IJCSN International Journal of Computer Science and Network, Volume 4, Issue 3, June 2015
ISSN (Online) : 2277-5420 www.IJCSN.org
Impact Factor: 0.417

Linear Regression Based Analysis to Find Traffic Prone Areas

[1]Aastha Khanna, [2] Tanvi Malhotra, [3] Sonal Meena, [4] Dr. Kalpana Yadav

[1,2,3,4] Department Of Information Technology, I.G.D.T.U.W,
Kashmere Gate, New Delhi 110006, India

Abstract - In recent years, traffic has emerged as a ubiquitous problem faced by thousands of commuters on daily basis. With the, ever-increasing number of vehicles emerging on road, the problem does not seem to fade away. It poses a strikingly major conundrum to a large number of people. Using the analysis we use Twitter as our Database and aim to find the traffic-prone areas (e.g. In Delhi) so that people are familiar with the areas, which are highly prone to Traffic. In this paper, we demonstrate how social media content can be used to predict real-world outcomes. In particular, we use tweets made from twitter handles to forecast traffic prone areas in a specified region. We further analyze those numbers using linear regression to find which area is more prone to traffic.

Keywords - *Twitter, Data Mining, Traffic Police, Linear Regression.*

1. Introduction

Social media has emerged as powerful source of information at such exponential rate that people can create, share and bookmark data with ease. Examples include Facebook, Twitter, Myspace and blogs. These social networking sites can be used as a form of collective wisdom [2]. Data Analytics is used in this project to help and predict the ever changing scenarios in an eclectic number of domains specifically traffic. We attempt to exploit this feature of Data Analytics in order to achieve the solution to a problem that has direct effect on thousands of people residing in Delhi. We aim to find the Traffic Prone areas in Delhi Region so that people can be cautious in the vicinity of those regions and be prepared in situations of a medical emergency.

In recent times, social network and social media platforms have been used as a source of information to detect real time events like Earthquakes, Storms, Fires, and Accidents etc. An event can be defined as a real-world occurrence that happens in a specific time and space.[8] Regarding traffic related events, people often share by means of an status Update Message Information about the current traffic situation around them while driving . For this reason, event detection from social networks is often employed with Intelligent Transport Systems (ITS s). ITS infrastructure incorporates transport network and vehicles and users. It also provides information about weather, traffic congestion or regulation and routes.

Deluge and enormous amount of data, with high variance, that traverses through the web presents to us with an opportunity to predict about particular outcomes without having to institute manual mechanism. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains.[5] Using Data Analytics and Data mining technique we focus on modeling and knowledge discovery for predictive rather than purely descriptive purposes. In statistical applications, some people divide data analysis into descriptive statistics, exploratory data analysis (EDA) and confirmatory data analysis (CDA). The focus of EDA is on discovering new features in the data and CDA on confirmation or falsifying existing hypotheses. We used predictive analytics to focus on application of statistical models for predictive forecasting or structural techniques to extract and classify information from textual sources, a species of unstructured data. All are varieties of data analysis.

Building on the features of Data analysis, we extracted the data from Twitter. We considered Twitter as our data repository because a) easily available fresh data source, b) Millions of people expressing their view in 140 character sequence, c) There are verified profile of people who express themselves and celebrities who influence millions of people, and d) Microblog analysis has become a part of every scientific journal these days.
Working majorly on big data, we would be using R language and Twitter API's along with Text clustering Algorithm in order to observe trends.

2. Related Work

Twitter Analytics is an area on a rise. We make use of Natural Language Processing to carry analytics on data from Twitter. Predictions using tweets is being used by a lot of Data Scientists these days. An important feature of Twitter is that it is real time. Also, it is available on mobile phones and on the web hence, anybody can tweet from anywhere at anytime. So, Twitter has become a popular data source for many these days. It can be used to detect real world events. Nate Silver is one such famous American statistician, who developed PECOTA[6] system in 2003 for forecasting Major League Baseball players' performance. This also inspired such similar projection systems for other sports.

Nate Silver gained much limelight when in 2012 United States presidential elections his predictions for winner of all 50 states and District of Columbia were correct.

Regressions' earliest form was method of least squares. Published by Legendre in 1805.

They define the following terms:
Dependent variable: Predicts an outcome variable
Independent variable: Predicts using set go these
Regression has many applications in predicting attributes and many statisticians have been using it for predicting many useful values. In March 1990, Orley Ashenfelter, a Princeton economics professor predicted price of wine without even tasting it. Linear regression was used for prediction in which different variables for live Average Harvest Rain, Temperature, etc and dependent variable price of the bordeaux wine.

3. Flow to Process

3.1 Extraction of Data

Creation of a twitter Application is an essential process of the whole extraction process from Twitter. Twitter apps are third party created applications that work with Twitter to enhance functionality and perform added features. For example, there are Twitter apps that allow Twitter users to organize the people they follow, search for people on Twitter, publish photos, and much more. Upon creation of a Twitter Application, the user is given with keys and access tokens which are used to authorize the user and the application for the extraction of data from a user on Twitter.

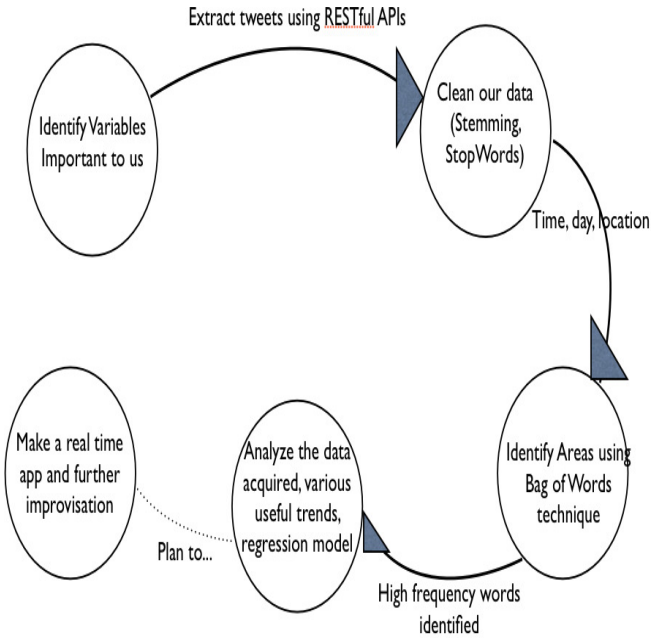


Fig.1 Process Flow Chart

3.2 Extracting Tweets from Twitter Handles Using Python

The goal of this research is an examination of both the distribution and predictive utility of features across varying contexts in Twitter. Accordingly, we must first gather a diverse set of data from Twitter. Of course, there are infinite choices for context, so for our analyses we focus on a carefully chosen subset, so that we can extract meaningful results. In this section we first describe a collection of tweets from Twitter using a twitter handle @dtptraffic and @traflinedel. This is followed by applying analytics on the data sets formed. Next, we discuss how we collected data using Twitter and the limitations involved

According to the Twitter's REST API Ver. 1.1, a specific developer can extract only 200 tweets in one go from some other user's timeline and handle. REST: Representational State Transfer APIs enable interaction between two softwares. At a time we can extract only 3200 tweets from a specific handle. By looping, and using Tweepy wrapper, we extracted useful tweets from twitter handles keeping the limitation of 3200 tweets at check. According to the REST API Ver 1.1, a tweet consists of a number of attributes like geolocation, favorited by, number of retweets, annotations, contributors, coordinates, created at, entities, favorite count, filter_level, id, id_str etc. Among the various attributes present before us, we selected only date, time, id, created at and text. We used

official twitter handles for the collection of tweets , so that only relevant and useful information can be extracted with some legitimacy.

Extraction of data from Twitter has been done using Tweepy packages and the extracted data has been stored in a csv(comma separated values) file. The .csv file is then used for further processing of information. Tweepy is open-sourced, hosted on GitHub and enables Python to communicate with Twitter platform and use its API. The current version of Tweepy is 1.13.

3.3 Cleaning of Data

Cleaning of data involves the removal of the invalid data points from the data set. Most of the data that is extracted from the twitter contains semantically wrong words and sentences as people tend to use short forms and texting lingo. This poses a major problem when data analysis has to be employed. Therefore, cleaning of data plays pivotal roles in the project. Cleaning involves basically two things a) Data points which are disconnected with the point of effect, b) Erroneous data points due to some external error because of some mistake in data collection or reporting.

Generally, cleaning also involves human judgment to decide which data points are valid and which are invalid but we are using the technique of removal of stop words which are not useful by using R . R is a powerful tool that can be used to clean the data, which has been extracted from twitter.

For instance, cleaning might involve detection of words like “delhi” and “DeLHi”. By cleaning it , we intend to convert it into one common case. Many words are frequently used but are only meaningful in a sentence. These are called StopWords. E.g. the, is, at, which. These words are not useful and irrelevant in the process of data analytics and act as impediment as they increase the size of the data frames. To improve the quality f results garnered from machine learning techniques, removal of these words is essential.

For the purpose of removing stop words from the data we use various packages like tm package, SnowballC package etc. It involves: a) Stemming of words, b) Cleaning of irregularities like converting all the text into the same case, c) Removal of Punctuations.

3.4 Frequency Calculation

On the basis of the data that has been collected from Twitter and further cleaned, Areas in Delhi are sorted and their frequency is calculated. Frequency calculation is a pivotal process in the whole process. The number of times an area occurs in the alert is the measure of how prone the area is to Traffic. As before, R is used for calculating the frequency of the areas by using DocumentTermMatrix (TermDocumentMatrix(x, control = list())

3.5 Regression Analysis

They should be numbered consecutively throughout the text. Equation numbers should be enclosed in parentheses and flushed right. Equations should be referred to as Eq. (X) in the text where X is the equation number. In multiple-line equations, the number should be given on the last line.

$$yi=\beta o+\beta 1*(xi)+\varepsilon i \quad (1)$$

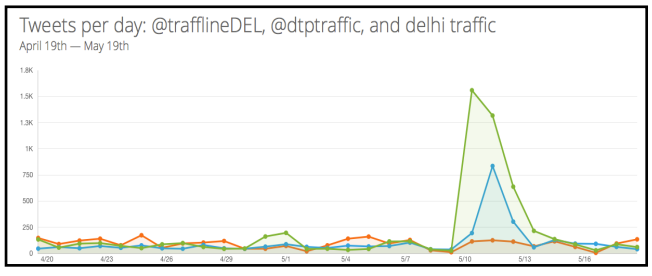
4. Implementation

The flow of implementation is as follows: Extraction of tweets (getting tweets from official authentic handles), Data Cleaning, Identifying the Major Regions, Carrying out various analysis & Building predictive linear model for the same.

4.1 Extraction of Tweets

As tweets form the basis of all our analysis this is an important step. These tweets are extracted by using Twitter’s REST API’s. From two officials, very active Delhi traffic police handles.

Extracting data from **Twitters Search API** it is necessary to make an application to get consumer key and authentication key which we need to input to our program. We then wrote a python script **ExtractTweets.py** to extract relevant traffic tweets using a **Tweepy** wrapper to call the *Search API* using the keys obtained from the app. In a single call we were able to retrieve around 200 recent tweets from a particular timeline, and at the maximum 3200 recent tweets. Twitter also provides a *Streaming API* to retrieve live streaming of relevant tweets.



Graph 1 Tweets per Day

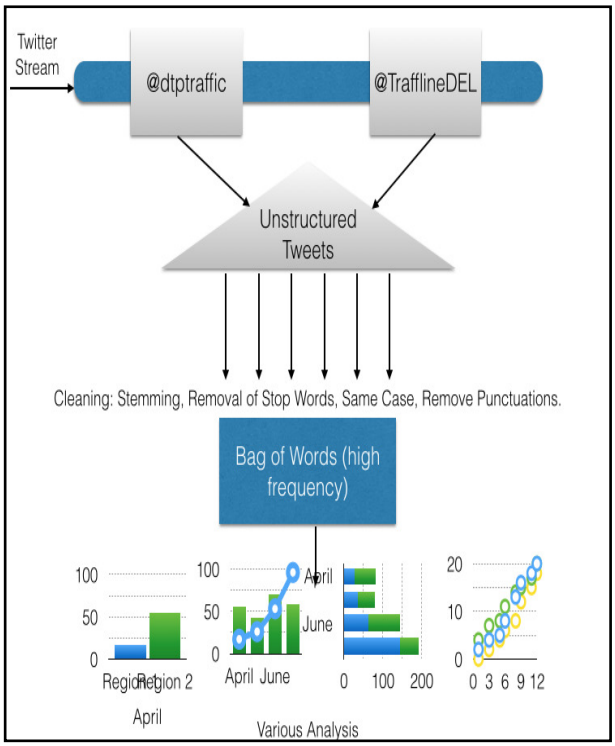


Fig. 2 Process of analyzing the tweets

4.2 Data Cleaning

But Tweets are inherently noisy and heterogeneous in nature. A study by Pearson Analytics states how half of the tweets are not conveying any useful information and are pointless [7]. So there is a need for cleaning data (unstructured tweets). So we use techniques like removal of stop words, removal of punctuations, stemming, making same case so that finding high frequency words from our clean corpuses is viable.

4.3 Identifying major regions

On carrying out the frequency analysis we find that the two most highly occurring words in our traffic related tweets are 'water logging' and 'breakdown'. So we removed the sparse entries from our high frequency words and added the bursty data to our set of stop words so as to obtain useful data i.e. to identify places in Delhi. Then we wrote a Python script **Places.py** to obtain geolocation of the places identified using Google's Map API. Using this API we get the longitude and latitude of the places that are there in the list in JSON format (many other formats available too).

4.4 Analysis

After obtaining the longitudes and latitudes we make a new data frame **DataFrame.csv** with all the useful variables for prediction like area name, frequency, longitude and latitude. And the **MasterTweets.csv** contain the tweets along with date and time. Using these two data frames we carry out number of analysis for observe some trends or patterns in the traffic patters in Delhi.

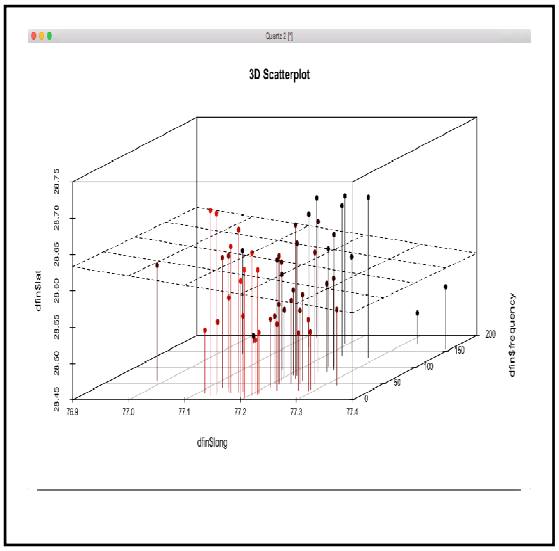
We first plot areas against frequency of traffic to identify regions of high traffic. We then mine our data to find how the traffic varies on weekends and weekdays.

4.5 Linear Predictive Model

We then use the linear regression technique to build a linear predictive model to identify traffic prone areas.We use linear regression for the same. We first use it against independent variable longitude only and then latitude only. We then take two independent variables both longitude and latitude.

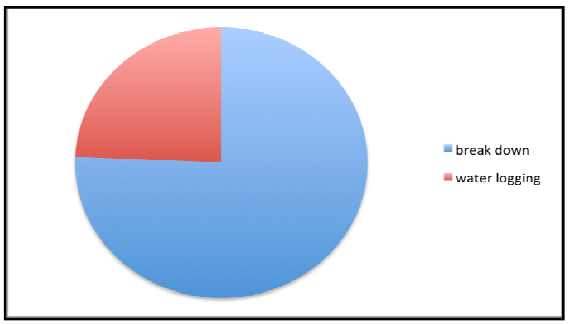
5. Conclusions

After applying the regression models and analyzing the data in detail, we realized that traffic prone is an essential problem faced by 1.8 crores commuters in Delhi on daily basis. To get a better view of the bigger picture in hand, we came up with the following results:



Graph 2 Longitude & Latitude v/s Frequency

The above diagram gives the pictorial representation of the areas in Delhi and the frequency of the traffic in that particular region. The x axis gives the longitude and the y axis gives the latitude. The convergent point of these two axes gives the location of the region. Now on the z axis we have the frequency of traffic in that region. The three dimensional graph gives the overview of the situation. Moreover , while perusing over the tweets extracted from twitter, we acknowledges a trend that their was two major reasons causing traffic in a particular region I.e.
a) Bus or Car Break down
b) Water logging
on further analysis, we came up with the following results:



Graph 3 Trend for traffic

Acknowledgments

We would like to thank the Twitter team for the ease of extraction of data they provide via their Stream API and Search API. We would also like to acknowledge the following packages used: tm by I. Fienerer , NLP by Kurt Hornik and Snowball C by M. B Valat.

References

[1] Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, Robert Baumgartner, Center for Complex Network and Sustem Research, Indiana University, Bloomington, “Web Data Exraction, Application and Techniques: A survey”

[2] Sitaram Asur, Bernardo A Huberman, Social Computing Lab, HP labs, “Predicting the future with Social Media”

[3] Narashima S Purohit, Meghana Bhat, Akshata B Angadi, Karuna C Gull, Department of Computer Science and Engineering, K.L.I.E.T, “Crawling through Web to Extract the Data from Social Networking Site- Twitter”,

[4] Adam Tsakalidis, Symeon Papadupoulus, Alexendra I. Ccristea, Yiannis Kompatsiaris, “Predicitng Elections for Multiple Countries Using Twitter and Polls”, 2015

[5] Bernardo A Huberman, Daniel M Romero and Fang Wu. Social Nteworks that matter: Twitter under the microscope. Jan 2009

[6] http://en.wikipedia.org/wiki/PECOTA

[7] http://www.pearanalytics.com/blog/2009/twitter-study-reveals-interesting-results-40-percent-pointless-babble/

[8] The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, III

[9] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[10] P.Ruchi and K. Kamalkar, “ET: Event from tweets,” in Proc. 22nd Int. Conf. Worl Wide Web Comput, Rio de Janeiro, Brazil, 2013