Case Study 1 22nd September 2016

Sentimental Analysis of Driverless cars using Twitter

Guanxiong Liu, Jiankun Bi, Abhishek Easwaran, Naveen Pothayath, Suchithra Balakrishnan

---

http://i4.coventrytelegraph.net/incoming/article8229408.ece/ALTERNATES/s1227b/JS52089875.jpg

Idea Conceptualization

Our basic idea was conceptualized when Uber took the world by storm after they announced that driverless cars would be on the roads in a few weeks in the city of Pittsburgh, Pennsylvania.

The United States' Federal Auto Safety regulators added to the buzz by officially announcing that they expected the nation's highways to become safer if the vehicles were driven by machines rather than humans.

Twitter was abuzz with a lot of people tweeting about the topic. What we noticed was that while there were positive tweets, there were a lot of negative tweets surrounding the same. People seemed skeptical about the idea that a car could be driven without a human behind the steering wheel, much like the unbelievable shock people felt when Ford Model A hit a top speed of 28mph.

We wanted to figure out what people were thinking about driverless cars. More than the positive aspects, we were keen on the negative sentiments that plagued people when they heard about driverless cars.

We wanted to identify if it was paranoia or something else that created a sense of skepticism among people.

We thought that we could use demographics, gender and age groups of the people who tweeted (and those who didn't) to get a fair idea of which section of people had what notion about this idea. However, due to Twitter's privacy policy, this bit could not be done.

Take a step back, we extracted a lot of data about what people were tweeting regarding driverless cars using various hashtags and we've put forth all the information.

The concept of driverless cars

Google summed it up in a nutshell. "Imagine if everyone could get around easily and safely, regardless of their ability to drive."

Self-Driving cars have the following modules:

- Sensors – Lasers, cameras and radars detect objects in the vicinity of the cars

- Shape – They are intended to be round in shape to maximize the field of view of the sensors
- Computer – Contains the software designed specifically for self-driving

https://www.google.com/selfdrivingcar/how/

The car draws a map of the area it is in, has sensors that constantly detect what is around it, has software that predicts what the objects would do next, and finally chooses the appropriate actions that the car must do.

Why self-driving cars matter

Cars are an integral part in today's world. Human driven cars have led to a staggering 1.2 million deaths due to accidents every year. So, it was important that someone identify the need to automate this industry since we know very well that computers, if 'trained' well, perform much lesser errors when compared to humans.

Additionally, people with disabilities can let go of their dependencies with self-driving cars.

And finally, the time spent agonizing over traffic while driving could instead be used productively.

Why did we choose to perform sentiment analysis?

Everyone has a different opinion about self-driving. We needed to know the sentiments of people with respect to self-driving cars. We segregated it into two parts – positive and negative. We also intended to find out the most popular tweets in our collection and identify why the tweets were popular. We then tried to identify the largest number of retweets and the most popular tweets and find out what they said and whether it was positive or negative.

The Data

Sampling Twitter Data with Streaming API about self-driving cars

```
search_results= twitter_api.search.tweets(q='driverless', count=100, lang='en')
```

As seen above, the keyword searched for was 'driverless'.

We collected a total of 393 tweets.

```
outfile = open('tweets.txt', 'w')
```

The code above shows that the file ‘tweets.txt’ was created and the tweets were stored into it.

Tweet Analysis and Tweet Entities with Frequency Analysis

1. Word Count

```
    top_word = PrettyTable(field_names=['Word','Count'])
```

Word	Count
driverless	240
RT	223
cars	132
Driverless	77
highway	61
first	58
car	57
Germany	54
code	54
cost	54
create	54
world’s	53
per	50
mile	46
@newscientist:	35
take	34
https://t.co/6qD632WC8s	31
https://t.co/dlzqr2fm45	31
#driverless	27
taxi	27

2. Most popular tweets

Count	Screen Name	Tweet Text
344	weknowwhatsbest	RT @weknowwhatsbest: Google scientists invented the driverless car and got the idea after watching the Obama presidency.
310	MarketWatch	RT @MarketWatch: Average cost to take a taxi in U.S.: $3.46 per mile Estimated cost for a driverless taxi: 35 cents per mile https://t.co/G…
189	CBCNews	RT @CBCNews: Driverless highway from Vancouver to Seattle proposed https://t.co/QPxmy8QRG5 https://t.co/g2YsZek7Bk
111	bishnoikuldeep	RT @bishnoikuldeep: Driverless cars are going to wipe out 4 million jobs https://t.co/aBFRGBNake via @bi_contributors
68	grescoe	RT @grescoe: Future's not about driverless cars. It's about carless drivers: in other words, cyclists, pedestrians, #straphangers. https://…
65	Forbes	RT @Forbes: A future of shared, fully electric, driverless cars on demand is closer to reality than it might appear. https://t.co/cvr4PX8D0p
59	SebastianThrun	RT @SebastianThrun: Amazing milestone for #selfdrivingcar: https://t.co/43djoGARdl
44	FerRomero_FREE	RT @FerRomero_FREE: Great question for future mobility: How will pedestrians negotiate with driverless cars? https://t.co/tjnJLILVTk https…
35	newscientist	RT @newscientist: Germany to create world's first highway code for driverless cars https://t.co/uKrMM5YbzU https://t.co/8DfMTzCGUK
35	newscientist	RT @newscientist: Germany to create world's first highway code for driverless cars https://t.co/dlzqr2fm45 https://t.co/6qD632WC8s

3. Most popular Tweet Entities and Top 10 User Mentions

Hashtag	Count	Mention User	Count
driverless	28	newscientist	39
TDE7	9	MarketWatch	23
Driverless	8	grescoe	15
cybersecurity	8	weknowwhatsbest	15
smart	7	PeterGleick	14
hacked	7	FerRomero_FREE	10
tech	6	chrisriddell	7
IoT	6	CurbedLA	6
AI	5	BV	6
autonomos	5	villeohman	6

Extracting all ‘friends’ and ‘followers’ of a popular user

We extracted the friends of a user names ‘DriverlessCaRR’. We chose this user since this user has the most appropriate tweets with respect to driverless cars.

The three tables are: Friends, Followers and Mutual Friends respectively.

Friend ID	Screen Name
40796368	kaybed1
223514649	Zeyawinhtut
292825731	smkayes
113802609	RachaelAnne_3
104746157	Sorafatul
235160347	rbenlee
1669662031	saikatsamanta18
52056156	Mwardi300
465888120	HP101201
473720203	ferryantonie
159912065	DataRecoverer
411747568	heavydutyrigs
136652675	mwahyyd
1094487468	NichakornBogy
419777782	alamashraf2020
302055316	anjan_91
111618328	UMANGLOVESLIFE
275031863	JenBDesjardins

Follower ID	Screen Name
400646160	follow_eiver
108813774	AutomotiveUX
1940136116	kt3_mame
142297445	RoboticsTrends
738259224	Itsyourfood
138419527	RitukarVJ
1642353530	sukyvokydoj
772873027436613632	optimizehicle
224441742	AitkenResearch
77271015	nealboudette
26548315	soledadobrien
9668532	emilmont
777440386750947328	dani_l_hogan
3948678562	manuel_cassar
544575063	bbaumgartner03
378623979	alvaro_arrue
776761808577331200	judytheredhead
776550863431200769	ChAs69672343

Mutual Friend ID	Screen Name
857645059	AutomotivePRBE
1256669196	wiemsoukeh
998318094	tc604
2338316309	FloreaMatei
2849882135	XclusiveAutoNJ
2692718616	MrDWilliamsBGHS
353665049	Biltac
21495840	gr8n8no
19578913	Verio
2373877795	Ancaster_Group
256655396	orilayeelay
35545125	FresnoBill
41295909	redsox54
54575144	carlisence
378019885	KevTynan10
2754117678	avtomaser
281337902	Suren_Fl
204955698	jacobyaaron
48906292	c230mike
546054204	AndreaRapelli
237846600	Ken_Simmons_NL
214237258	xavirocal
268247115	TrendLineNews
135290960	WaldorfPress
1902665809	HaysRedmaro
87654481	althiha
2592276564	ArmandJessie
2979217499	manueliasperez

Sentiment Analysis of Driverless Cars

```
def mongoDB_connect():
    client = pymongo.MongoClient()
```

Firstly, we stored our tweets into MongoDB since it’s designed for unstructured data like json respond from twitter api. The code above shows our connection with MongoDB.

We used the following keywords:

```
key_word = ['driverless',
            'AutonomousCaRR',
            'googlecar',
            'autonomouscars',
            'selfdriving',
            'futuristiccars',
            'driverlesstechnology',
            'SelfDrivingUber',
            'UGV']
```

We extracted a total of 5754 tweets for the keywords.

Two concepts were used here:

1. PMI Value
   The PMI value, also called Pointwise Mutual Information, is a mathematical function that gives people a general idea of how "close" two words could be by measuring the co-occurrence probability divided by their own probability to show up. The detailed function is showed in below.

$$PMI(w1, w2) = \log\left(\frac{P(w1 \wedge w2)}{P(w1)P(w2)}\right)$$

2. Semantic Orientation Value
   The basic idea of Semantic Orientation is introduced in Turney's paper. A brief description is that given a positive and negative word pool, we can calculate PMI from a word to these two pools. By calculating the over-all distance from these pools, we can have a orientation of it. The advantage of Semantic Orientation is that once we have general positive and negative word pools, we don't have to get any training data for particular input. Therefore, it's a kind of unsupervised classification method. The general function of Semantic Orientation is also listed here.

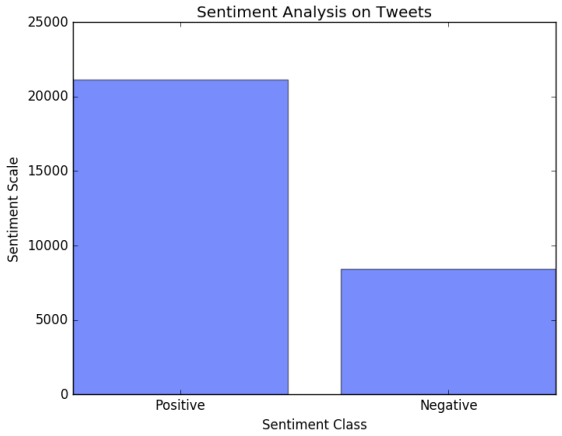
$$SO(w) = \sum_{w' \in P^+} PMI(w, w') - \sum_{w' \in P^-} PMI(w, w')$$

By using the previous mentioned methods, we calculated the semantic orientation value and came up with a list of the 10 most positive and negative words.

Word	Semantic Orientation Value
autonomous	102.595997287
day	74.5923936053
auto	67.1979537487
car	63.0891322622
amp	61.3199661459
ford	58.5628261828
a-ugv	58.1485382108
cross	53.6277545396
autos	52.4839898106
going	51.8391172464

Word	Semantic Orientation Value
apple	-56.1334657501
could	-55.9001934622
insurance	-52.5995170403
bad	-46.2039325324
freedom	-46.0580240552
also	-45.6672164765
piss	-42.0580240552
idea	-41.6016465549
chrisl2185	-40.303136553
cops	-39.357584337

We then plotted a graph of positive and negative values based upon their sentimental scale. We added the semantic orientation values of positive and negative tweets and came up with the numbers.



As expected, a lot of the words did not make sense but we did get insights into what people actually felt. For example, we found out that people were talking negatively and their tweets contained the words like insurance, freedom, bad ~ idea and cops.

One surprising inclusion was Apple which topped the charts. We intended to find out more and dug deeper by analyzing tweets which contained ‘Apple’ and our keywords. Apart from that, we did analysis of Uber, Tesla, Google too. Our results are described graphically below.

Visualizations

Apple

There's a sense of negativity surrounding Apple's lack of progress in the project and as we can see, words like 'strategy', 'rethink', 'bumps' and 'slips' have made their way here indicating that Apple needs to probably rethink its strategy.

Google

Google on the other hand doesn't seem to have that much of negativity air surrounding its project and people is also talking about the possibility of combine google map with its driverless car.

We haven't included results of Ford, Tesla and Uber because their wordclouds indicate either positive or neutral sentiments and hence are not a part of our analysis.

Limitations while performing Sentiment Analysis

We primarily targeted three areas:

1. Demographics
2. Age
3. Gender

What we intended to do was as follows:

1. Identify the geo-location of people tweeting negatively about driverless cars
2. Get information about their age and segregate further into age groups
3. Get their gender and then segregate again
4. Companies could then use this information to try and change peoples' perspectives by using their PR team. For example: If males in Worcester in the age group of 40-50 have skepticism about the safety of driverless cars, then a particular company could use that bit of information to try and target that audience and change their perspective of self-driving cars.

Unfortunately, Twitter classified that as private information and we could not get that information using Twitter Streaming API.

Additionally, even if the API did work, we cannot verify the authenticity of the information provided by people. Given a lot of disadvantages, it is rather difficult to gather accurate information from twitter.

Conclusion

We set off with the intention of identifying negative tweets surrounding driverless cars. From the negative topic that people concerned, we could find out the critical problem that need to be solved. Based on the result, we came to a conclusion that the general perception of people regarding this were issues like insurance, cops, freedom etc. Hence, companies can get some light from these results and treat them as business needs to address these issues and attempt to change peoples' perspectives.