1, D = 1, Q = 1; Seasonal period = 28). Although this model had the highest RMSE, it is essential to note that it considers both trends and seasonality, which is important for capturing complex patterns in data.

## PROPHET

The Prophet model, developed by Facebook (Taylor and Letham, 2018), performed second best with an RMSE of 0.114 for all test data. This model was improved using hyperparameter tuning, considering the critical events as cutoff periods. Those days are key vaccine dates and distributions that occurred during the period in the study.

The best params selection reflected the model’s ability to adapt to trend changes and seasonality, indicating its flexibility and versatility.

This model introduces two main components for producing forecasts. The first component is a model developed over many iterations of forecasting various data at Facebook, and the second component is a system for measuring and tracking forecast accuracy. It uses (Harvey and Peters, 1990) a decomposable time series model with three components: trend, seasonality, and holidays. Described in the following equation:

y(t) = g(t) + s(t) + h(t) + ϵt

|  |
| --- |
| **Figure 4:** Models comparison all test data |

# Discussion

|  |  |  |  |
| --- | --- | --- | --- |
| RMSE | **One week** | **One month** | **All** |
| ***LR*** | 0.073 | 0.081 | 0.088 |
| ***Prophet*** | 0.093 | 0.101 | 0.114 |
| ***SARIMA*** | 0.123 | 0.133 | 0.144 |

***Table 2:*** *(\*) RMSE comparison*

This study comprehensively analyses public sentiment towards COVID-19 vaccines, expressed through tweets from May 2020 to May 2021. It was leveraging the shared Twitter data stream from archive.org and applying time series forecasting and sentiment analysis from a different perspective.

When it comes to Big Data, it is important to select carefully the correct tooling and software. In this study, using Hadoop for building a Data Lake, has played a significant role in withstanding over 3.4 million tweets and files.

Access to this raw data for analysis purposes requires transformation, reshaping, and persistence for future analysis. Reviewed the options and confronting benchmark results for the database in scope, Redis, Mongo DB and Cassandra DB. MongoDB was selected as an adequate database. It integrates with Spark, which has also played two important roles: the data ingestion process to bring data from Hadoop into MongoDB and the analytical role for machine learning models. All of them, accompanied by the sentiment labelling process, that was facilitated by RoBERTa.

Without undermining traditional statistical methods, trends and seasonality were associated with pre-selected cutoff points from important positive and negative news/events during the pandemic. While this started as a requirement for Linear Regressions to understand those trends and seasons, it turned out to be a runner in the final results, resulting in the best of the three models. Also, influences decisions on the bests seasonality period for SARIMA, and cutoffs days as cross-validation periods and hyperparameter tuning of the prophet model.

The findings underscored the superiority of the Linear Regression model in forecasting accuracy, as evidenced by the lowest RMSE of 0.088. This model incorporated features such as lag periods with aggregate sentiment scores and the daily retweet rate. In addition, SARIMA and Prophet, despite having higher RMSE results, delivered valuable insights due to their adept handling of complex data patterns to incorporate trends and seasonality.

The top three features contributing to the performance of the Linear Regression model were ‘sentiment\_score\_lag\_1’ (coefficient: 0.344), ‘retweet\_rate’ (coefficient: 0.15), and ‘sentiment\_score\_lag\_28’ (coefficient: 0.096). This suggests that recent sentiment and the retweet rate were significant influencers in predicting future sentiment.

Understanding public sentiments and successfully predicting them has enormous implications for helping public health policymakers, as well as understanding public sentiment through such data can guide strategies to boost vaccine acceptance and uptake. It is also worth noting that simpler models like Linear Regression, despite their ease of use and implementation, can also bring accurate predictions and perform as well, if not better, than their more complex counterparts.

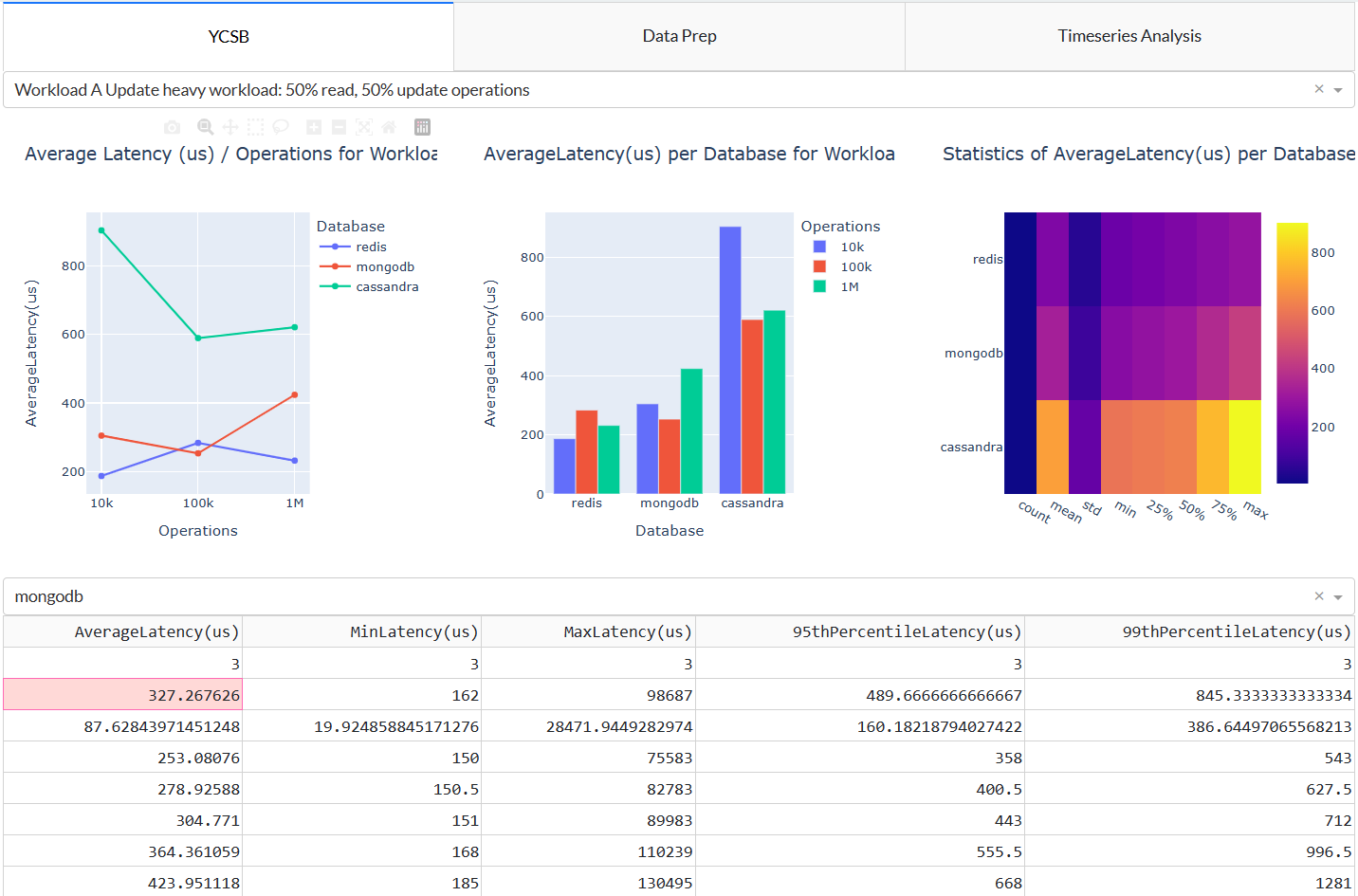
# References

1. Bastian, H. (2020) ‘Covid-19 Vaccines With “Minor Side Effects” Could Still Be Pretty Bad’, 21 July. Available at: https://www.wired.com/story/covid-19-vaccines-with-minor-side-effects-could-still-be-pretty-bad/.
2. Beasley, D. (2020) ‘AstraZeneca suspends leading COVID-19 vaccine trials after a participant’s illness’, 8 September. Available at: https://www.reuters.com/article/us-health-coronavirus-astrazeneca-idUSKBN25Z392.
3. Cooper, B.F. *et al.* (2010) ‘Benchmarking cloud serving systems with YCSB’, in *Proceedings of the 1st ACM symposium on Cloud computing*, pp. 143–154. Available at: https://courses.cs.duke.edu/fall13/cps296.4/838-CloudPapers/ycsb.pdf.
4. Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman (2021) ‘The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation’, *PeerJ Computer Science*, 7, pp. e623–e623. Available at: https://doi.org/10.7717/peerj-cs.623.
5. EDUCB (2023) ‘EDUCB - Redis Persistence’, *EDUCB - Redis Persistence*. Available at: https://www.educba.com/redis-persistence/ (Accessed: 14 May 2023).
6. Harvey, A.C. and Peters, S. (1990) ‘Estimation procedures for structural time series models’, *Journal of forecasting*, 9(2), pp. 89–108.
7. Ledford, H., Cyranoski, D. and Van Noorden, R. (2020) ‘The UK has approved a COVID vaccine — here’s what scientists now want to know’, 3 December. Available at: https://www.nature.com/articles/d41586-020-03441-8.
8. Liu, Y. *et al.* (2019) ‘RoBERTa: A Robustly Optimized BERT Pretraining Approach’.
9. M. Pirani *et al.* (2022) ‘A Comparative Analysis of ARIMA, GRU, LSTM and BiLSTM on Financial Time Series Forecasting’, in *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*. *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, pp. 1–6. Available at: https://doi.org/10.1109/ICDCECE53908.2022.9793213.
10. Metaxas, P. (2017) ‘Retweets indicate agreement, endorsement, trust: A meta-analysis of published Twitter research’, *Computer Science Department: Wellesley College* [Preprint]. Available at: http://cs.wellesley.edu/~pmetaxas/WorkingPapers/Retweet-meaning.pdf.
11. Nikolopoulos, S. (2023) ‘Sentiment Analysis using Twitter API and RoBERTa model’, 9 February. Available at: https://www.odbms.org/2023/02/sentiment-analysis-using-twitter-api-and-roberta-model/.
12. Pfizer Inc (2020) ‘Pfizer and BioNTech Announce Positive Early Data From Clinical Trial of Omicron BA.4/BA.5-Adapted Bivalent Booster in Individuals 18 Years and Older’, 13 October. Available at: https://www.businesswire.com/news/home/20221010005742/en/.
13. Samal, K.K. *et al.* (2019) *Time Series based Air Pollution Forecasting using SARIMA and Prophet Model*, *ITCC 2019: Proceedings of the 2019 International Conference on Information Technology and Computer Communications*, p. 85. Available at: https://doi.org/10.1145/3355402.3355417.
14. Taylor, S.J. and Letham, B. (2018) ‘Forecasting at scale’, *The American Statistician*, 72(1), pp. 37–45. Available at: https://doi.org/10.7287/peerj.preprints.3190v2.
15. The Guardian (2020a) ‘Moderna: First human trial results raise hopes for coronavirus vaccine’, 18 May. Available at: https://www.theguardian.com/society/2020/may/18/first-human-trial-results-raise-hopes-for-coronavirus-vaccine.
16. The Guardian (2020b) ‘“Stay home”: Varadkar announces sweeping two-week lockdown’, 27 March. Available at: https://www.theguardian.com/world/2020/mar/27/stay-home-varadkar-urges-irish-in-drastic-lockdown.
17. WHO (World Health Organization) (2020) ‘Timeline: WHO’s COVID-19 response’. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline#!
18. Wilkins, A.S. (2018) ‘To Lag or Not to Lag?: Re-Evaluating the Use of Lagged Dependent Variables in Regression Analysis’, *Political Science Research and Methods*. 2017/05/03 edn, 6(2), pp. 393–411. Available at: https://doi.org/10.1017/psrm.2017.4.

**END**

# Appendix

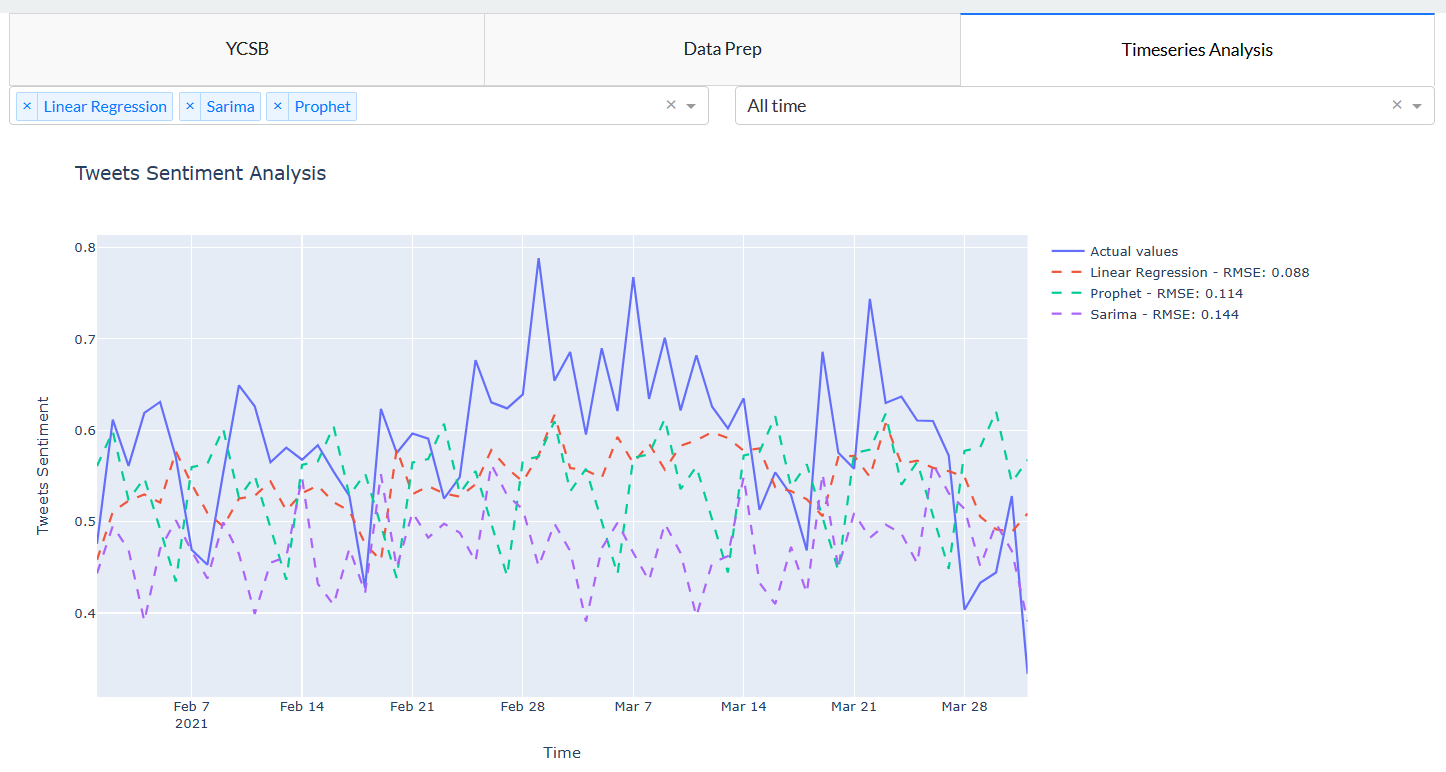
# Appendix I – Dashboard YCSB



# Appendix II – Dashboard Data Preparation



# Appendix III – Dashboard Time Series Models



# Appendix IV – Keywords

|  |  |
| --- | --- |
| **Hashtags** | **Keywords** |
| #VaccineResearch | Coronavirus |
| #Pandemic | Vaccination rate |
| #Sinovac | Johnson & Johnson |
| #Moderna | Immunisation |
| #VaccineDistribution | Public health |
| #Immunology | Vaccine |
| #VaccineSupplyChain | Moderna |
| #AntibodyTests | AstraZeneca |
| #Janssen | GlaxoSmithKline |
| #VaccineHesitancy | Vaccine mandate |
| #Vaccine | Booster shot |
| #BoosterShot | Vaccine hesitancy |
| #VaccineManufacturing | Novavax |
| #VaccineNationalism | Vaccine passport |
| #Novavax | Pfizer |
| #COVID19 | Sinovac |
| #AstraZeneca | Vaccine rollout |
| #DiagnosticTests | Vaccination |
| #COVIDVaccine | Pandemic |
| #GetVaccinated | COVID-19 |
| #VaccineTrials | BioNTech |
| #PCRTests | COVID vaccine |
| #Vaccination | Sinopharm |
| #Pfizer | Sanofi |
| #VaccineStorage |  |
| #HerdImmunity |  |
| #Coronavirus |  |
| #Sinopharm |  |
| #VaccineMandate |  |
| #BioNTech |  |
| #VaccineSafety |  |
| #Sanofi |  |
| #VaccineEfficacy |  |
| #SerologyTests |  |
| #JohnsonAndJohnson |  |
| #StaySafe |  |
| #VaccineProduction |  |
| # Immunisation |  |
| #VaccineRollout |  |
| #PublicHealth |  |
| #ClinicalTrials |  |
| #GSK |  |
| #VaccineDevelopment |  |
| #VaccinePassport |  |

# Appendix IV – YCSB Logs

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **WORKLOAD** | **A** | | | **B** | | | **C** | | | **D** | | |
|  | **10k** | **100k** | **1M** | **10k** | **100k** | **1M** | **10k** | **100k** | **1M** | **10k** | **100k** | **1M** |
| **Redis** |  |  |  |  |  |  |  |  |  |  |  |  |
| **Mongo** |  |  |  |  |  |  |  |  |  |  |  |  |
| **Cassandra** |  |  |  |  |  |  |  |  |  |  |  |  |
| **WORKLOAD** | **E** | | | **F** | | |  | | | | | |
|  | **10k** | **100k** | **1M** | **10k** | **100k** | **1M** |
| **Redis** |  |  |  |  |  |  |
| **Mongo** |  |  |  |  |  |  |
| **Cassandra** |  |  |  |  |  |  |