MWSUG 2016 - Paper AA06

Analyzing sentiments in tweets for Tesla Model 3 using SAS Enterprise Miner and SAS Sentiment Analysis Studio

Tejaswi Jha, Oklahoma State University, Stillwater, Oklahoma
Praneeth Guggila , Oklahoma State University, Stillwater, Oklahoma

ABSTRACT

Tesla Model 3 is making news in the history of automobiles as never seen before. The new electric car already has more than 400,000 reservations and counting. We carried out a descriptive analysis of sales of all Tesla models and found that the number of reservations till date are more than three times sales of all previous Tesla cars combined. Clearly there is a lot of buzz surrounding this and such buzz influences consumers' opinions and sentiments and which in turn lead to bookings. This paper aims to summarize findings about people's opinions, reviews and sentiments about Tesla's new car Model 3 using textual analysis of tweets.

For this, we used the live streaming data from Twitter over time and studied its pattern based on the booking timeline. We have been collecting data from March 2016 when the interest of people in this model spiked suddenly. A sample of 1000 tweets was analyzed from a total of about 10,000 collected via fumes. We used the SAS® Enterprise Miner and SAS® Sentiment Analysis Studio to evaluate key questions pertaining to the analysis such as following. What features do people think about? What are the factors that motivate people to reserve Tesla? What factors are discouraging them?

INTRODUCTION

Tesla motors introduced their first electric sport car – 'Tesla Roadster' in 2008. Telsa motors became popular with their second electric car 'Model S', a fully electric luxury sedan which became the world's second bestselling plug-in car after Nissan Leaf. Tesla Motors has sold almost 140,000 electric car worldwide after its first release. The new model Tesla model 3 is released by the company at an affordable price for lower income consumers. The 'Tesla Model 3' is an all-electric four-door compact luxury sedan which was unveiled in March and the deliveries are planned for the end of 2017. According to the company officials, within a week of the unveiling, 325,000 Model 3 reservations were made, more than triple the number of Model S,Tesla had sold by the end of 2015. These reservations represent potential sales of over US$14 billion [6]. As of May 15, 2016, Tesla had taken about 373,000 reservations [7]

In US, there are five cities which have more than 50% of total reservations of electric cars including Tesla Model 3 as seen below. The reservations for Tesla Model 3 is continuously at rise in these cities [8]

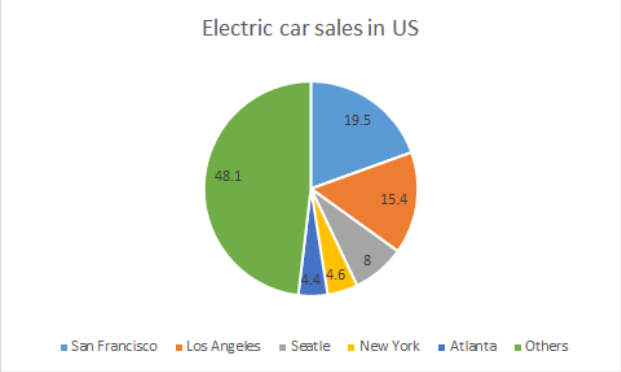

Figure 1.1- Electric car sales in US – top cities

On looking into the interest of people in Tesla Model 3 based on their searches on google, it was found that the interest is particularly high in the cities shown below.

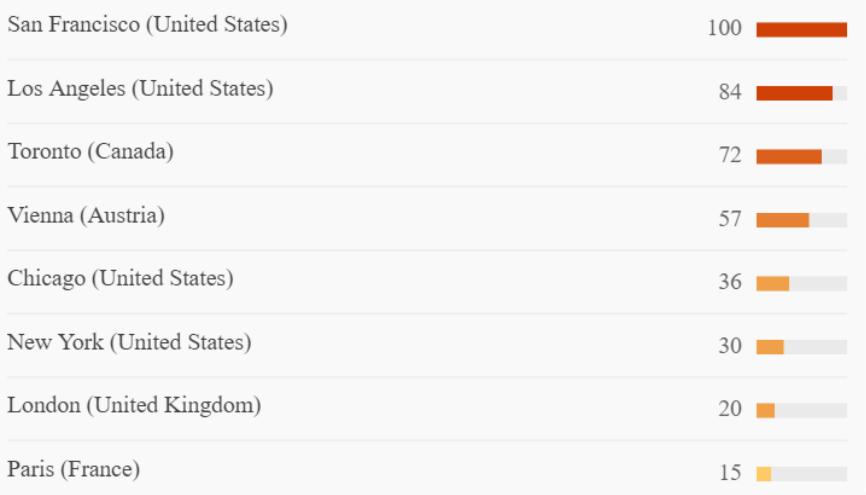

Figure 2.2- Top cities for tesla model 3. Web Search. Worldwide, Past 90 days

Wouldn't it be great if we could find the reasons behind such high reservations by understanding people's opinion about the latest Tesla car – Tesla Model 3.

We can find out what topics are people talking about in the tweets Also, overall sentiment of the tweet corpus can also be analyzed by sentiment analysis. This would help us understand what the positive as well as negative sentiments which are expressed by the users via tweets. This analysis would help us quantify how people feel about the new electric car by their sentiments expressed via tweets.

DATA PREPARATION

The primary source of data for this research was the twitter feeds posted by users. Tweets posted around the time of announcement of Tesla Model 3, were collected using FLUME which uses the exposed twitter API to save tweets. The process flow for the analysis can be seen in the figure below-

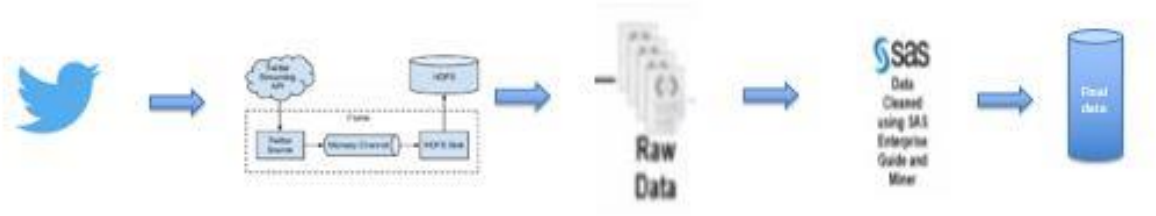

Figure 3 – Preparation of data

FLUME saved the tweets on the Memory Channel (MEM) after which the data is sent to the Hadoop Distributed File System (HDFS) sink, and then to the HDFS. HDFS is used to store data collected from social networking sites. Since the data is in an semi-structured format (.json), it is converted into a structured format (.CSV) using Python. Over a period of 18 weeks, 10,000 tweets were collected under the handle #TeslaModel3. 90% of the tweets obtained under this handle were tweeted within 60 days after Tesla Model 3 unveiled.

METHODOLOGY

The data was portioned into two stratified samples (training and Validation). The training data was used to build the model and the validation data to text the accuracy of the model. This provided an honest assessment of the models built. Then, a sample of Twitter feeds were classified into positive and negative categories and this sample was used to train the statistical models in the sentiment analysis studio which was later used to classify the remaining data.

Once the tweets were collected and converted into SAS dataset, there were two analysis carried on the data-

1. Creating text clusters, text topics and concept links to identify meaningful tweets and understand association between the terms.
2. Generating text rules based on text clusters.

ANALYSIS 1- CREATING TEXT CLUSTERS, TEXT TOPICS AND CONCEPT LINKS

In the SAS Enterprise Miner, the file Import node, Text Parsing node, Text Filter node, Text Cluster node and Text Topic node were connected in a flow as seen below.

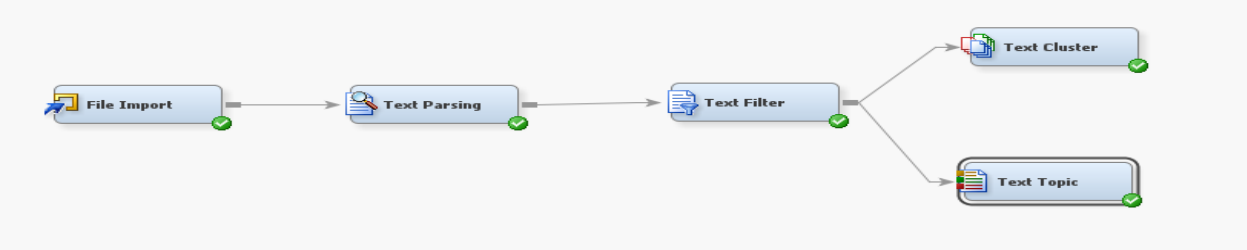

Figure 4 – Text mining process

Detailed description of node settings, node functions and node results

File Import

After converting json file into excel file using python, file import node is used in SAS enterprise miner for importing the data. File import node is used to convert the external data files like spreadsheets and database tables into a format that SAS recognizes as a data sources. We imported sample of 1000 tweets into SAS for analysis.

Text Parsing

The text parsing node parses the data set containing tweets in order to quantify the words used in the tweets. It analyzes and filters same words and special characters. It creates the term by document matrix. Generally, terms are single words considered along with their synonyms/stems, multi word phrases, parts of speech etc. From the result of text parsing node, we could find that the most frequent words were teslamodel3, elonmusk, electric, wait etc. There were a lot of misspelt words and short form of words in the tweets (like k instead of ok, ryt instead of right) which were taken care of in the next step of filtering.

Text Filter

Text filter node helped us to restrict the number of terms in the tweets by removing similar words that are not useful for our analysis. For running this node we provided a complete English dictionary containing terms and synonyms. In the property panel, the frequency weighing was default and term weight was set to mutual Information

We enabled the spell check option which suggested potential synonyms. We manually looked into some words and treated them as synonyms such as “innovation”, “breakthrough”, “revolution”,” disruptive” etc. This node created a compact set of meaning texts.

Concept Links

Concept links are a type of association analysis between the terms used which help to understand the relationship between words. They can be viewed in the interactive filter viewer. We generated concept links to answer two questions regarding people’s choice -

1) Why an electric car?

The answer to this question was found in the concept link of electric car as seen below-

Concept link of Electric car

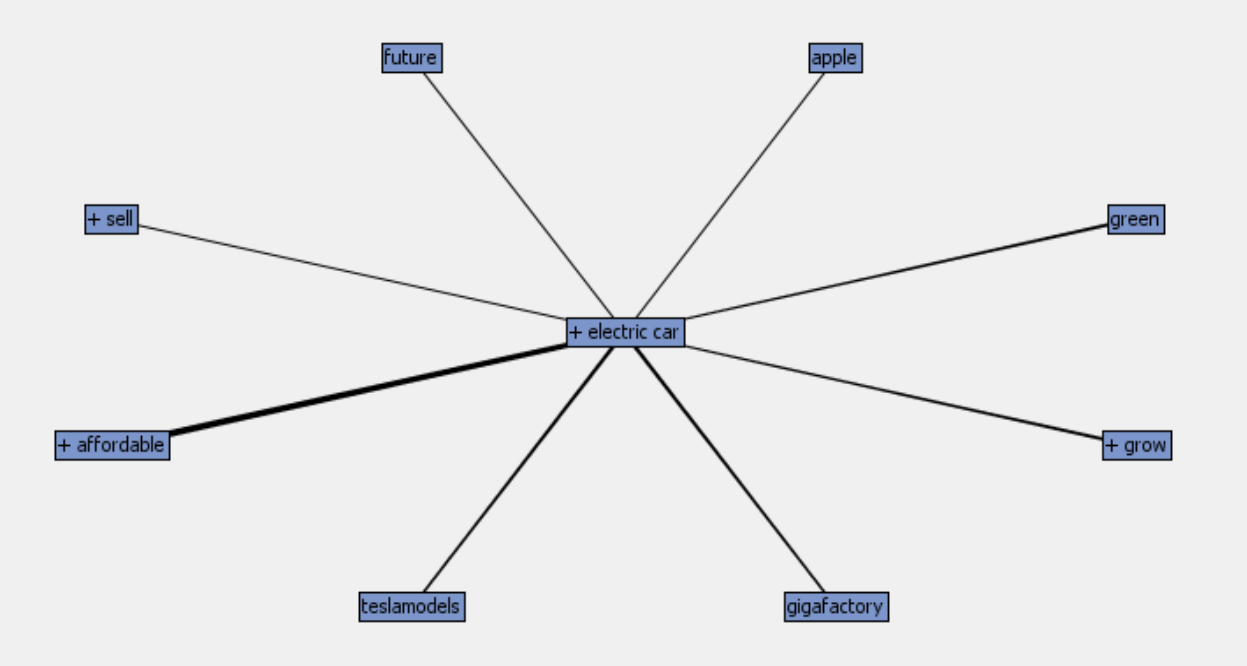

Figure 5- Concept link of Electric car

The link shows the term electric car to be analyzed in the center and the terms that it is mostly used with as links. The width of the link here is directly proportional to the strength of association of the term with electric car. Also, it is showing how many times the two terms co-exist in a tweet regarding Tesla Model 3 In this concept link electric car is strongly related to term affordable, tesla model and gigafactory when compared to other terms future, sell, grow, apple and green. Most tweets show that with minimized styling, basic model of TeslaModel 3 is the most affordable electric car.

2) Why a Tesla Model 3?

The answer to this question was found in the concept link of TeslaModel3 as seen below-

Concept link of Tesla Model 3

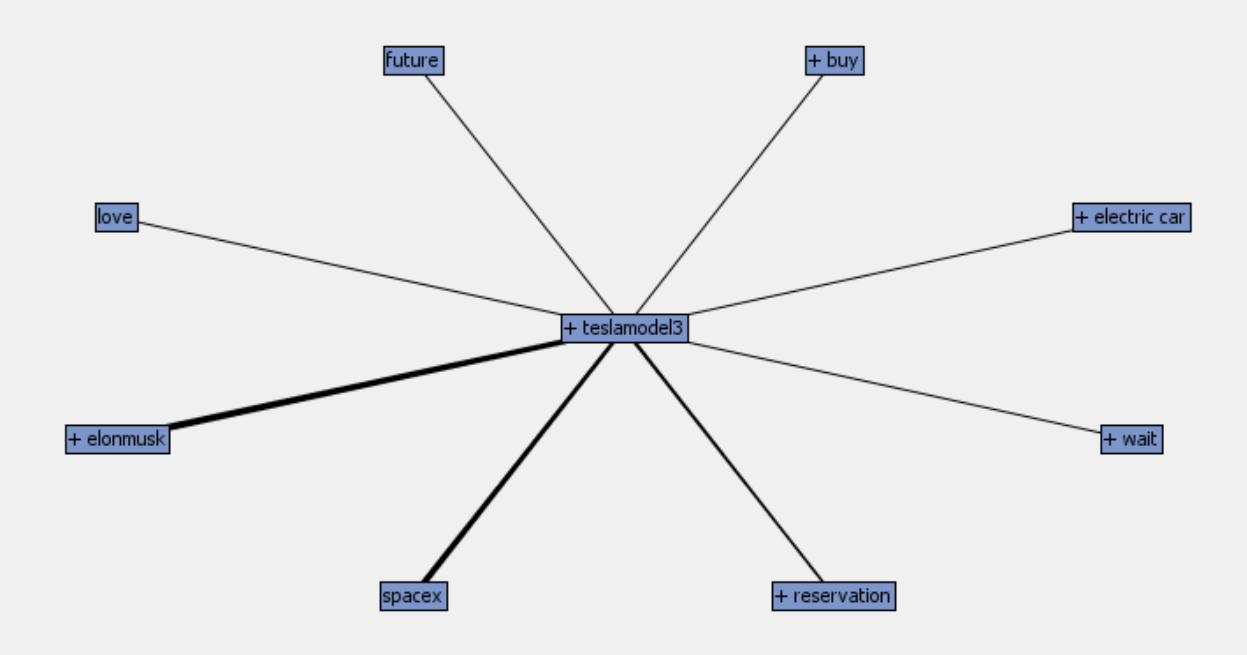

Figure 6- Concept link of TeslaModel3

The most frequently associated words with Tesla Model 3 is Elon Musk and SpaceX. An ardent fan base of Elon Musk was found in twitter with most positive tweets about him, TeslaModel3 and his other venture SpaceX. The tweets are drawing parallels between technology used in TeslaModel3 and SpaceX. Users are comparing the car's interior with spaceship interiors.

The links clearly show the mindset of users regarding TeslaModel3. The tweets show association of this car with words like future, reservation, love, wait etc.

Text Cluster

The text cluster node in Enterprise Miner groups the similar terms in the dataset together. In this case four clusters are generated and all of them are well separated from each other as seen in Figure 7. The pie chart (Figure 6) shows the distribution of the cluster frequencies for the four prominent clusters. The frequencies are well distributed amongst all four clusters with cluster# 3 showing a little high frequency than the rest.

Figure 7 Cluster Frequency

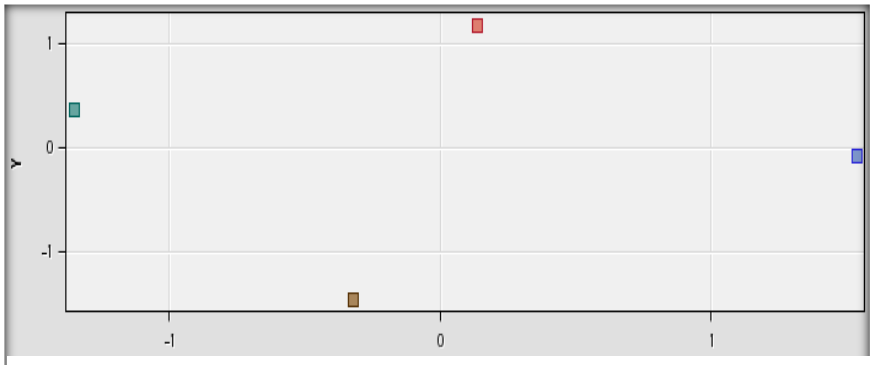

Figure 8 Distance Between Clusters

List of clusters generated

The four clusters generated from the text cluster node can be seen as below. The words in each of these clusters clearly tell a story regarding increasing reservations, anticipation and wait period for this car.

Table 1 - Distribution and explanation of text clusters

Cluster ID	Descriptive terms	Percentage	StoryLine
1	+video, + cost, +electricvehicle, +success, +Pre-orders, +Priority	20%	This cluster is a group of terms that describe the factors related to the success of tesla model 3 car
2	+tax break, +subsidy, +government, + huge, +buyer	20%	This cluster group describes about the subsidy benefits of the car
3	+reservation, +Plant, +factory, +elonmusk, +wait	33%	This cluster is group of terms related to production of tesla model 3 car
4	+affordable, +sell, +year, +stock, +autonews, +price, +autopilot, +interior	26%	This cluster is group of terms that related to describing the factors associated with the features and market value of tesla model 3 car

Text Topic

Next we connected the Text Topic node to the Text Filter node which enabled us to combine the term into topics so that we can analyze further. Using text topic node and by care carefully selecting terms, 8 user topics were defined as shown below.

Table 2 – Text Topics

Topic ID	Topic terms	Explanation
1	+electric car, +market, +affordable, +teslamodel3, +future	This topic is a group of terms that describe the Tesla Model3 car
2	+spacex, +elonmusk, +great	This topic talks about Elon Musk ,his qualities and ventures.
3	+elonmusk, +teslamodel3, love	This topic is about love for Tesla Model3
4	+reservation, +company, +electric car, +elonmusk	This topics is a group of terms that talks about reservation of the car and brand value
5	+wait, +love, +excite, +deposit, +look	This topic is group of terms that talks about excitement and waiting period of the Tesla Model3 car
6	+buy, +stock, +pay, +teslamotors	This topic is group of terms that talks about stocks and market shares of Tesla.
7	+reserve, +gm, +Chevy bolt, +pre-order	This topic is comparing Tesla Model3 with its competitors like general motors and Chevy Bolt.
8	+teslamodelx, +tech, +prototype	This topic is group of terms that compares Tesla Model3 car with other Tesla cars.

ANALYSIS 2 – SENTIMENT ANALYSIS

Statistical model

We used the SAS® Sentiment Analysis Studio to build a statistical model with a stratified sample of 426 tweets. 80% of the tweets were used to train the model. On the train data, when we run the statistical models, the Smoothed Relative Frequency and CHI Square Model gave the best results.

The result of this model can be seen as below.

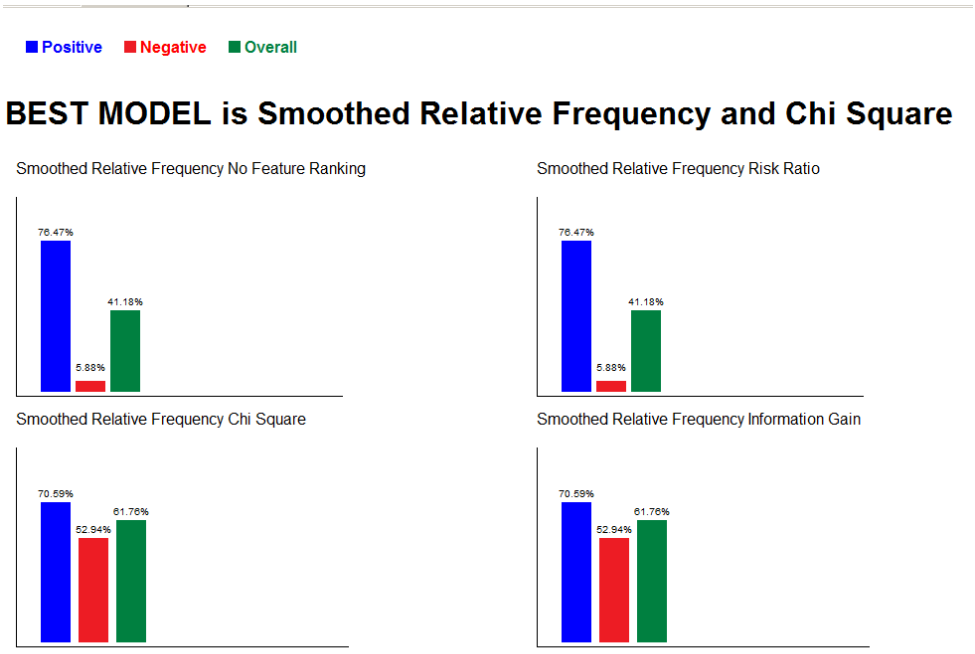

Figure 9 – Model Comparison

The positive precision is doing better than the negative precision with overall precision of 61.7%.

Chi-square is a feature ranking algorithm that basically classifies the features of the document based on its frequency and importance and uses it to build a model.

As there is a difference in the size of documents, in order to attain correct length of document and number of feature words per document, the smoothed relative frequency algorithm performs text normalization. Next we brought the test data in to test model accuracy.

Test Results

We tested for a small data set containing 50 positive tweets and 50 negative ones.

Test for positive tweets

Figure 10 – Test for positive tweets

The model correctly predicted the positive directory with 80% precision of positive tweets.

Test for negative tweets

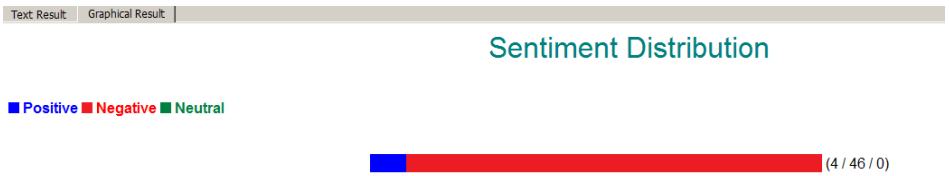

Figure 11 - Figure 12 – Test for positive tweets

The model correctly predicted the negative directory with 92% precision of negative tweets.

The statistical model did a very good job of predicting the tweets as positive and negative. In order to see what terms were classified by the model as positive and negative, we built a rule based model in Enterprise Miner.

Rule Based Model

In order to build the rule based model, the flow of nodes in Enterprise miner is shown as below. All the nodes like input data node, Data partition, Text Parsing and Text Filter have the same settings of their properties as before in the cluster analysis.

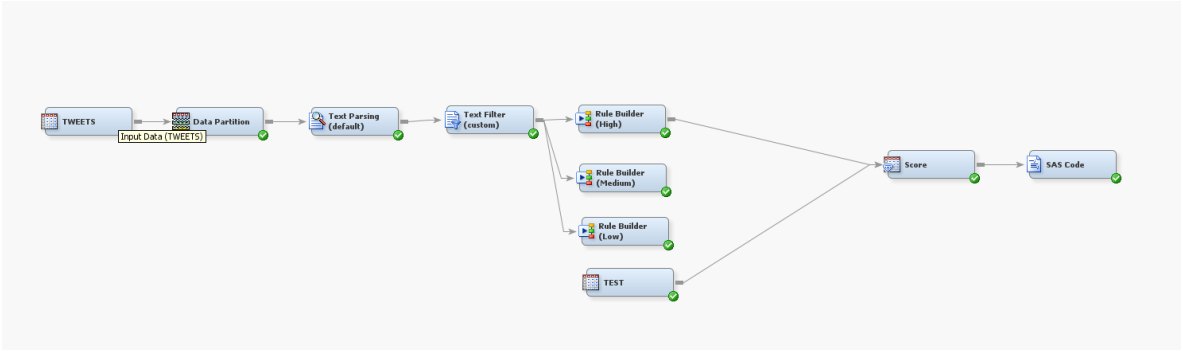

Figure 13 – Rule based model

Text Rule Builder Node

The “Text Rule Builder” node generated an ordered set of rules.
It identified the different subset of terms that describe our target sentiment (positive or negative).
The same rules predict the sentiments in the test data.

The Rules indicate the presence or absence of one or a small subset of terms (for example, “term1” AND “term2” AND (NOT “term3”)). Only those tweets match the rule which contain at least one occurrence of term1 and of term2 but no occurrences of term3.This set of derived rules creates a model that is both descriptive and predictive.

When categorizing a new document, the model will proceed through the ordered set and choose the target that is associated with the first rule that matches that document. The text rule builder node were added to the flow with different settings to get comparable results. The rule based categorizer automatically generates ordered set of rules to describe and predict the target variable.

The text rule builder node is run with different settings of generalization error, purity of rules and exhaustiveness (High, medium and Low). The setting with high generalization error, purity of rules and exhaustiveness gave the best results with lowest misclassification rate.

Fit Statistics	Statistics Label	Train	Validation
ASE	Average Square...	0.006274	0.003375
DIV	Divisor for ASE	2544	642
MAX	Maximum Absol...	0.745558	0.719774
NOBS	Sum of Frequen...	848	214
RASE	Root Average S...	0.079208	0.058097
SSE	Sum of Squared...	15.96087	2.166925
DISF	Frequency of Cl...	848	214
MISC	Misclassificatio...	0.029481	0.042056
WRONG	Number of Wron...	25	9

Figure 14- Rule Based Model Fit Statistics

The validation misclassification rate was found to be 4% which is slightly higher than desired, but considering the challenges with analyzing limited content in tweets, we used this model to score.

Improving the model

To further improve the efficiency of the model, we manually checked the 'change target values' of the rule builder node to see if any reviews were classified incorrectly.

Change Target Values-WORK.TRCHANGE

Text	Data Partition	Target Variable	Original Target	Predicted Target	Why Classified	Posterior Probability	Assigned Target
@WillOremus Most car pollution is produced during the manufacturing process, long before it gets driven. #TeslaModel3	Training	F2	NEGATIVE	POSITIVE	car & ~charge & ~buy & ~battery & teslamodel3	96.4%	NEGATIVE
For #TeslaModel3 Buyers, the Waiting Could be the Hardest Part http://www.thetruthaboutcars.com/2016/04/teslas-volume-is-growing-but-the-waiting-could-be-the-hardest-part-for-model-3-buyers/... #AutoNew	Training	F2	NEGATIVE	POSITIVE	car & ~charge & ~buy & ~battery & teslamodel3	96.4%	NEGATIVE
A discussion about Tesla's quality and battery degradation https://youtu.be/glPgqLlqv5c #TeslaModel3 #Tesla	Training	F2	POSITIVE	NEGATIVE	battery	46.9%	POSITIVE
Over 47M miles driven on Autopilot, the more you drive, the more we'll learn	Training	F2	POSITIVE	NEGATIVE	autopilot	40.0%	POSITIVE
I just realized the incredible traffic analysis possible when >300k cars on the road have AutoPilot hardware #TeslaModel3	Training	F2	POSITIVE	NEGATIVE	autopilot	38.5%	POSITIVE

Figure 15 – Editing target variables

The highlighted tweet in figure 14 has a negative touch but was originally classified as positive. We changed the classification of the tweet from positive to negative in the source file. Similarly, some tweets were wrongly classified by the rules, whereas for some, the original target was set incorrectly. After making all required changes the model was run again and this time the misclassification rate came down to 2.7%

After improving the model, going ahead, we checked the rules that were built to understand how the rules are classified as positive and negative.

Positive Rules

The positive rules contain terms like elon musk, autopilot, battery, charge and buy. The precision of positive rules ranges between 100 to 99%

Some of the positive rules given by this node are shown below-

POSITIVE	5 week	100.0%	26.26%	41.59%	100.0%	23.41%	37.94% 36/36	10/10
POSITIVE	6 elonmusk	100.0%	29.82%	45.94%	100.0%	27.80%	43.51% 39/39	13/13
POSITIVE	7 tsla	100.0%	32.76%	49.35%	100.0%	29.76%	45.86% 43/43	8/8
POSITIVE	8 elon	100.0%	35.21%	52.09%	100.0%	32.20%	48.71% 29/29	6/6
POSITIVE	9 teslamodel3 & ~car & ~autopilot & ~battery	99.71%	83.07%	90.63%	98.29%	83.90%	90.53% 585/587	148/151
POSITIVE	10 car & ~charge & ~buy & ~battery & teslamodel3	99.34%	92.39%	95.74%	98.42%	91.22%	94.68% 98/101	21/21

Figure 16 – Positive Rules

Negative Rules

Some of the negative rules given by this node are shown below

NEGATIVE	12	crap	100.0%	20.90%	34.57%	100.0%	11.11%	20.00%	14/14	2/2
NEGATIVE	13	accident	100.0%	28.36%	44.19%	100.0%	27.78%	43.48%	7/7	3/3
NEGATIVE	14	crapy model	100.0%	35.82%	52.75%	100.0%	33.33%	50.00%	5/5	1/1
NEGATIVE	15	insanely	100.0%	41.79%	58.95%	100.0%	33.33%	50.00%	4/4	0/0
NEGATIVE	16	hate	100.0%	47.76%	64.65%	100.0%	33.33%	50.00%	4/4	0/0
NEGATIVE	17	battery	97.30%	53.73%	69.23%	75.00%	33.33%	46.15%	4/5	0/2
NEGATIVE	18	long	95.12%	58.21%	72.22%	75.00%	33.33%	46.15%	3/4	0/0
NEGATIVE	19	delivery & ~teslamotors	95.56%	64.18%	76.79%	75.00%	33.33%	46.15%	4/6	0/0
NEGATIVE	20	buy & car	92.31%	71.64%	80.67%	77.78%	38.89%	51.85%	5/11	1/1
NEGATIVE	21	deal	90.91%	74.63%	81.97%	77.78%	38.89%	51.85%	2/4	0/0
NEGATIVE	22	waiting	88.14%	77.61%	82.54%	77.78%	38.89%	51.85%	2/4	0/0
NEGATIVE	23	late & ~tesla	84.62%	82.09%	83.33%	80.00%	44.44%	57.14%	5/8	1/1

Figure 17- Negative Rules

The negative rules contain the terms like crap, accident, late, hate, waiting etc. Some terms like autopilot and battery were categorized as both positive and negative. The precision of negative rules is from 100 to 65%.

Score Node

For scoring we used a Score Node and then connected it to the dataset containing all the tweets. We then used a SAS Code node to export the scored dataset.
After this, we used this model to score the test data set which had 300 tweets (150 positive and 150 negative)

The model showed 143 observations classified as negative and 157 are classified as positive.

CHALLENGES IN ANALYZING SENTIMENT IN TWEETS

- Some tweets with different related topics with Tesla included a # for teslamodel3.In such cases, though the model categorizes them as positive or negative, the results are not actually useful as they do not make business sense.
- Some tweets are very short; it is difficult to categorize them as positive or negative.
- In some tweets, the users want to express their sentiments with a link to a website including a # for teslamodel3.So the actual sentiment of the user are not clear by the tweet alone. Such tweets have to be ignored for sentiment mining.

CONCLUSION

The large amount of information contained in microblogging web-sites makes them an attractive source of data for opinion mining and sentiment analysis. This research sheds light on the views and opinions of twitter users about Tesla Model 3 and found that the there is a very strong liking among the users for this car. There were very few tweets with negative sentiments on Tesla Model 3. Twitter does not represent the entire spectrum of electric car fans or haters. As a future scope of analysis, other sources of people's opinions can be added to the tweet corpus such as articles in car magazines or editor columns. A richer data set will improve the model.

REFERENCES

1. http://support.sas.com/documentation/cdl/en/emag/65762/PDF/default/emag.pdf
2. http://www.sas.com/en_us/software/analytics/text-miner.html
3. Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS® by Goutam Chakraborty, Murali Pagolu, Satish Garla. 3) Sentiment Analysis and Opinion Mining by Bing Liu (May 2012). 4) SAS Institute Inc 2014. Getting Started with SAS® Text Miner 13.2. Cary, NC: SAS Institute Inc
4. Sharat Dwibhasi, Dheeraj Jami, Shivkanth Lanka, Goutam Chakraborty, 2015, "Analyzing and visualizing the sentiment of the Ebola outbreak via tweets "

5. Chakraborty, Goutam and Pagolu, Murali. 2014 “Automatic Detection of Section Membership for SAS® Conference Paper Abstract Submissions: A Case Study” Proceedings of the SAS Global Forum 2014 Conference, 1746-2014, Washington, DC.: SAS Institute Inc.

6. *Baker, David R. (2016-04-01).* "Tesla Model 3 reservations top 232,000". San Francisco Chronicle. Retrieved 2016-04-02. Tesla Motors had sold 107,000 Model S cars by the end of 2015

7. *Cole, Jay (2016-05-18).* "Tesla, Musk Plan $2 Billion Stock Sale To Build Model 3, 373,000 People Reserved". *InsideEVs.com*. Retrieved 2016-05-18.

8. http://wheels.blogs.nytimes.com/search/detroit/page/23/?_r=0

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Tejaswi Jha
Oklahoma State University
Phone:4057622618
Email:tejaswi.jha@okstate.edu

Praneeth Guggilla
Oklahoma State University
Phone:405-780-5330
Email:praneeth.guggilla@okstate.edu