International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4s
DOI: https://doi.org/10.17762/ijritcc.v11i4s.6323
***Article Received**: 01 January 2023 **Revised**: 04 February 2023 **Accepted**: 16 February 2023*

Real-Time Streaming Analytics using Big Data Paradigm and Predictive Modelling based on Deep Learning

J Ruby Dinakar[1], Dr Vagdevi S[2]
[1]VTU Research Scholar, Faculty of CSE
PES University
Bangalore, India
Email:rubydinakar@gmail.com
[2]Professor & Head AIML dept
City Engineering College
Bangalore, India
Email:vagdevi04@gmail.com

Abstract— With the evolution of distributed streaming platforms analysing humongous time series data, which is streamed continuously from IoT devices become lot easier. In most of the IoT networks the data are in motion or in data centre/cloud. It is possible to process this data in real time similar to edge devices using the big data framework. In data intensive applications predictive analytics require more resources to perform complex computations. Apache Flink framework is capable of performing real time streaming of schema less data and scales very high in distributed environment with low latency, it is used to collect and store the data in the cloud. This work suggests a suitable environment to collect, transport, preprocess and aggregate the data stream to perform predictive analytics using deep learning models. Deep learning automatically extracts features and builds models after training, it has the potential to solve problems that can't be solved by conventional machine learning models. Therefore, the use of algorithms based on deep learning is recommended for forecasting temporal data. Also, we discuss a number of different deep learning forecasting models and analyse the performance of different deep learning forecasting models in order to determine which one is the effective model for single step, multi step and multi variant methods based on error functions with respect to streamed sensor data.

Keywords- Apache Flink, stream processing, real time streaming, streaming analytics, predictive modelling, deep learning

I. Introduction

Digitization of information plays a significant role in advancement in data processing and analyzing techniques. Streaming data are highly prone to change, and the data stream as a whole is frequently inconsistent and lacking in detail. The characteristics of the data stream present a number of challenges, some of which are: 1. Elasticity - The amount of streaming data is growing on a daily basis and is expected to continue this trend. Stream processing systems are required to dynamically adjust to the load in order to achieve and sustain a desired level of service quality. There is a possibility that stream data sources will not always transmit large amounts of data. 2. Data Volume and Diversity of Data - Data streaming is concerned with massive amounts of data that are continually updated in real time. In data streaming, it is not uncommon for problems to arise such as the loss of data or the corruption of data packets. The data that is received in a stream is frequently heterogeneous and comes from a variety of applications and sources. Because of the nature of this data, managing it presents a challenge to the systems that handle data streaming and processing. 3. Timeliness - Stream data loses its value over time, so it's important to access it as soon as possible. Examining the data while it is in a usable state allows the system that is streaming and processing the data to function properly. A system that has high performance and can tolerate failure is required with respect to the time-sensitive nature of stream data. 4. Fault Tolerance - The processing of streams happens continuously and in real time. It is not possible to make a copy or retransmit the data correctly while it is being streamed. At no point should either the performance or availability of the system be compromised. Even if one part of the system fails to function properly, the rest of the processing system should continue to function normally.

In a distributed environment collecting data from various sensors from different IoT devices impose a challenge for ingesting the related data coming from multiple sources as a single stream of data continuously for performing analytics on the collected data in an instance. When the data is huge in size it is very difficult to note the key insights in the data. Streaming data analytics becomes most useful when multiple data streams are combined from different types of sensors. For example, in a hospital various vitals are measured for a patient including blood pressure, body temperature, heart rate, respiratory rate etc. These different types of data come from different devices, but when combined and analyzed it provides a valuable insight about a person's health status at any point of time. However, combining this live stream data with historical data provides a powerful context and promotes more insights into the current condition of the patient. The Stream processing engine Apache Flink helps to ingest data from multiple sources combined into a single stream based on the timestamp as an input for analytics operations. This data can be stored as JSON objects in the cloud storage using mongo dB which supports complex data types.

Artificial intelligence is the development of machines that can perform tasks that normally require human intelligence. As a subfield of artificial intelligence, machine learning enables machines to learn typically from data without being explicitly programmed. Artificial neural networks are used in deep learning, a subfield of ML. ANNs take their inspiration from the brain's structure and function and are able to learn on their own. Instead of being given explicit instructions on how to solve a problem, ANNs are taught to "learn" models and patterns. Deep learning neural networks are a potential tool for time series data forecasting. Recent developments in the field have demonstrated that these networks outperform conventional methods such as regression for predicting future values. Pollution of the air is a form of environmental hazard that is responsible for the majority of the health problems that occur in urban areas. The problem has been made worse by a number of factors, including automobiles, emissions from factories, and the combustion of fuel in vehicles. In this study, related work on applying various deep learning models to air pollutant data is investigated, as well as data preparation and performance of CNN, Simple RNN, and LSTM are studied for single time step, multiple outputs, and multiple time steps methods.

II. RELATED WORK

Many recent researches focus on performing streaming analytics on sensor data in real time to gain useful insights about the data. Apache Spark process data in real time as micro batches whereas Flink process the data in real time as soon as the data arrives. Zheng T Y et.al., proposed Stream Cube as a new incremental technology to process big data. They have implemented an intelligent data processing system using AI techniques. The framework is based on batch processing [1]. The authors [2] have discussed how to create value through streaming analytics, various use cases for real time streaming, challenges, opportunities and leading vendors for streaming. It also listed the difference between batch processing and streaming. A.A. Hassan et.al.,[3] provided the overview of various big data architectures Lambda, Kappa, Delta, big data processing tools Map reduce, Spark and Storm. Kalajo et.al [4] reviewed the issues related to load balancing, fault tolerance, scalability and heterogeneity during stream processing. They compared all the available tools and techniques in big data processing. Soumaya et.al [5] discussed the Lambda and Kappa architectures in detail. They also compared the existing technologies like Spark, Map reduce and Storm. They proposed a new five layered architecture which includes presentation, filtering, storage, real time processing and integration but there are no validation techniques to justify the architecture.

A large number of researchers have focused their efforts in recent years on developing deep learning predictive models based on time series data. Using deep learning models, numerous authors have proposed a variety of techniques for applying analytics to streaming data. Huang and his fellow workers.al [6] Comparisons were made between a variety of machine learning approaches. Validation was also performed on the CNN-LSTM model's capability of accurately forecasting PM2.5 concentrations under realistic conditions. The authors [7] proposed the most common time-series forecasting architectures, basic neural network building blocks, and instructions for making one-step-ahead forecasts in their work. They also discussed how the newly developed hybrid deep learning models combine statistical and deep learning components in order to outperform pure methods in both categories. Specifically, they discussed how this was accomplished. The use of deep neural networks is investigated in this study [8], which focuses on the definition and classification of time series. Investigations are conducted using both conventional approaches to time series and ANN methods. The advantages and disadvantages of each potential solution are then taken into consideration. We will discuss the key performance indicators that are used to evaluate the accuracy of forecasting models.

Kang et al. [9] exploited big data analytics, machine learning models, and various other techniques in order to compare and contrast the most recent studies on the evaluation of air quality. In addition to this, it investigates possible areas for future research, as well as problems and requirements for the future. Heydari et al. [10] have discussed a method that is capable of producing more accurate and reliable forecasts for industrial air pollution. The use of a multi-verse optimization metaheuristic method to integrate long short-term memory and predict NO2 and SO2 by optimizing the LSTM's hyper parameters is the most important contribution that this paper makes. This method was described in the introduction. This study aims to predict PM2.5 pollution, which is one of the most dangerous disease-causing pollutants in the world. The bidirectional long-range forecasting method will be used to accomplish this goal. Torres et al. [11] have discussed the practical aspects that must be taken into consideration in order to successfully apply deep learning to time series. Some examples of these practical aspects include setting the values for hyper-parameters and selecting the best frameworks. A deep learning framework with an extended

DOI: https://doi.org/10.17762/ijritcc.v11i4s.6323
***Article Received**: 01 January 2023 **Revised**: 04 February 2023 **Accepted**: 16 February 2023*

temporal sliding long short-term memory model was proposed by Wenjing et al. [12]. By employing a multi-layer bidirectional long short-term memory (LSTM) as well as hourly historical PM2.5 concentrations, meteorological data, and temporal data, the model integrated the ideal time lag to realize sliding prediction. This was accomplished by using the data to realize sliding prediction.

III. Methodology

The process of streaming in real time is broken down into several stages. The simplified steps are shown in figure1. At first, the data that is generated comes from a number of different sensors, and it is gathered by collection agents. The data that have been collected are pre-processed by having any redundant information removed from them. The next step is to investigate the unstructured data in order to find patterns. After undergoing preliminary processing, the raw data are then converted into feature vectors that can be used in predictive modelling. These features are saved and kept in the cloud so that analytics can be performed on them. The single step, multi-step, and multivariant forecasting methods all make use of deep learning-based forecasting models, which are applied to the features. The outcome of the performance review is presented via the analytics dashboard in the form of comparative charts.

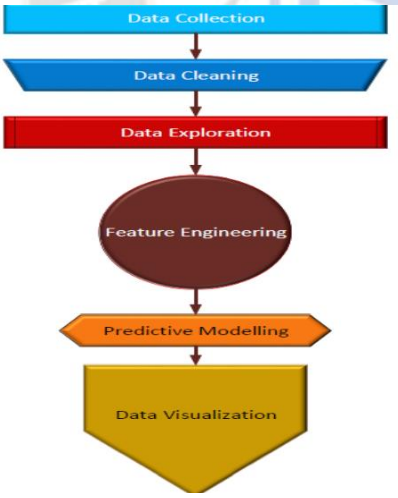

Fig 1: Streaming analytics process overview

IV. Implementation

An experimental setup test bed is created to show the process of real time streaming analytics. The environmental telemetric are taken as input. To generate the data a mote collects sensor data once an hour, and some of the pollutants that it monitors include CO, NO, O3, NO2, NOx, PM2.5, C6H6, PM10, and NH3. Arduinos are used as nodes, and a Raspberry Pi is used as a gateway and also it works as a collection agent. The data is published using the distributed message broker Apache Kafka into various topics in multiple partitions. There are three brokers. One of them is elected as a leader using polling and the messages are written in to the respective topics. To overcome the failures the data is replicated and the Kafka broker logs all the streaming information in case any broker fails, zookeeper elects the next leader and using the logs it can recover from failures. The data are streamed to Apache Flink using Kafka consumer. A variety of data from several streams are filtered based on sensor id and timestamp. Also, redundant data are eliminated by considering the topic and timestamp. The various streams are combined into a single stream and connected to a single one for a particular timestamp. The raw data are transformed to feature vectors and sampling has been done on the data. The analytic task based on deep learning forecast models are evaluated to forecast the required pollutants according to different methods. The pictorial view of the process is given in figure2.

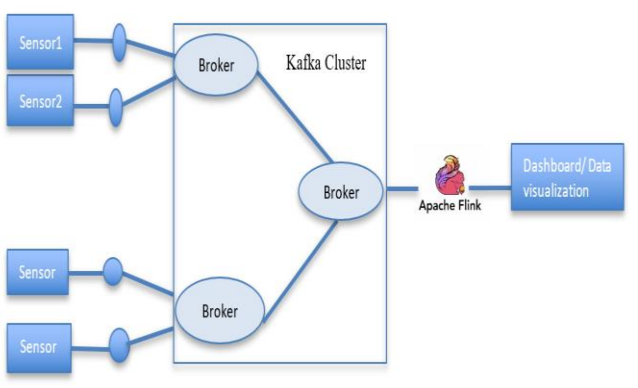

Fig 2. Streaming analytics

V. Experimental results

The purpose of the test bed was to make hourly and daily predictions regarding the amount of various pollutant present. For each one-hour forecast, the moving average was computed by applying the formula to the data from the preceding six hours. The forecast for the following day was based on the data from the previous twenty-four hours. A large number of models are validated using single-step, multi-output, and multi-step methods, respectively. The mean absolute error was chosen to serve as the criterion for the calculation of the model's overall accuracy. A straightforward moving average is used for the base model. All of these models the baseline, the linear model, Dense, CNN, and RNN with LSTM are investigated for the single step. The mean absolute error value of the linear model is significantly lower when compared to that of other models. The RNN with residual connection in a multi-output model has a smaller number of outputs than other types of RNNs because the model is trained to minimize the error value. The autoregressive model known as LSTM did very well when applied to a multistep model. The standard deviation of the absolute value is 0.1628. In our implementation, deep learning models have generally performed well, with the exception of single-step forecasting.

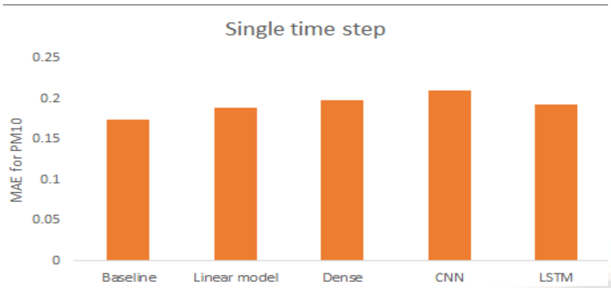

Fig3. Plot for comparing PM10 pollutant mean absolute error for single time step method

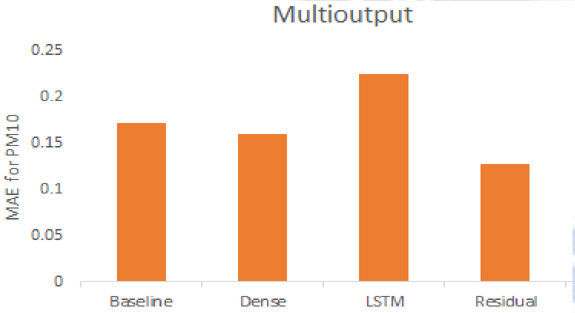

Fig4. Plot for comparing PM10 pollutant mean absolute error for multi output method

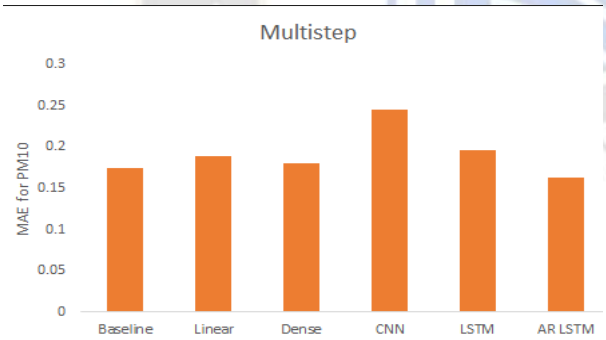

Fig5. Plot for comparing PM10 pollutant mean absolute error for multi step method

VI Conclusion

Because of digitization, a substantial quantity of data is produced by means of all of the connected devices. Streaming analytics relies heavily on the utilization of big data platforms, which has been a significant contributor to the field's growth. In order to achieve a balance in human health and life, it is now necessary for us to conduct data analytics in real time. This is necessary so that we can gain a deeper understanding of the ecosystem. In this paper, we discussed related work on real-time data streaming utilizing a variety of deep learning models on air pollutant data. Additionally, we looked at the mean absolute error rate on streamed sensor data utilizing Apache Flink. During this experiment, data was being continuously streamed in an atmosphere with limited space. This work can be extended to a more polluted and high-traffic environment in order to stream data continuously. In order to forecast pollutant levels, deep learning models make use of historical data. According to the findings of the study, the straightforward linear model was effective in providing accurate immediate forecasts of a single time or step. When compared to other models, the performance of deep learning models was superior in multi-step and multi-output models. The models were trained to produce results with a significantly lower rate of error. Calculating efficiency also makes use of other evaluation criteria, such as root mean square error (RMSE). Methods of performance tuning can also be utilized to achieve the goal of increased productivity.

References

[1] Kolajo, T., Daramola, O. & Adebiyi, A. Big data stream analysis: a systematic literature review. J Big Data 6, 47 (2019). https://doi.org/10.1186/s40537-019-0210-7

[2] Namiot, Dmitry. (2015). On Big Data Stream Processing. International Journal of Open Information Technologies. 3. pp 48-51.

[3] Fernandes, Eliana & Salgado, Ana Carolina & Bernardino, Jorge. (2020). Big Data Streaming Platforms to Support Real-time Analytics. 426-433. 10.5220/0009817304260433.

[4] David J. Hill, Barbara S. Minsker," Anomaly detection in streaming environmental sensor data: A data-driven modeling approach", Environmental Modelling & Software, Volume 25, Issue 9,2010, pp 1014-1022, ISSN 1364-8152, https://doi.org/10.1016/j.envsoft.2009.08.010.

[5] Huang C-J, Kuo P-H. A Deep CNN-LSTM Model for Particulate Matter (PM2.5) Forecasting in Smart Cities. Sensors. 2018; 18(7):2220. https://doi.org/10.3390/s18072220

[6] Lim, B, Zohren S. 2021 Time-series forecasting with deep learning: a survey. Phil. Trans. R. Soc. A379: 20200209. https://doi.org/10.1098/rsta.2020.0209

[7] Mahmud, Amal & Mohammed, Ammar. (2021). A Survey on Deep Learning for Time-Series Forecasting. 10.1007/978-3-030-59338-4_19.

[8] Kang, Gaganjot et al. "Air Quality Prediction: Big Data and Machine Learning Approaches." International journal of environmental science and development 9 (2018): 8-16.

[9] Heydari, A., Majidi Nezhad, M., Astiaso Garcia, D. et al. Air pollution forecasting application based on deep learning model and optimization algorithm. Clean Techn Environ Policy (2021). https://doi.org/10.1007/s10098-021-02080-5

[10] Jeya, S., & Sankari, L. (2020). Air Pollution Prediction by Deep Learning Model. 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). doi:10.1109/iciccs48265.2020.9120

[11] Torres JF, Hadjout D, Sebaa A, Martínez-Álvarez F, Troncoso A. Deep Learning for Time Series Forecasting: A Survey. Big Data. 2021 Feb;9(1):3-21. doi: 10.1089/big.2020.0159.

[12] Wenjing Mao, Weilin Wang, Limin Jiao, Suli Zhao, Anbao Liu, Modeling air quality prediction using a deep learning approach: Method optimization and evaluation, Sustainable Cities and Society,Volume65,2021,102567,ISSN2210-6707, https://doi.org/10.1016/j.scs.2020.102567.

[13] Apache Flink. https://flink.apache.org/.

[14] Apache Kafka https://kafka.apache.org/