*Digital Health and Informatics Innovations for Sustainable Health Care Systems*
*J. Mantas et al. (Eds.)*


*doi:10.3233/SHTI240824*

# Deep Learning-Based Prediction of Daily COVID-19 Cases Using X (Twitter) Data

Nourhan AHMED [a], Khansa SAEED [a] , Jeevitha Lora RODRIGUES [a], Maha NAEEM [a], Andrea CORREA [a] , Chairungroj SANABBOON [a], Shareh ROSTAM NIAKAN KALHORI [b,c,1] and Thomas M. DESERNO [b]
[a] *Information Systems and Machine Learning Lab, Department of Mathematics, Natural Science, Economics and Computer Science, Institute of Computer Science, University of Hildesheim*
[b] *Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Braunschweig, Germany*
[c] *Department of Health Information Management, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran*
ORCiD ID: Nourhan AHMED https://orcid.org/0000-0002-2689-7069, Khansa SAEED https://orcid.org/0009-0006-0541-9806, Jeevitha Lora RODRIGUES https://orcid.org/0009-0004-3419-9975, Maha NAEEM https://orcid.org/0009-0003-2387-5272, Andrea CORREA https://orcid.org/0009-0009-8463-5719, Chairungroj SANABBOON https://orcid.org/0009-0008-1759-6118, Shareh Rostam Niakan Kalhori https://orcid.org/0000-0002-7577-1200, Thomas Deserno https://orcid.org/0000-0003-3492-4407

**Abstract.** Due to the importance of COVID-19 control, innovative methods for predicting cases using social network data are increasingly under attention. This study aims to predict confirmed COVID-19 cases using X (Twitter) social network data (tweets) and deep learning methods. We prepare data extracted from tweets by natural language processing (NLP) and consider the daily G-value (growth rate) as the target variable of COVID-19, collected from the worldometer. We develop and evaluate a time series mixer (TSMixer) predictive model for multivariate time series. The mean squared error (MSE) loss on the test dataset was 0.0063 for 24-month Gvalue prediction when using the MinMax normalization with recursive feature elimination (RFE) and average or min aggregation method. Our findings illuminate the potential of integrating social media data to enhance daily COVID-19 case predictions and are applicable also for epidemiological forecasting purposes.

**Keywords.** COVID-19, TSMixer, deep learning, predictive models, social network

## 1. Introduction

The COVID-19 pandemic has catalyzed extensive research aimed at understanding its dynamics, forecasting its spread, and mitigating its impact [1]. Several studies utilize statistical and machine learning models to forecast the dynamics of cumulative COVID-19 cases across different countries [1,2]. Kumar et al. employ auto regressive-integrated moving average (ARIMA) and seasonal auto regressive-integrated moving average

[1] Corresponding Author: Sharareh Rostam Niakan Kalhori, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran, Fax: +982188983517; E-mail: sharareh.niakankalhori@plri.de.

(SARIMA) models to forecast COVID-19 trends for the top 16 countries. By selecting optimized model parameters and evaluating metrics, they highlight the exponential rise in confirmed and recovered cases in several countries [3]. Similarly, Chaurasia & Pal utilize ARIMA and regression models to predict worldwide COVID-19 death cases revealing a decline in death cases after May 2020 [4]. In contrast, Sardar et al. proposed an autoregressive-modeling framework based on machine learning (ML) and statistical methods to forecast confirmed COVID-19, finding ARIMA as the most suitable model [5]. However, these methods are limited to huge datasets [2] like tweets. Due to the temporal nature, high volume, nonlinearity, and variety of social media data, advanced methods like TSMixer are necessary for time-series analysis. TSMixer combines the strengths of linear models with the flexibility of nonlinear approaches. It enhances predictive capabilities by integrating multi-layer perceptrons (MLPs) and using deep learning techniques like normalization and residual connections. TSMixer effectively handles cross-variate information by alternating MLP applications between time and feature domains, inspired by the MLP-Mixer architecture from computer vision. This method captures temporal dependencies and cross-variate insights [6]. Hence, this study examines TSMixer for predicting the daily-confirmed COVID-19 cases based on G-values in conjunction with COVID-19-related keywords extracted from X social network.

## 2. Material and Methods

### *2.1. Data Sources*

Our experiment involves two distinct datasets: (i) the G-values (growth rate) of daily-confirmed COVID-19 cases and (ii) globally collected X tweets in three months, from January 1 to March 31, 2021. We take the instructions for data collection from X for academic and research purposes [2]. We structure the data based on date-time (year˙month˙day hour˙minute˙second) in the format 2021-01-03 00:00:57 and encompass other variables too. Similar to the World Health Organization (WHO) reporting COVID-19 cases weekly[3], we design and aggregate the data of COVID-19 tweets in a weekly format. This approach captures the typical disease progression cycles more effectively since COVID-19 symptoms and the resultant testing and reporting often follow a weekly cycle due to work and lifestyle patterns. The G-value data consists of 90 data points aligned with the X-driven dataset. Another two-year dataset covers a broader temporal scope, spanning from 01.01.2021 to 31.12.2022, comprising a total of 730 data points (Fig.1).

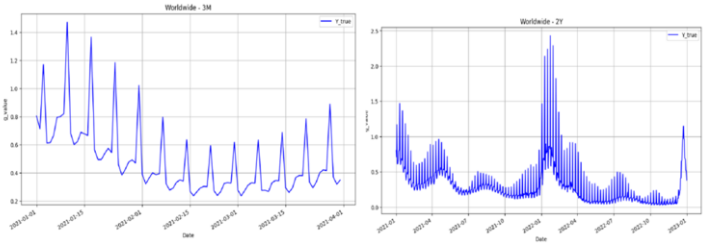


**Figure 1.** G-values on three months (left) and two years (right)

[2] https://developer.twitter.com/en/use-cases/do-research/academic-research
[3] https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---1-september-2023

### 2.2. Data Preparation and Experimental Steps

Fig. 2 presents the detailed implementation of the word-embedding pipeline, which includes merging tweets for each day, translating words into embeddings, and using aggregating approaches to produce a 500-dimensional vector for each day. The text-cleaning task is crucial for a variety of reasons, including noise removal, improved model performance, allowing it to focus on significant patterns and correlations in the data, and standardization, which ensures that it has a consistent format. For data processing, managing dataset columns and rows, lemmatization, and emojis removal, we employ the Python libraries pandas, Spacy, Natural Language Toolkit (NLTK), and demoji, respectively.

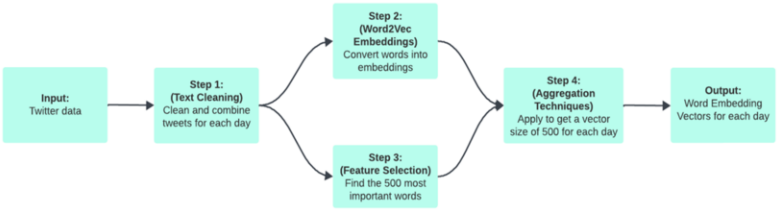


**Figure 2.** Pipeline of preparation and word embedding

We clean-up the data collected from X for two main purposes: frequency analysis and G-value prediction. The frequency analysis generates a list of keywords and calculates their frequency (number and percentage) over time (per day). For this purpose, we filter the data to retain only relevant columns such as tweet content, creation timestamp, language, author ID, and engagement metrics. We clean the tweet content, convert it to lowercase, and remove spaces, emojis, hashtags, links, tags, numbers, and other special characters. We further apply tokenization, stop-word removal, and lemmatization. Furthermore, we calculate and add the column "freq count".

For G-value calculation, the copious amount of available X data serves as a valuable resource for analyzing and quantifying the collective emotional responses of the general population toward COVID-19 on a global scale. One critical metric in this analysis is the proposed parameter Gvalue (7), which corresponds to the global growth rate in the total number of confirmed COVID-19 cases on a rolling-forward daily basis. This parameter, denoted as $G_t$, can be mathematically expressed as:

$$G_t = \left(\left(Y_t - Y_{\{}t-1\}\right)/Y_{\{}\,t-1\}\right) * 100\%$$

where $Y_t$ and $Y_{\{}t-1$ represent the global number of confirmed COVID-19 cases at time $t$ and $t-1$ (the previous day), respectively. For the G-value prediction, we prepare the tweets as input for different models and aggregate the final records by date, resulting in a single record containing the combined tweet words generated on that date. Following the cleansing, we keep the records for one day per file. Afterward, we train Word2Vec embedding to rebuild the linguistic contexts for the words and then, we create the model using the Python library Gensim. Hyper-parameters such as min count, window, and vector size customize the model for the dataset. We ignore words with frequencies below the threshold of 100 from the Word2Vec model. Implying that the algorithm considers two words before and two words following the current word, we set the window size to two and the vector size to 500. We train the Word2Vec model with the tokenized X words merged every day. After training, the model retrieves word vectors from the X corpus that we use for the subsequent steps of Gvalue predictive model development. For feature selection, we apply Kbest and recursive feature elimination (RFE). Following

feature selection, we apply aggregation techniques such as sum, min, max, and average to purify meaningful information from the selected word vectors. For developing the predictive model through an 80/20 train-test split, we use a time series mixer (TSMixer) (6) for multivariate data in a time-series forecasting task developed in Python and several key libraries, including TensorFlow, Pandas, and NumPy. The training loop runs for 100 epochs, with early stopping implemented to prevent overfitting. We conduct experiments alongside analyses of normalization methods including MinMaxScaler and StandardScaler. We consider evaluation metrics including mean absolute error (MAE) and mean squared error (MSE) loss.

## 3. Results

COVID-19, corona, virus, vaccine, work, and death are among the 20 words with the highest count generated from this dataset (Table 1). The transformation of this text-based data into a frequency dictionary results in a file containing 1,167,386 and 47,548 distinct words daily before and after thresholding (frequency > 100) respectively.

**Table 1.** The top 20 words with the highest frequency in the text of the tweets

| | Word | Sum | | Word | Sum |
|---|---|---|---|---|---|
| 1 | Covid | 20,874,100 | 11 | Health | 1,144,499 |
| 2 | Vaccine | 4,161,461 | 12 | Test | 1,027,660 |
| 3 | People | 2,782,086 | 13 | Mask | 988,618 |
| 4 | Case | 1,888,316 | 14 | Vaccination | 892,263 |
| 5 | Death | 1,848,497 | 15 | Virus | 867,263 |
| 6 | Coronavirus | 1,767,408 | 16 | Week | 856,331 |
| 7 | Year | 1,639,695 | 17 | Corona | 846,460 |
| 8 | Day | 1,417,993 | 18 | Today | 835,362 |
| 9 | Time | 1,393,081 | 19 | Work | 797,244 |
| 10 | Pandemic | 1,376,676 | 20 | Country | 778,136 |

The most notable outcomes reveal a minimum MSE of 0.0176 and MAE of 0.0602 of the TSMixer model, through the implementation of StandardScaler normalisation, RFE in conjunction with the min aggregation technique for three-month Gvalue prediction (Fig.3). However, for 24-month Gvalue prediction, MinMax normalization with RFE feature selection and ave or min aggregation method outperforms other settings as there is MSE of 0.0063 and MAE of 0.068 (Fig.4). In both Fig. 3 and 4 the right graphs show the higher performance result.

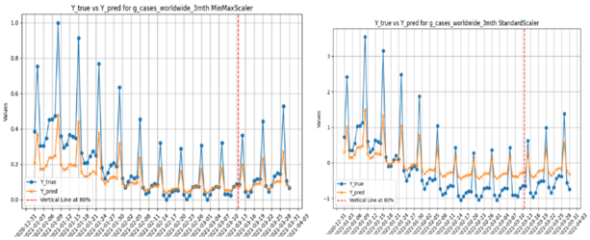

**Figure 3.** G-value prediction by TSMixer for three months with MinMaxScaler (left) and StandardScaler (right)

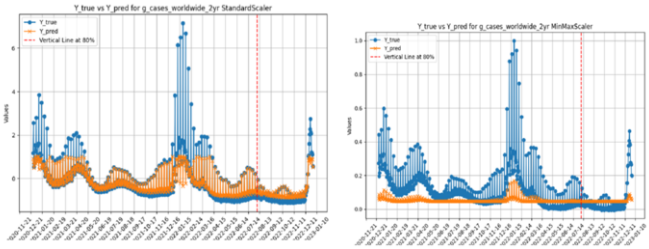


**Figure 4.** G-value prediction by TSMixer for 24 months with StandardScaler (left) and MinMaxScaler (right)

## 4. Discussion and Conclusions

The study demonstrates the prediction of COVID-19 cases by utilizing X social network data alongside deep learning approaches. TSMixer emerged promising for 3 and 24-month data through various experiments based on feature selection, aggregated vectors, and normalisation mechanisms. Results show different performances across various settings although the aggregation methods performed comparable. For such large data extracted from X and G values, the normalized-MinMax setting and RFE consistently outperform in terms of lower error. Although the X platform provides high volumes of data, it may not fully represent broader society, potentially limiting the data generalizability. While social media platforms are valuable for real-time insights, caution is needed when interpreting related findings to truly understand COVID-19 activity.
The study provides a promising avenue for generating accurate and timely predictions of COVID-19 cases. We understand the role of social media in epidemiological prediction providing insights for public health authorities and policymakers aiming to strengthen their forecasting capabilities during global health crises.

## References

[1] Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data in brief. 2020;29:105340.
[2] Chew AWZ, Pan Y, Wang Y, Zhang L. Hybrid deep learning of social media big data for predicting the evolution of COVID-19 transmission. Knowledge-Based Systems. 2021;233:107417.
[3] ArunKumar K, Kalaga DV, Kumar CMS, Chilkoor G, Kawaji M, Brenza TM. Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA). Applied soft computing. 2021;103:107161.
[4] Chaurasia V, Pal S. COVID-19 pandemic: ARIMA and regression model-based worldwide death cases predictions. SN Computer Science. 2020;1(5):288.
[5] Sardar I, Akbar MA, Leiva V, Alsanad A, Mishra P. Machine learning and automatic ARIMA/Prophet models-based forecasting of COVID-19: Methodology, evaluation, and case study in SAARC countries. Stochastic Environmental Research and Risk Assessment. 2023;37(1):345-59.
[6] Chen S-A, Li C-L, Yoder N, Arik SO, Pfister T. Tsmixer: An all-mlp architecture for time series forecasting. arXiv preprint arXiv:230306053. 2023.
[7] Thomas P. J-value assessment of how best to combat COVID-19. Nanotechnology Perceptions. 2020;16(1):16–40-16–40.