Motor Tweet Vogue: Analysis from Twitter Data about Cars

ABSTRACT

MTV (Motor Tweet Vogue) is an application for analyzing the current trends in car brands based on Twitter data, which is collected using "Twitter streaming API". MTV shows a visual representation of people's general interest in cars from each of the fifty states in the U.S. It visualizes the most popular car brands in each state. It then identifies which car brands in those states are positive and negative in terms of popularity. MTV also visualizes the popularity of each brand in the United States.

INTRODUCTION

Twitter is one of the most popular social networking sites. Through that medium people express their interests, views and opinions. It is for that reason the tweets on Twitter resemble the opinions of those in the general public. In our present time, a great deal of analysis is being done through the data acquired from Twitter. However, in the specific area of car trends data from Twitter is not explored as extensively. Even though Twitter provides general trends, it does not provide trends in particular areas of interest—finding the popularity of a specific car brand. MTV plays a significant role in identifying the trends in car models. The analyses on tweets about cars explore the current trends in cars in a particular state and the changing interests in those states.

Since Twitter does not provide older tweets, we collected tweets using Twitter Streaming API. Twitter streaming API has a limit of 180 calls per 15 min window. For our dataset, we used tweets that were gathered over a period of a month and used that to identify the trends. The Twitter "track" parameter was used for retrieving tweets about cars. It identified tweets by using pattern matching. For the identification process, we had to manually input all the car brands and car models released in the end of 2014 and 2015. The program was configured to run the entire day with enough interval time to avoid the issue of "Query Limit". To access the Twitter Streaming API, "Tweepy" from the Python library was used. The Twitter responses were in JSON format and it included all the details about a tweet and the public profile of the author. We collected the tweet text, time of tweet, location, geo co-ordinates and the retweet count for each of the tweets in our analysis.

Data cleansing had to be performed on the collected data so that any and all erroneous data was removed. In order to identify the state of each tweet, we used either geocode or the location value associated with the tweet. We divided entire dataset into two: one with geo coordinates and other with the location value. Reverse geo-coding was performed on the data with Geo co-ordinate values to identify the location. Direct mapping was performed to identify which state the tweets were sent from; the location value was also used. After the state had been identified, each tweet was mapped to the corresponding state code.

Since the data collection through streaming API uses pattern matching for the tweet collection, car entity recognition was a challenge. Many of the car brands do not have unique names. There are several car brands that have names that refer to animals or cities: Jaguar and Austin. To identify a real trend in cars it was important to identify tweets that referred to cars rather another entity of the same name. To ensure that car entities were identified, Alchemy API was used. The application interface uses natural language processing and machine learning algorithms to identify entities and performs sentiment analysis on text data.

Once the entities were identified, the corresponding sentiment values were calculated using Alchemy API. Alchemy can identify the sentiment of an entire tweet and the sentiment of the entities in the tweet and the relevance of each entity in the tweet. Alchemy API has a query limit of 30000 per day. Once the polarities of the tweets were identified, we calculated the total popularity of each model in each state. Any positive and neutral polarity added to the popularity while negative polarity limited it.

MTV visualizes four car trends. First, visualization defines the general interest of cars in each state by calculating the percentage of total tweets in each state to the total tweets in United States. It also defines the most popular car brand in each state. Second, visualization draws the top five positive popular car model in each state. Third, one identifies the top five brands with negative polarity. Fourth, visualization plots the top five popular states of each model.

Implementation of MTV is divided into five phases:

- Data Collection through Twitter Streaming API.
- Data cleansing and removal of erroneous Records.
- Geo coding and Reverse Geo coding to identify the state of each tweet.
- Entity Recognition and Sentiment Analysis of tweets.
- Visualization of tweet data.

IMPLEMENTATION

Data is collected through Twitter streaming API. We have collected 50 GB of tweets in 30 days period. Data was collected through all day using "Tweepy" python package. Data collected in every 300 second were saved to a file and a 50 second time out provided before the next request to Twitter Streaming API.

Data cleansing has been performed on the collected data and removed all erroneous data. In order to identify the state of each tweet we used either geocode or the location value associated with the tweet. We divided entire dataset into two, one with geo coordinates and other with location value. Reverse geo coding was performed on the data with Geo co-ordinate values to identify the location. A direct mapping has been performed to identify the state of tweets with location value. After the state has been identified, each tweet was mapped to the corresponding state code.

Since the data collection through streaming API uses pattern matching for the tweet collection, car entity recognition is a challenge. Most of the car names are not unique, there might be other entities as well with the same name. To identify the real trend in cars, it is important to identify tweets about car entities. Car entities were identified by using Alchemy API. Alchemy API uses natural language processing and machine learning algorithms to identify entities and for performing sentiment analysis on text data.

Once the entities were identified, the corresponding sentiment values were calculated using Alchemy API. Alchemy can identify sentiment of the entire tweet and the sentiment of the entities in the tweet and the relevance of each entity in the tweet. Alchemy API has a query limit of 30000 per day. Once the polarity of tweets were identified, we calculated the total popularity of each model in each state. Positive and neutral polarity adds to the popularity while negative polarity limits the popularity.

MTV visualizes four car trends. First visualization defines the general interest of cars in each state by calculating the percentage of total tweets in each state to the total tweets in United States. It also defines the most popular car brand in each state. Second visualization draws the top five positive popular car model in each state. Third one identifies the top five brands with negative polarity. Fourth visualization plots the top five popular states of each model.

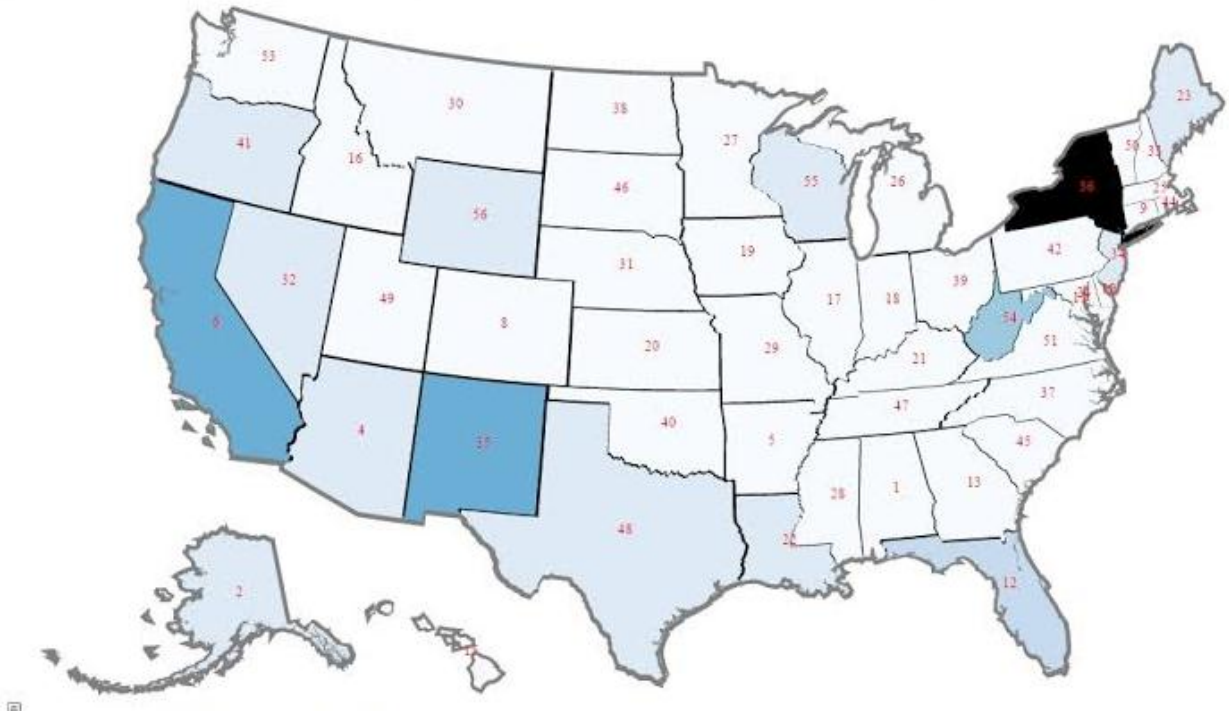

Fig:1 General Interest in cars

ALGORITHM

Popularity count: Popularity measure will be calculated by the count of car entities using the tool Stanford Named entity Recognizer and the count for each brand in each state will be calculated.

Sentiment analysis: Sentiment analysis will be performed in three phases.

Phase1: All emoticons will be replaced with the meaning using an emoticon dictionary which is having all emoticons and its meaning by using Wikipedia.

Phase2: All words will be assigned a pleasant score by using the DAL dictionary, which has score for each word ranging from 1 to 3.

Phase3: Using Naïve Bayes classification, sentiment analysis will be performed on all the tweets.

Trends: Trends will be calculated by using the popularity count together with positive and negative polarity. Positive polarity will be added to the polarity count and negative polarity will be deducted from the polarity count.

EXPERIMENTS

We have conducted several experiments for the entity recognition. We used NLTK, Stanford Named Entity recognizer and Alchemy API for the Entity Recognition. Both NLTK and Stanford Entity recognizer didn't provide satisfactory result. The accuracy result of these two tools were less than 20 percentage for identifying car entities while alchemy API showed 80 percentage accuracy in identifying car entities. Accuracy of Sentiment analysis is more or less similar with NLTK and Alchemy API.

We manually identified Car entities in 50 tweets and compared the results with NLTK, Stanford Named Entity Recognizer and Alchemy API.

	Total Tweets	Alchemy	NLTK	Named Entity Recognizer
Entity recognition	50	40	10	10
Sentiment Analysis	50	45	44	

Alchemy API provided satisfactory result in terms of Entity Recognition and Sentiment analysis. So we used Alchemy API for identifying the entities and for performing sentiment analysis on tweets with car entities.

CHALLENGES

- Since twitter does not provide data from previous years a comparison of trends in cars over a significant period of time was made impossible. For a comparison of that nature, data must be collected over an extended period of time. Therefore we limited the nature of MTV only to compare the current trends.
- Lack of location information in tweet was a challenge. Since we are comparing trends in each states of United States, location information is important. And if we are taking only data with location information, that will not be suffice to derive a pattern. We overcome this challenge to an extent by collection 50 GB of tweets and we got enough tweets with location information to derive a pattern.
- Geocoding and Reverse geocoding for a large amount of data was another challenge. We resolved that challenge by taking the latitude and longitude range of each state and mapped each geo coordinate value to the corresponding state. For location to map to the correct state we used all cities in a state in US and mapped it to corresponding state.
- For Entity recognition and Sentiment analysis, we used Alchemy API. Alchemy API has a query limit of 30000 per day. So we had to process data in small amounts. If the calls exceeds the query limit, we have to process the data again else we miss some relevant data.

- Rendering data for visualization on a web page was another challenge .To establish correlation of web page dimensions and data volume to plot was difficult. Plotting data on a Map was another challenge. These were solved by using D3.js.

CONCLUSION

Social networks are good sources for analyzing popular trends in many areas. MTV successfully identified the trending car brands in United States. It clearly defined a pattern change in customer interest towards car brands in different states.