Chapter 11 Big Data Analytics

Introduction

Lot of traffic related data of different modalities gets generated in transportation setup, especially in urban regions and smart cities. The traffic cameras installed at different junctions capture traffic videos and send them to central servers. There are cameras with additional processing capabilities that can count the number of vehicles at different time instances, and also systems that can detect speed, headway etc. Often users report traffic-related incidents or scenarios through social media. All these data can be helpful to generate several insights about current and future traffic scenarios including but not limited to data-driven identification of traffic bottleneck, seasonality in traffic volume, route planning, and also identifying and alerting about current or future traffic incidents.

1: Collecting Traffic Related Data for Processing

The traffic related big data can be of various modalities such as video (feeds from traffic cameras), stream of numbers (such as counts reported by traffic counters), text and image (traffic-related posts by users and city governance). Such data can be stored in appropriate storage for further analysis.

2: Analyzing Traffic Data for Better Insight

As part of the project, traffic cameras were set up at different parts of the city. The locations of the cameras are shown in Figure 11-1. The cameras additionally gave information about the number of vehicles that crossed the junctions at specific intervals, vehicle speed, length of the vehicles etc. This information was extracted and analyzed.

For example, Figure 11-2 shows the variation of speed and the variation of vehicle length at two junctions, as captured by the cameras. We see that the average speed drops down after 10 am, which is generally true due to work schedules of offices and businesses. The speed decreases further around 3 pm, which is generally not a busy hour. However, we see from Figure 11-3 that the number of larger vehicles (possibly heavy vehicles) increase slightly around that time. These vehicles are generally slower due to the load. This mixed traffic leads to a slowdown in the speed. The plots also corroborate the fact that vehicle speeds are generally higher

Figure 11-1 The region considered for the study related to traffic counts

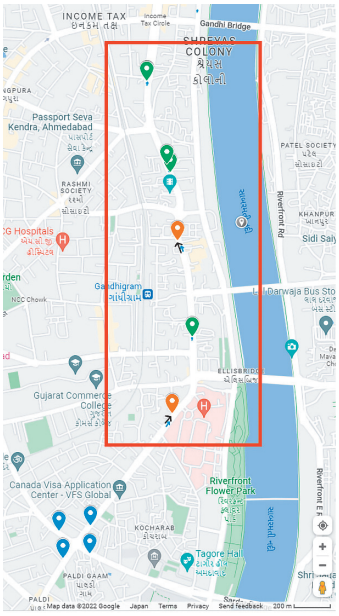


Figure 11-2 Average Speed Heatmap

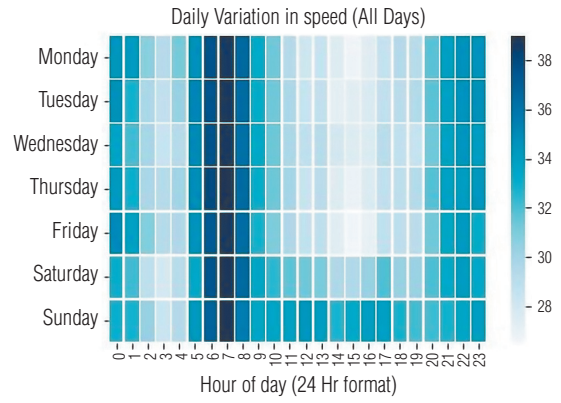


Figure 11-3 Length of vehicles

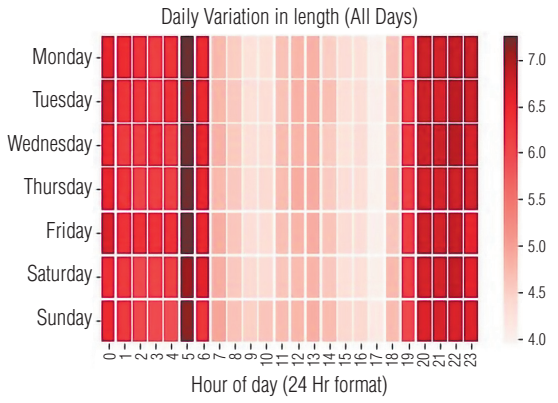


Table 11-1 Traffic Count Prediction

Model	RMSE	MAPE
LSTM	16.44	33.38%
Conv LSTM	15.73	23.60%
Social LSTM	15.45	20.97%
Deep Transformer	15.43	20.98%

Figure 11-4 Root Mean Squared Error vs actual counts

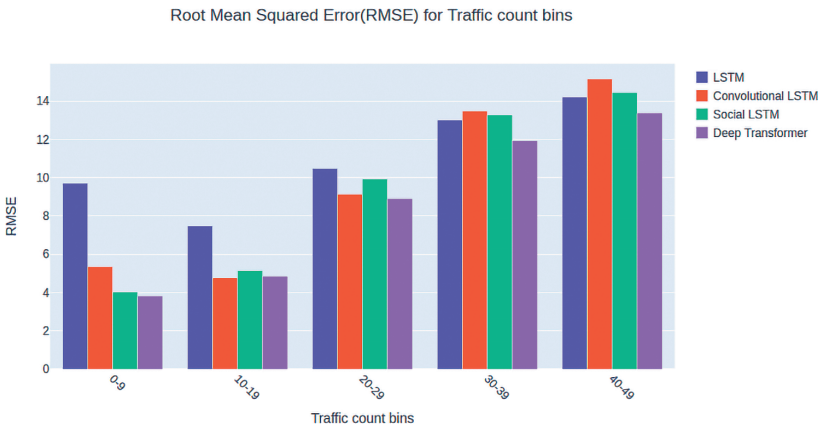
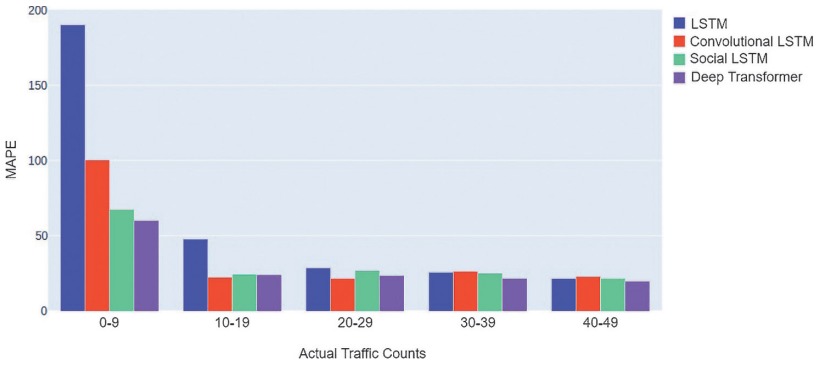


Figure 11-5 Percentage Error vs Actual Count



on weekends due to lesser traffic on the road. This may also indicate the necessity of deploying traffic personnel on road to regulate the traffic at these hours.

3: Predicting Future Traffic

Appropriate forecasting of future traffic can help in predicting congestions, which can in turn help in making various decisions. For example, based on this information civic bodies can decide on deployment of traffic personnel, adjust signal timings, and individuals can make appropriate routing decisions. Such actions reduce congestion and the waiting/plying time of vehicles on the road, thereby reducing the amount of emission. However, traffic forecasting, which deals with the prediction of traffic at different future timesteps based on past traffic information, is an extremely challenging task due to the highly complex traffic patterns. The traffic forecasting problem has often been posed as a time series prediction problem, where the input is the traffic volume (e.g. vehicle count) at the different past timestamps (say 7 am, 7:15 am, 7:30 am etc.) and the output is the expected traffic volume at that junction at some future timestamp (e.g. 9 am). These traffic counts can be obtained from the traffic counters (either from the traffic cameras directly) or by analyzing the traffic videos captured from different junctions.

Time series modeling can be performed using different techniques. One can use the ARIMA based statistical models for the same or use the recent modeling techniques involving LSTMs and stacked-LSTMs. These techniques only consider the past counts for the future predictions. However, the traffic volume is also impacted by the spatial layout of the junctions, and traffic from one junction is influenced by (and itself influences) traffic at other junctions. Hence advanced methods, that consider spatial information also while making the future traffic predictions can give better predictions.

To verify the possibility of making good predictions for future traffic volume, a case study was performed based on traffic data collected from the Ahmedabad city. The performance of different models for this prediction task are shown in Table 11-1. Here, RMSE stands for Root Mean Square Error and MAPE stands for Mean Average Percentage Error. Since both are error metrics, lower values indicate better performance.

From the above comparison, we see that the social LSTM that considers the spatial information, and the deep transformer model that considers higher order interaction between counts can predict the traffic volume with more accuracy. The column plots shown in Figure 11-4 and Figure 11-5 indicate that as the count of actual vehicles increases, the misprediction count (RMSE) also increases slightly, however the percentage error comes down considerably.

Figure 11-6 Traffic Count Prediction

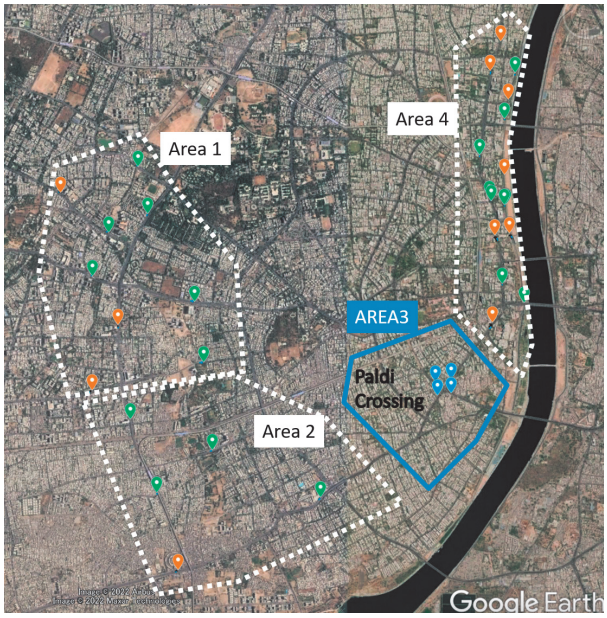
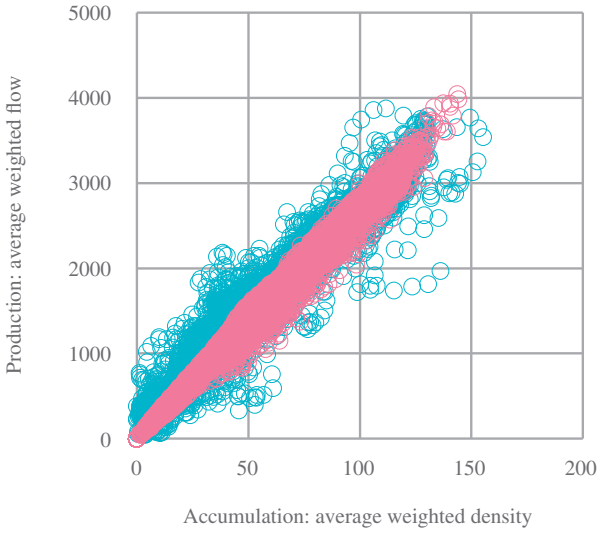


Figure 11-7 Estimated MFD indicators



These observations mean that by employing appropriate methods, it is possible to predict the volume of future traffic on the roads. This can help in better preparedness of intelligent decision making for predicted high congestion levels at future timestamps.

Predicting Area Traffic Condition

Macroscopic Fundamental Diagram or MFD is one of key performance indicator to express area traffic states. MFD expresses the interaction between area traffic states which connecting the total number of cars on the road at any given time (the accumulation) with the rate at which

trips reach their destinations (the output). The area is set as group of detectors in Figure 11-6. Paldi area (Yellow boundary) is a targeted area of estimation form other three area. Both MFD indicators were estimated with high accuracy in comparison with observed data form traffic detectors in Figure 11-7.

MFD is available for local government and its traffic agency to manage and control traffic in the viewpoint of area. Dynamic Inflow restrictions can be applied the areas where traffic is concentrated and it can be said that traffic can be dispersed spatially, leading to efficient use of the road network as stock.

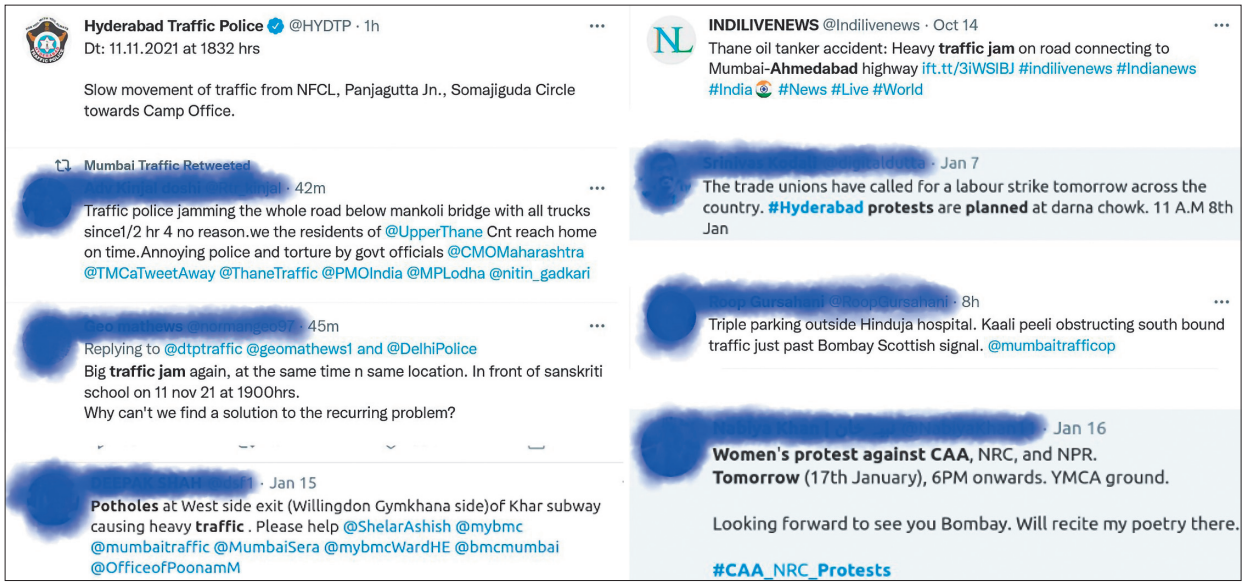
Using Social Media Data for Identifying Traffic Incidents

Often road traffic congestion can be induced by incidents or events like road accidents, infrastructure damages, rallies, protests, adverse weather events, disabled vehicles, roadway debris, daily rush hours etc. Detection of these incidents on time or ahead of time wherever possible will help the traffic authorities to alleviate road congestion problem, and can help the commuters as they can pre-plan their trip accordingly.

As mentioned earlier, several advanced devices such as loop detectors, GPS probe vehicles, cameras etc. installed on transportation network can help detect traffic congestion. But, due to rapid growth of transportation networks the cost of procurement, installation and maintenance of these sensors also increases. On the other hand, with more and more people joining social media and posting information about real information from the ground, obtaining traffic related information from social media has become a possibility. Such information would complement the information which can be obtained from the hardware devices in the traffic network.

It is observed that civic authorities as well as general people or groups often publish traffic related data in social media sites like Twitter. The Twitter handles of Traffic Police of several major Indian cities are highly active and give periodic updates and alerts about traffic flow. Users also report about congestions, road blockades, diversions. Interestingly, it has also been observed that often reports or news about future incidents like rallies, processions, marches etc. that can affect traffic are posted in social media well ahead in time. The impact of such events on the traffic can be detected by the cameras, loop detectors and other hardware sensors only at the time of the impact. Active consumption and

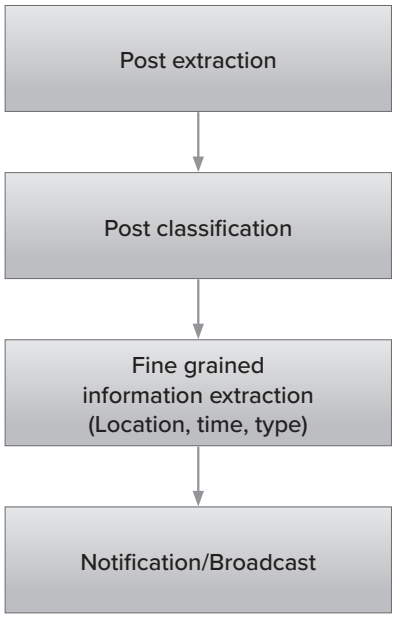
Figure 11-8 Sample tweets about traffic scenario or incidents



filtering of social media data can thus complement the information given by the hardware sensors.

As it can be seen from the tweets shown in Figure 11-8, many posts contain information about current situation, situation due to recent past incident and also about future events that can impact traffic. Often these posts contain specific location information to get an idea of which region will be impacted. Timely extraction of such posts, and then identification of the nature of impact, time, and location can be helpful for traffic planning.

Figure 11-9 Workflow of a system for utilizing social media posts for traffic information extraction



As part of the project, a system was developed to consume social media posts and firstly determine whether they are about traffic incidents. Important fine-grained information from the posts were extracted. The same information can then be broadcast of notified to users. A brief workflow of the system is shown in Figure 11-9.
For example, consider the following tweets (real tweets extracted from social media):

1. "All are invited to attend the rally for filing Nomination Papers of Dr Bhagwanth Rao BJP candidate of Hyderabad parliament Constituency on 25/ 3/ 2019 at 9am from Bhagyalaxmi temple Charminar"
2. "#Pedestrians account for 40% of all #road #traffic deaths in #Bengaluru. It is critical to design our #streets for #pedestrians, save lives - @akhilasuri talks to @ChristinMP_TOI"
3. "Surgery of Tumor 23.5 + kg at Gujarat Cancer Research Institute:GCRI Ahmedabad #oncology #surgery #oncosurgery #tumor #Cancer #India #Gujarat #Ahmedabad @GCRI_1972 @civilhospamd @PariseemaD @bonestumor @MoHFW_GUJARAT"

Among the above tweets, the first tweet is related to traffic incidents. The third one is not related to traffic incident. Although the second tweet belongs to traffic policy, it is not related to traffic incident. A well trained machine learning system is able to classify such contents. Also from the fist tweet, specific information like event type="rally", date="25-Mar-2019", Time="9 am", and location="Bhagyalaxmi temple Charminar" can be extracted.

Mobility Performance Report

The mobility performance is able to be provided by Big Data Annalistic. The department, related of traffic management within existing resources, collect, analyze, and summarize highway congestion data

and make it available upon request to regional transportation planning agencies, congestion management agencies, and transit agencies. The mobility performance by the department with big data shall be opened not only these management agencies but also to the public. The citizen is ability to know their current traffic state and environmental impact so that they have an opportunity to change their mobility for better mobility of themselves and sustainable choice for better environmental impact.

A lot of countries and region has been trying to make a mobility performance report. As a example of California department of Transportation, United State of America, data is collected in real-time from nearly 40,000 individual detectors spanning the freeway system across all major metropolitan areas of California. These big data opened the two ways:Website and reports. The website, called Performance Measurement System (PeMS), is available for related public agencies, private-traffic information suppliers, researchers/engineers and citizen. The accessibility to the big data is enhanced. On the other hand, annual mobility performance report is legislatively mandated by Government Code in the case of U.S.A. The department of transportation, CA, issues annual and quarterly reports, bottle neck mapping and et al through the website.

The big data and mobility performance should be recognized as shared wisdom toward smart transport and city in the viewpoint of both public agency and citizen.

1: Motivating more people to use public transport by providing timely information

Smart systems that can help people to take transport related decisions can be helpful for various aspects. For example, on-demand information regarding public transport schedule or current running status, information about congestion along different routes, accurate estimate of travel time along junctions, information about planned or unplanned events or incidents along traffic route etc. can be helpful for people to take transport/commute related decisions. Although this information can be obtained from multiple sources, this information is scattered over different special purpose applications and websites. People tend to use multiple mobile based applications or websites to get transit schedules for different modes of transport. Online schedules in many instances are clunky, involving multi-step input attributes to gain a simple answer. Many have pdf-based schedules or separate links to each metro/rail line. In some instances, there are separate sites for bus and rail.

Going to each individual source according to the exact information need, and forming an appropriate instruction by filling out input fields in a web-based interface or app (e.g. for finding train schedule or running status), or searching in a map (e.g. finding nearby points of interest), or going through large text to figure out (e.g. prior information about traffic diversion in a route, pdf-based schedule) requires lots of effort from the user. It is desirable to streamline these different types of activities and bring them under one common interface where users can seek information in a natural way through conversations. Due to the ease of interactions, such conversational systems will make the task of information seeking easier for everyone, and especially for non-tech-savvy users. Moreover, it will save the user from taking multiple steps (opening the app/website, entering specific search query in the form expected by the service etc.). Getting timely information about public transport services can help towards better adoption of public transport. Efficient route finding and alerts regarding traffic related incidents can help people from taking appropriate routes thereby reducing congestion and traffic delays. Hence, development of such AI-powered conversational systems for smart transportation can be beneficial for implementation of smart cities and urban transportation systems.

Towards this goal of developing a AI-powered conversation system for transport queries, suitable dataset is required. Unfortunately, there are no datasets for this purpose that fits the necessity of a multilingual country like India, where people prefer communicating in regional languages. Often in a single utterance there are both regional language and English language words. Handling such code-mixed utterances in a conversation requires special attention.

As a part of the project, we have created a dataset and also an experimental system for a transport domain conversational system. We call the dataset as mTransDial. The system can address user queries belonging to multiple categories such as place-search, train-search, distance-search, traffic etc. A brief description of the query intents are provided below:
Place_search: It contains queries for searching nearby Points of Interest like petrol pumps, toll plazas, shops, restaurants etc.
Train_search: It contains queries to find bus/train/metro from a source to destination place.
Distance_search: It consists of queries to find distance from a source to a destination, and time required to reach destination.
Traffic: Queries about traffic information and travel alert on a particular place/city are added under this intent.
Greeting: This intent is added to wake up the dialog assistant and has queries like “hi”, “hello” etc.
Thank you: This intent is for the dialog system to understand that the user has got the answer and the dialog system can check the need for further assistance.
Goodbye: Queries in this intent are to end the conversation.

Table 11-2 A few sample queries from dataset and the corresponding English translations

Intent	Query from the dataset	Query in English
Train_search	please mujhe kal mehadeepattanam se somajeeguda tak kee sabhee buses dikhaen	Please show me all the buses from mehadeepattanam to somajeeguda for tomorrow
Place_search	mujhe batao ki sabase kareebee coffee shop kahaan hai	tell me where is the nearest coffee shop
Distance_search	Begamapet jaane mein mujhe kitana time lagane vaala hai	how long is it going to take me to get to Begumpet
Traffic	kya husain saagar ke raaste mein traffic halka hoga	will traffic be light on the way to hussain sagar
OOS	bank kab tak khula hai	how long is the bank open until
Greeting	hellooo	hello
Thank you	dhanyawad	thank you
Goodbye	theek hai byee	okay bye

OOS: These are the out-of-scope queries and are added in the dataset so that our domain specific dialog system is fault tolerant to out of domain queries.

A few sample queries from the dataset, from the above intents are shown in Table 11-2. We expect the system to receive queries where non-standard spellings are used (like plz for please, tym for time etc.).

To assist in the development of road transport domain dialog system, we created a prototype that shows the usability of mTransDial in this domain. To develop this prototype we use Rasa framework, an open-source machine learning framework to develop conversational chatbots. This framework has two components (a) Rasa NLU responsible for understanding the user intent and extracting entities and (b) Rasa Core which decides the set of actions to be performed based on the previous user inputs.

We used a subset of the English data to build this prototype. Based on Rasa NLU, we annotated this data subset with the following entity information: Source_loc, Destination_loc, Near_loc, point_of_interest. The Source_loc and Destination_loc entities are to be extracted if the query is of Train search or Distance_ search intent class. For Place search intent we extract Near_loc,

Figure 11-10 Sample response for train search query

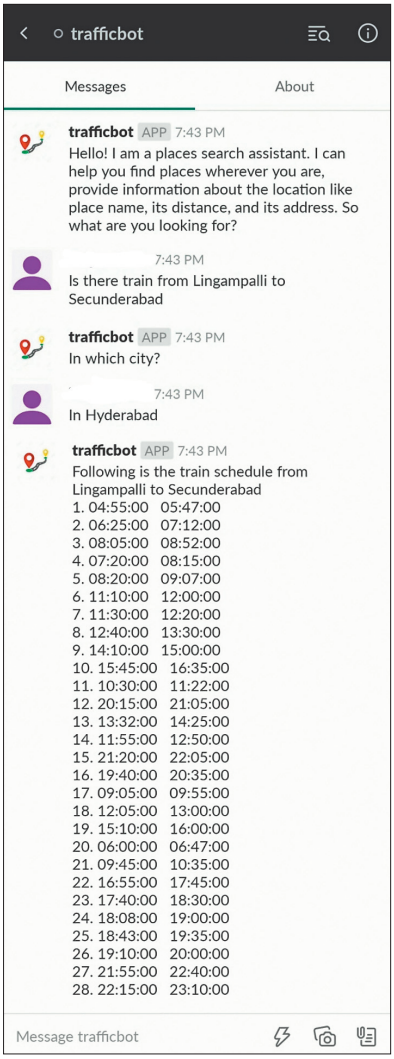


Figure 11-11 Sample response for place search query

point_of_interest entities to search for the nearby point of interests around the location provided by the user. From Rasa NLU pipeline we used CRF entity extractor to extract the relevant entities from the intents.

After intent classification and extraction of entities we performed relevant sets of action based on the intents. If the classified intent is Place_search or Distance_search we call an external map-based API (here, MapmyIndia [https://www.mapmyindia.com]) to find the points of interest near to the Near_loc extracted or to find distance between Source_loc and Distance_loc. If the classified intent is Train_search we use GTFS (General Transit Feed Specification, a common format for public transportation schedules and associated geographic information) to fetch the relevant bus/metro/train schedule. Figures 11-10 and Figure 11-11 show the responses of the system for train search and place search queries. It can be seen that the system can handle user queries posted in natural language and provide appropriate responses.

Conclusions

In this chapter we assess the use of big data analysis for understanding traffic flows and demonstrate how this can be used for providing appropriate intervention suggestions for decision making at multiple levels. It is understood that lowering the congestions will lead to lesser carbon emission. Making people better informed about the current and future traffic can make them plan accordingly. Being properly informed about public transportation will also increase their trust on the transport system and may motivate them to adopt public transportation. At a higher level, the city authorities can use insights from big data analytics for civic and traffic planning which can lead to both short- and long-term benefits to the traffic infrastructure, and congestion handling.