Traffic Analysis by Using Random Forest Algorithm Considering Social Media Platforms

Kovuru Sridevi , T.Ganesan , B V S Samrat , S Srihari

Abstract— Nowadays social media is an important platform for communication. People started utilizing social media platforms like Twitter, Facebook, WhatsApp, google+, Instagram, etc. People keep updating live incidents through these platforms. The live updates include weather details, traffic details at various places and current events in cities. In major metropolitan cities, traffic is a burning issue. People have to wait for quite a while in heavy traffic. Under such conditions, people started updating the traffic details and congestion details through these platforms. In this paper, social media data is used for analysis and prediction of traffic in the intervals of one hour. A web page has been created for making the estimation data available for users across the globe. Random forest algorithm is used for estimation of traffic based on the traffic congestion level three hours before the estimation period, the day of the week, whether the day is a holiday or not. 88% accuracy is achieved using this model. This model also presents an alternative route by comparing predicted traffic by this model in all the possible routes and suggesting the best possible route with minimum traffic for end users convenience.

Index Terms— Data mining, Prediction, Random Forest Algorithm, Social Media, Traffic Analysis.

1. INTRODUCTION

According to the recent reports on social media analysis, out of 7.5 billion people across the globe, 3.19 billion people are active social media users i.e., 42%. This demographics and statistics show that social media plays a major role in updating the current events in the world. This is taken as an advantage and used to report traffic updates to the world. This will help people who are not using social media but are in great need of the updates. Traffic is a major field that needs concern because of the tremendous increase in population in cities and a large increase in the use of automobiles as opposed to using public transport. People not only update their profile about personal life but also focus on social events. These social media updates contain unformatted text, which should be processed to get information. The major concept involved in this paper is data mining. Machine learning algorithms are also used so as to process data and predict the traffic at different places and at different times.

The continuous urbanization has left the cities authorities with numerous issues, mainly in view of traffic control. Using the model in this paper, one can avoid traffic congestion by taking alternative routes when a route found congested. To reduce these problems, technology can be utilized effectively. This paper suggests a model which will help people in all aspects of traffic. To accomplish this task, data extraction is needed from social media sites and the application of some processing methods to get data useful to the end user. The processing techniques used are data mining and machine learning. These are the most important techniques to analyse the data.

Processing the data is needed to make it available to apply data analysis techniques directly. This is the place where data mining techniques come into the picture. Data mining is nothing but examining huge data sets and extracting the required information. After getting the data in the required form, machine learning methods are applied to make predictions.

Huge data about traffic is available on social media. By processing that data, it is converted into the desired form to apply further analytics models. Here the goal is to predict the traffic based on the traffic congestion level three hours before the estimation period, the day of the week, whether the day is a holiday or not. To get these results machine learning technique is applied. Machine learning is a field of artificial intelligence that uses statistical methods to give computer systems the ability to learn with the help of data, without any explicit programming. There are many machine learning techniques used for prediction, like Naïve Bayes, random forest, support vector machine, k-nearest neighbour etc. Here, random forest algorithm is used to predict the traffic in the intervals of one hour after comparing the results with other machine learning algorithms.

Random forest is a supervised learning algorithm. It is a learning method for regression, classification and other tasks. This algorithm works by creating a group of decision trees at the time of training and output the class that is either mean or mode of the individual trees. There are two platforms widely used by people for analysis. They are WhatsApp and Twitter. When WhatsApp is considered, the work is implemented by getting the data from a group chat history which is solely

Revised Version Manuscript Received on March 08, 2019.

Kovuru Sridevi, Computer Science and Engineering Department, KLEF, Vaddeswaram, Guntur Dt, Andhra Pradesh, India. (e-mail: k.sridevi.1998@gmail.com)

T.Ganesan, Computer Science and Engineering Department, KLEF, Vaddeswaram, Guntur Dt, Andhra Pradesh, India. (e-mail: tganesanit@gmail.com)

B V S Samrat, Computer Science and Engineering Department, KLEF, Vaddeswaram, Guntur Dt, Andhra Pradesh, India.

S Srihari, Computer Science and Engineering Department, KLEF, Vaddeswaram, Guntur Dt, Andhra Pradesh, India.

meant for traffic updates posted by different members of the group. Then the data required is extracted followed by application of various analytics.

When Twitter is considered as the base platform for work, extraction of tweets is done by using Twitter API or Twitter Archiver. This paper used Twitter Archiver for extracting tweets. Using this application, tweets are based on user tweets, hashtags, geotagged tweets, advanced search or brand mentions. After extracting the data, it is used to apply analytics as per requirement.

By applying these analytics, the goal is to help people finding the traffic in the desired location easily and get rid of traffic problems. Users are provided with all kinds of data required to make a proper estimation of the journey. This will be very helpful for people who don't use social media. This data is made available to everyone by creating web pages for an easy user interface. These pages can be accessed by anyone from anywhere across the globe to know the traffic details without the need to disclose their identity. The group which is considered works only on traffic updates from various areas. This will fetch data from desired locations.

The rest of the paper is presented as – section II gives the related work about this concept. Section III presents the proposed system. Section IV is about results and discussion. The paper is concluded in section V and future scope in section VI.

II. RELATED WORK

Article[1] deals with real time traffic monitoring and analysis model. Here, it is dealt with classifying the tweets as either traffic related or non-traffic related. Support vector machine algorithm is used for classification as it is a binary classification. Binary classification is used since here there are only 2 alternatives either traffic or non-traffic. Here 95.75% accuracy is achieved. It is also used to discriminate the reason for traffic as, by an external event or not. For this multiclass classification is used and 88.89% accuracy. It considered user messages as Status Update Message (SUM). Article[12] also used the SVM technique.

Articles[2],[13] also deals with classifying a tweet into traffic event or non-traffic event. Here, natural language processing and support vector machine algorithms are used. Natural Language Processing is used for analysing the tweets and support vector machine is used for classification. Hence, the system is capable of detecting traffic data.

Article[3] is all about creating a web service and an android app. Web service here uses HDFS which is Hadoop dynamic file subsystem database. Here, natural language processing is used for classification. The author didn't use binary classification as in articles[1],[2]. The author used multi-class classification in order to classify tweets so as to recognize traffic, non-traffic due to the crash or congestion and also traffic due to many external events like earthquakes, tsunami, social events etc. The android application will help to take a route from source to destination and find if the traffic is heavy or not.

Article[4] deals with applying text mining concept. This also as in articles[1],[2] deals with assigning a label as traffic tweet or not. Here the method used is the same as in articles[1],[2] which is the support vector machine. By using binary classification 75% accuracy and by using multi-class classification 88.89% accuracy is achieved.

Article[5] deals with creating an android application which should be accessed by logging in with username and password to get the path from source to destination, see traffic and choose an alternative path if required. This paper has used natural language processing for text mining. The speciality with this system is it will suggest an alternate path if the traffic in the original path is heavy.

Article[6] deals with the same concept as in articles[1],[2]. That is using natural language processing for text mining and support vector machine for binary classification into traffic or non-traffic event. If there is a heavy traffic in any of the paths, then it suggests an alternative path. It is built on SOA architecture. It also does multi-class classification. This paper deals with the accident, traffic jam, vehicle breakdown alerts also.

Article[7] works on the same concept as in the above articles. It categorizes the tweets into traffic or non-traffic tweets. Here also, natural language processing is used. This system provides the cause of the traffic. An Android application is also developed for this purpose. Here, the k-means clustering algorithm is used. Article[15] also used the clustering method to find if a tweet is traffic related or not. Euclidean distance method is employed to calculate the similarity between the tweets.

Article[8] is a survey work on this topic. Here, the author retrieved the data, cleansed the data and normalized it. Later applied Naïve Bayes algorithm on the data to classify the tweets into traffic or non-traffic data.

Article[9] works on the prediction of traffic congestion condition. Here, C4.5 decision tree model is used for prediction. Tweets from some particular accounts are considered based on geo-tagged and traffic- related keywords and made a web application which is accessible by everyone and can be updated every 30 minutes. Article[14] also works on prediction and achieved 96.31% accuracy using the k-fold method.

Article[10] works on the same concept as in articles[1],[2],[3] which is classifying whether a tweet is traffic related or not using support vector machine. The semi-supervised approach is used to achieve this goal.

Opinion mining is nothing but sentiment analysis. That means, to find if a tweet is either positive or negative or neutral. In the article[11], the author has used tweets from Twitter related to traffic and categorized as positive or negative. 87.15% accuracy is achieved.

Now, this paper analyses if a tweet is positive or negative or neutral by the application of sentiment analysis. Moreover, this predicts the traffic and suggests an alternative route by considering predicted traffic by this model. This makes the paper unique.

III. PROPOSED SYSTEM

This model deals with certain steps. They are:

1. Collection of tweets(data)
2. Processing the data
3. Analysing the data
4. Prediction Model
5. Alternative route suggestion

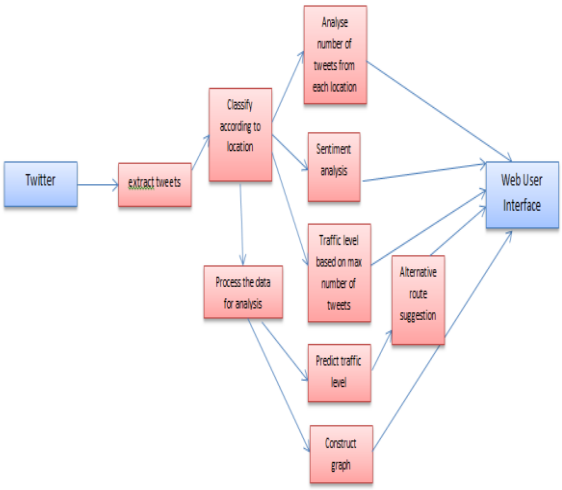

Figure – 1 – Steps involved in the system

1. Collection of tweets(data):

In order to achieve the above-said goals, at first, extraction of tweets from Twitter or chat history from WhatsApp is needed.

From Twitter:

Twitter archiver is used to create a new google sheet with add-on and extract tweets based on hashtag search. Or use Twitter API by creating consumer key, consumer secret key, access token and access token key. By using these keys, tweets are extracted from Twitter.

Tweets are extracted based on hashtags, user tweets, geotagged tweets, advanced search or brand mentions.

After getting the tweets into an excel sheet, different analytics methods are applied to the data to achieve the above goals.

From WhatsApp:

A group is created in WhatsApp which deals with traffic updates. Members of that group keep updating the traffic details by posting a message in the group. By going into options of the chat, it is exported to mail. From there file can be downloaded and processed to get the data in the desired format.

2. Processing the data:

The process begins with the excel sheet of tweets with all the details. Then extract only required fields into another excel sheet. The fields extracted are tweet and tweet location. And then, the tweets are classified based on the location of the tweet and extract date, day and time. Each sheet in excel represents a location and tweets automatically get into the sheet based on location by using a python code. By executing that code, all tweets are organized in sheets as per their location.

3. Analysing the data:

Analysing the data includes a certain number of tasks:

1. Sentiment analysis of the collected tweets
2. Classifying the traffic level based on the highest number of tweets on that particular route at a particular level
3. Plot an analytics graph on different areas traffic levels during different times of the day
4. Analyse the number of tweets from each place

Task-1- Sentimental analysis of the collected tweets:

Sentiment analysis is the process of identifying and categorizing opinions expressed in a text to determine the attitude of the writer, whether it is a positive, negative or neutral opinion about the topic. Here, sentiment analysis is done by considering heavy traffic as a negative sign, slow traffic as a positive sign and normal traffic as a neutral sign. Here, to perform sentiment analysis, the Naïve Bayes algorithm is used. It is a popular algorithm for classifying text. It is based on the Bayes theorem. Naïve Bayes algorithm has a high success rate in sentiment analysis when compared to other algorithms. This algorithm is based on the bag-of-words model. With this model, a word of the document is checked if it is in a positive words list if so, the total score is added with 1. If the word is in negative words list, 1 is subtracted from the total score. If the word is not present in the training set, Laplacian smoothing is applied, that is using '1' instead of the conditional probability of the world.

S.NO	SENTIMENT	EXAMPLE WORDS
1	Positive	Slow, fine, good, fantastic, nice
2	Negative	Heavy, hate, terrible
3	Neutral	Normal

Table – 1 - Sentiment analysis

Task-2-Classifying the traffic level based on the highest number of tweets on that particular route at a particular level:

This task is based on the analysis of the majority. For example, if 10 tweets are considered between source and destination, if 6 out of 10 represents 1 traffic rate, that rate is considered as the final traffic rate. If the model gets 10 tweets from a route, out of which 7 are indicating heavy traffic, 2 are indicating normal traffic and 1 tweet is indicating slow traffic, then it is considered as heavy traffic on majority criteria. A python code is employed to implement this in real-time.

Task-3-Plot an analytics graph on different areas traffic levels during different times of the day:

This is a graphical view of the data available with the model. Here, an 'R' code is used to plot the traffic for each hour at each place. This will give a nice view of the data available. This will help to analyse the data since pictorial representation matters a lot.

Task-4-Analyse the number of tweets from each place:

Regarding this task, the usage of social media by people at different locations is analysed. By counting the number of traffic-related tweets at each place, it is known how the people are engaged with the social media platforms. Here, a python code is used to accomplish this task.

4. Prediction model:

This paper developed a prediction model for traffic. One year data is gathered from December 30th, 2017. Now, by using this data, traffic is predicted for the next day. The data include the date, day, month, and holiday or not followed by 24-hour data in hourly intervals.

This is used to train the model in 70-30 ratio for train and test set. This model predicts the traffic one hour prior. The constraints taken into consideration for traffic prediction are whether it is a holiday or not, day of the week and three hours traffic ahead of the target hour. This will predict the traffic in advance.

The dataset is kept updated hourly and try to predict the hourly traffic so that it will help people not waste their time by spending hours in traffic areas. Random forest algorithm is used here and achieved 93.42% by using the random forest algorithm.

Random forest algorithm will add randomness to the model, at the time of growing the trees. It will search for the best feature among the random subsets. Random forest considers the average of the results of all decision trees generated.

Here, the test features are taken into account, used rules of each decision tree to predict the outcome and stores the predicted output. Then calculated the votes for each target and take the high voted target as the final prediction for the random forest algorithm.

Three other algorithms are also used for comparison. They are C4.5, k nearest neighbour, support vector machine algorithms. By comparing these 4 algorithms this paper concludes that random forest algorithm best suits this model.

5. Alternative route suggestion:

When the user searches for a route from source ‘A’ to destination ‘B’ at a particular time, instead of just predicting the traffic and showing a single route with the predicted traffic, traffic is predicted using this model in all the possible routes from the given source to destination and compare the results. After comparison, the best route is suggested from source ‘A’ to destination ‘B’ having minimum traffic. By doing this, two problems of the users are solved. The first problem is when the route required is predicted with heavy traffic, users have to wait until the traffic is cleared. But now user need not wait instead go with an alternative route. The second problem is the burden on the end user to search for an alternative route. This is reduced by showing all the possible routes with predicted traffic. Now, users have the choice of selecting the route. This model shows the route with minimum traffic.

Finally, a web user interface is developed for all the users. This model is highly beneficial because this is made available for all the users across the world and people can access it without disclosing their identity.

IV. RESULTS AND DISCUSSION

This model is implemented in real-time traffic and achieved an accuracy of 86% on a comparison of real-time traffic with this model results.

1. Collection of tweets(data):

Almost 95% tweets are extracted from Twitter. The loss of 5% is due to unclear tweets. Unclear tweets include noisy tweets and irrelevant tweets.

2. Processing the data:

The data is classified based on location and a dataset is made ready for analytics. Date, day, month, holiday or not and 24-hour data are included. A python code is used and applied text mining process.

3. Analysing the data:

Task-1- Sentimental analysis of the collected tweets:

Naïve Bayes algorithm is used for sentiment analysis.

S.No	Tweet	Sentiment Analysis Result
1	Koti to Secunderabad traffic is heavy	Negative
2	Secunderabad to Charminar traffic level is slow	Positive
3	Uppal to Kukatpally traffic is normal now	Positive

Table – 2- Sentiment analysis results

According to the above-cited results, 80% accuracy is achieved.

Task-2-Classifying the traffic level based on the highest number of tweets on that particular route at a particular level:

Here, the result is shown as the maximum voted traffic level.

```
koti-secunderabad ---heavy traffic

1 1 0
secunderabad-charminar ---heavy traffic

2 2 0
mehdipatnam-secunderabad ---normal traffic

0 0 1
secunderabad-uppal ---heavy traffic

1 0 0
uppal-kukatpally ---normal traffic

0 0 1
>>>
```

Figure – 2 - Maximum voted traffic level considered

Here, among the three columns first one is indicating heavy traffic level, the second column is indicating slow traffic level and the third column is indicating normal traffic flow. If the same number of votes for heavy and slow, or heavy and normal are encountered, heavy is considered to avoid risk. If the same number of normal and slow votes, then normal traffic level is considered.

Task-3-Plot an analytics graph on different areas traffic levels during different times of the day:

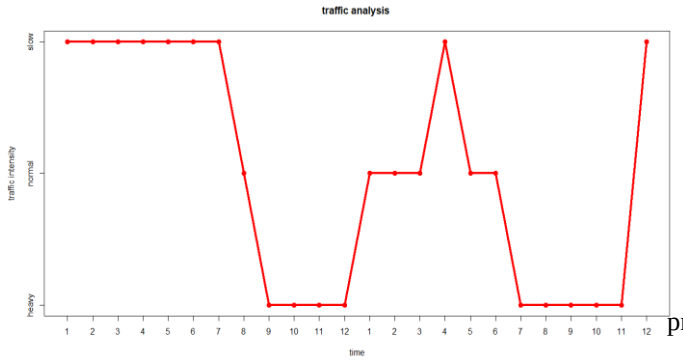

Figure – 3 – 24-hour traffic data graph

A route is taken and plotted the 24-hour traffic levels. This is implemented for any other route also in the same way.

Task-4-Analyse the number of tweets from each place:

The number of tweets is calculated from each place. Here, three cities are considered namely Hyderabad, Vijayawada, and Chennai. The count of tweets is displayed from each of the above-described cities.

```
no. of tweets from hyderabad: 9

no. of tweets from vijayawada: 7

no. of tweets from chennai: 7
```

Figure – 4 - Tweets count analysis

4. Prediction model:

Here Random forest algorithm is used. The results are as follows:

```
[1] slow
Levels: heavy normal slow
```

Figure – 5 - Traffic Prediction

Confusion Matrix and Statistics

Prediction	Heavy	Normal	Slow
Heavy	0	0	0
Normal	0	51	2
Slow	0	3	20

Table-3-Results of prediction

Accuracy : 0.9342

Accuracy

Algorithm	Min	1st qu.	Median	Mean	3rc	
KNN	0.7500000	0.8160920	0.8469388	0.8443536	0.869565	0.913943
SVM	0.7888889	0.8505747	0.8762887	0.8754557	0.903263	0.979381
RF	0.8152174	0.8453608	0.8681319	0.8672697	0.882978	0.936842

Table-4-Accuracy measures

Accuracy is the percentage of correctly classified instances out of all the instances in the data.

Kappa

Algorithm	Min	1st qu.	Median	Mean	3rd qu.	Max
KNN	0.2425249	0.3463471	0.4427422	0.4332139	0.5073628	0.6274038
SVM	0.2936345	0.4756839	0.5620789	0.5594167	0.6233538	0.9049951
RF	0.5782093	0.6323706	0.6936027	0.6938403	0.7339674	0.8610095

Table-5-Kappa measures

Kappa is accuracy in classification. It is more useful for problems which have an imbalance in classes.

When this is compared with real-time traffic, 86% accuracy is seen.

5. Alternative route suggestion :

This model takes the source and the destination input from the user. Then find all the routes from source to destination and predicts the traffic in all the routes found. Later compares the traffic levels in all the routes and finally suggests the route with minimum traffic.

Finally, the webpage view of our system is:

Figure – 6 - Web page view – 1

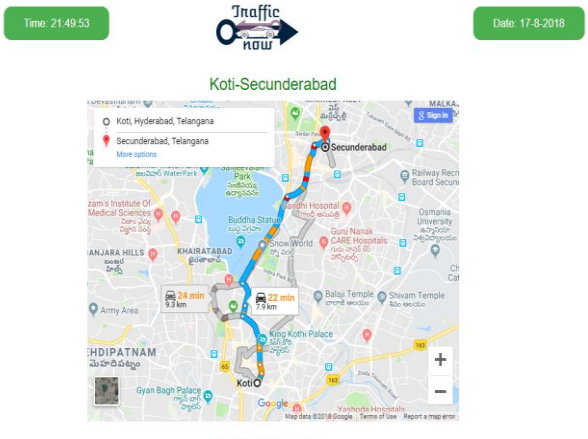

Figure – 7 - Webpage view-2

Here, the traffic is updated in the database by a python program and database is connected with a webpage by using PHP code and displays the data from the database.

The interface also presents the live updates from twitter. The tweet as soon as it reaches the twitter is taken and processed to update the traffic live status in that particular location mentioned in the tweet.

Here, the route of the desired source and destination is shown. This will help people to know the route before starting the journey. This will help to decide another alternate route when the traffic is heavy.

V.CONCLUSION

This paper proposes a system for real-time traffic analysis and prediction. Analysis of the Twitter data includes extraction of tweets from the Twitter using Twitter API or Twitter archive, then classifying the tweets based on location. This paper also deals with applying sentiment analysis for the tweets to classify it either into positive or negative or neutral tweets. Then, preference is given to the traffic level, which is tweeted by the majority of people. A system is created to predict traffic based on date, day and holiday or not data along with three hours of traffic data prior to the targeted hour. Graphs are plotted based on 24-hour traffic in a route. Finally, this paper analyses the number of active participants in each place based on the number of tweets from that place. This model uses random forest algorithm for prediction. This algorithm is more useful as it is applicable to both classification and regression (prediction). This algorithm has given a great accuracy of about 86% when compared to real-time traffic. Alternative routes are suggested for a given source and destination. Live status is also given through this model. This will help people in many ways.

VI. FUTURE SCOPE

This can be extended to a level by which traffic can be predicted based on external factors like –if some festival is being celebrated or if there is a procession or a political meeting, or an event, or earthquake, or natural calamities. In addition to suggesting the minimum traffic route, this can be extended to a level at which suggestions are based on the minimum traffic route with least distance. It can be extended to a level that traffic prediction also considers wrong routes, shortcuts and non-popular routes in the city into account.

ACKNOWLEDGMENT

We would like to thank T.Ganesan sir for guiding and helping us in our work. We would also like to thank everyone who helped us and supported us.

REFERENCES

1. Elonora D' Andrea, Pietro Ducange, Beatrice Lazzerini, IEEE, and Francesco Marclloni, Member, IEEE, "Real-Time Detection of Traffic From Twitter Stream Analysis", IEEE Transactions on intelligent transportation systems, vol.16, no.4, August 2015.
2. Mr.Pujari Aditya, Mr.Taware Ashish, Mr.Lakade Pankaj, Mr.Wagh Mahesh, Prof. Nale R.K, "Tmap:Traffic Detection using Tweet Analysis", International Journal for Research in Applied Science & Engineering Technology(IJRASET), volume 5 issue III, March 2017.
3. Sandeep G Panchal, Prof. R.S.Apare, "Real Time Traffic Detection using Twitter Tweet Analysis", International Journal of Engineering Trends and Technology(IJETT), volume 47, number 8 may 2017.
4. Sweety Kumari, Firdos Khan, Shekh Sultan, Ruchita khandge, "Real-Time Detection of Traffic From Twitter Stream Analysis", International Research Joural of Engineering and Technology(IRJET), volume-3, issue-4-2016.
5. Mrs.Kavita Sawant, Miss. Shital Pawar, Miss. Poonam Jadhav, Mrs. Sayali Vidhate, Mrs. Nirasha Bule, Mrs. Snehal Patil, "Traffic Detection from Real Time Twitter Stream Analysis and Navigation System", Intrnational Journal of Engineering Science and Computing(IJESC), volume 7 Issue no. 5
6. Sayali Dhanawad, Rucha Kulkarni, Shraddha Raut, Prof. D.S. Gogawal, "Twitter stream Analysis for traffic detection in real time", International Journal of Advance Engineering and Research Development, volume 4, issue 4, april-2017.
7. Harshita Rajwani, Srushti Somvashi, Anuja Upadhye, Rutuja Vaidya, Trupti Dange, "Dynamic Traffic Analyzer using Twitter", International Journal of Science and Research (IJSR), Volume 4 issue 10.
8. Sandeep G.Panchal, Ravindra S.Apare, "A survey on Traffic Detection from Twitter Tweets Analysis", International Journal of Science and Research (IJSR), volume 5 issue 12.
9. Sakkachin Wongcharoen and Twittie Senivongse, "Twitter Analysis of Road Traffic Congestion Severity Estimation, 13th international joint conference on computer science and software engineering (JCSSE).
10. Chaudhari Bhavesh N, Dalvi Rajat R, Divate Karishma A, Narkhede Charulata T, Prof. Handore Sonali A, "Monitoring System for Traffic Analysis Using Twitter Stream", International Journal on Recent and Innovation Trends in Computing and Communication, volume 4, issue 10.
11. K.Boopalan, C.Nalini, A.Rajesh, "Mining Opinions about Traffic Status Using Twitter Messages", International Journal of Civil Engineering and Technology (IJCIET), volume 8, issue 2.
12. Dr Manna Sheela Rani Chetty, Mahesh Dhanvi Yamarthi, Sai Mounika Varikuti, Manohar Sai Jasti, "Instantaneous Monitoring of Road Traffic From Twitter Stream Analysis", International Journal of Pure and Applied Mathematics, Volume 115, No. 8.
13. B Sureesha, V Priyadarshini, "Monitoring and Analysis of Dynamic Traffic Analyzer using Twitter", International Journal of Computer Science and Technology, volume 7, issue 4.
14. Arief Wibowo, Edi Winarko, Azhari, "Predicting the Road Traffic Density Based on Twitter Using The TR-P Method", International Journal of Computer Science and Network Security, Volume 17, issue 8.
15. Supriya Bhosale, Sucheta Kokate, "Traffic Detection Using Tweets on Twitter Social Network", International Journal of Advance Research in Computer Science and Management Studies, volume 4, issue 7.