Student Name(s): Shailendra Khadka Yadav, Syed Ali Mutahir and Sanjeev Kumar

Student Number(s): 500718414, 500797615, 500775128

Course Title: Social Media Analytics

Assignment #	**1**
Due Date	**3rd April, 2017**
Group # (if applicable)	**8**

I hereby certify that I am the author of this document and all sources used in the preparation of this assignment have been cited in accordance with Ryerson's Code of Student Conduct directly or paraphrased in the document. Sources are properly credited according to accepted standards for professional publications. I also certify that this paper was prepared by me (all group members if it is a group paper) for this purpose.

Social Media Topic Mining: Twitter data of Honda and Toyota

Shailendra Khadka Yadav 500718414
Syed Ali Mutahir 500797615
Sanjeev Kumar 500775128

DS8006 Social Media Analytics

Ryerson University

April, 2017

1. Summary

The goal of this project is to apply techniques from social media analytics and text mining to analyse online tweets of the people about the Honda and Toyota cars. We collected the twitter data of Honda and Toyota cars with the help of Twitter API and then performed the data mining from this collected data to do the analysis. The data mining and analysis was focused on customers talk about the brand, car insurance, customer satisfaction with the brand, competitors marketing strategies and issues related to these two brands. With the help of this social media mined data, companies are able to compare their brands with the other company's brands to increase the performance of their brands and hence increase their business.

Group Members and Responsibilities:

Shailendra Khadka Yadav: Data collection, Literature review, brand issues data mining, coding and analysis.

Syed Ali Mutahir: Data collection, customer's feedback data mining, coding and analysis.

Sanjeev Kumar: Data pre-processing, competitor's data mining, coding and analysis.

Name	Research work	Code	Report Writing
Shailendra Khadka Yadav	**R**	**R**	**R**
Syed Ali Mutahir	**R**	**R**	**A**
Sanjeev Kumar	**A**	**R**	**R**

RACI Chart

A: Accountable

R: Responsible

2. Problem Statement and Dataset Selection

These days, to increase business it is compulsory for every organization to do research on itself and also competitor's brand and do marketing accordingly. It becomes easy with the help of social media. Social media not only provides the overview of industry but also helps in marketing of the targeting customers. Social media helps the organizations to understand following type of questions and this is the purpose of this project.

- What do consumers say and hear about my brand?
- What are the most talked about product attributes in my product category and competitor's product?
- What are the most talked about insurance attributes in my product category and competitor's product?
- What are the competitors doing to excite the market?
- What issues are most important to resolve in all models?

We collected 50,000 tweets for Honda and also 50,000 tweets for Toyota from Twitter and saved it in .csv format.

3. Previous work / Literature review

With the invention of social networks, such as Facebook, Twitter, LinkedIn and Google+ in the past, and its ever-growing nature, there is a massive amount of unprocessed data in the web [5, 7]. The analysis of these massive data can be utilized to gain new knowledge from it [7]. The so obtained knowledge can be utilized to make better decisions in various areas such as business, politics, sports and more.

In this globalized age of information, people are inclined to put their opinions in the form of likes, posts, comments, shares or reviews on various topics happening in the world spontaneously [3]. It can be of great importance to find out the topics that are trending at the moment [3]. Thus, topic mining is a great area of research with many research works being already done in it [3, 6].

The research papers that were consulted while doing this work are listed in the references. Paper cited [1] presents managerial insights into how social media analytics can improve automotive quality management. Paper cited [2] discusses about the hot topics of online social network mining in helping researchers to solve the challenges that exist in social network mining. Paper cited [3] discusses about topic mining in the Twittersphere, with an in-build update mechanism based on time slices and dynamic vocabulary. Paper cited [4] discusses on how research on topic models can be conducted for short text scenarios and demonstrates that the effectiveness of trained topic models can be highly influenced by the length of the documents. Paper cited [5] discusses about how to obtain information from available big data in social networks. Paper cited [6] discusses about heterogeneity of data sources and introduce a novel topic modeling framework designed to handle heterogeneous sources. Paper cited [7] performs empirical analysis of the various data mining techniques such as clustering, classification, etc. for social network websites.

4. Data Collection & Pre-processing

In this project, we collected the data from twitter social media platform with the help of Twitter Search API and use R programming language for Analysis and visualization. The search key words were “Honda” and “Toyota”. The data collected contains 50,000 tweets of Honda and Toyota consumers each. There were 16 attributes in each twitter data. This data was further mined by creating some categories like body type of the car, insurance, customer satisfaction, competitors marketing strategies, “Issues” for analysis part of our project. Body type tweets were used as classification based on six categories i.e. “Hatchback”, “Wagon”, “SUV”, “Crossover”, “Coupe” and “Sport”. Customer satisfaction tweets were mined based upon “Happy” tweets and “Sad” tweets. Data related to competitors marketing strategies was collected using “discount”, “offer”, “lease” and “finance” key words in the available tweets. Some tweets were related to the customers, problems of Honda and Toyota were also mined by using the search key word “Issues”.

5. Analysis

This project is divided into 6 analysis parts.

a) The first part is used to mine the data based on the body types of the cars as shown in figure 1. On the basis of collected data, we divided the cars in six categories (“hatchback”, “wagon”, “suv”, “crossover”, “coupe” and “sport”). From the data and bar graph of this figure, we analysed that, most of the people are taking about crossover models of both Honda and Toyota. Its mean market for this model is high and both the companies should focus more on this segment.

b) The second part explains about the no. of people worried about the insurance of the cars as shown in figure 2. In this we tried to find out the people tweets who were talking about the insurance which might be the worried factor, because if the insurance is high for some type of vehicle then buyer will be more effected, which result, it will effect negative on the sale. Vice versa if a company has less insurance it would be the strength of company and will be helpful in the marketing and ultimately increase the sale. Also if we compared the number of tweets geographically, we analyze that this insurance problem is more in North America as compared to other parts of the world.

Currently we only tried to see a big picture of the company, but it can increase up to the each product of the company. We tried to find the insurance talk about each category but due to less tweets we were unable to get the values in most of the products.

c) The third part tells about the satisfactions of customers related to their brand as shown in figure 3. This type of grapy is most important for any organization because it is a feedback to the company directly from his/her client’s satisfaction. This graph is also useful in today’s world, that there is no need to get the feedback from customers but calling them, make them fill some survey or go to web site for their experiences.

In this we tried to find the happy and sad tweets of the customer about their car. If supposed that if a tweet has a happy face :) then it is happy tweet and that person is satisfying with his/her car while if it has :(then the person is unhappy. So when we checked the graph it showed that there are more tweets on happy and sad in case of Honda but if check the percent of happy and sad tweets then we can see that Toyota has less then Honda. It shows that Honda customers do more tweets then Toyota but Toyota customers are more satisfied as compared to Honda.

Similarly we have more tweets then it is also possible that which car has more satisfaction. And which type has worst satisfaction. So if a company checks this graph can easily understand the feedback of the customers.

d) The fourth part gives us information about the competitor's marketing strategies as shown in figure-4. From the data we analyzed that people are taking about discount, offer, finance and lease of the cars and tweets are more for Honda cars. After drawing the graph, we analyzed that these tweets are more in Honda. It shows that Honda is more aggressive in marketing by providing more discount, finance, lease facilities to more customers to increase their sale.

e) The fifth part puts light on the issues or problems which are faced by customers of Honda and Toyota brand as shown in figure 5. This graph is also extremely important to know the standing of the production manufacturing, because if there are more issued in there, more customers will be unhappy, which ultimately effect on the company reputation.

Here we tried to find out the issues in both Honda and Toyota, and we came to the point that Honda has quite high percent of issues as compare to Toyota. It also shows that we should go deep into the root cause of the problem by finding the car type and then issue type that is causing the issue but due to less data we can't go deep into it.

f) The sixth part shows us the geographical locations of Honda and Toyota tweets as shown in figure6. Most of the tweets were not having geographical location attributes. So, we deleted the null values from the data and added the available geographical locations of the tweets of both Honda and Toyota on the world map. From the map we analyzed that people are tweeting from all over the world but more tweets are from North America which clearly indicate that Honda and Toyota customers are more in these area. This graph is also helpful in marketing the locations and also allocating the budged accordingly.

g) Screenshots of the resulting network visualization are included in appendices.

6. Conclusions

In the conclusion, we can say that social media is playing very important role in any business and the result of this "Social Media Topic Mining" project is very useful for the cars manufacturing companies to increase their business. These companies can keep eyes on competitor's business and marketing strategies and can give more benefits to the targeting customers by giving good offers, services, coordinating with insurance companies to decrease the insurance premium and most importantly by taking customers feedback and resolving their issues related to cars. These companies can also see the geographical locations of the people doing tweets and can make advertisement and marketing plans accordingly. A report from Forbes website shows that, 28 percent car buyers discuss or communicate a recent purchase experience by using social media. 38 percent of consumers report they'll purchase a car after consulting social media. 84 percent of all automotive shoppers are on social media with 24 percent using the networking site as a resource for purchasing their last vehicle. People use social media to express themselves and share the things they love, so cars are a natural fit. While the power of social media will always lie in the person-to-person connections between users, it's impossible to forget about how these networks are being monetized and sustained by successful brands and advertisers.

As an individual, I have learnt about working on various modules of the social media project independently. I have experienced the power of social media in business by understanding the statistics behind it with the help of visualization of the data.

As a member of a group I have leaned how to integrate all the modules of the project as a single entity to produce a real-time useful product and experienced the teamwork culture.

7. References

[1] AS Abrahams, J Jiao, GA Wang and W. Fan. “Vehicle defect discovery from social media”. ELSEVIER.

[2] G Nandi and A Das. “Online Social Network Mining: Current Trends and Research Issues”. International Journal of Research in Engineering and Technology.

[3] JH Lau, N Collier and T Baldwin. “Online Trend Analysis with Topic Models: #twitter trends detection topic model online”.

[4] L Hong and BD Davison. “Empirical Study of Topic Modeling in Twitter”.

[5] M Adedoyin-Alowe, MM Gaber and F Stahl. “A Survey of Data Mining Techniques for Social Network Analysis”.

[6] R Ghosh and S Asur. “Mining Information from Heterogenous Sources: A Topic Modeling Approach”.

[7] SGS Fernando, MdGaparMdJohar and SN Perera. “Empirical Analysis of Data Mining Techniques for Social Network Websites”. COMPUSOFT, February-2014, Volume-III, Issue-II.

[8] https://www.forbes.com/sites/drewhendricks/2015/04/09/a-look-at-how-car-brands-are-effectively-using-social-media/#63e14d8e17a0

Appendices

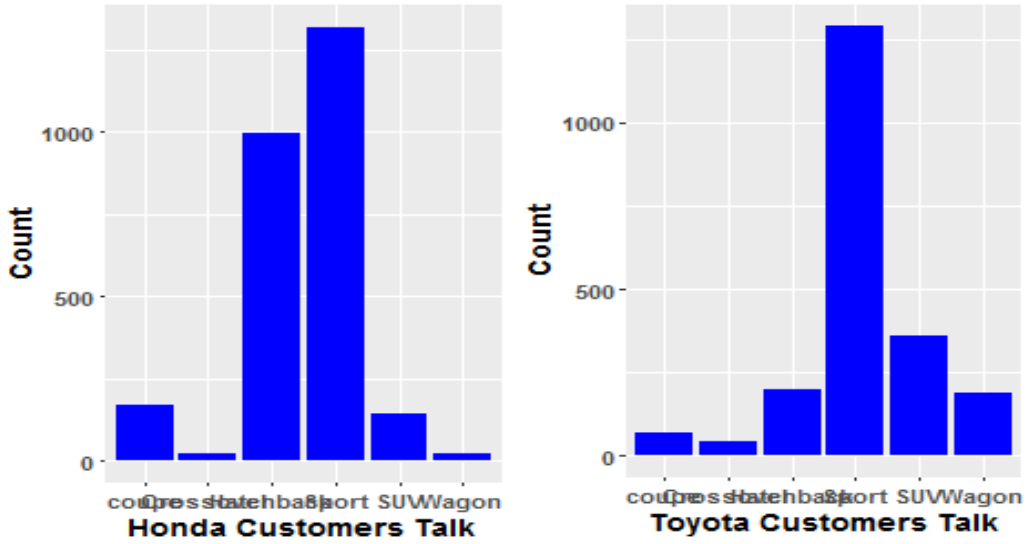

Figure 1

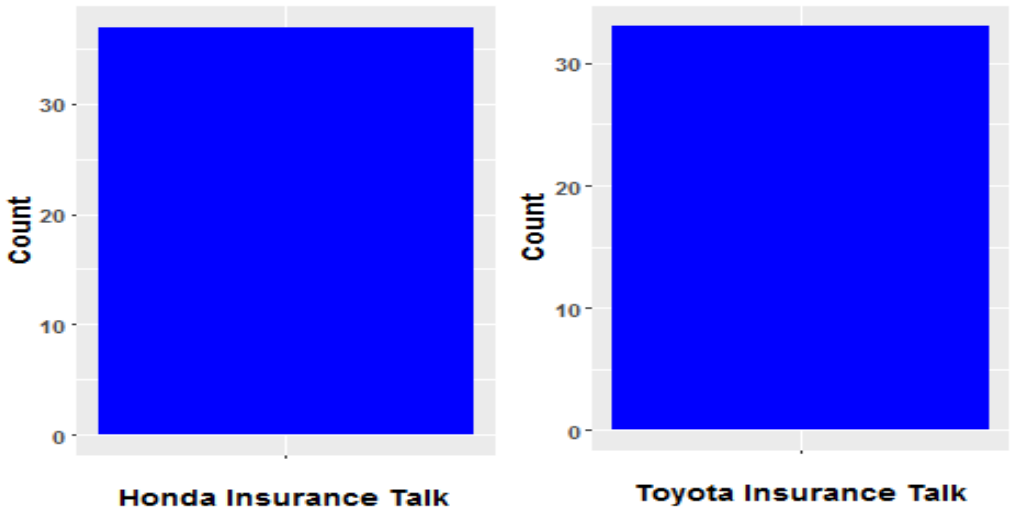

Figure 2

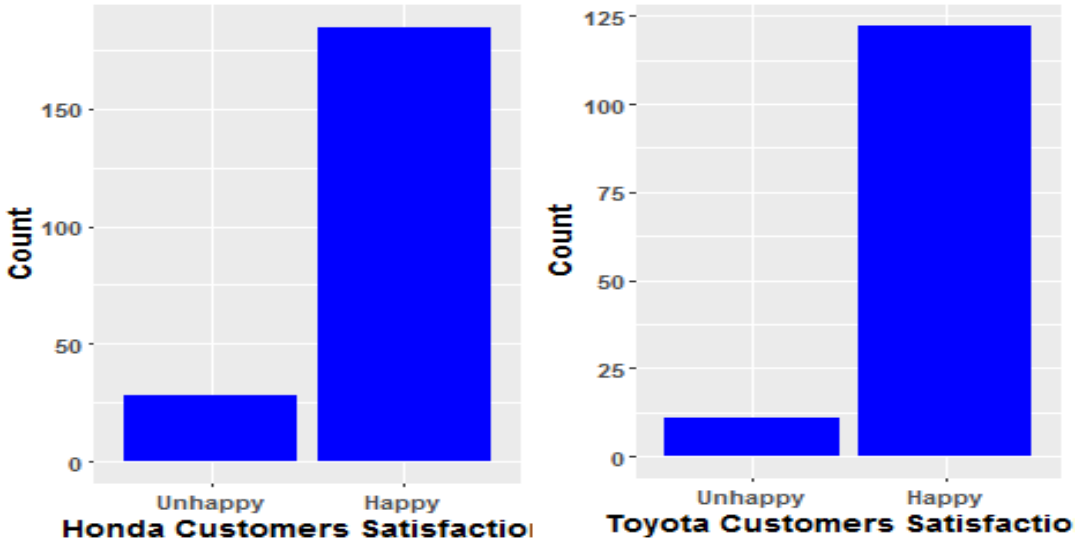

Figure 3

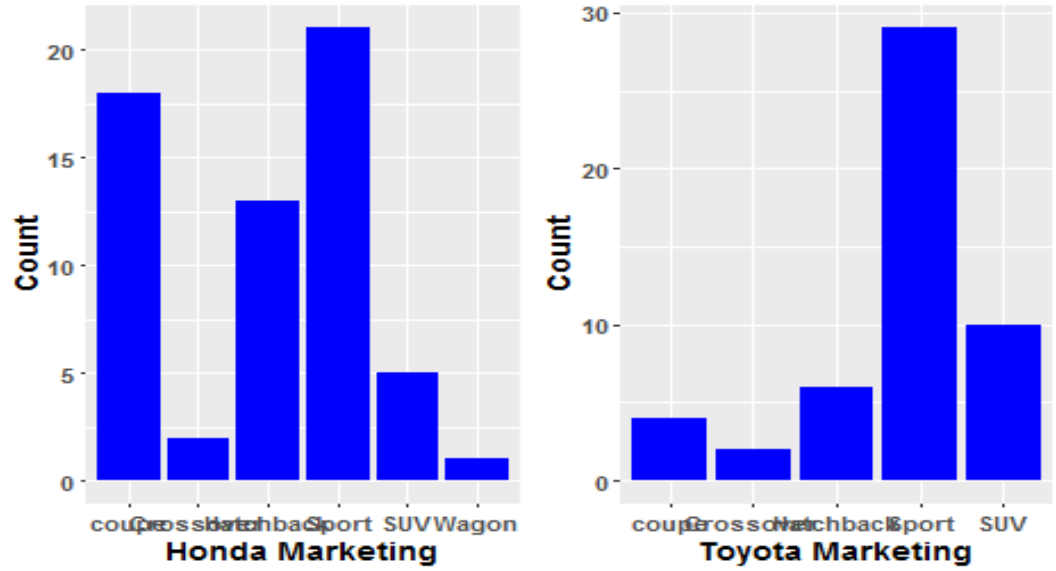

Figure 4

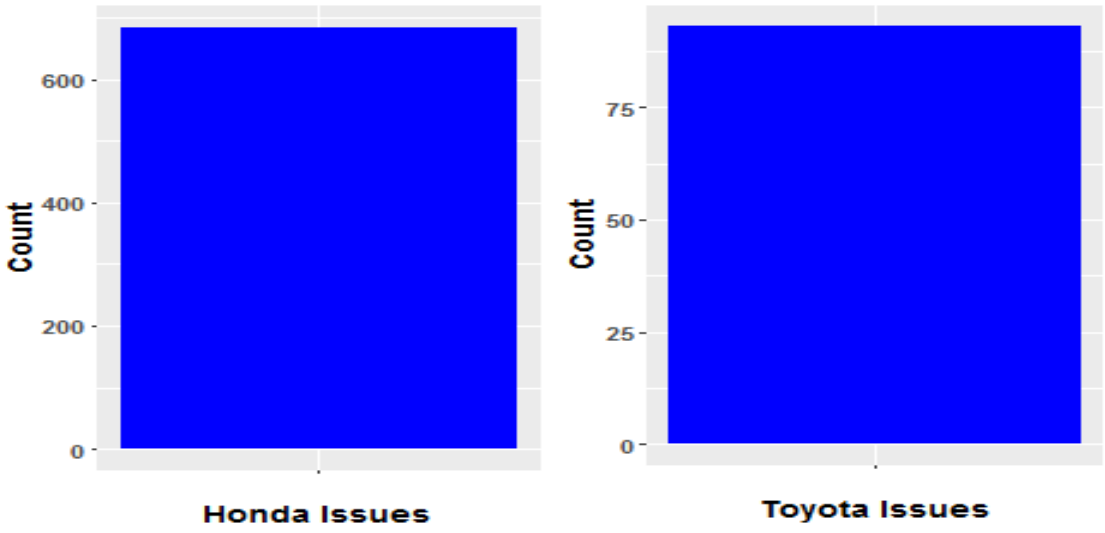

Figure 5

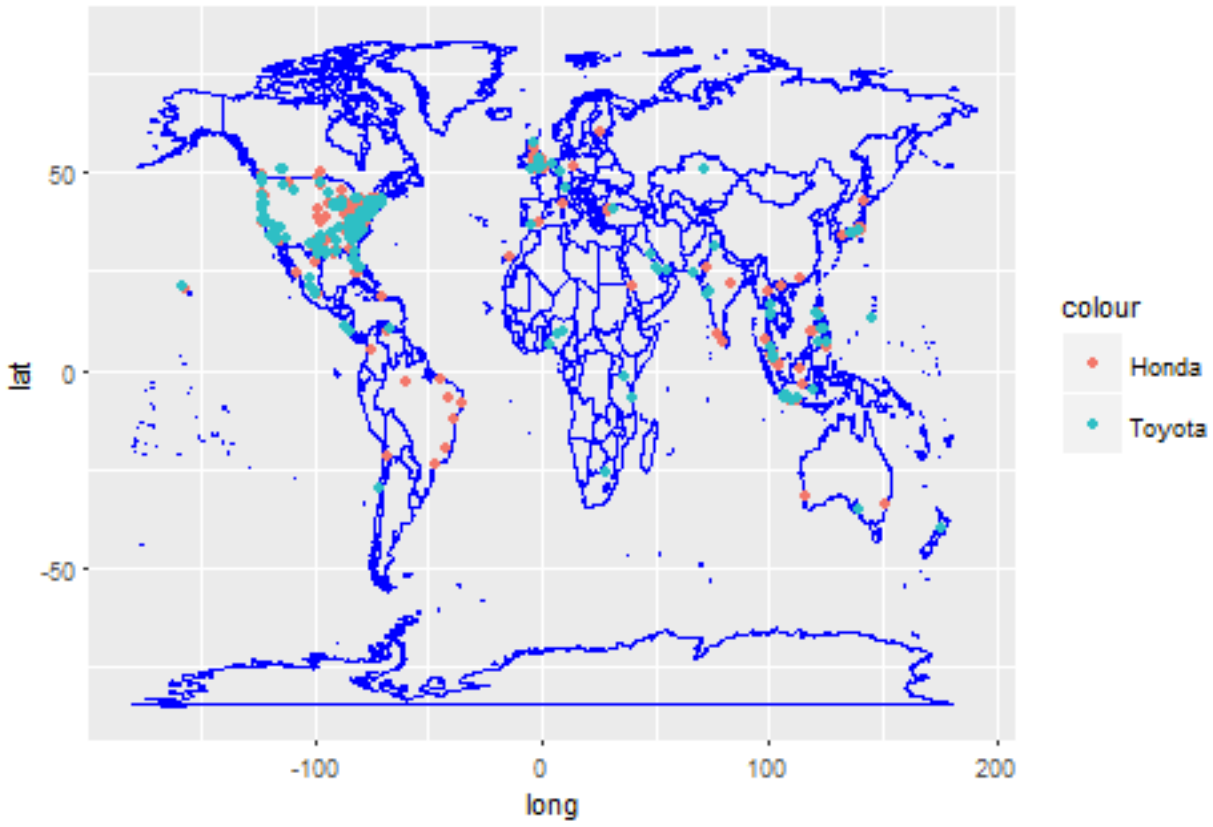

Figure 6

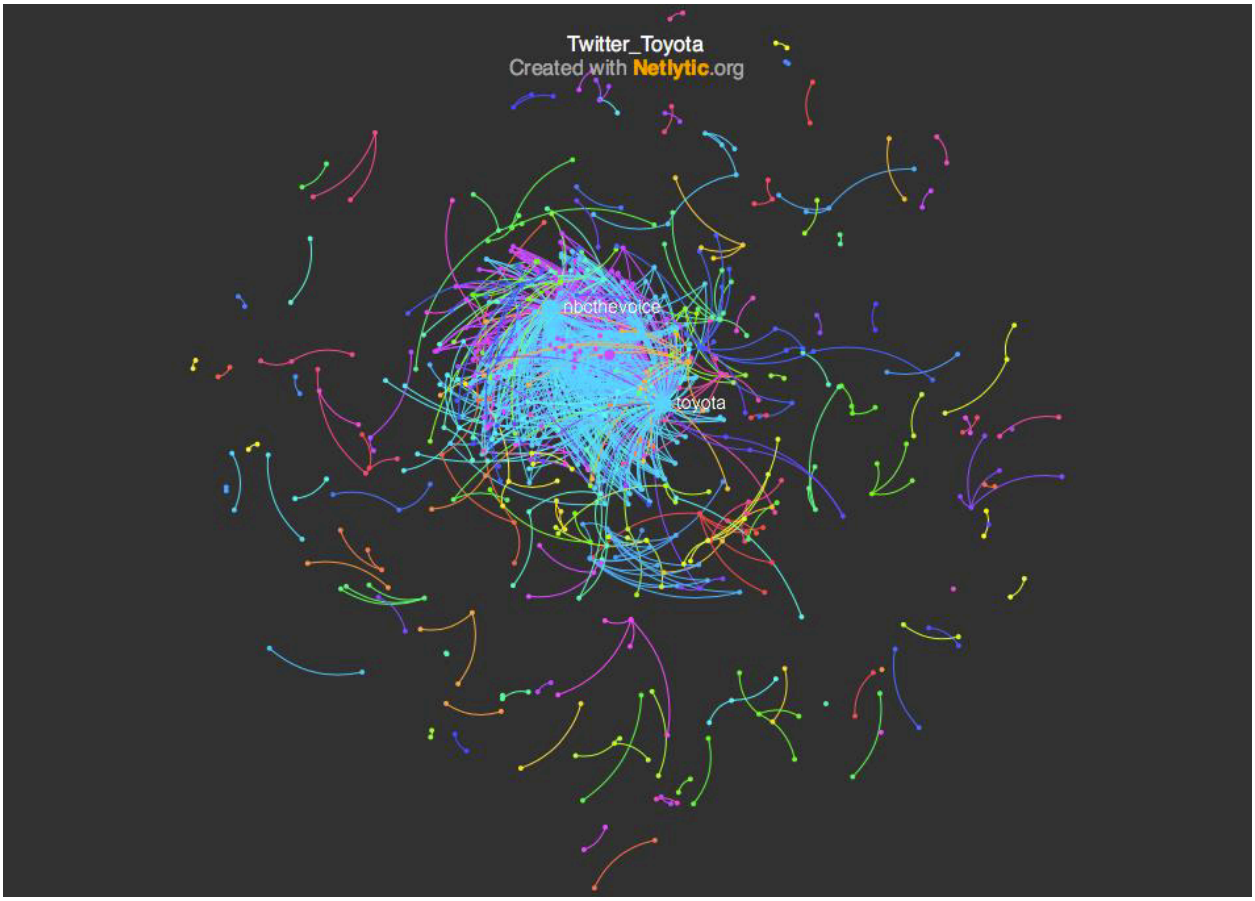
Twitter_Toyota
Created with Netlytic.org
nbcthevoice
toyota

Figure 7

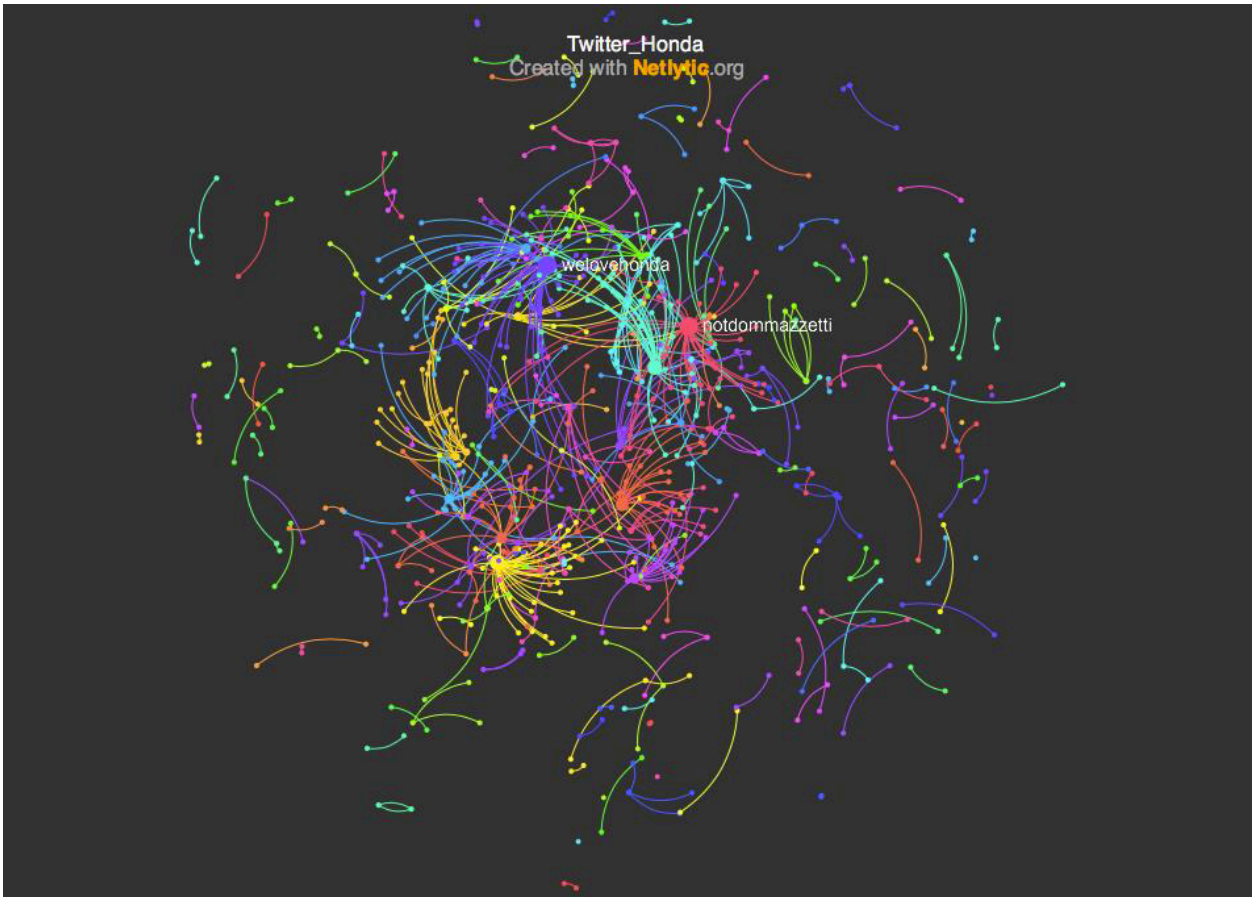
Twitter_Honda
Created with Netlytic.org
welovehonda
notdommazzetti

Figure 8