Sentiment Analysis of Opinions about Self Driving Cars

Nachiket Kawitkar, Swapneel Deshpande, Dr. Goutam Chakraborty, Dr. Miriam McGaugh
Oklahoma State University

INTRODUCTION

“From 2020, you will become a complete backseat driver”, says the Guardian. The manufacturers are claiming that the Self driving cars will revolutionize motoring. However, few wonder if the greatest danger for these cars is that they will be ‘too safe’ to drive. While few automakers believe that newest technology and added features in these cars will potentially save 30,000 lives a year. The only obstacle to that is convincing the customer to give up the control of their car and hand it to a computer. Recently, Tesla made its cars semi-autonomous, not only did the newer version of their cars have an autopilot feature but also the tens of thousands of the existing customer cars became way better with an overnight update. However, the company recommended keeping hands at the wheel at certain times when this autopilot version would like the human to take control of the car. While recently, Uber will be allowing its customers to call for its newer self-driving car cabs from their mobile phones. This feat makes Uber only one of the very few companies who achieved this milestone in the car and ride sharing market. A lot is being said about these self-driving cars online. A lot of comments, concerns, statements, suggestions can be found online, few are positive, negative or even neutral. So making a complete analysis of how people are taking this new technology currently is indeed challenging.

Imagine an analysis of comments and reviews on the internet about self-driving cars, in order to understand what exactly the customers have been liking or disliking about this newest technology currently in the market. Utilizing text mining, we can locate the terms that have been used most frequently in regards to the self-driving cars and check how they are affecting the customer decision. We can further analyze the relation in between these terms and thus gauge customer satisfaction or dissatisfaction towards this futuristic technology. Sentiment mining can help us figure out why or why not this technology a hit or a failure amongst the current generation. Companies can use this analysis to improve the current self-driving cars and even design targeted marketing campaigns towards their customer base to further make larger revenue. On the other side the customers can use this analysis for gauging how are their peers dealing with this futuristic self-driving cars, as investing in these autonomous cars currently is a very pricey decision. So we believe our analysis will help both of the parties make a calculated decision on moving forward with this futuristic dream.

DATA ACCESS

The data contains sentiments of people voiced about self-driving cars obtained from http://www.crowdflower.com/data-for-everyone. The data set contains these reviews about the self-driving cars composed of 7,156 observations and 9 variables. We have considered 2 variables Sentiment and Text for our analysis and decided to drop variables containing usernames, date, time, location etc.

DATA DICTIONARY

Variable	Role	Description
Sentiment	Target	This field represents sentiments classified as positive, negative or neutral
Text	Text	This variable represents the actual comments posted regarding self-driving cars

METHODOLOGY

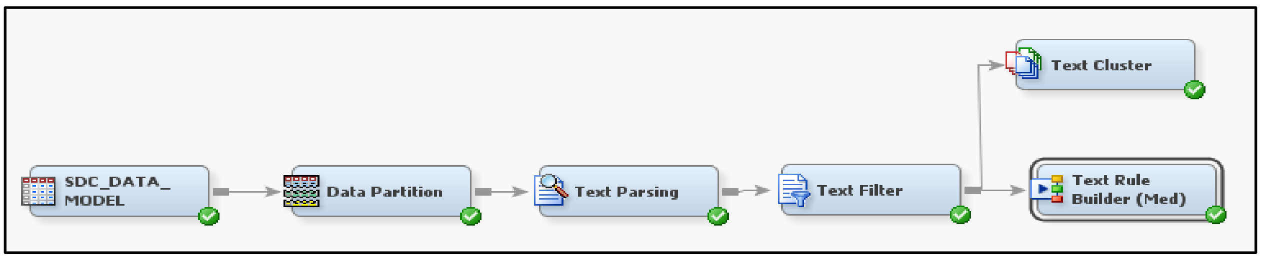


Figure 1: Text Mining Process Flow

Data partition

Training	Validation
70%	30%

Text parsing

The SAS® dataset after partitioning was attached to a text parsing node. In order to clean the unstructured data a few modifications have been made to it. Utilizing the properties panel following settings have been made. Alongside the default options abbr., prop., num. and parts of speech have been ignored. Further the find entities option has been set to standard. In order to not have repetitive terms as well as consider one word or term as a whole, the detect different parts of speech option is set to NO.

To comprehend the most frequently occurring terms and count the occurrence of these words in different documents, a term called frequency document matrix has been used from the text parsing node. An analysis of rarely used terms is also done via this process. The most helpful in exploration and modeling process are the terms which are moderately used in these sentiments.

Cars, self-driving, driving, google car, future were some of the most frequently used terms which make sense as google was the first one to come into the market to come up with a self-driving car and its autonomous car had created quite a buzz online. The terms ‘aa’, ‘aaa’, ‘i’, ‘u’, ‘a’, ‘rt’ were kept by the text parsing node which later have been eliminated using text filter node.

Terms

Term	Role	Freq	# Docs	Keep
google ...		534	499	Y
ââ ...		573	487	Y
driverless ...		441	427	Y
ì ...	Miscellaneous Proper Noun	518	379	Y
+ driverless car ...	Noun Group	382	375	Y
âââ ...		362	353	Y
+ self-driving car ...	Noun Group	248	246	Y
google car ...	Miscellaneous Proper Noun	239	237	Y
+ google car ...	Noun Group	231	228	Y
ã ...		257	212	Y
+ drive ...		201	198	Y
future ...		149	146	Y
+ want ...		136	135	Y
û ...	Miscellaneous Proper Noun	153	130	Y

Figure 2: Text Parsing Output

Text filtering

The text filter node is further added to the text parsing node as it provides the functionality to eliminate the least frequent and irrelevant terms by using the interactive filter option in the properties panel. To correct the misspelled words the spell check option is enabled in the text filter properties panel as shown below, ‘provlem’ to ‘problem’, ‘bwest’ to ‘best’, ‘automous’ to ‘autonomous’ and so on as shown in figure 3. Manually grouping of terms with same meanings is done using the interactive filter. The terms car, automobile, van, vehicle, etc. are grouped together and represented as term ‘car’ as shown in figure 4.

EMWS1.TextFilter_spellDS

Parent # Docs	Term	# Docs	Parent
6.0	changeist	1.0	changes
24.0	provlem	1.0	problem
4.0	compite	1.0	compete
24.0	transportnation	1.0	transportation
15.0	bwest	1.0	best
79.0	automous	1.0	autonomous
10.0	crzy	1.0	crazy

Figure 3: Spell Check in Text Filter

	TERM	FREQ	# DOCS	KEEP	WEIGHT
⊟	car	3748	3371	✔	0.015
	automobile	5	4		
	van	4	4		
	automobiles	4	4		
	vehicles	65	63		
	cars	1588	1489		
	vans	2	2		

Figure 4: Synonym grouping

Concept links

Concept links provides an overview of association of the term at the center with the other terms in the document. The strength of association between the linked terms is shown by the width of the link. In the concept link below, the term 'hit' is associated with human, road, pedestrian, accident, etc. On further exploring the term 'accident', it is discovered that the terms like car crash and death are also associated with the parent term. All these terms are closely associated with driving hazards caused during car driving.

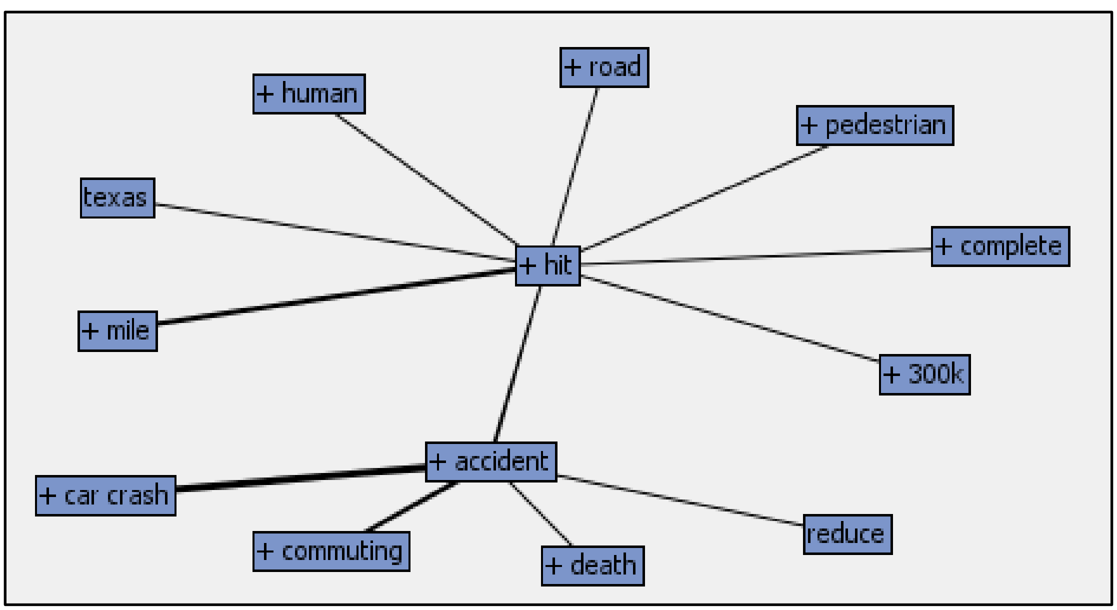


Figure 5: Concept link for 'hit'

Below you can see a concept link showing higher strength of association between terms 'google', 'steer wheel' and 'wheel'. This echoes the push google had made to get rid of the steering wheel of its self-driving car back in 2014. However, its test model had a steering wheel due to rules set by California DMV which requires a steering wheel on all test vehicles so that the driver can take over in case of a failure.

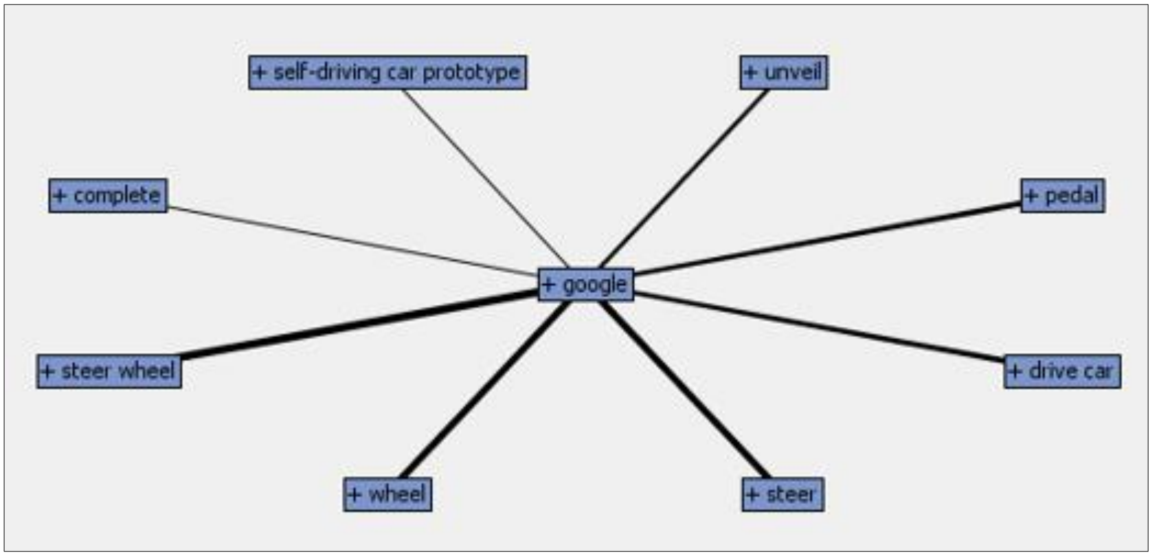


Figure 6: Concept link for 'google'

The concept link below shows stronger association between words 'ride' and 'drinking'. There are a lot of sentiments being shared about self-driving cars solving the problem for drinking and driving, which has been a major problem in the United States. According to a survey by the National Traffic and Highway safety administration every day in America, another 28 people die as a result of drunk driving crashes.

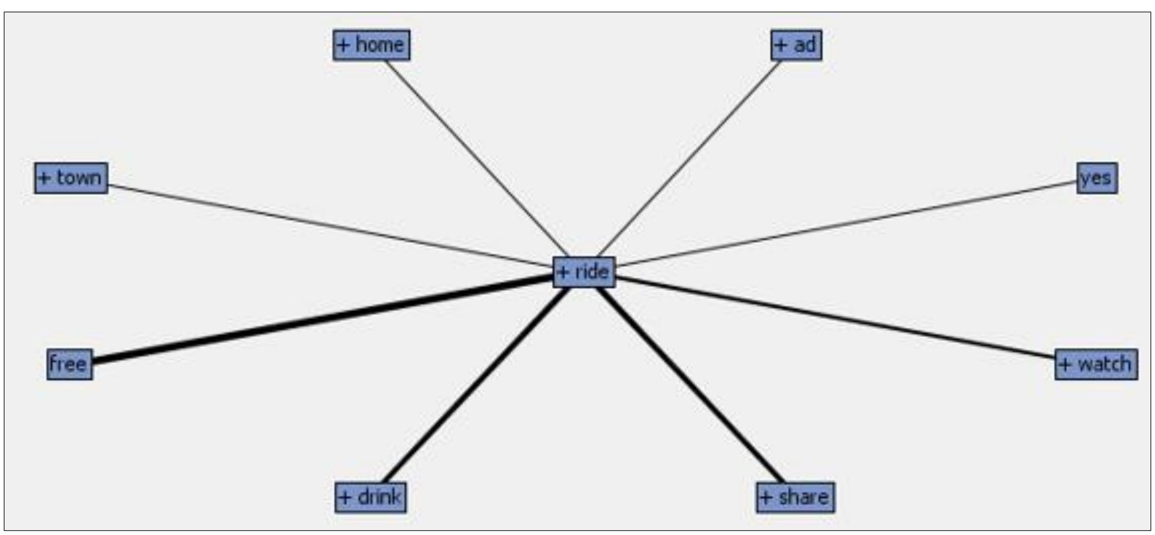


Figure 7: Concept link for 'ride'

Below concept link shows stronger association between words ‘accident’, ‘car crash’ and ‘300k’. As we all know that after an accident bodily injury to others split liability limits on our personal and commercial automobile insurance policies that reads $100,000 per person/$300,000 per accident. A lot of sentiments have been shared on whether the self-driving cars will change the way the insurance industry works and is the $300,000 limit on insurances after accidents going to change?

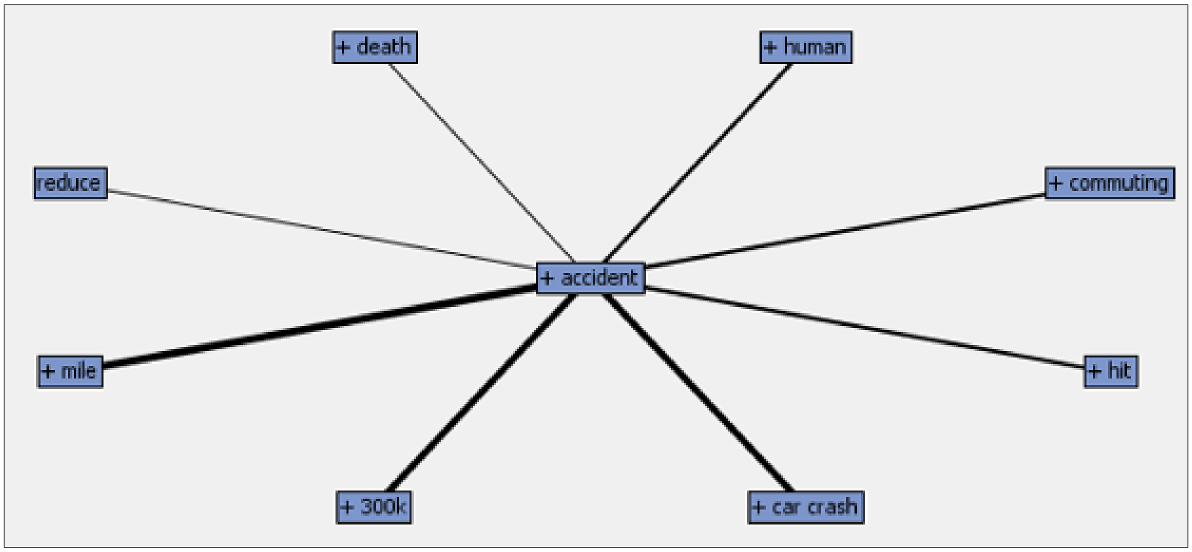


Figure 8: Concept link for ‘accident’

The concept link below shows stronger association between words ‘commute’, ‘morning’ and ‘hurry’. A large number of times everyone loved to speed up their commute to work in the morning and reach just in time before that meeting. A lot of sentiments have been shared which suggest that self-driving cars will help fasten the morning office commute.

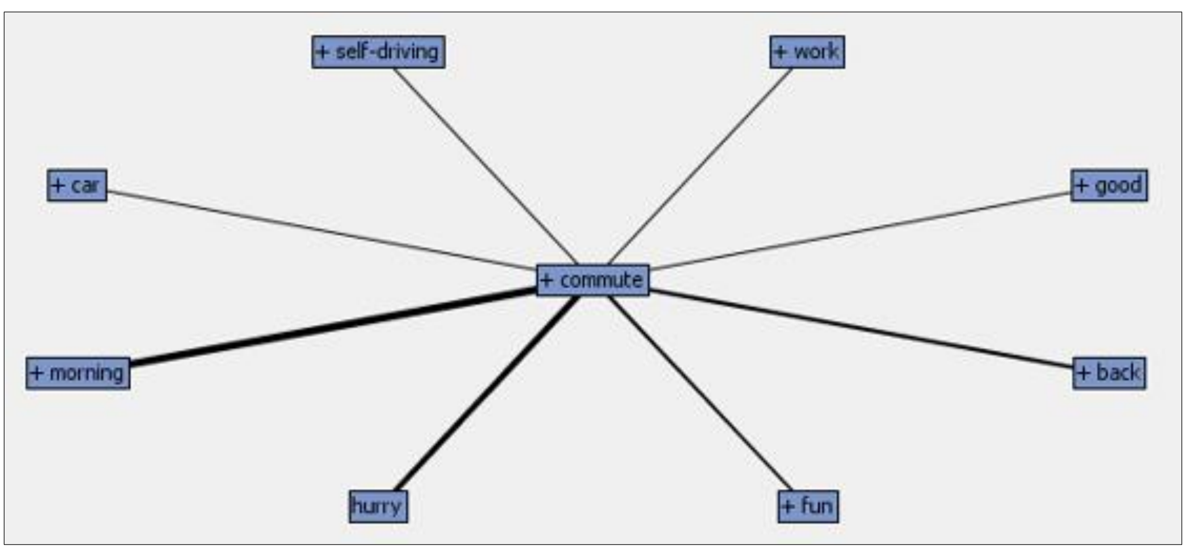


Figure 9: Concept link for ‘commute’

Text clustering

After filtering the data by using the spell check and synonym grouping options, Text Cluster node is used for the grouping of terms belonging to a certain topic. Using the Expectation-Maximization Cluster Algorithm, 7 clusters are obtained having well distributed frequencies except for cluster 4. The below figure shows the 7 clusters formed are well separated in 2 dimension space satisfactorily.

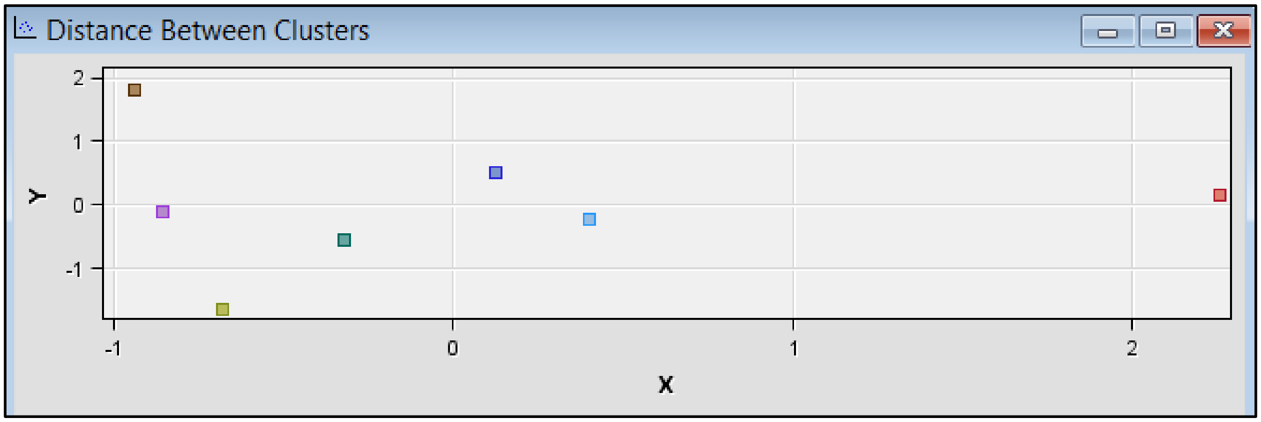


Figure 10: Distance between Clusters

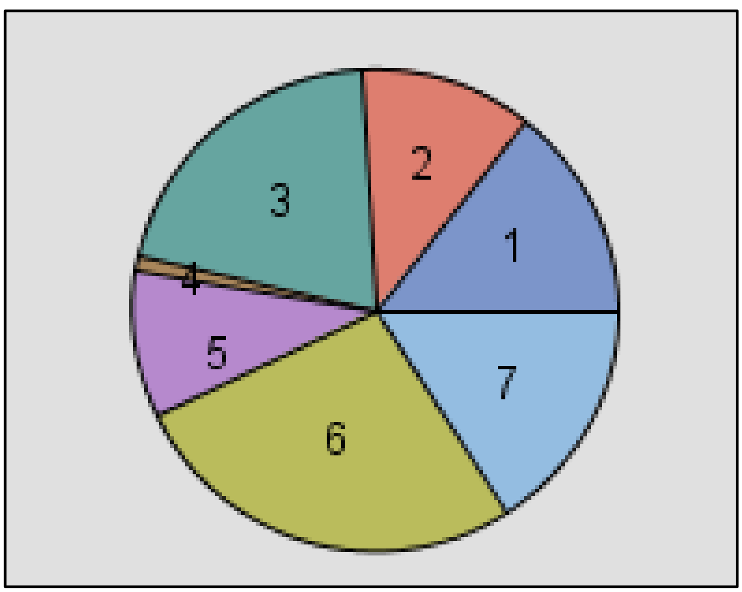


Figure 11: Distribution of Cluster Frequencies

Above pie chart shows the cluster frequencies. Cluster table describes different clusters as shown in figure below. Cluster 7 shows the excitement in people while Cluster 6 talks about technology changes and Cluster 4 describes the happiness of work done on this innovation.

Cluster ID	Descriptive Terms	Frequency	Percentage
1	+drive +want +wait +wheel +cool +'steer wheel' +steer ...	549	14%
2	+driving +'drive car' +spot +google +car +future +uber ...	459	12%
3	+driver +day +great +hope +look +self-driving +road ...	791	20%
4	+love +big +innovation +fun +live +design +work ...	41	1%
5	+accident awesome +life +idea +mile +human traffic ...	377	10%
6	+car +'self-driving car' +self-driving +technology +verge +google +test ...	1056	27%
7	+people +good +robot +amaze +excite +driving +'drive car' ...	610	16%

Figure 12: Descriptive terms in Clusters

RULE BASED MODEL

After filtering the data we have added a Text Rule Builder node and utilized different settings in properties panel. The generalization error, exhaustiveness and purity of rules are set to low, medium and high in three separate nodes.

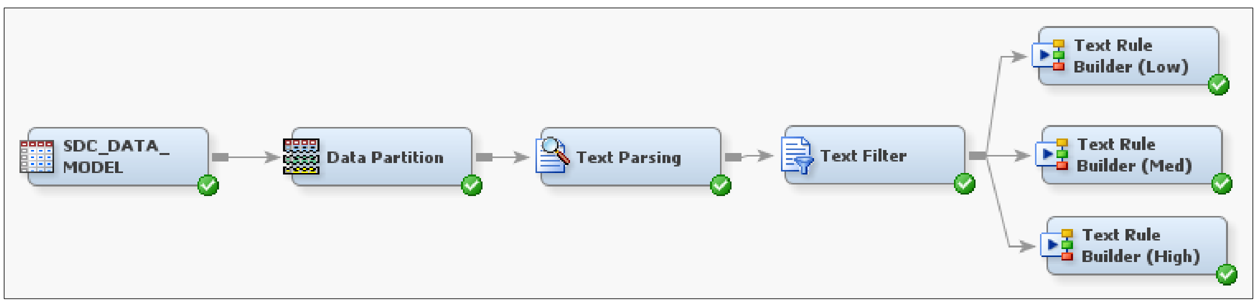


Figure 13: Rule Based Methodology

Text Rule Builder node with settings as Medium was the better than other two based on the lowest misclassification rate. For the validation data, the misclassification rate was 32.47%.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
sentiment		_ASE_	Average Squar...	0.053057	0.049024
sentiment		_DIV_	Divisor for ASE	11649	5007
sentiment		_MAX_	Maximum Abs...	0.721158	0.714892
sentiment		_NOBS_	Sum of Freque...	3883	1669
sentiment		_RASE_	Root Average ...	0.230342	0.221414
sentiment		_SSE_	Sum of Square...	618.0638	245.4632
sentiment		_DISF_	Frequency of C...	3883	1669
sentiment		_MISC_	Misclassificati...	0.294875	0.324745
sentiment		_WRONG_	Number of Wro...	1145	542

Figure 14: Fit Statistics for Rule Based Model

SENTIMENT ANALYSIS

SAS® Sentiment Analysis Studio gives a quick overview of classification of the opinions into positive and negative. Keeping 20% of the data aside, a statistical model is built which uses 80% of the remaining data for training and 20% for validation purpose. With 70.44% overall precision, Smoothed Relative Frequency and No Feature Ranking model is chosen as the best model. The size of the document and words per document varies from one document to other. Using the text normalization method, the length of the document is kept consistent and this is achieved using Smoothed Relative Frequency algorithm.

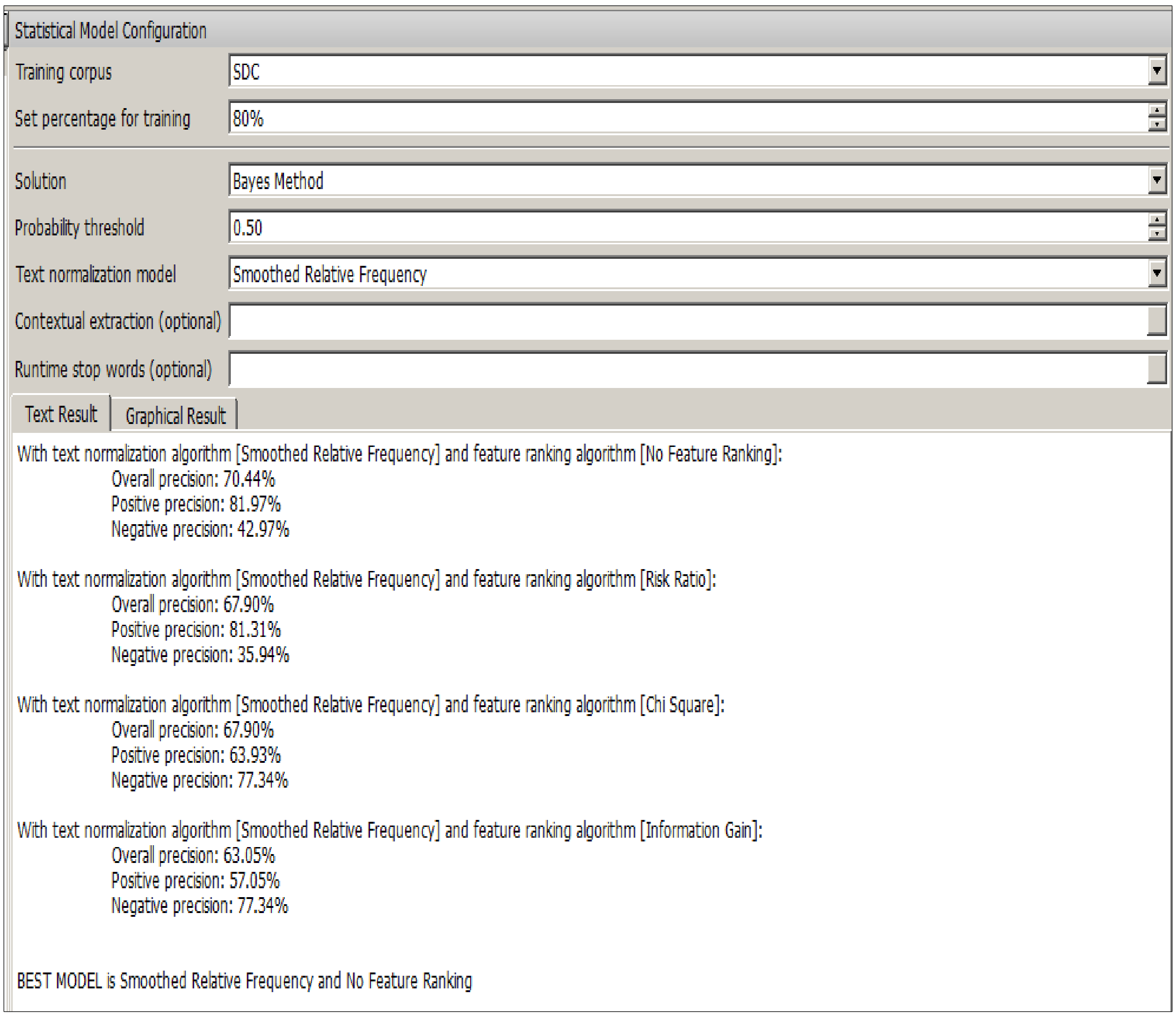


Figure 15: Text results of Statistical Model

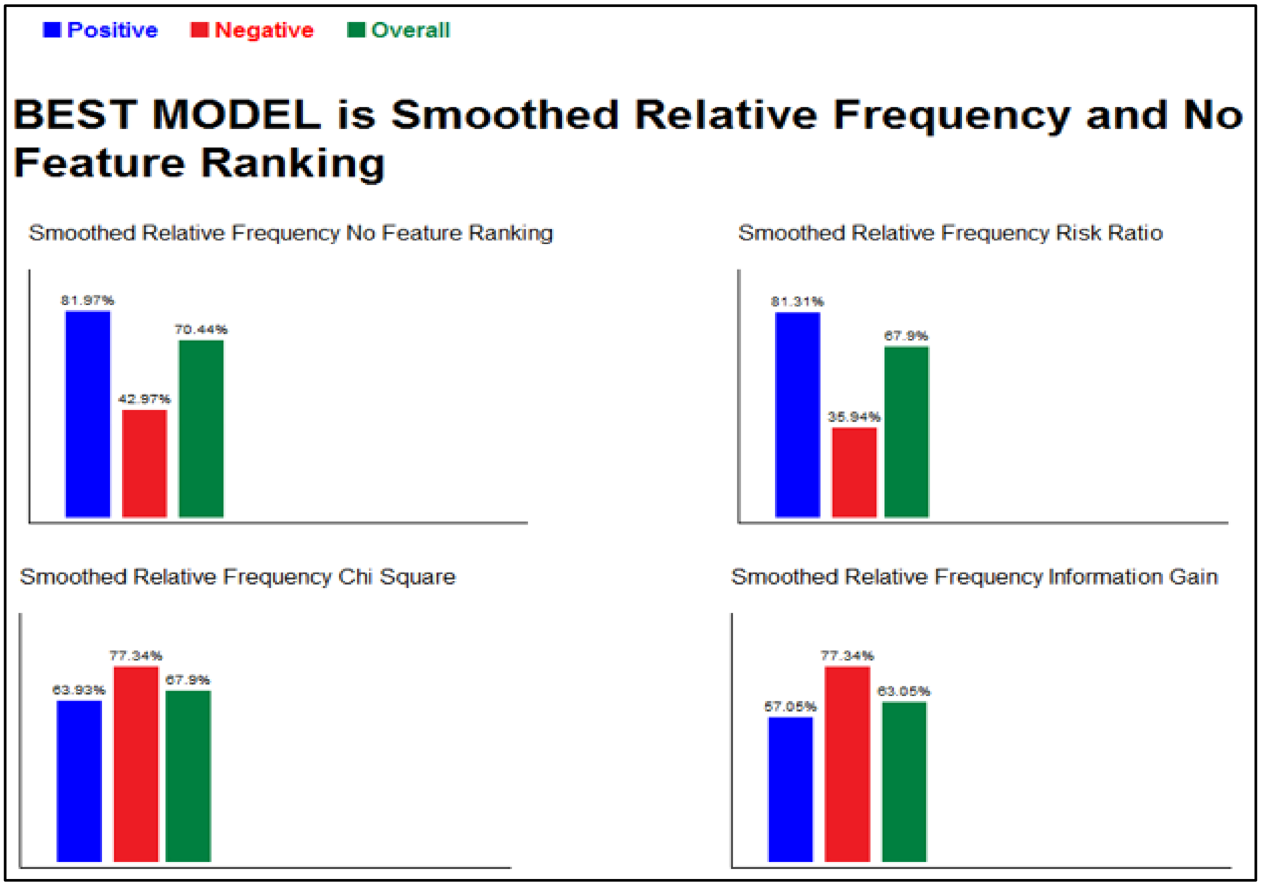
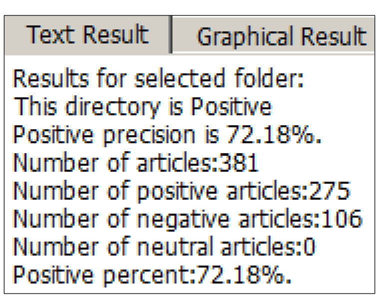


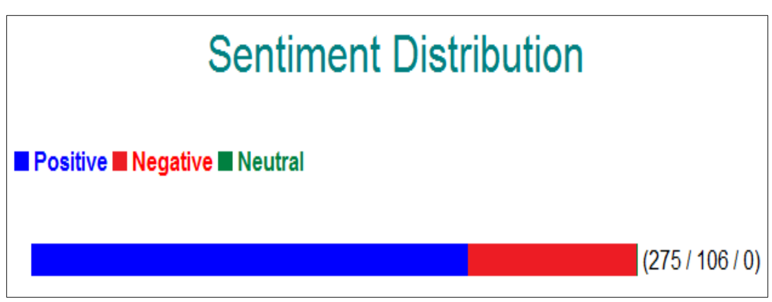
Figure 16: Graphical results of Statistical Model

Model Testing

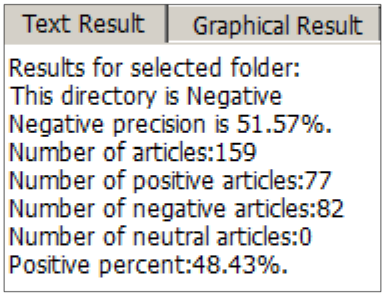
The 20% of the data for testing purpose which produces the below results for positive and negative opinions respectively.

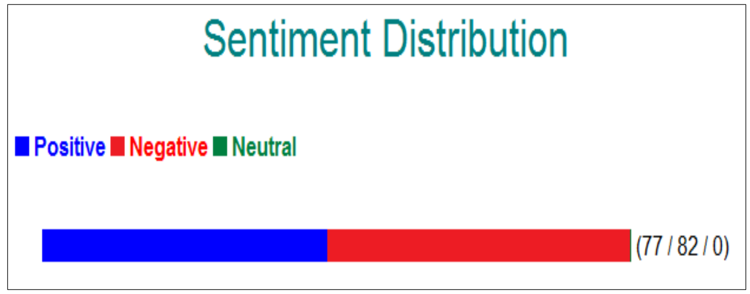
Testing Positive Reviews





The model correctly identifies the directory as Positive with 72.18% positive precision.





The model correctly identifies the directory as Negative but has a lower negative precision as compared to the positive directory.

CONCLUSION

Using SAS® Sentiment Analysis Studio, the reviews of any text online can be quickly classified into a positive or negative sentiment. A quick summary can be generated which reflects the sentiments of the person writing this opinion. Such analysis can be extremely helpful to the audience that depends on others opinions before they make any purchase, especially in case of newer technologies like self-driving cars as they are a considerable investment.

Online Opinions give insights into the people's expectations from this newer car technology. This information can be leveraged by the auto makers to include different functionalities in their products and shape their marketing campaigns to cater to the needs and expectations of their customers. Depending on how often they are utilized together, a relationship can be defined in-between terms using concept links. For example the term like "accident", "car crash" "Commuting" are strongly associated with "car". This indicates the fear about safety of these newer and technologically advanced self-driving cars.

REFERENCES

- Guizzo, Erico. "How google's self-driving car works." *IEEE Spectrum Online, October* 18 (2011).
- Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS® by Goutam Chakraborty, Murali Pagolu, Satish Garla
- SAS® Institute Inc. 2014, Getting Started with SAS® Text Miner 13.2. Cary, NC: SAS® Institute Inc.

ACKNOWLEDGEMENT

We wish to express our sincere thanks to Dr. Goutam Chakraborty and Dr. Miriam McGaugh for their valuable guidance.

CONTACT INFORMATION

Nachiket Kawitkar
Oklahoma State University
Phone: 405-762-3719
Email: nachiket.kawitkar@okstate.edu

Nachiket Kawitkar is a Master of Science student in Business Analytics at Spears School of Business, Oklahoma State University. He has been working as a Risk Analyst Intern on financial analytics projects with Elevate Credit Services, Fort Worth since June 2016. He is a SAS® Certified Base Programmer, SAS® Certified Statistical Business Analyst, SAS® Certified Advance Programmer and also a SAS® Certified Predictive Modeler. He has presented a poster at the Analytics Experience, 2016 at Las Vegas and has a paper presentation at the SCSUG conference, 2016 at San Antonio along with his co-author.

Swapneel Deshpande
Oklahoma State University
Phone: 405-714-1241
Email: swapned@okstate.edu

Swapneel Deshpande has a Masters in Environmental Engineering and is currently a Master of Science student in Business Analytics at Spears School of Business, Oklahoma State University. He has worked as an Analytics Intern on Health and Safety Project's at ISN Software Corporation, Dallas Texas and is currently working as a Graduate Associate – Tech and Analytics at Oklahoma State University. He has presented a poster at the 2016 Western Users of SAS® Software Conference, San Francisco followed by Analytics Experience, 2016 at Las Vegas and has a paper presentation at the SCSUG conference, 2016 at San Antonio along with his co-author.

Dr. Goutam Chakraborty
Oklahoma State University
Email: goutam.chakraborty@okstate.edu

Dr. Goutam Chakraborty is Ralph A. and Peggy A. Brenneman Professor of Marketing and Director of Master of Science in Business Analytics program at Oklahoma State University. He is also founder of SAS® and OSU data mining certificate and SAS® and OSU marketing analytics certificate at Oklahoma State University. He has published many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has over 25 Years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.

Dr. Miriam McGaugh
Oklahoma State University
Email: miriam.mcgaugh@okstate.edu

Dr. Miriam McGaugh is Clinical Professor in Business Analytics program at Oklahoma State University. She was a Community Health Epidemiologist at Oklahoma State Department of Health for almost 15 years. During that period she used to be a Lecturer at Spears School of Business, Oklahoma State

University and played a big role in teaching and encouraging students to achieve honors like SAS® Certified Base Programmer and SAS® Certified Advance Programmer.