Professor Changseung Yoo

Ruite Xu 260621929

Junkang Zhang 260867086

Chris Zong 260872947

Junfeng Yang 260696954

Zhentian Jiang 260658190

Yuan Li 260866398

Introduction

Luxury car is among the fastest growing industries in the next five years, and it is estimated at approximately 12% of CAGR[1]. In this report, we will look at how social influencer can make impact on luxury car industry on social media, more specifically on Twitter. By implementing social network analysis, we will have better understand of how attributes of interests from influencers can play important roles in advertisements of different luxury cars and followers' reactions and attitudes towards influencer's tweets. Moreover, performing graph analysis using Google Vision can investigate relative hashtags and mentions on the corresponding pictures, and this can help to do later analysis on attributes. Overall, this report aims to leverage on social influencers to make more impactful campaigns and advertisements on Twitter.

Methodology

To apply both of social media network analysis and graph analysis, our work was separated to two parts:

Part 1: Social media network analysis

Data Collection

After chose topic: luxury car, we scrapped 37000 tweets using Twitter API in Python based on hashtag *#luxurycars*, information includes user information, Retweet Count, Text of tweet, etc.

---

[1] Global Luxury Cars Market 2018-2022. (n.d.). Retrieved from https://www.technavio.com/report/global-luxury-cars-market-analysis-share-2018

Data preprocessing

Considering of processing power of lap top and the future output of NetworkX, we random subsampled 5000 from 37000 tweets we scraped.

Then we extract User screen name, Mention, Type of tweet, Mention user's screen name, followers count, friends count, listed count, favorites count, statuses count from original data (since original data are messy with multiple type, like json-like string, dictionary-like string, etc.)

Attributes	Rename
Retweet_Count	retweets_received
followers_count	follower_count
friends_count	following_count'
listed_count	listed_count'
Degree	network_feature_1
Betweeness	network_feature_2
Closeness	network_feature_3

Figure 1

Note: See the file "df3.csv" for details.

Any retweet, mention or reply should result in an arrow from the person retweeting to the person retweeted, mentioned or replied to. We created a three-column CSV file as follows: If @XYZ retweets a tweet by @ABC, then put the following in the CSV file like in assignment 1 part 2. (see figure 2)

User_name	Mention_name	Type
Flea_Breeland	Flea_Breeland	Tweet
24Roger_S	24Roger_S	Tweet
JohnnyIcePicks	JohnnyIcePicks	Tweet
McClain_on_NFL	ChronBrianSmith	Tweet
973espn	MikeGillShow	Tweet
mikesicehouse1	mikesicehouse1	Tweet
TannerPhares	TwitVI	RT
davedabbah	TalkToTheRamble	RT
dfs_jordan	OTHeroics1	RT
charliessports	charliessports	Tweet
VegasKillers	TheSharkWins	RT
Nasteedunx	Nasteedunx	Tweet
LineStarNBA	LineStarNBA	Tweet
MMS_Picks	MMS_Picks	Tweet
Bigree40	briksnchips	RT
tonyrhartley	tonyrhartley	Tweet
flerp20	flerp20	Tweet
maddydeleon790	SportsWatch1	RT

Figure 2

Note: See the file "df_final.csv" for complete version.

Then we used NetworkX for the three column CSV file and generated the network graph (see Figure 3) and calculated three features: Closeness, Betweenness, Degree.

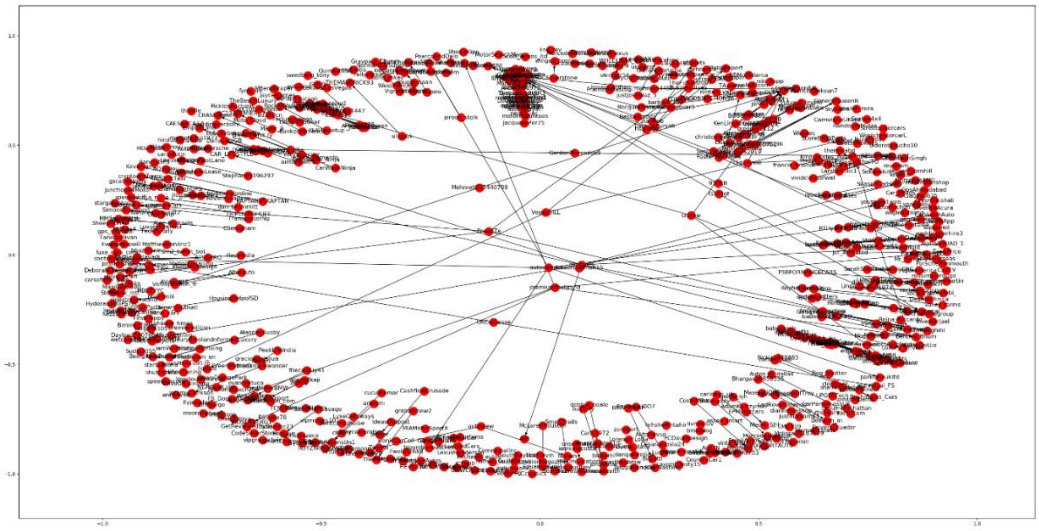

Figure 3

According to the Gini coefficients calculated in assignment 1, the weights are assigned to the attributes in figure 6 after normalizing the Gini coefficients. (see figure 4)

Attributes	Weights
retweets_received	0.1018
follower_count	0.0950
following_count	0.0975
listed_count	0.4387
network_feature_1	0.0906
network_feature_2	0.0988
network_feature_3	0.0776

Figure 4

Normalize the data "df3.csv" first, then get each user's score using formula:

$$Score = \sum_{i=1}^{7} weight_i * attribute_i$$

Here is the chart following (figure 5) is the top 10 influencers:

user_name	score	related
bsindia	9.52119716	no
LegendPorsche	5.64597805	yes
Linda_U_W_A	3.86788343	no
alonethoughts	3.70072231	no
TheLuxuryRep	3.57274364	no
HerdadeFozdaRe	3.18327703	no
WereBillionaire	2.56973074	yes
ATSOPRO	2.40786826	no
VegasBiLL	2.40529075	yes
knowledgelover1	2.29452286	no
_d_a_n_i_e_l__b	2.08587235	no

Figure 5

Note: Please see the file "top100_user.csv" for the complete top 100 influencers.

We have found top 100 social influencers in the network, based on hashtags. Considering the size of networks, we have manually searched on top 100 to ensure the relativity with car mentions, and there are 58 out of these top 100. After doing this, we have finalized selections, top 10 out of these 58 car-labeled influencers for further picture analysis.

Part 2: Graph analysis

Data Collection

After getting top 10 influencers from the 58 luxury-car-related users from the top 100 influencers in the network, we scrapped max 50 tweets using Twitter API and selenium in Python for all the top 10 users, information includes URL of pictures, # of comments/likes, and # of retweets in their respective tweets.

Data preprocessing

We used Google Vision to label and tag those picture we have scrapped with URL and then created engagement score which is equal to 0.5 * # of likes + 0.3 * # of comments + 0.2 * # of retweets. Based on engagement score, we have selected top and bottom 25 tweets. So that by applying LDA later, we could see which topic positive/negative influence has on comments/likes and retweets.

Modelling and LDA

Google Lable Model:

We first decided that the top 30 percent to be labeled as is_better = 1 (means it is better), and the lower 70% be labeled as is_better = 0 The model predict engagement with image labels (text) as predictors has a precision score about 66.9%, and detail will be found in (see appendix: attribute 1) in attribute. The confusion matrix is as follows (figure 6):

	Pos	Neg
True pos	33	36
True neg	34	73

Figure 6

Also, we extracted top words with negative and positive effect, they are as follows:

Positive:

sports car, automotive design, design, automotive vehicle, vehicle, land vehicle, car

Negative:

life, photography, venue stadium, device, technology, nature, nose design, graphics, parallel, darkness, turquoise, travel

Latent Dirichlet allocation:

In natural language processing, latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics. LDA is an example of a topic model.

In LDA, each document may be viewed as a mixture of various topics where each document is considered to have a set of topics that are assigned to it via LDA. This is identical to probabilistic latent semantic analysis, except that in LDA the topic distribution is assumed to have a sparse Dirichlet prior.

Each document is assumed to be characterized by a set of topics. This is like the standard bag of words model assumption and makes the individual words exchangeable.

By applying LDA, we grouped tags into 5 topics (figure 7):

topic 1	topic 2	topic 3	topic 4	topic 5
racing	vehicle	architecture	car	font
technology	car	building	vehicle	design
device	automotive	sky	land	text
motorsport	design	water	sports	graphic
product	land	tree	automotive	logo
formula	luxury	property	design	line
motorcycle	wheel	dog	supercar	graphics
electronic	motor	estate	performance	brand
advertising	tire	area	coupe	blue
food	exterior	home	classic	watch

Figure 7

(table of top 10 words according to their scores for each topic)

As mentioned before, based on engagement score, we have selected top and bottom 25 tweets. So that by applying LDA later, we could see which topic positive/negative influence has on comments/likes and retweets from the following result (figure 8):

Topics	topic 1	topic 2	topic 3	topic 4	topic 5
Average Topic Weights for the top quantile	0.1023746	0.1860688	0.1898478	0.4412757	0.0804331
Average Topic Weights for the bottom quantile	0.1114156	0.2121149	0.1474005	0.1014118	0.4276571
Difference	-0.009041	-0.026046	0.0424472	0.3398639	-0.347224

Figure 8

The main difference as we can see from the table above is that topic3 and 4 are the top 2 topics which have the highest positive topic weights difference, while the other 3: topic 1, 2 and 5 have the highest negative topic weights difference. Based on difference, our recommendation will be mentioned in next section.

Recommendation

Based on analysis, we recommend luxury cars to proceeds through two strategies, brand awareness and brand perceptions, by adjusting influencers network profiles and content designs. At first, luxury car brands can try to find top influencers, who we have discovered in the social network analysis, to post related contents, such as pictures and hashtags, and make more social attraction and feedbacks, such as comments, likes, and retweets, from followers. By offering monetary rewards and other benefits, such as test-drive and discount on certain models, luxury car brands can obtain influencers on twitter, and this can increase the brand awareness. Secondly, luxury car brands should adopt topic 3 and 4 from LDA analysis, and they emphasize more on car functions and associations and are environment and background conditions related. For instance, luxury cars showing more performance and off-road capability are more likely to attract social feedbacks, and

car photos taken in background of great architectures and properties may increase levels of arousal from followers in the social network.

Limitation

1. It is hard to decide the coefficient of # of likes/comments/retweets in calculating engagement score, we may be a little subjective.
2. To use LDA, we must assume that each document is characterized by a set of topics, which in fact, could not be guaranteed every time.
3. As mentioned in 'Data collection' and 'Data preprocessing' sections, even though we follow the most usual way to find top influencer, due to the messy social media environment, it is not possible to find only user who is really related to a certain topic, since some general Twitter user like some official accounts of big journal company may also be mentioned in lots of tweets, so we have to manually label if a user is related to our topic of interest or not. For example, in our case, the top 1 influencer: *bsindia* , is the official account of Business Standard, it does not really related to luxury cars, but still be labeled as the top 1 influencer.
4. It is hard for us to analyze video in some tweets, so we only include picture, text and social media analysis.

Appendix:

```
0.6697247706422018
The precision for this classifier is 0.6697247706422018
The recall for this classifier is 0.6822429906542056
The f1 for this classifier is 0.6759259259259258
The accuracy for this classifier is 0.6022727272727273

Here is the classification report:
              precision    recall  f1-score   support

       False       0.49      0.48      0.49        69
        True       0.67      0.68      0.68       107

   micro avg       0.60      0.60      0.60       176
   macro avg       0.58      0.58      0.58       176
weighted avg       0.60      0.60      0.60       176


Here is the confusion matrix:
[[33 36]
 [34 73]]

The top 10 most informative features for topic code True:
sports car automotive design design automotive vehicle vehicle land vehicle land vehicle car car
vehicle

The top 10 most informative features for topic code False:
life photography venue stadium device technology nature natural nose design graphics parallel darkness
turquoise travel
```

Attribute 1