Real Time Traffic Prediction Based On Social Media Text Data Using Deep Learning

B. Mounica and K. Lavanya*

School of Computer Science and Engineering, VIT University, Vellore, India
E-mail: premkumarmounica@gmail.com; lavanya.k@vit.ac.in
**Corresponding Author*

Received 25 June 2021; Accepted 26 August 2021;
Publication 29 October 2021

Abstract

Due to urbanization Traffic management is one of the major issues in contemporary civic management, considering this circumstance traffic analysis is turning into the need of the present world. Text data generated by Twitter, Facebook and other social media platforms can be used for traffic management. Big data helps in traffic prediction and traffic analysis of advancing metropolitan zones. Constant traffic investigation requires preparing of information streams that are produced persistently to increase fast experiences. To measures stream information at a fast rate advancements on high figuring limit is required. Social media text data can be processed by using batch processing and stream processing with big data architecture through Spark and Hadoop framework. In this paper big data architecture is proposed for real time traffic text data analysis. In architecture Spark and Kafka are used in combination. Kafka helps in pipelines text data used in conjunction with spark stream processing engine. Big data architecture using Spark, Kafka with ability for processing and preparing huge measure of information, have settled the serious issue of handling and putting away constantly streaming data. The traffic information from Twitter API is streamed. In The proposed model pointed toward ensemble neural network model to reduce the variance

Journal of Mobile Multimedia, Vol. 18_2, 373–392.
doi: 10.13052/jmm1550-4646.18211

in results for better prediction foreseeing traffic stream text data by incorporating Spark and Kafka that will be of an extraordinary incentive to the public authority for traffic management and analysis.

Keywords: Ensemble neural network model, streaming data, big data architecture, real time traffic analysis, social media data.

1 Introduction

Intelligent Transportation Systems Consumes and generate data. As technology and population are seeing an exponential growth so volume, variety, velocity (VVV) in data volume plays a major role in big data for storing images, videos, text data and processing it with high speed with velocity, as the data is structured example relational table, unstructured example images and video files, semi-structured example Json, XML file as variety of data that is generated and consumed in big data, therefore ITS is expected to generate, consume data on a Big data scale. ITS has various data sources like tabular enterprise data, unstructured data in social media ete, ITS sources tabular data from Systems such as Passenger Ticketing System and later is used in analysis. Toll gates collect data from RFID tags and sensors in Vehicle as part of sensor data collection in ITS. User comments from Social media sites generate data for traffic analysis can be a rich source of ITS data. Broad exploration is being done to place this information into great use. Enormous data generated from different sources emerges with big data technology. Examination on the best way to utilize the accessible information with Artificial Intelligence is accomplishing interesting and essential outcomes.

As per 2019 the TomTom Traffic index API covers 416 cities from 57 countries providing free traffic information for helping traffic related issues. It says Bangalore congestion level is 71% and worlds first rank, peek time of road congestion is Friday 7pm to 8pm. As per world health organization totally around 1.35 million people dies with road accident and 50 million injured with accidents so technology helps to decrease the count of accidents and deaths [10]. This based on streaming data framework using Apache Spark viably takes care of unstructured and real time information. It is multiple times quicker than Hadoop Map Reduce framework [14, 15]. The basic components of Hadoop:

- HDFS: Hadoop Distributed File System is a storage area, providing very high collective bandwidth among the cluster.

- YARN: It is a platform for managing compute resources in the clusters and scheduling the user applications.
- Map Reduce: MapR educe is a programming framework for processing and generating big data sets with parallel, distributed algorithm on cluster.the three operations performed are map, shuffle and reduce operations.

Spark works with different in-fabricated libraries, segments that incorporate Spark core API, SQL, spark Streaming, spark MLlib and spark GraphX. Figure 1 speaks on Spark environment [12, 13].

- Spark Core API: Fundamental functionalities in Spark is based on Spark Core. It fuses Resilient- Distributed-Datasets a primary function of Spark's primary deliberations and is liable for task booking, issue recuperation, in-memory calculation and also memory on the board. It is the establishment for handling huge datasets.
- Spark SQL: Spark SQL is a unit of Spark which gives SQL interface for Spark to work with organized or semi-organized information and for executing queries using SQL.
- Streaming: Streaming is a Spark core API which is answerable for high volume-throughput, adaptable and issue open minded stream preparing continuously streaming information streams acquired from information sources for streaming example are, Kafka, kinesis and Flume.
- MLlib: MLib is a Spark engine library that encourages usage on Machine Learning, Deep Learning calculations and make Machine Learning adaptable. This is a Dataframe based bundle of Spark rather than utilizing RDD.
- GraphX: Graphx is nothing but a Spark API for controlling, working and executing diagrams and chart equal calculations.

Basically there are four different categories in NoSql databases [33]:

(a) Key value based:

The data stored in the tables are maintained with key value pairs for unique identification of that particular record and retrieval of data is based on that key. Hashing function is performed on key value pairs. Example for key value based are Redis, Riak.

(b) Document based:

The data is stored in form of documents by using Xml or Json. In document based data model application logic to write in to databases is easy but reading

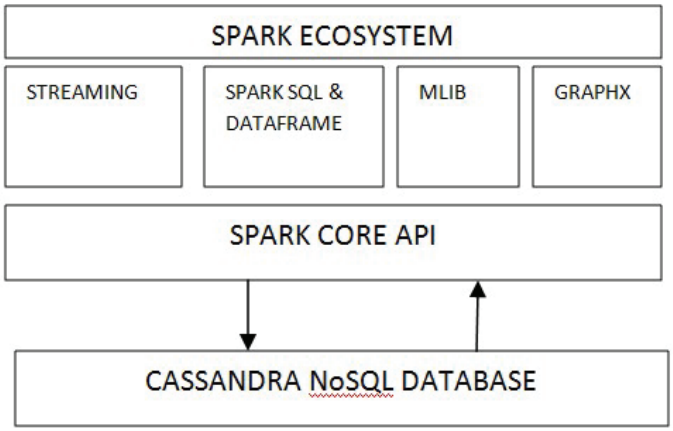

Figure 1 Apache Spark ecosystem using Cassandra Nosql Database.

of data is time consuming because the data is stored in form of documents or files. In document based the query engine is very powerful to facilitate queries to fetch the data while reading. The internal structure of the database has to be used for extracting the Meta data. This type is recommended when data to be stored is having major percentage of semi structured and unstructured data. Example for document based data bases are MongoDB, CouchDB.

(c) Columnar based:

In this data is stored in columns and each row will have a row id to retrieve any single record from column. These data bases are highly efficient for aggregate functions, online processing websites, timestamp based data collection. The related columns store under one column family. Example for columnar database is Cassandra, Hbase.

(d) Graph based:

Graph based databases store the data in form of graphs with nodes and connections for better understanding and heterogeneity. Example for graph based is Neo4J. In this paper Apache Spark ecosystem with Cassandra NoSQL Database [32] architecture used.

On account of ongoing traffic examination, not just new but information that is dynamic is constantly being created through IoT gadgets that get shipped off on information streaming application, for example, an outsider API from where Spark streaming gets ceaselessly streaming information for preparing and performing continuous analysis. The information through

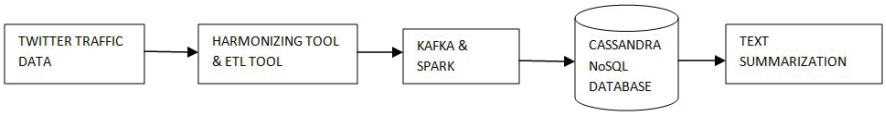

Figure 2 Big data architecture for Traffic Management.

the API be pushed into the flash application utilizing a distribute buy in informing stage known as Kafka. Kafka is an adaptable, quick and issue lenient streaming stage that utilizes ongoing information pipelines which has low dormancy and high volume throughput which can be coordinated with dispersed stream handling systems, for example, Spark for constant ingestion and preparing of information streams [10, 11].

The work proposed can be actualized by incorporating Apache Spark along with Apache Kafka and Cassandra NoSQL database alongside the utilization of profound neural network outfit model learning for effectively performing continuous traffic investigation and figure and. This framework is created utilizing different MLlib in SPARK to deal with streaming information consecutively and incrementally post which SparkSQL for executing SQL related quires on them.

2 Related Work

Big data sources related to intelligent transportation systems and deep learning models like Reinforcement, Supervised, unsupervised Learning techniques also ITS applications related to big data and challenges discussed [1]. ITS big data architecture is designed on real time traffic analysis based using batch processing Hadoop distributed file system (hdfs)with Hadoop ecosystem, Kafka is used for streaming data feeds [2]. Proposed Correlation methods in traffic data analysis, historical data and streaming data together with a combination of framework with a clustering model based on SAGA-FCM i.e simulated annealing and genetic algorithm to overcome drawbacks of fuzzy - means algorithm (FCM) traditional approach [3]. Proposed Software-Defined Internet of Vehicles architecture for Internet of Vehicles is proposed to reduce the number of protocols in real time query services [4]. The design of Big data architecture for intelligent transport system is proposed for dynamic toll charging, Spark is used for near real analysis and stream processing and data is stored in Cassandra Nosql database for toll generation [5]. on IoT based traffic data collected from network is processed with hdfs and map reduce and also traffic mirror method is proposed by Collecting network traffic with

graph based model proposed [6]. Using Spark framework for analysis and the data is collected from sensors proposed a model to forecast the traffic state of road congestion with levels of high to low [7]. Proposed model based on ensemble neural network for recurring, nonrecurring congestion with CNN and LSTM algorithms for improving the performance of traffic prediction [8]. Research on probability analysis using big data related technology for online processing of semistructured data and also proposed AsterixDB model proposed for semistructured data compare with Spark streaming and Cassandra NoSQL database for processing data [9]. In this article real time traffic analysis on unmanned aerial vehicles based video is performed by Recognition and tracking using Mask -RCNN to resolve occurrence of segmentation issues in machine learning, motor vehicle count and speed calculation. Estimation of arithmetical space for pixels measure in the image, speed computation of pedestrians and reasons behind the images calculated [29]. In this article traffic analysis using social media data using natural language processing is proposed. In this paper traffic congestion is calculated and compared between two places using Twitter data and Face book tags. Steps followed for processing is data extraction, data storage and analysis [30].

3 Data Collection

For any assessment collecting and building a dataset is a fundamental requirement. social media sites trade over the collected data in significant volumes. Sites like Face book, twitter etc are a major source for procuring passive data. This work has used tweet data set from twitter for analytical work. Tweet dataset is procured through tweet API and direct purchase. Reasons for purchase and KYC details need to be given for purchase from twitter. Twitter generates the following Consumer key, Consumer secrete key, Access token key, Access token secrete key. User's get authenticated through these keys for downloading tweets. Hash tags related to traffic are used for data extraction from tweeter, the keywords used are Tollgate, Road traffic, Road accidents, Road work, Traffic, Accidents, Road closed, Congestion, Tollgate, Vehicle.

Twitter shares the data set as JSON file which is converted to CSV format for ease of use, Null value removal which are invalid qualities influence accuracy in ML algorithms. Data set is extracted with meaning full attributes like ID, Latitude, Longitude, Reweet, Tweet etc, point co ordinates in tweets data i.e. latitude and longitude are converted to location through reverse

geo-coding through geopy install, this gives user tweet street name, state, country etc.

4 Analytical Tools and Techniques

4.1 Approaches to Data Processing and Analysis

Data is processed in large volumes through application of spark engine operations deriving insights on various analytical perspectives like predictive, descriptive and diagnostic analysis on data inputs.

4.1.1 Data processing types in spark

- Data processing in batch

Data when collected over a period of time and processed in large volumes through spark engine is called Batch data processing in spark. A time interval is fixed based on various business parameters and data collected and processed in batches of this time interval. The time interval between the batches of data processed is a key factor in the efficiency with which data can be Batch Processed.

- Real-time or stream processing

When continuous stream of data is subjected to real-time processing through spark engine it's called Stream processing or Real time processing of data. Though Streams of data are continuously taken for input there is a time or data interval that is applied for logical segregation for processing of data. Spark stream processing is useful when time sensitive decisions are critical like traffic analysis, weather modelling etc. Real time processing when combined with time series analysis helps to identify trends and patterns in seasonal weather between previous years and current year, comparing traffic growth patterns in highways, bridges across time can be effectively taken up for better management of assets [17].

4.1.2 Neural networks

Neurons are individual processing units in computation terminology which when stitched together as a common group to process a specific pattern becomes Neural Network.

Neural networks acts in close capacity to human brain, the network learns and process information over a period time based on the input data fed. Neural networks learn based on weighted parameters associated within the neurons

and the data fed to it, neural networks widely used to forecast modelling complex applications used in image recognition, communication systems, supply chain optimisation etc [18].

Types in neural network

Deep neural networks: These types of networks have layer upon layers of neural networks – an input neural layer and subsequent output neural layers with many thousands of nodes called neurons interconnected in each layer. Back-propagation neural networks: these are neural networks which are trained with an optimization technique to back-propagate error and adjust weights across neural layers thus reducing the error in the next learning. Feed-forward Neural Networks (FNNs): are modelled is a reverse of Back-propagation neural networks where error is not fed back to the neural layers in the network. For this work, we apply multiple types of neural networks to accurately predict the traffic.

- Ensemble Learning:

When multiple algorithms on learning are combined in machine learning process for improved prediction performance it's called Ensemble machine learning [19]. Base models or the individual models are joined to form an Ensemble machine learning model. Advantage of using Ensemble model is not one model is used as primary fit which reduces errors in the event the model is a misfit and thereby improving performance as each base model is assigned weight age for their output and then averaged out. having said this the overall ensemble model may not be better than an individual best fit model. A normal ensemble method is a classifier that sums up the output of the base models used and projects the output with a major vote of the base models. One more popular approach is the weighted model approach of the base models used in the ensemble model for prediction [20–22].

Other types of ensemble techniques used widely are

- Bagging or Bootstrap aggregation where the model is trained with multiple datasets and results averaged out, this approach is applied on all the base models in the ensemble model, this helps in reduction of variance of prediction and over fitting of the model on the applied data sets.
- Boosting is another technique where the models are subjected to sequential learning to reduce bias

The performance of ensemble model primarily depends on based models in use.

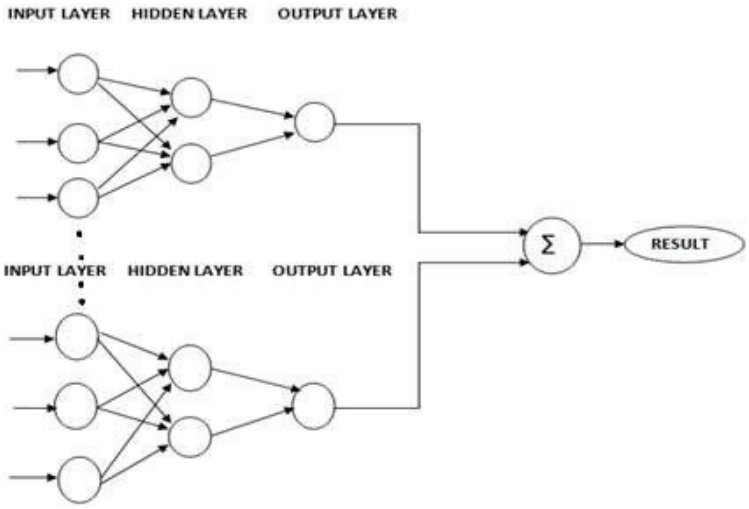

Figure 3 Neural network ensemble learning.

When ensemble model is applied for neural network, each neural network differ with each in following ways [23–25]:

- Hidden layers and nodes in each of base neural networks differ from each other ie the neural network architecture is different for each of the base model.
- Initial conditions like the weight age and learning rate are different for each of the neural network leading to difference in training conditions of every base models applied.
- Training methodology of each base model is different with respect to batch size applied so as to get dissimilar result out of each model applied.
- Each of the base models is subjected to different training data by applying random sampling approach.
- Different base models have different training algorithms for training the model and optimization of results.

5 Proposed Model

Twitter API data which is procured through steaming using apache Kafka is then utilised for traffic analysis and traffic conditions prediction. The model analyses and predicts using Spark engine with neural network ensemble and store in a NoSQL like CASSANDRA.

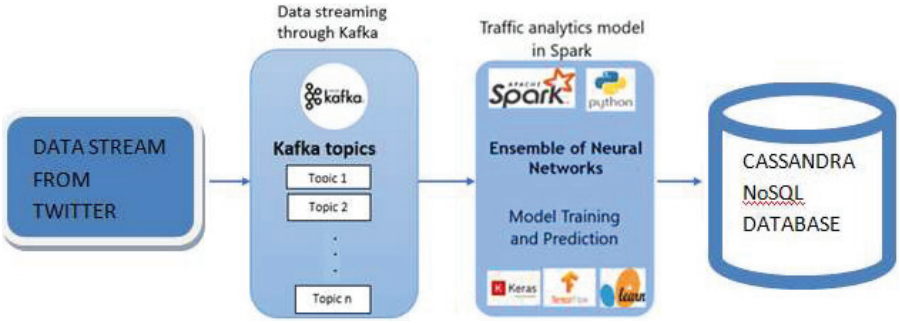

Figure 4 Proposed system architecture.

Data Streaming from twitter, Data is continuously streamed from source using Apache Kafka streams. Kafka uses 'publish-subscribe' model to manage real-time data volume. Kafka makes use of an API that allows consumption of real time events. Kafka has topics defined to categorize data which producers publish and consumers subscribe, Apache zookeeper is used for coordination between the Kafka components. Producer API is written in Python and data is streamed by execution of the same and published as a topic than can be consumed by a similar python based consumer API. The data from the consumers is integrated with streaming module developed in spark that analyse the streams for prediction.

Analytics model using neural network ensemble is used in which The Data is consumed by Spark engine and is processed by a model that makes use of ensemble methodology of neural networks in predictive analysis [26–28]. These neural networks are trained for better accuracy for the final output. In process Base neural networks creation: Effectiveness in ensemble model is enhanced when the base model in neural network is varied substantially across parameters like architecture, initial conditions, training algorithm, data etc. Deep neural networks differing in hidden layers are used for model training. The layers hidden in deep neural determine the performance output of a model. There are no specific rules for the number of layers hidden that need be used in network, but is based on the problem statement. For this work diverse base networks are applied for base model creation of the ensemble model.

Over fitting which causes a small error when applied on a trained data set but on applying the same on untrained data set the error margin is huge which is caused as the model is not able generalize new data. By applying

regularization techniques during data training over fitting error is eliminated to a large extent. Regularization assigns smaller weights on the model and model is stabilized by addition of penalty on function of cost proportionate to given weights. Additional ways in regularizing approach may also include using dropout, imposition of constraints on weight of range and activation based model penalization. Regularization technique applied for this work adds penalty and reduce the weights of the model by additional parameters on the networks leading to smaller weights across the network.

In Data Storage source data for the train and test data is finally stored within Cassandra NoSQL database having more efficiency to store unstructured columnar data. Data from the streamed API using Kafka is loaded in Cassandra database and is used as a sink, using a Cassandra-Kafka sink connector. We can write events from Kafka to Cassandra using the Cassandra Sink. The connector translates the value from the Kafka Connect SinkRecords to JSON before inserting the rows using Cassandra's JSON insert capability. In Cassandra, the task expects pre-created tables. While the data forecasted is stored into the storage system after process from Spark [34].

The performance metrics [31] used as follows:

1. Mean Square Error (MSE) is to measure model quality at time of prediction with the real time data. A actual dataset $t_i = \{\mathrm{t1}, \mathrm{t2}, \mathrm{t3...tn}\}$ and corresponding observation values are $t'_i = \{t'_1, t'_2, \ldots, t'_n\}$, the values matches between the actual and the predicted observation given by with the difference of error.

$$\text{Mean Square Error} = \frac{1}{n}\sum_{i=1}^{n}(t_i - t'_i)^2 \tag{1}$$

   Where n is the total number of observations, t_i is the i^{th} actual observation and t'_i is the i^{th} predicted value.
2. Root Mean Squared Error (RMSE) also called as stranded deviation the t_i residuals equal to the t'_i value, better fit of regression line. Mean Squared Error means the square root of average used for loss function of the squared differences with the sum of actual, predicted values as shown in given formula,

$$\text{Root Mean Squared Error} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - t'_i)^2} \tag{2}$$

Table 1 Neutral network model performance comparison study

Metric	Single Neural Network	Ensemble Neural Network
MSE	5.7	4.773
RMSE	2.389	2.176
R2 score	0.823	0.847

3. R^2 score is coefficients of determination that variance portion is calculated with dependent and predicted independent variable given by,

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (t_i - t'_i)^2}{\sum_{i=1}^{n} (t_i - \bar{t})^2} \tag{3}$$

$\sum_{i=1}^{n} (t_i - t'_i)^2$ is sum of residuals square and $\sum_{i=1}^{n} (t_i - t)^2$ is total sum of the squares, where x is the actual residual value and x' is the predicted residual value and $\overline{x}$ represents the mean.

For ensemble model results estimate, single neural model is compared with ensemble neural network performance and results are generated based on the test. Table 1 shows the results with the metric performance for errors in Mean Square. Root Mean Squared Error and R2 score of single neural network and ensemble neural model. As shown in table, the model exhibits lower score for RMSE, MSE and much higher score for R^2 score with single neural network. From the results MSE, RMSE performance better for the ensemble neural model compared the single neural network model and also table values says R^2 score is close to 1 which implies model performs superior. Thus, we can conclude with the results ensemble in multi neural networks had induced an enhanced performance compared with single model.

Figure 8 highlights the plot for predicted performance of single neural network over a 7 day period vs traffic flow actual. Time of a particular day is represented by x axis and Volume of vehicles per lane for a given day is represented in y axis.

Figure 9 shows best fit line graph on predicted data values. The dataset used for the below work is https://www.kaggle.com/mounicapremkumar/traffic-analysis-twitter-dataset. The attributes listed in dataset are Tweet, Geo_location, Latitude, Longitude, ID, Created _at, Retweet_Count, Coordinates. In the experimental results concentrated attributes are Tweet, Created _at, Geo_location based on graphs generated. Traffic dataset by using Tweet binder https://tweetbinder.intercom-attachments-2.com/i/o/277710248/eb4bafe82db57dee755f493c/Traffic+Los+Angeles+Sample.xlsx, Twitter

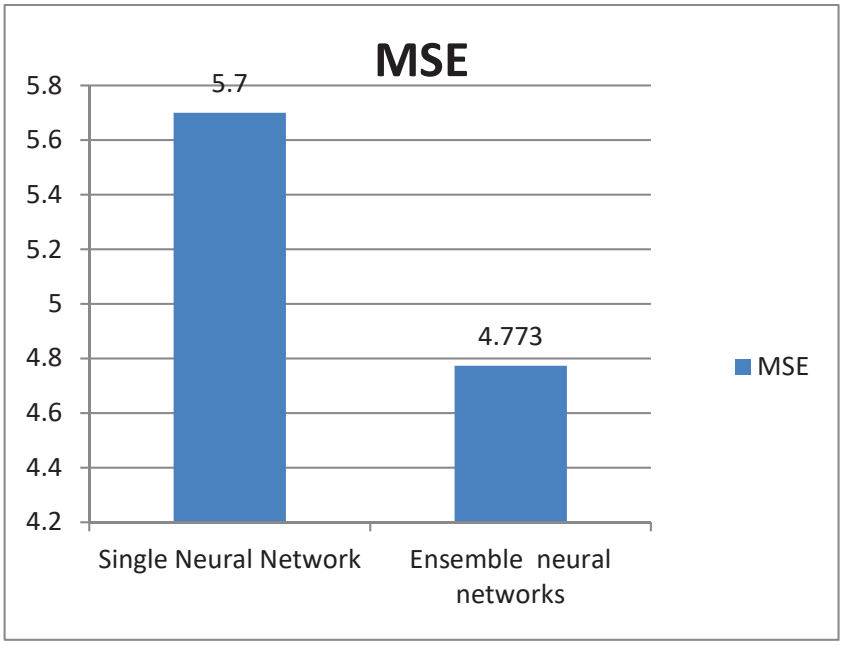

Figure 5 MSE values of single and ensemble of multiple neural networks.

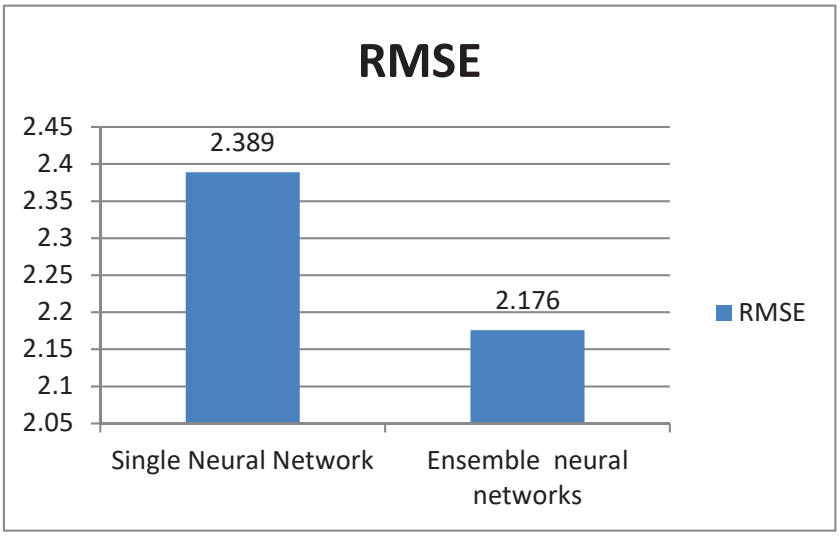

Figure 6 RMSE values of single and ensemble of multiple neural networks.

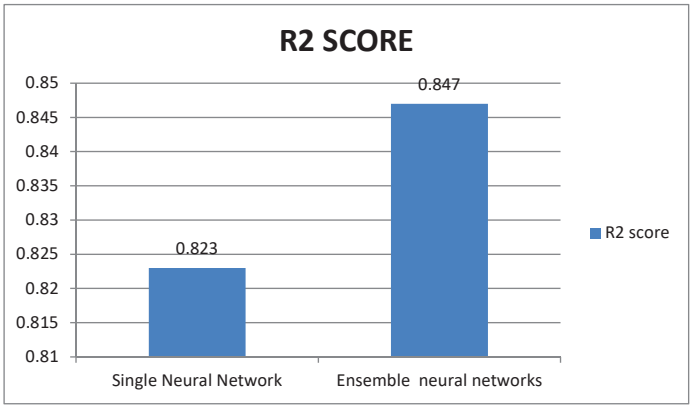

Figure 7 R^2 values of single and ensemble of multiple neural networks.

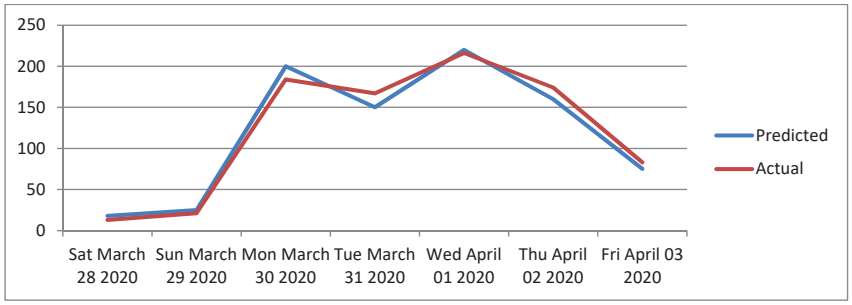

Figure 8 The plot for predicted performance of single neural network over a 7 day period of traffic flow actual.

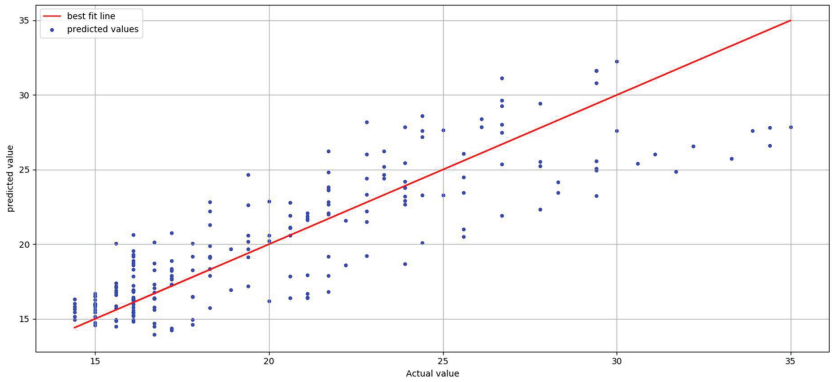

Figure 9 Best fit line vs prediction of traffic flow for a period of 10 days by the ensemble model.

dataset related to traffic from data.world website and dataset link is shared below https://data.world/chanalytics/2017-sxsw-twitter-traffic.

6 Conclusion

Performance metrics of ensemble neural network results are better compared with singe neural network. Through the usage of Apache Spark a forecast model for traffic analysis is implemented using ensemble techniques in machine learning model combined with multiple levels of deep neural networks capable of real-time prediction of traffic flow on roads. With enhanced accuracy, through apache Kafka streaming data management services along with spark engine for processing traffic data performance improvements were observed in multiple level deep neural networks when the same was compared with individual neural networks for prediction of vehicular traffic. Application of hyper parameter optimization like batch size, rates in dropouts, epoch numbers etc is also expected to give further improvement. Differing combinations in levels of deep neural networks can also be applied for enhancing prediction performance.

References

[1] L. Zhu, F. R. Yu, Y. Wang, B. Ning and T. Tang, "Big Data Analytics in Intelligent Transportation Systems: A Survey," in IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 1, pp. 383–398, Jan. 2019, doi: 10.1109/TITS.2018.2815678.

[2] Amini, S. et al. "Big data analytics architecture for real-time traffic control." *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)* (2017): 710–715.

[3] Hua-pu Lu, Zhi-yuan Sun, Wen-cong Qu, "Big Data-Driven Based Real-Time Traffic Flow State Identification and Prediction", Discrete Dynamics in Nature and Society, vol. 2015, Article ID 284906, 11 pages, 2015.

[4] X. Wang, C. Wang, J. Zhang, M. Zhou and C. Jiang, "Improved Rule Installation for Real-Time Query Service in Software-Defined Internet of Vehicles," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 2, pp. 225–235, Feb. 2017

[5] Merlin Mathew (merlinthemathew@gmail.com), Aishwarya Balakrishnan (aishu9298@gmail.com), Bikky Kumar Goit (navabryt@gmail.com), Mounica. B (premkumarmounica@gmail.com), "Dynamic Toll Charge using openCV and Spark", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 3, pp. 33–37, May–June 2019.

[6] L.U. Laboshin, A.A. Lukashin, and V.S. Zaborovsky. 2017. The Big Data Approach to Collecting and Analyzing Traffic Data in Large Scale Networks. Procedia Comput. Sci. 103, C (March 2017), 536–542.

[7] Prathilothamai M., Lakshmi, A. M. Sree, and Viswanthan, D., "Cost effective road traffic prediction model using Apache spark", Indian Journal of Science and Technology, vol. 9, 2016.

[8] G. Kothai, E. Poovammal, Gaurav Dhiman, Kadiyala Ramana, Ashutosh Sharma, Mohammed A. AlZain, Gurjot Singh Gaba, Mehedi Masud, "A New Hybrid Deep Learning Algorithm for Prediction of Wide Traffic Congestion in Smart Cities", Wireless Communications and Mobile Computing, vol. 2021, Article ID 5583874, 13 pages, 2021.

[9] W. Liyong and P. Vateekul, "Improve Traffic Prediction Using Accident Embedding on Ensemble Deep Neural Networks," 2019 11th International Conference on Knowledge and Smart Technology (KST), 2019, pp. 11–16.

[10] G Guerreiro, P Figueiras, R Silva, R Costa, R J Goncalves, "An architecture for big data processing on intelligent transportation systems. An application scenario on highway traffic flows", 2016 IEEE 8th International Conference on Intelligent Systems (IS), 2016.

[11] P Figueiras, G Guerreiro, R Costa, Z Herga, A Rosa and R J Goncalves, "Real-Time Monitoring of Road Traffic using Data Stream Mining", IEEE International Conference on Engineering, Technology and Innovation, 2018.

[12] A. I. Maarala, M. Rautiainen, M. Salmi, S. Pirttikangas and J. Riekki, "Low latency analytics for streaming traffic data with Apache Spark," 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, 2015, pp. 2855–2858.

[13] B. Zhou, J. Li, X. Wang, Y. Gu, L. Xu, Y. Hu, L. Zhu, "Online Internet traffic monitoring system using spark streaming," in Big Data Mining and Analytics, vol. 1, no. 1, pp. 47–56, March 2018.

[14] Biem, E. Bouillet, H. Feng, A. Ranganathan, A. Riabov and O. Verscheure, "Real-Time Traffic Information Management using Stream

Computing," IEEE Data Engineering Bulletin, vol. 33, no. 2, pp. 64–68, 2010

[15] M. M. Rathore, H Son, A. Ahmad, A. Paul, G Jeon, "Real-Time Big Data Stream Processing Using GPU with Spark Over Hadoop Ecosystem", International Journal of Parallel Programming, vol. 46, Issue 3, pp. 630–646, 2018.

[16] B. Yadranjiaghdam, S. Yasrobi and N. Tabrizi, "Developing a Real-Time Data Analytics Framework for Twitter Streaming Data", IEEE International Congress on Big Data (BigData Congress), 2017, pp. 329–336.

[17] A. I. Maarala, M. Rautiainen, M. Salmi, S. Pirttikangas and J. Riekki, "Low latency analytics for streaming traffic data with Apache Spark," 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, 2015, pp. 2855–2858.

[18] Yu, L., Wang, S., and Lai, K.K. "Credit risk assessment with a multistage neural network ensemble learning approach", Expert Systems with Applications, 2008, 34, pp. 1434–1444.

[19] X. Qiu, L. Zhang, Y. Ren, P. N. Suganthan and G. Amaratunga, "Ensemble deep learning for regression and time series forecasting," 2014 IEEE Symposium on Computational Intelligence in Ensemble Learning (CIEL), 2014, pp. 1–6.

[20] Yu, L., Lai, K.K., and Wang, S. "Multistage RBF neural network ensemble learning for exchange rates forecasting". Neurocomputing, 2008, 71, pp. 3295–3302.

[21] W. Liyong and P. Vateekul, "Improve Traffic Prediction Using Accident Embedding on Ensemble Deep Neural Networks," 2019 11th International Conference on Knowledge and Smart Technology (KST), 2019, pp. 11–16.

[22] Lai, K. K., Yu, L., Wang, S. Y., and Zhou, L. G. "Credit risk analysis using a reliability-based neural network ensemble model", Lecture Notes in Computer Science, 2006, 4132, pp. 682–690.

[23] Weissenbacher D, Sarker A, Klein A, O'Connor K, Magge A, Gonzalez-Hernandez G. "Deep neural networks ensemble for detecting medication mentions in tweets", Journal of the American Medical Informatics Association, 2019, Vol. 26, Issue 12, pp. 1618–1626.

[24] Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., and Muller, P. "Deep Neural Network Ensembles for Time Series Classification", 2019 International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1–6.

[25] Marushko, E.E., Alexandr Doudkin. "Ensembles of neural networks for forecasting of time series of spacecraft telemetry", Optical Memory and Neural Networks, 26(1), 2017, pp. 47–54.
[26] L.N. Araujo J.T. Belotti, Taalves, Tadano, H. Siqueira. "Ensemble method based on Artificial Neural Networks to estimate air pollution health risks", Environmental Modelling and Software, 123, 2019.
[27] Liu, Y., Yao, X. "Ensemble learning via negative correlation", Neural networks, 12(10), 1999, pp. 1399–1404.
[28] Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J. and Wozniak, M. "Ensemble Learning for Data Stream Analysis: a Survey", Information Fusion 37, 2017, pp. 132–156.
[29] H. Zhang, M. Liptrott, N. Bessis and J. Cheng, "Real-Time Traffic Analysis using Deep Learning Techniques and UAV based Video," 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2019, pp. 1–5, doi: 10.1109/AVSS.2019.8909879.
[30] Shekhar, H., Setty, S., and Mudenagudi, U. (2016). Vehicular traffic analysis from social media data. 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI).
[31] Chicco D., Warrens M.J., Jurman G. 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput. Sci.
[32] Barata, M. and Bernardino, J. Cassandra's Performance and Scalability Evaluation. DOI: 10.5220/0005980101270134 In Proceedings of the 5th International Conference on Data Management Technologies and Applications (DATA 2016), pages 127–134.
[33] Gupta, A., Tyagi, S., Panwar, N., Sachdeva, S. and Saxena, U. "NoSQL databases: Critical analysis and comparison," *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*, 2017, pp. 293–299.
[34] https://www.sciencedirect.com/topics/computer-science/apache-spark

Biographies

B. Mounica is currently working as an Assistant Professor in New Horizon College of Engineering, Bangalore and Research scholar from School of Computer Science and Engineering (SCOPE) in VIT University, Vellore. She completed M.Tech in Computer Science and Engineering from JNTU University, in the year 2012. She also received B.Tech degree in Information Technology from the Anna University, India, in 2005. Her current research interests include Data analytics, Data Mining and Warehousing, Machine Learning, Big data.

K. Lavanya is currently working as an Associate Professor in the School of Computer Science and Engineering(SCOPE) in VIT University, Vellore. She received her Ph.D. degree in Computer Science and Engineering from VIT University, Vellore, on August 2015 [July 2011–August 2015]. She completed ME in Computer Science and Engineering from VIT University, Vellore, in the year 2011. She also received BE degree in Computer Science and Engineering from the Anna University, India, in 2005 Her current research interests includes Computational Intelligence, Data Science, NoSQL databases, Data Mining and Warehousing, Machine Learning.