Paper 500-2013

Opinion Mining and Geo-Positioning of Textual Feedback from Professional Drivers

Mantosh Kumar Sarkar, Goutam Chakraborty

Oklahoma State University, OK, USA

ABSTRACT

The widespread adoption of mobile applications has tremendously expanded the scope of obtaining timely customer feedback. While many companies collect feedback from their customers via mobile apps, often they restrict their analysis to numeric data and ignore analyzing customer feedbacks and sentiments from textual data because of perceived difficulties associated with analyzing text data. In this paper, we analyze customer feedbacks by professional drivers sent via a mobile app that the drivers use to locate a store, check their reward points etc. The drivers are customers of a retail & energy company which offers a variety of services such as gas stations and convenience stores and also offer amenities such as food from national restaurant chains, trucking supplies, showers and RV dump stations. The company experts currently manually classify these textual feedbacks into positive and negative groups. We demonstrate how SAS® Text Miner can be used to automatically generate and summarize topics from positive and negative feedbacks. In addition, we demonstrate how SAS® Sentiment Analysis studio can be used to build rules to predict customers' sentiments automatically so that expert's time can be used for more strategic purposes. Finally, we also show how feedbacks with positive and negative sentiments can be geo-positioned on the US map via JMP® scripts for providing a better visualization of sentiment distribution.

OVERVIEW

Life as a professional driver is often very challenging. These professionals stay away from their homes for extended periods of time and that can be very tough on them and their families. Truck stops, where professional drivers can relax, get supplies and charge up, act almost as second homes to these drivers. Therefore, it becomes very important for truck stop companies to maintain proper standards of services like showers, food and fuel so that the drivers are satisfied with the amenities at the truck stops. Truck stop companies therefore like to routinely track and analyze professional drivers' feedbacks to address their concerns. However, truck drivers are a tough group of customers to get feedback from because they do not like to fill-in surveys via mail or answer questions by phone or email. In this research we discuss a truck stop company that allowed truck drivers to send in free-from comments via mobile apps and how analyses of such free-form comments with the help of SAS® Text miner and SAS® Sentiment Analysis Studio provide deeper insights into what these drivers are thinking about.

DATA AND COMPANY OVERVIEW

A leading retail & energy company which has more than 280 truck stops located in more than 39 US states has kindly agreed to provide the data for this paper. The company wishes to remain anonymous. It offers fuel, fast food and convenience store services to its customers. In addition it offers additional amenities such as food from national restaurant chains like Subway, Arby's and Carl's Jr, as well as trucking supplies, showers, and RV dump stations. The company has won several industry specific excellence awards for its outstanding services.

The feedbacks considered for this paper were obtained via a mobile app which is used by professional drivers to locate a store, find nearby fuel stations and check loyalty program reward point balances. These feedbacks are free-form comments sent by professional drivers who can mention anything they feel like at the point when they access the app. Presently, company experts are classifying these feedbacks into positive and negative by manually reading and analyzing them.

METHOD

USING SAS® TEXT MINER

To get deeper insights into the topics the professional drivers are talking about, SAS® Text Miner is employed. Fig.1 depicts the process workflow of various nodes that are used. After data cleaning and validation 2,335 feedbacks were usable for this research. We decided to keep the data and analysis separate for positive and negative comments.

SAS® Text Miner can easily handle different data formats such as textual files, web pages etc. for analysis. Data collection step is followed by parsing of the data [3]. Text parsing is used to create a dictionary of all the words which appeared in the feedbacks. Text parsing also performs sentence identification, parts of speech and determines stemming words. This step is considered to be the most important step in text mining process. It is used to quantify the terms used in the dataset followed by text filter node. Text filter node reduces the number of parsed terms to most

valuable and relevant terms. Standard English dictionary is used to spell check the words. An interesting output of filter node is concept links. This helps in identifying relationship between terms. Fig. 5 shows concept link diagram of word “shower” obtained from negative feedbacks. Thickness of the links between words shows strength of association between terms. Here the term “Showers” is highly associated with terms key, water, toilet, shower head, clean, sink, towel and hair. Digging deeper in the actual comments related to these terms, we find that customers are complaining that showers are not clean, water is cold, sink is dirty, key to shower rooms does not work properly, there are not of enough towels available and hairs are all over sink.

The output from text filter node is then fed to the Text Topic node, which generates 25 multi-term topics by default. Given the data size, 25 topics are deemed too many. By trial-and-error, only 13 multi-term topics were retained for negative dataset and only 3 for positive dataset by changing “Number of Multi-term topics” in the property panel (as shown in fig.3). So, it seems in this data set the number of topics for complaints far outnumber the topics for praise which is consistent with prior consumer research findings that people are more likely to complain than compliment.

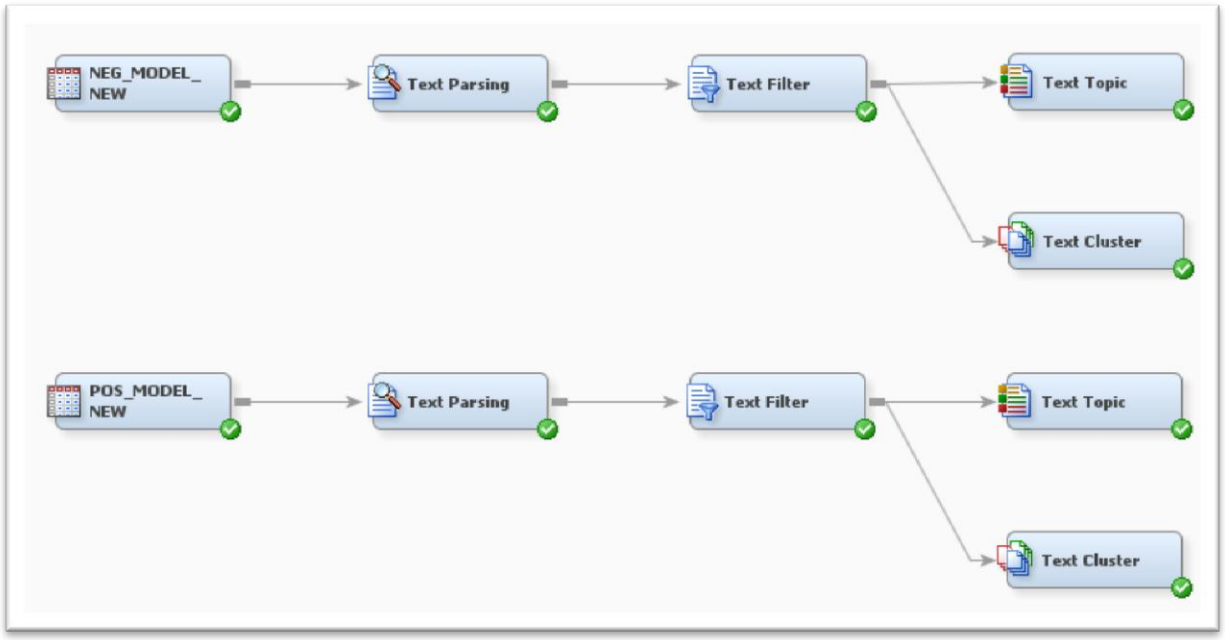


Fig.1. Text mining process of feedbacks

SAS® Text Miner is also used for clustering of positive and negative feedbacks separately. Clustering resulted in 12 clusters of fair sizes for negative comments. Results revealed top factors that customer complained about were dirty shower rooms, trash all over the parking lot, rude employees, etc. These results are similar to what were obtained via topic mining. For positive comments, clustering resulted in 3 groups and their interpretations are similar to the topics found for the positive comments.

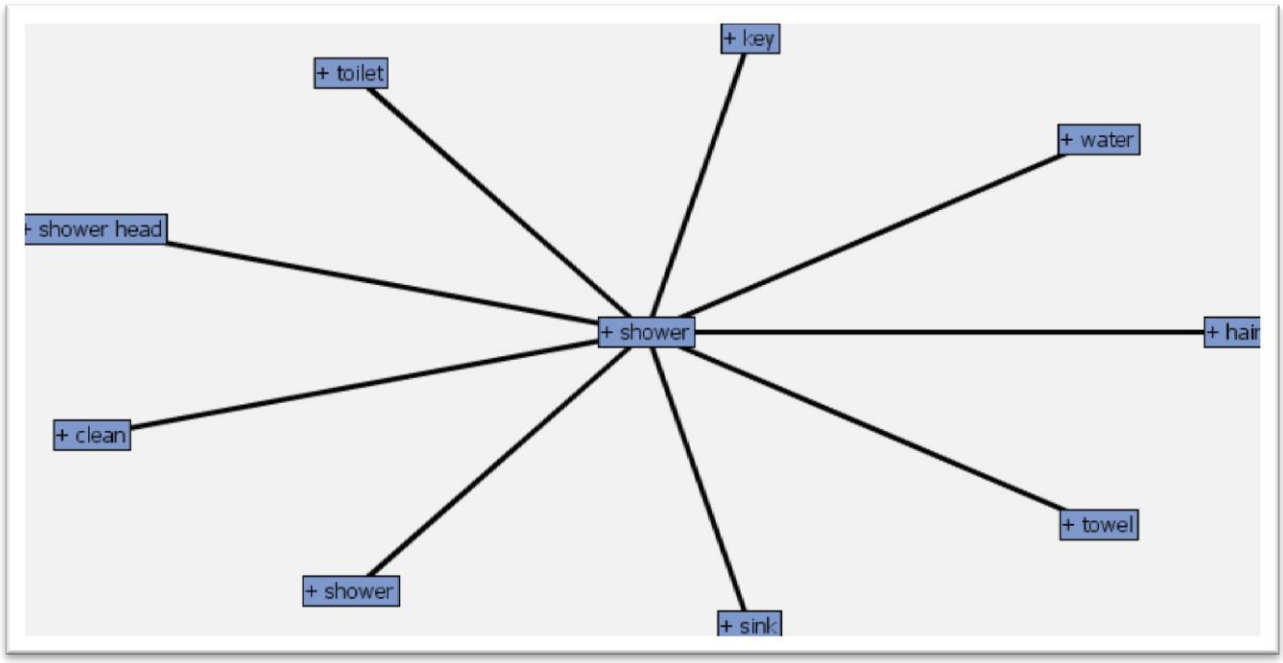


Fig.2. Concept link diagram for term “shower”

Text-topic node(-ve comments)

General	
Node ID	TextTopic
Imported Data	
Exported Data	
Notes	
Train	
Variables	
User Topics	
Term Topics	
Number of Single-term To	0
Learned Topics	
Number of Multi-term Top	13
Correlated Topics	No
Results	
Topic Viewer	
Status	
Create Time	10/18/12 12:36 PM
Run ID	ea03ad8c-5eb9-4f2c-b7be
Last Error	
Last Status	Complete
Last Run Time	11/8/12 8:32 PM
Run Duration	0 Hr. 0 Min. 4.45 Sec.
Grid Host	
User-Added Node	No

Text-topic node(+ve comments)

General	
Node ID	TextTopic2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
User Topics	
Term Topics	
Number of Single-term To	0
Learned Topics	
Number of Multi-term Top	3
Correlated Topics	No
Results	
Topic Viewer	
Status	
Create Time	10/31/12 3:14 PM
Run ID	891dd8f5-1e0b-4dd6-b470
Last Error	
Last Status	Complete
Last Run Time	11/8/12 8:40 PM
Run Duration	0 Hr. 0 Min. 3.99 Sec.
Grid Host	
User-Added Node	No

Fig.3 Properties panel of Text-topic node for positive feedbacks and negative feedbacks

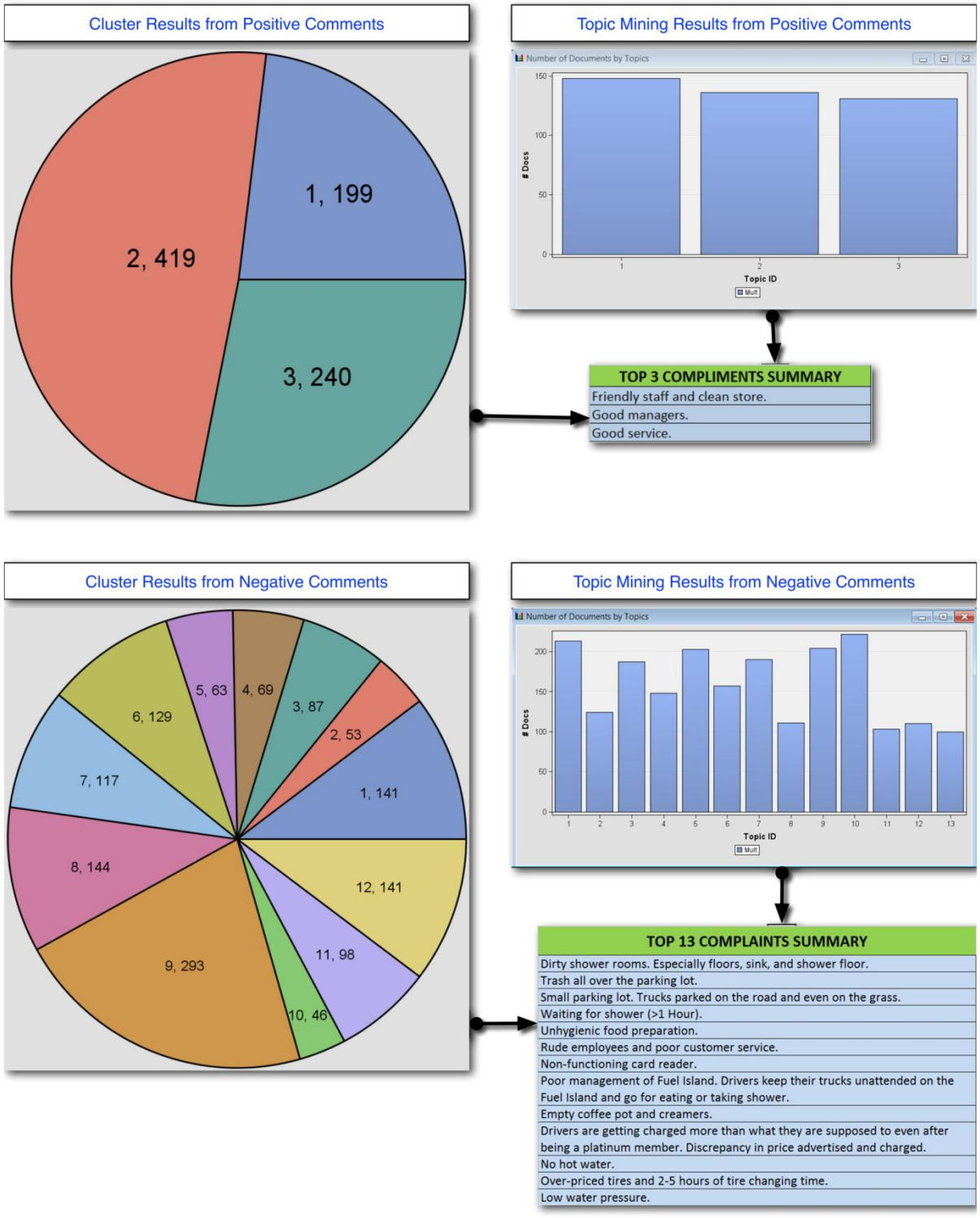
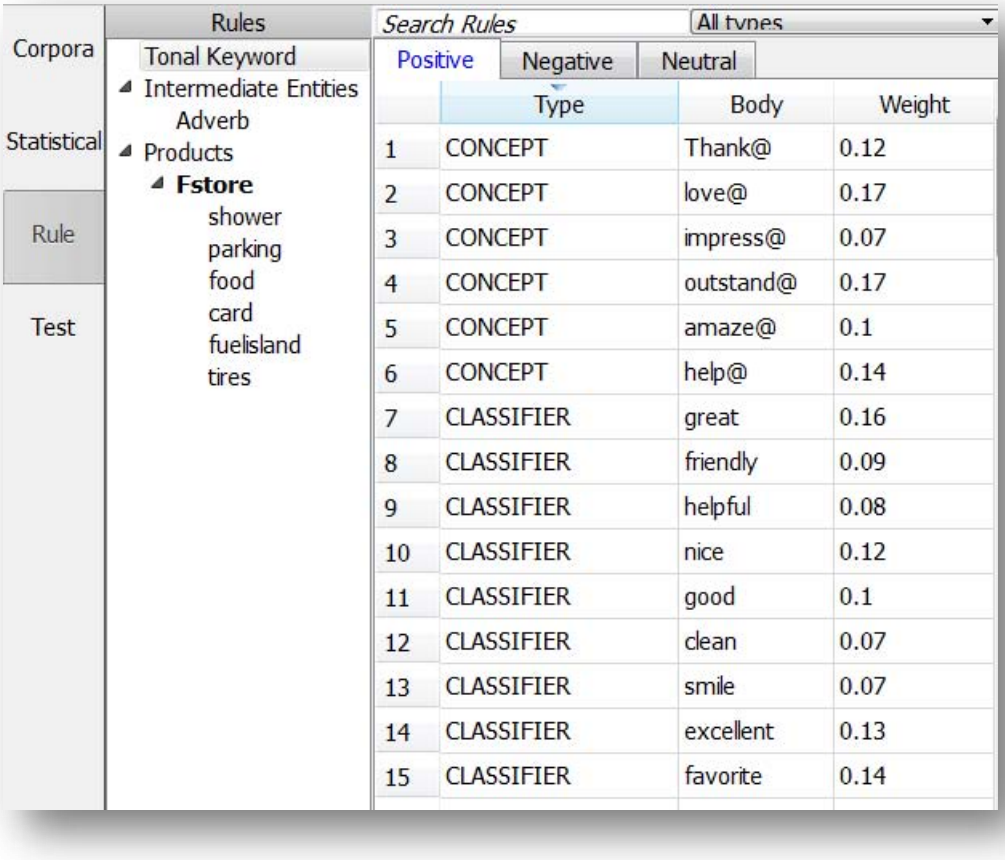


Fig.4. Summary of results from SAS® Text Miner

RULE BASED SAS® SENTIMENT ANALYSIS STUDIO MODEL

SAS® Sentiment Analysis studio is employed for classifying the feedbacks into positive and negative sentiments and categorizing them into features. SAS® sentiment analysis studio has three different types of models: statistical model, rule based model or a hybrid model. Hybrid model is a combination of statistical model and rule based model. In practice statistical model is very important to get started with the process of sentiment mining and provides a baseline model that can be set up quickly. Using it as a starting point, a rule based model can then be built where each lexicon rule can be modeled and analyzed over and over again. A successful rule based model often can explain results very intuitively for managers who do not like to read detailed statistical analysis. Rule based model uses natural language processing for analyzing each text feedbacks and calculates sentiment weight of each positive sentiment as well as negative sentiment.



	Type	Body	Weight
1	CONCEPT	Thank@	0.12
2	CONCEPT	love@	0.17
3	CONCEPT	impress@	0.07
4	CONCEPT	outstand@	0.17
5	CONCEPT	amaze@	0.1
6	CONCEPT	help@	0.14
7	CLASSIFIER	great	0.16
8	CLASSIFIER	friendly	0.09
9	CLASSIFIER	helpful	0.08
10	CLASSIFIER	nice	0.12
11	CLASSIFIER	good	0.1
12	CLASSIFIER	clean	0.07
13	CLASSIFIER	smile	0.07
14	CLASSIFIER	excellent	0.13
15	CLASSIFIER	favorite	0.14

Fig. 5. Rule based model showing few positive rules

BUILDING STATISTICAL MODEL

Statistical model requires basically two different kinds of feedbacks: positive and negative. For this paper, 10% of data was held back for testing. The balance 90% of the data was used for modeling. In the modeling data, we used 80% of data for training and remaining 20% of data is used for validation purpose. Smoothed Relevancy Frequency text normalization algorithm is used to build this model [1].

The statistical model built classified 136 feedbacks correctly as negative and remaining 5 feedbacks as positive in the test data (as shown in fig 6). On the surface it is an excellent result! But when we extracted the rules from the statistical model via import learned features, most of the rules were difficult to understand. For example, numbers were classified as negative sentiment rules and weight of term “bad” was same as the weight of the term “worst”. The list was also very long. Same situation was experienced during the testing of positive directory with pre-classified positive feedbacks by company experts.

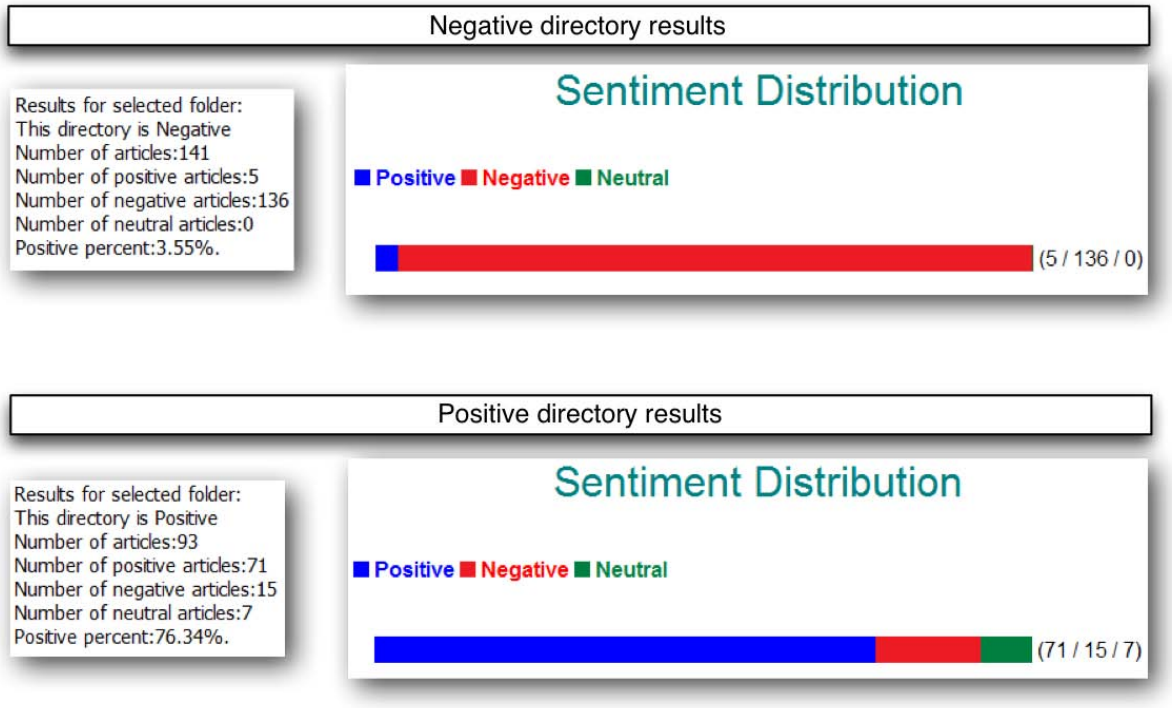


Fig.6 Testing results from statistical sentiment model

As shown in fig.6, statistical model correctly classified 71 articles out of 93 articles as positive.

WRITING RULES FOR RULE BASED MODEL

In order to understand what words were classified into positive and negative and add features, we chose rule based sentiment analysis model. Initially 200 negative and 200 positive rules were imported as learned features from the built statistical model. All the rules which did not make any sense or very vague were removed. Some of these excluded rules consist of those including numbers and words having neutral sentiment.

In this paper, we use CLASSIFIER rules, CONCEPT rules and PREDICATE_RULES rules. CLASSIFIER rules are used to match a term or a phrase [1]. CLASSIFIER rules contained all words or strings that appeared in documents which were used to build the model. For example word “good” was assigned a weight of 0.10 on a scale of 0 to 0.19. All the words with their sentiment weights were included in a section called “Tonal Keyword”.

To ensure all the word forms of a word get a match, @ symbol is used with a word, which changes the type of rule from CLASSIFIER to CONCEPT rule.

Which is more positive “good” or “very good”? What if we could assign more weight to adverb + word automatically? To address these questions, a new entity called “adverb” was created. This new entity contained words such as “always”, “really”, “very” etc. A rule was added which could take benefit of these adverbs as follows:

```
PREDICATE_RULE (ORDDIST_5, "_a{_def{Adverb}}","_b{_def{TonalKeywordPositive}}") 0.27
```

According to this rule if an adverb comes before a positive keyword included in Tonal Keyword section within a gap of 5 words, then it will gain more sentiment weight. Note that weight assigned in this case is 0.27.

On performing text mining of the feedbacks, we have found the features that these professional drivers were talking about. Features that were included are as follows.

I. Shower
II. Parking
III. Food
IV. Card(Loyalty card)
V. Fuel island
VI. Tires

Since we already mentioned all the positive and negative keywords in Tonal Keywords section, a simple CONCEPT rule was used to classify feedbacks under these features. For example, to classify positive feedbacks for the feature shower following rule was used.

```
CONCEPT_RULE  (SENT,"_c{_def{TonalKeywordPositive}}","_def{Fstoreshower}")       1
```

The above rule means that if the word “shower” or equivalent to shower appears in a comment with positive words as defined under positive section of Tonal Keyword, then it will be considered a match.

PERFORMANCE OF RULE BASED MODEL

On applying these rules on the negative test directory containing 141 feedbacks, 112 feedbacks were classified as negative,19 feedbacks as positive and 10 as neutral which is very good. As we carefully read all the comments in negative directory, we found that experts actually misclassified some of the mixed comments which the rule based model was pointing out as neutral! We found that most number of complaints were regarding shower followed by food, parking, fuel island, tires and card. Fig.7 depicts the exact number of feedbacks that were classified under each of the features.

Rule based model did a great job in classifying positive feedbacks. In the test directory containing 93 feedbacks, 81 feedbacks were classified correctly as positive which resulted in accuracy of 87.10 %(as shown in fig.8).

Food received highest numbers of positive feedback followed by showers, parking, tires and card.

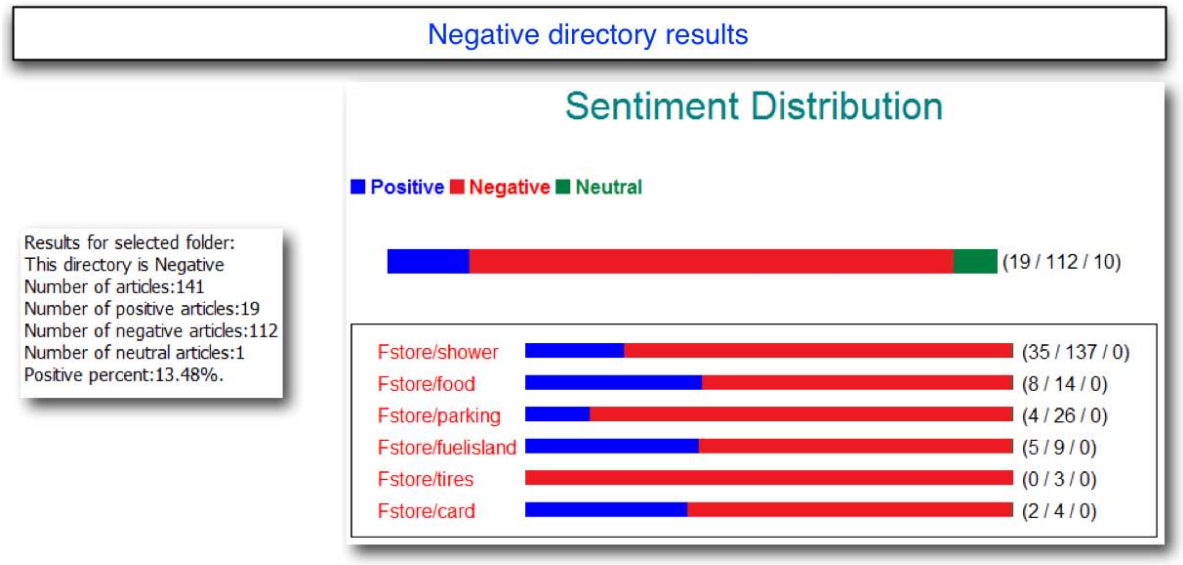


Fig.7. Testing results of negative feedbacks from Rule-based sentiment model

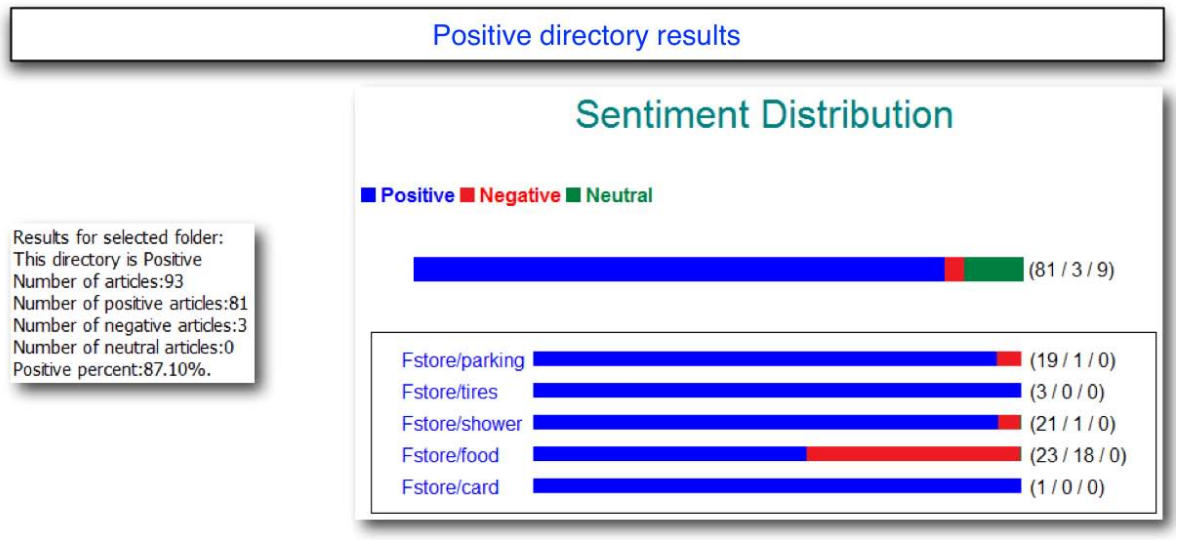


Fig.8. Testing results of positive feedbacks from rule based sentiment model

GEO-POSITIONING OF SENTIMENTS

Wouldn't it be great if we could map and show how these positive and negative sentiments are distributed across the continental United States?

To answer this question, JMP® scripting language has been used to plot these feedbacks on the US state map and results were very appealing to the managers of the retail and energy company. We identified top highways and store locations where most number of complaints or compliments were generated. To keep the anonymity of truck stop location, actual store location information has been omitted. The code to generate this geo-location map is shown in the appendix. All the highways have been color coded using default color code of JMP® Pro 10, so as to distinguish them on the US map.

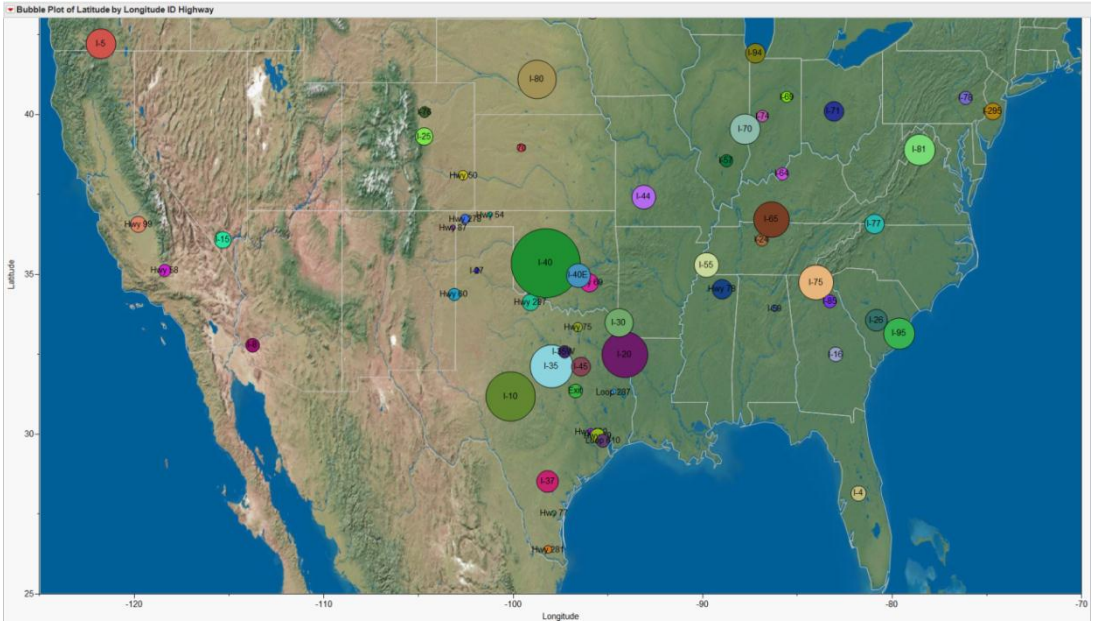


Fig.9. Geo location of negative feedbacks on US state map

As shown in fig.9, fig.10, the highway which received the most number of complaints as well as compliments was I-40. Perhaps this reflects the most commonly traveled routes by the professional drivers who are customers of this company.

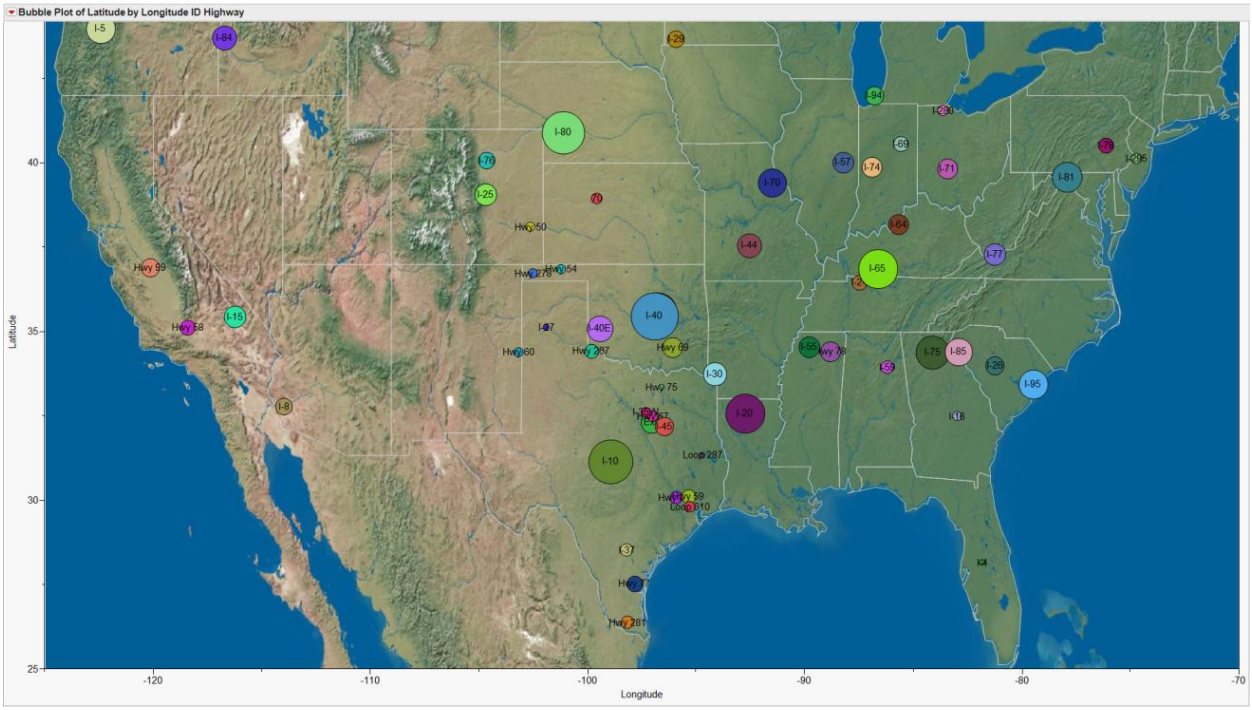


Fig.10. Geo location of positive feedbacks on US state map

CONCLUSION

Analyzing reviews from customers can provide insightful information which in turn helps companies to improve their quality of service and help in differentiating their services from their competitors.

Prompt response is the key to win customers loyalty. This can only be done when there is an automatic system in place which can immediately classify customer's feedbacks into positive and negative and provide appropriate response. Once the rules have been developed, it will be easy to set up an automatic scoring system for future comments that can be classified in real time and appropriate management interventions can happen. The geo-positioning customer's feedbacks and finding out the store locations where these feedbacks are generated will also help the company to narrow down a particular location and resolve the issues quickly. This will surely help the company stay ahead of its competitors.

REFERENCE

[1] SAS® Institute Inc. 2011. *SAS® Sentiment Analysis Studio 1.3: User's Guide*. Cary, NC: SAS® Institute Inc.
[2] SAS® Institute Inc. 2012. JMP®® 10 Scripting Guide. Cary, NC: SAS® Institute Inc.
[3] "Introduction to Text Miner." In "SAS® Enterprise Miner Help." SAS® Enterprise Miner 6.2 . SAS® Institute Inc., Cary, NC

CONTACT INFORMATION

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

Mantosh Kumar Sarkar, Oklahoma State University, Stillwater OK, Email: Mantosh.sarkar@okstate.edu

Mantosh Kumar Sarkar is Master's student in Management Information Systems at Oklahoma State University. He is a BASE SAS® 9 and a certified SAS® predictive modeler using Enterprise Miner 6. He has worked as a session coordinator for Statistics and data analysis at SAS Global Forum 2012, Orlando. In May, 2012, he received his SAS® and OSU Data Mining Certificate.

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

Dr. Goutam Chakraborty is a professor of marketing and founder of SAS® and OSU data mining certificate and SAS® and OSU business analytics certificate at Oklahoma State University. He has published in many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired M2007 data mining conference. He has over 25 years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.

APPENDIX
JMP CODE FOR GENERATING GEO-POSITIONING MAP

This code plots every feedback on US map provided it has location information. Following code assumes location information such as “Latitude”, “Longitude” and “Highway”.

```
Bubble Plot(

	X( :Longitude ),
	Y( :Latitude ),
	Coloring( :Highway ),
	ID( :Highway ),
	Bubble Size( 19.05 ),
	Title Position( 0, 0 ),
	SendToReport(
		Dispatch(
			{},
			"1",
			ScaleBox,
			{Min( -125 ), Max( -70 ), Inc( 10 ), Minor Ticks( 1 ),
			Rotated Labels( "Horizontal" )}
		),
		Dispatch(
			{},
			"2",
			ScaleBox,
			{Min( 25 ), Max( 45 ), Inc( 5 ), Minor Ticks( 1 ),
			Rotated Labels( "Horizontal" )}
		),
		Dispatch(
			{},
			"Bubble Plot",
			FrameBox,
			{Frame Size( 835, 523 ), Background Map(
				Images( "Simple Earth" ),
				Boundaries( "US States" )
			), Grid Line Order( 3 ), Reference Line Order( 4 ),
			DispatchSeg(
				ShapeSeg( 1 ),
				{Line Color( {204, 204, 204} ), Fill Color( "None" ),
				Missing shape fill( 2147483647 ), Missing value fill( -
14540253 )}
			)}
		)
	)
);
```