Volume: 10 Issue: V Month of publication: May 2022

DOI: https://doi.org/10.22214/ijraset.2022.42800

Web Traffic Time Series Forecasting Using ARIMA Model

Vrushant Tambe[1], Apeksha Golait[2], Sakshi Pardeshi[3], Rohit Javeri[4], Prof. Gajanan Arsalwad[5]
[1, 2, 3, 4, 5]*Department of Information Technology, Trinity College of Engineering and Research, Pune*

Abstract: Web traffic prediction is a major concern since it has the potential to produce severe snags in the working of websites. It is one of the most difficult tasks to make predictions about future time series values, so been a hot topic for research. The increase in web traffic may encounter a crashed site or very slow loading time. Such disturbances may cause many disturbances for the users, consequently decreased users rating of the site and user move to another site that affects the business. We have implemented a forecasting model to predict web traffic. ARIMA model is used for Web traffic time series forecasting. We have used some of the features like page name, date visited, and the number of visits for prediction with higher accuracy.

Keywords: Web traffic prediction, ARIMA model, Time series forecasting, Data Collection and Feature Understanding.

I. INTRODUCTION

People who work for web service providers need to know how much traffic a web server is getting, because if they don't, customers might have long waited and leave the site. However, this is a difficult task because it requires making accurate predictions about how people will act based on their randomness. In this article, we show how to build an architecture that takes source data and uses it to make predictions about how many people are going to see a given page at a given time. Depending on the website's response, web applications handle HTTP GET requests, media apps spread content based on what the user wants, and so on and so forth Request time will have a big impact on how the end-user sees the quality of the service. A lot of people have left a lot of platforms because they took too long to respond. However, the response time is the time between when the application receives the request and when it sends back the answer. This is called the response time. This can't be taken away. In the case of web services, the response time is too long for customers to expect. Developers have been able to figure out when the response time is too long, known as web congestion. A time series with the dates and number of page views make sense for the problem. The purpose of this research is to design a forecasting model to predict web traffic based on the certain features like page name, visited date and the number of visits for pages for a year. As more people gain access to the internet around the world, the increase in traffic to practically all websites have become unavoidable. The increase in website traffic could bring a slew of issues, and the company that is able to deal with the variations in traffic the most effectively will emerge.[7]As most people have experienced a crashed site or a very slow loading time for a website when there are a lot of people using it, such as when various shopping websites may crash just before festivals as more people try to log in to the website than it was originally capable of, causing a lot of inconveniences for the users and as most people have encountered a crashed site or a very slow loading time for a website when there are a lot of people using it, such as when various shopping websites may crash just before festivals as a result, it's possible that users will give the site a lower rating and instead use another site, lowering their business. As a result, a traffic management approach or plan should be implemented to limit the danger of such disasters, which could jeopardise the company's existence. Until recently, there was no need for such tools because most servers could handle the traffic influx. However, the smart phone era has increased demand to such a high level for some websites that businesses have been unable to respond quickly enough to maintain the inconsistent customer service level.

II. LITERATURE SURVEY

During the construction of the prediction model, the system successfully rebuilt the existing model and added new features, resulting in increased model efficiency. New features were used in various combinations.

1) For capturing weekly, monthly, quarterly, and yearly page popularity, use the median of specified window length in each time series as an independent feature.
2) Golden ratio-based median of medians of variable time frame windows.

To determine the importance of each feature, the study [1] analysed the obtained results and compared the ac curacies in various cases. Next, we'll try to figure out how to tweak parameters in an existing model to get better results. Study wanted to find the most suitable forecasting model based on time-series which helps us to forecast future traffic data when there is enough dataset is provided.

Having this goal in mind, study began to search for models based on prediction, which would enable us to predict the value data. However, upon more research, we found that it, not a prediction but rather forecasting, after which we focused on that. Study in [2] came across so many timeseries forecasting models that it made our work both tedious and fun at the same time. Paper proposed a time series forecasting technique to predict internet traffic based on past values using past values. Many forecasting techniques like ARIMA are used extensively in literature for making forecasts, but it is useful mostly for a time series which is linear in nature. On the other hand, neural networks like RNN are very useful in forecasting time series which are nonlinear in nature. Proposed technique uses Discrete Wavelet Transform and using a high pass filter and a low pass filter producing linear and nonlinear parts for the time series. The proposed technique [3] clearly outperforms ARIMA and RNN. And because of the simplicity of the technique, it can be easily employed at data centres. The paper [4] put forward a new engineering approach to prediction of campus network exit-link traffic trend. And it predicts that EPTS can have following effect in network traffic forecasting if having enough historical data. Web Traffic Time Series Forecasting

a) To predict network exit-link traffic trend based on historical network traffic data, so we can layout the network resource planning in advance.

b) It is easy to implement and its computing complexity is acceptable.

The paper [5] compares the traffic flow forecast effects of the LSTM network, BPNN model and ARIMA model on time series captured at a single point. The proposed LSTM network can accurately predict the traffic flow based on the relatively stable time series under normal conditions. However, the traffic system on roads is stochastic and complex, and often affected by abnormal factors like bad weather, traffic accident and large events.

TITLE	PUBLICATION AND AITHOR	TECHNICAL DETAILS
Web Traffic Prediction of Wikipedia Pages	2018IEEE International Conference on Big Data (Big Data) [1] -Navyasree Petluri, Eyhab Al-Masri	In the process of building prediction model, System successfully rebuilt the existing model and added new features to observe improvements in efficiency of model. Applied new features in different combinations 1) Median of specified window length in each time series as an independent feature for capturing weekly, monthly, quarterly and yearly page popularity 2) Median of medians of variable time frame windows based on golden ratio. Study analysed the Obtained result and compared the accuracies in different cases, to know the importance of each feature. As a next step, we will try to work on how to tune parameters in existing model to Observe better results.
Traffic Forecasting using Time-Series Analysis.	2021 6th International Conference on Inventive Computation Technologies (ICICT) [2] - Mohammmad Asifur Rahman Shuvo, Muhtadi Zubair Afsara Tahsin Purnota, Sarowar Hossain, Muhammad Iqbal Hossain	Study wanted to find the most suitable forecasting model based on time-series which helps us to forecast future traffic data when there is enough dataset is provided. Having this goal in mind, study began to search for models based on prediction, which would enable us to predict the value data. However, upon more research, we found that it, not a prediction but rather forecasting, after which we focused on that. Study came across so many time- series forecasting models that it made our work both tedious and fun at the same time.

Predicting Computer Network Traffic: A Time Series Forecasting Approach Using DWT, ARIMA and RNN	2018 Eleventh International Conference on Contemporary Computing (IC3) [3] - Rishabh Madan, Partha Sarathi Mangipudi	Paper proposed a time series forecasting technique to predict internet traffic based on past values using past values. Many forecasting techniques like ARIMA are used extensively in literature for making forecasts, but, it is useful mostly for a time series which is linear in nature. On the other hand, neural networks like RNN are very useful in forecasting time series which are nonlinear in nature. Proposed technique uses Discrete Wavelet Transform and using a high pass filter and a low pass filter producing linear and nonlinear parts for the time series. The proposed technique clearly outperforms ARIMA and RNN. And because of the simplicity of the technique, it can be easily employed at data centres.
An Engineering Approach to Prediction of Network Traffic Based on Time-Series Model	2009 International Joint Conference on Artificial Intelligence. [4] -Fu-Ke Shen, Wei Zhang, Pan Chang	The paper put forward a new engineering approach to prediction of campus network exit-link traffic Trend and it predicts that EPTS can have following effect in network traffic forecasting if having enough historical data. 1) To predict network exit-link traffic trend based on historical network traffic data, so we can layout the network resource planning in advance. 2) It is easy to implement and its computing complexity is acceptable.
Traffic Flow Forecast Through Time Series Analysis Based Deep Learning	2020 IEEE Access [5] - Jianhu Zheng, Mingfang Huang	This paper compares the traffic flow forecast effects of the LSTM network, BPNN model and ARIMA model on time series captured at a single point. The proposed LSTM network can accurately predict the traffic flow based on the relatively stable time series under normal conditions. However, the traffic system on roads is stochastic and complex, and often affected by abnormal factors like bad weather, traffic accident and large events.

III. RELATED WORK

Experts have dismissed neural networks (NNs) as non-competitive all throughout years, and NN aficionados have proposed a slew of new and sophisticated NN architectures, many of which lack solid empirical assessments when compared to simpler univariate statistical methods. Many time series prediction competitions, such as the M3, NN3, and NN5 competitions, backed up this theory [18– 20]. NNs have been labelled as unsuitable for forecasting as a result. The poor performance of NNs in the past could be due to a variety of factors, one of which is that the individual time series were frequently too short to be simulated using advanced methods. Alternatively, the time series characteristics may have changed over time, resulting in even longer time series containing insufficient relevant data to fit a complex model[7,8]. As a result, when using complicated methods to represent series, it's vital that they're the right length and come from a somewhat stable system. Furthermore, NNs are generally chastised for being closed systems. As a result, forecasting specialists have always opted to employ simpler statistical methods [9]. However, we are currently living in a huge data environment. Over the years, businesses have accumulated a large amount of data that provides valuable insight into their business processes. In the context of time series, big data does not automatically imply that each time series contains a lot of information. Rather, they frequently suggest that a given field has a large number of related time series. Univariate prediction algorithms that evaluate individual time series independently may not be reliable in this case.

They become inaccessible in the situation of large data, because a single model can learn from several similar timeseries at the same time. Furthermore, even more advanced models, such as neural networks (NNs), benefit as much as feasible from access to large amounts of data [10,11]. The Recurrent Neural Networks (RNN) play a role in this new field of growing scientific interest in the NN. Results never seen previously in the field of language and time series analysis are beginning to be realised with this new form of neural networks specialising in the sequence prediction problem [18].

RNNs, on the other hand, have major memory issues, which were remedied when the LSTM was introduced into the research field. In addition to the RNN's regular hidden state, this new type of RNN has a new internal memory (cell state). This makes it easier to prevent vanishing or exploding gradient difficulties when training LSTMs [19]. LSTM can be utilised in a predictive setting since time series have seasonality components. If a monthly time series exhibits yearly seasonality, for example, the value of the same exact month the previous year is more useful in predicting the value for the next month. Suilin et al. accomplished a fantastic job on the Kaggle challenge for Wikipedia's web traffic estimate, which included this concept. [20]. Despite the fact that this dataset has been widely utilised for the prediction of time series related to online traffic, it has not been employed in the construction of the LSTM with minimal data since the researchers believe that alternative models, such as ARIMA, are more efficient in these situations. Other study Mathematics 2021, 9, 421 4 of 21 TSF has resulted in more complex prediction algorithms that do not account for the seasonal component. [21]. For multivariate forecasting issues, Qin et al. presented an RNN. Different weights are allocated to the various driving data in this model based on how important they are in contributing to the forecast at each time stage. The authors compared this model to ARIMA, NARX RNN, Encoder Decoder, Attention RNN, Input Attention RNN, and the Dual Stage Attention RNN in order to validate it [22]. The model presented by Qin et al. has been studied more recently. The authors claimed in [23] that their model can cope with the spatial temporal series' fundamental properties. Researchers have been stacking RNN or LSTM to reach the needed outcome in FTS challenges in recent years. However, with FTS prediction models in time series with little data, we discovered a gap in the literature [24]. We suggested a supervised architecture based on LSTM that is trained through distributed data parallelism and follows the Downpour technique due to a gap in the state of the art in the prediction of time series with limited data.

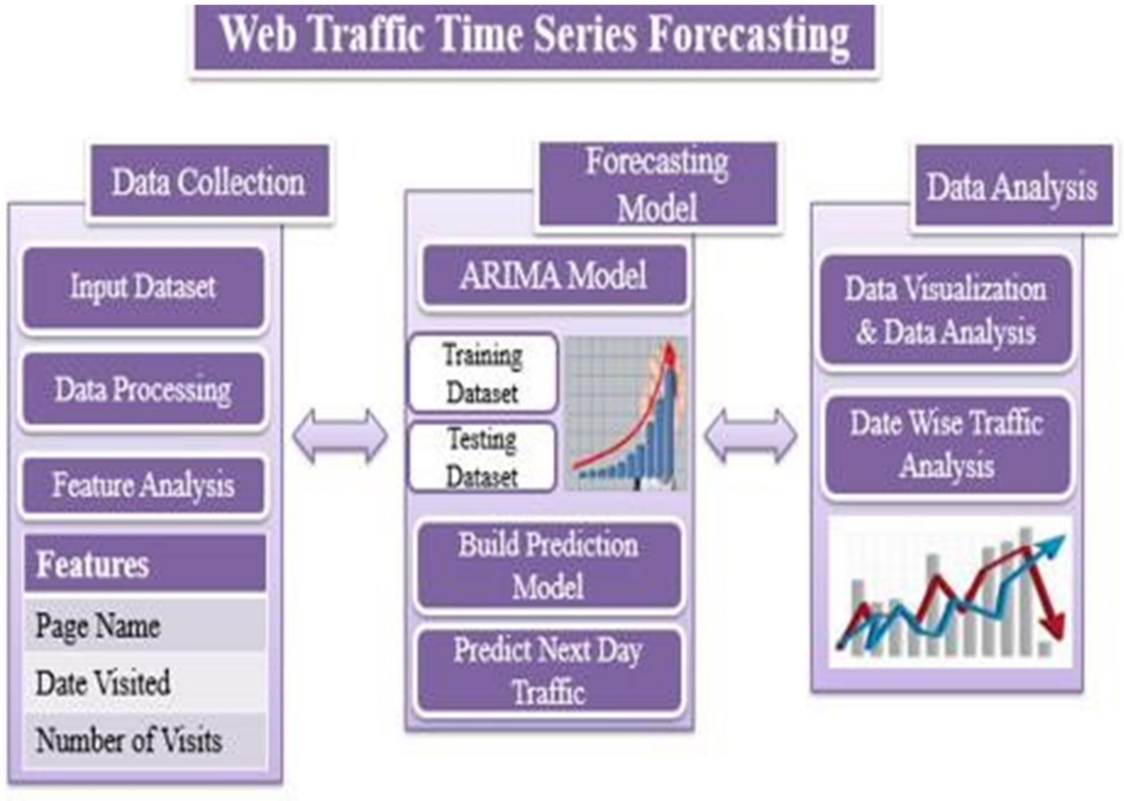

Figure 1: System Architecture

IV. ARIMA MODEL

ARIMA (Auto regressive Integrated Moving Average model) is a statistical analysis technique that uses time series data to better understand or forecast future trends. An autoregressive integrated moving average model is a type of regression analysis that determines how strong one dependent variable is in comparison to other changing variables. The purpose of the model is to anticipate future securities or financial market movements by looking at the discrepancies between values in a series rather than actual values. The full model can be written as,

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q}$$

Where, yt' is the differences series (it may have been differences more than once).

The "predictors" on the right-hand side include both lagged values of yt and lagged errors.

A standard notation would be ARIMA with p, d, and q, where integer values substitute for the parameters to indicate the type of ARIMA model used.

The parameters can be defined as:

1. p - order of the auto regressive part
2. d- degree of first difference involved
3. q - order of the moving average part

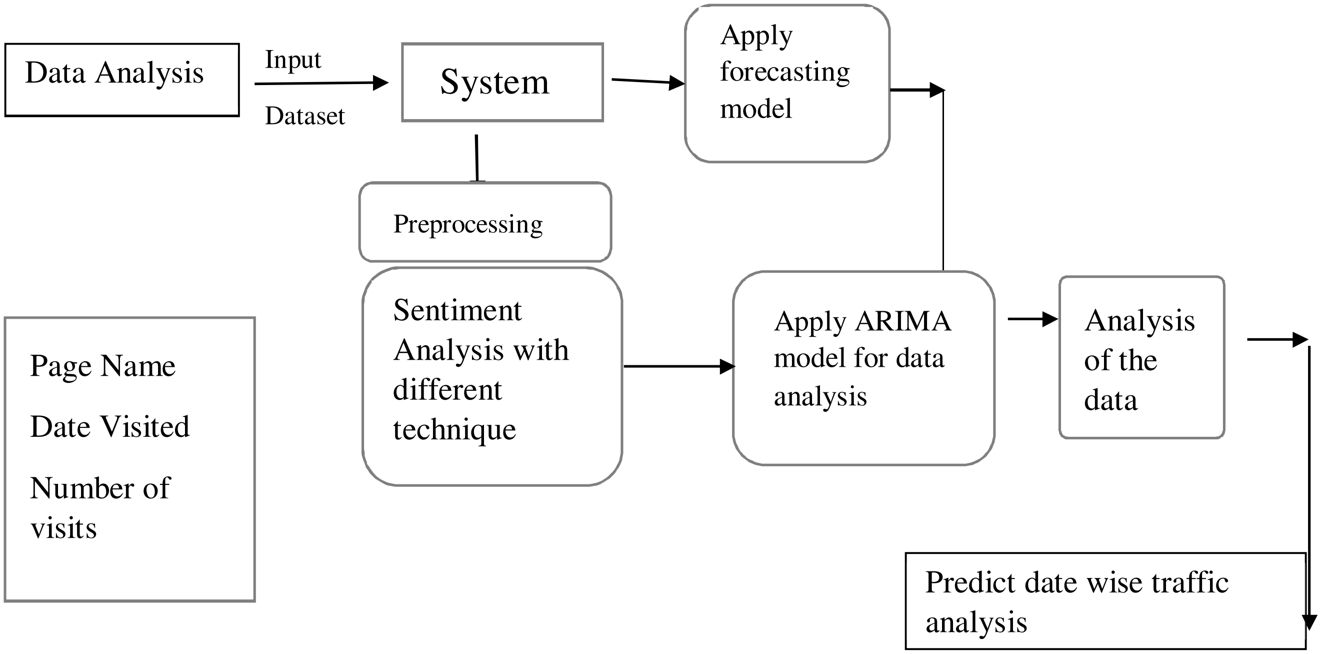

Figure 2: DFD Level 2

V. RESULT AND EVALUATION

A. Tools and Technologies Used

1) Jupyter Notebook

a) The Jupyter Notebook is an open-source web application that you can use to create and share documents that contain live code, equations, visualizations, and text.

b) Jupyter Notebook is maintained by the people at Project Jupyter.

2) Databases

The database basically used for user storing user details like Username and Password.

The tool used for db functionalities was MYSQL GUI Browser.

B. Limitations

Server machine should be on all the time.

TEST ID	TEST CASE NAME	TEST CASE DESCRIPTION	EXPECTED RESULT	ACTUAL RESULT
1	To verify login name and password	Enter valid login name and password	System should display the homepage	Same as expected
		Enter valid login name and password	System shows the error message	Same as expected
2	Data collection	The web traffic dataset is downloaded from the kaggle	Use pre-processing to clean the data	Same as expected
3	Feature extraction	Extract the feature from the data	Features for predicting the web traffic are extracted	Same as expected
4	Time series analysis	Find the underlying trends and patterns in the data	The pattern in time series data is observed	Interpret and integrate the pattern with other
5	Build prediction model	Model is trained using ARIMA	Predict the future web traffic	Same as expected

Figure 3: Test Cases

VI. CONCLUSION

Our research's primary objective is to develop a consistent forecasting model for predicting the future traffic of Wikipedia pages. To validate our prediction model, we use ARIMA model on Web Traffic Time Series Forecasting dataset. We have trained the data with this model using features like page name, visited date and the number of visits for pages for a year to predict the future web traffic.

REFERENCES

[1] Navyasree Petluri and Eyhab Al-Masri, "Wikipedia Page Traffic Prediction," **2018** IEEE International Conference on Big Data (Big Data).

[2] Mohammad Asifur Rahman Shuvo, Muhtadi Zubair, Afsara Tahsin Purnota, Sarowar Hossain, and Muhammad Iqbal Hossain, "Traffic Forecasting Using Time-Series Analysis," 6th International Conference on Inventive Computation Technologies, **2021**. (ICICT).

[3] Partha Sarathi Mangipudi and Rishabh Madan, "Predicting Computer Network Traffic: A Time Series Forecasting Approach Using DWT, ARIMA, and RNN," **2018** Eleventh International Conference on Contemporary Computing (IC3).

[4] Fu-Ke Shen, Wei Zhang, and Pan Chang, "An Engineering Approach to Network Traffic Prediction Using a Time-Series Model," International Joint Conference on Artificial Intelligence, **2009**.

[5] Jianhu Zheng and Mingfang Huang, "Traffic Flow Forecasting Using Deep Learning and Time Series Analysis," IEEE Access, **2020**. P Montero-Manso.

[6] Montero-Manso, P.; Athanasopoulos, G.; Hyndman, R.J.; Talagala, T.S. Fforma: Featurebased forecast model averaging. Int. J. Forecast. **2020**,36, 86–92.

[7] Rangapuram, S.S.; Seeger, M.W.; Gasthaus, J.; Stella, L.; Wang, Y.; Januschowski, T. Deep state space models for time series forecasting. Adv. Neural Inf. Process. Syst. **2018**,31, 7785–7794.

[8] Tealab, A. Time series forecasting using artificial neural networks methodologies: A systematic review. Future Comput. Inform. J. **2018**,3, 334–340.

[9] Tyralis, H.; Papacharalampous, G. Variable selection in time series forecasting using random forests. Algorithms 2017,10, 114.

[10] Chen, W.C.; Chen, W.H.; Yang, S.Y. A big data and time series analysis technology-based multi-agent system for smart tourism. Appl. Sci. **2018**,8, 947.

[11] Boone, T.; Ganeshan, R.; Jain, A.; Sanders, N.R. Forecasting sales in the supply chain: Consumer analytics in the big data era. Int. J. Forecast. **2019**,35,170–1801

[12] Chen, D.; Gao, M.; Liu, A.; Chen, M.; Zhang, Z.; Feng, Y. A Recurrent Neural Network Based Approach for Web Service QoS Prediction. In Proceedings of the **2019** 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 25–28 May 2019; pp. 350–357.

[13] Zhou, K.; Wang, W.; Huang, L.; Liu, B. Comparative study on the time series forecasting of web traffic based on statistical model and Generative Adversarial model. Knowl.-Based Syst. **2020**, 213, 106467.

[14] Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. The M4 Competition: 100,000 time series and 61 forecasting methods. Int. J. Forecast. **2020**, 36, 54–74.

[15] Yang, Y.; Lu, S.; Zhao, H.; Ju, X. Predicting Monthly Pageview of Wikipedia Pages by Neighbor Pages. In Proceedings of the **2020** 3rd International Conference on Big Data Technologies, Qingdao, China, 18–20 September 2020; pp. 112–115. Mathematics **2021**, 9, 421 20 of 21

[16] Bojer, C.S.; Meldgaard, J.P. Kaggle forecasting competitions: An overlooked learning opportunity. Int. J. Forecast. **2020**.

[17] Fry, C.; Brundage, M. The M4 Forecasting Competition-A Practitioner's View. Int. J. Forecast. **2019**.

[18] De Gooijer, J.G.; Hyndman, R.J. 25 years of time series forecasting. Int. J. Forecast. **2006**, 22, 443–473.

[19] Madan, R.; SarathiMangipudi, P. Predicting computer network traffic: A time series forecasting approach using DWT, ARIMA and RNN. In Proceedings of the 2018 Eleventh International Conference on Contemporary Computing (IC3), Noida, India, 2–8 August **2018**; pp. 1–5.

[20] Le, P.; Zuidema, W. Quantifying the vanishing gradient and long-distance dependency problem in recursive neural networks and recursive LSTMs. arXiv **2016**, arXiv:1603.00423.

[21] Suilin, A. kaggle-web-traffic. **2017.** Available online; https://github.com/Arturus/kaggle-web-traffic/ (accessed on 19 November**2018**).

[22] Cinar, Y.G.; Mirisaee, H.; Goswami, P.; Gaussier, E.; Aït-Bachir, A.; Strijov, V. Position-based content attention for time series forecasting with sequence-to-sequence rnns. In Proceedings of the International Conference on Neural Information Processing, Guangzhou, China, 14–18 November **2017**; Springer: Cham, Switzerland, **2017**; pp. 533–544.

[23] Qin, Y.; Song, D.; Chen, H.; Cheng, W.; Jiang, G.; Cottrell, G. A dual-stage attention-based recurrent neural network for time series prediction. arXiv **2017**, arXiv:1704.02971.

[24] Liang, Y.; Ke, S.; Zhang, J.; Yi, X.; Zheng, Y. Geoman: Multi-level attention networks for geo-sensory time series prediction. In Proceedings of the 2018 International Joint Conference on Artificial Intelligence (IJCAI 2018), Stockholm, Sweden, 13–19 July **2018**; pp. 3428–3434.

[25] Smagulova, K.; James, A.P. A survey on LSTM memristive neural network architectures and applications. Eur. Phys. J. Spec. Top. **2019**, 228, 2313–2324.

[26] Miyaguchi, A.; Chakrabarti, S.; Garcia, N. ForecastingWikipedia Page Views with Graph Embeddings. **2019**. Available online: http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26647399.pdf (accessed on 30 November 2020).

[27] Wunnava, V.P. Exploration of Wikipedia traffic data to analyze the relationship between multiple pages. Master's Thesis, University of North Carolina, Chapel Hill, NC, USA, May **2020**.

[28] Bai, S.; Kolter, J.Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv **2018**, arXiv:1803.01271.

[29] Srinivasan, A.; Jain, A.; Barekatain, P. An analysis of the delayed gradients problem in asynchronous sgd. In Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May **2018**.

[30] Dean, J.; Corrado, G.; Monga, R.; Chen, K.; Devin, M.; Mao, M.; Le, Q.V.; Mao, M.Z.; Ranzato, M.; Senior, A.; et al. large scale distributed deep networks. Adv. Neural Inf. Process. Syst. **2012**, 25, 1223–1231.

[31] Talyansky, R.; Kisilev, P.; Melamed, Z.; Peterfreund, N.; Verner, U. Asynchronous SGD without gradient delay for efficient distributed training. In Proceedings of the International Conference on Learning Representations (ICLR 2019), New Orleans, LA, USA, 6–9 May **2019**.

[32] Tian, C.; Ma, J.; Zhang, C.; Zhan, P. A deep neural network model for short-term load forecast based on long short-term memory network and convolutional neural network. Energies **2018**, 11, 3493.

[33] Liu, Y.; Guan, L.; Hou, C.; Han, H.; Liu, Z.; Sun, Y.; Zheng, M.Wind power short-term prediction based on LSTM and discrete wavelet transform. Appl. Sci. **2019**, 9, 1108.

[34] Liu, Z.; Yan, Y.; Hauskrecht, M. A flexible forecasting framework for hierarchical time series with seasonal patterns: A case study of web traffic. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July **2018**; pp. 889–892.

[35] Shelatkar, T.; Tondale, S.; Yadav, S.; Ahir, S. Web Traffic Time Series Forecasting using ARIMA and LSTM RNN. In Proceedings of the ITM Web of Conferences 2020; EDP Sciences: Ulis, France, **2020**; Volume 32, p. 03017.

[36] Petluri, N.; Al-Masri, E. Web Traffic Prediction of Wikipedia Pages. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle,WA, USA, 10–13 December 2018; pp. 5427–5429. Mathematics **2021**, 9, 421 21 of 21

[37] Du, S., Pandey, M., & Xing, C. Modeling Approaches for Time Series Forecasting and Anomaly Detection. Technical Report. 2017. Available online: http://cs229.stanford.edu/proj2017/final-reports/5244275.pdf (accessed on 30 November **2020**).

[38] Ragno, R.; Papa, R.; Patsilinakos, A.; Vrenna, G.; Garzoli, S.; Tuccio, V.; Fiscarelli, E.; Selan, L.; Artini, M. Essential oils against bacterial isolates from cystic fibrosis patients by means of antimicrobial and unsupervised machine learning approaches. Sci. Rep.**2020**, 10, 1–11.

[39] Ieracitano, C.; Paviglianiti, A.; Campolo, M.; Hussain, A.; Pasero, E.; Morabito, F.C. A novel automatic classification system based on hybrid unsupervised and supervised machine learning for electrospun nanofibers. IEEE/CAA J. Autom. Sin. **2020**, 8, 64–76.