Paper 104-2012

Classification of Customers' Textual Responses via Application of Topic Mining

Anil Kumar Pantangi, Goutam Chakraborty, Oklahoma State University, Stillwater, OK, USA

ABSTRACT

Customer satisfaction surveys play a vital role in monitoring business performances in most industries. Typically, customer satisfaction surveys are a blend of closed-end questions (numeric responses) and open-ended questions (textual) leading to a collection of structured and unstructured data. Textual comments are generally cumbersome and time consuming to summarize and analyze. Sometimes, businesses use experts to rate unstructured data to understand the nature and the valence of customers' responses. In this paper we illustrate the application of Topic Mining, using SAS® Text Miner; to first categorize customers' responses collected via a survey of customers of a B2B (business-to-business) company. Then, we use the topics to build a predictive model to automatically classify the responses into negative versus non-negative categories. The best predictive model had a sensitivity of 93% and a misclassification rate of 12% in the validation data. In addition to generating excellent predictions, the topics identified in this analysis also revealed valuable insights about the operational performances of the B2B Company.

INTRODUCTION

Every company today is trying to acquire new customers and retain current customers. In order to retain current customers, many companies use customer satisfaction surveys as a tool to get feedback about their products and services as perceived by their current customers.

Customer surveys typically contain closed-end questions that generate structured numeric data and open-ended questions or comments that generate unstructured textual data. Survey analysis has been so far confined mostly to structured quantitative data analysis. According to Dobson (2010), many companies fail to analyze the patterns in the textual data. However, textual comments and responses provided in these surveys contain a wealth of information. When these textual comments are further explored using text mining techniques, they have the potential to increase the predictive and/or explanatory ability of models. [4]. Kleij and Musters (2003) in their research article demonstrated how text analysis of open-ended survey responses can complement preference mapping. [5].

Text mining is a complex process owing to the unstructured nature of the data and the difficulties associated with understanding and processing of language. [3]. Many researchers have published work related to applications of text mining in various domains. These applications mainly fall under the general categories of text categorization, information retrieval and measurement. In recent years text mining is also being used for discovering trends in textual data and improving the performance of predictive models [5].

In order to reap benefits from the wealth of information that may be present in textual data, the users' comments in surveys first have to be interpreted. [1]. In order to get a good understanding of the voice-of-customer, often a manager is interested in first grouping customer comments based on the valence (positive, negative, neutral) of these comments. In many companies this is typically done by an expert who reads each customer's textual response and then classifies each response into a category. Of course, rating by a single expert is subjective and may be biased. To avoid such subjectivity, companies sometimes use multiple experts. However, if multiple experts are used, there is a possibility of disagreement among experts that need to be resolved via mutual discussion. Regardless of whether a company uses a single or multiple experts, it is obvious that the whole process can be very tedious, time consuming and perhaps prone to many different errors and biases.

In this paper, we illustrate a methodology to classify customer responses automatically using topic mining. We have used topic node available in text mining tab of SAS Enterprise Miner® to extract topics present in the data. The topics are used for analyzing the text data and for knowledge discovery. Prior to the topic analysis, textual comments are parsed and filtered. These topics are further used as inputs in predictive models to classify each comment.

SURVEY DETAILS

In this study we used data collected from a survey of US customers of a B2B (business-to-business) company, who wish to remain anonymous. The company uses the survey data to track and analyze customer satisfaction with the services provided by the company.

The survey contains questions leading to both structured and unstructured data. Numerical responses include perception of company's products and service, satisfaction with the company, etc. The textual responses include customers' comments. However, for the purpose of this paper, we have considered primarily the textual comments. The data used for this analysis consists of 205 unique comments from customers. These comments describe customer's experiences on various factors related to supplier's products and services.

For the purpose of this research, each comment was rated independently by three experts into two categories:

negative versus non-negative (given the small sample size, we chose to combine positive and mixed as one category for this research). Any conflicts in the ratings among experts were resolved via mutual discussion among the experts.

METHODOLOGY

TOPIC MINING

As explained earlier, the text data used for this analysis is customers' responses in the comments section of the survey. The text topic node in the SAS Enterprise Miner 7.1 was used to extract topics from this text data. The topic node was preceded by Text Parsing and Text Filter nodes (Figure 1 below displays the process flow):

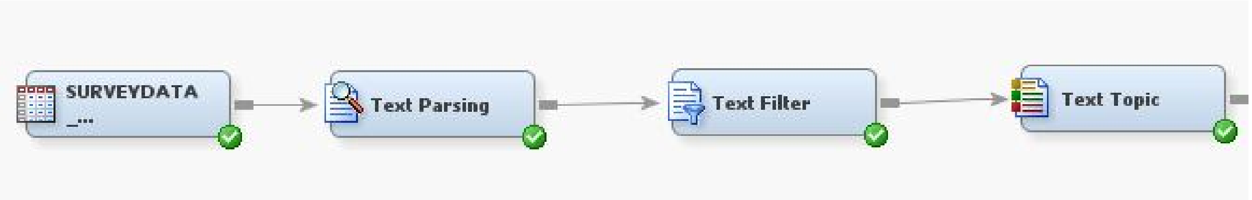
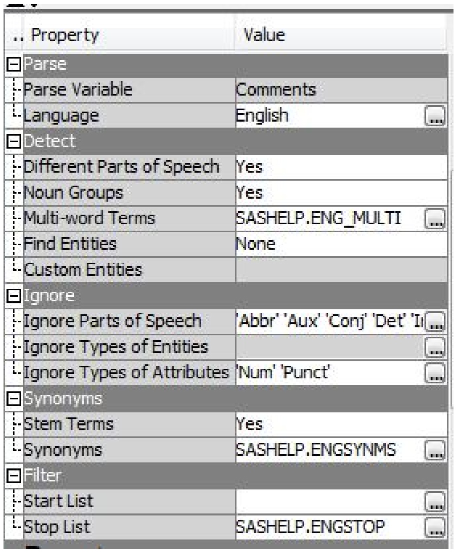


Figure 1: Process flow of Topic Analysis

The Text Topic node performs cluster analysis to group the documents and summarizes the collection by identifying "topics". This is a computer-intensive task because the node uses singular-value-decomposition (SVD) in the background to capture information from a sparse term-by-document matrix. The node can be configured to identify single-term topics or multi-term topics in the data. The properties of the node should be decided carefully based on the size of the document collection.

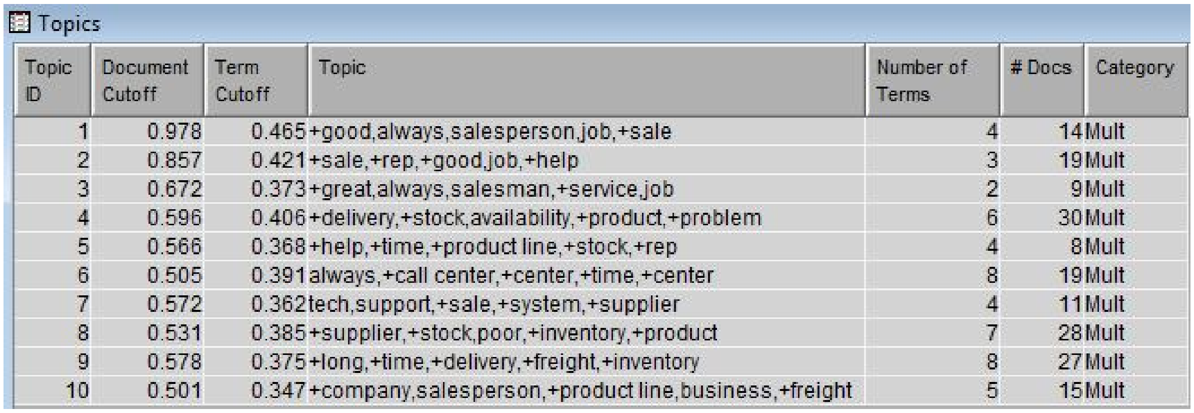
Text topic node also enables an analyst to create the topics of interest using groups of terms identified in text parsing. The objective of creating a list of topics is to establish combinations of words that are interesting to analysts. Default stop list available in SAS Enterprise Miner was used in this research. Figure 2 shows the properties used for the text parsing node.



Property	Value
⊟Parse	
Parse Variable	Comments
Language	English
⊟Detect	
Different Parts of Speech	Yes
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULTI
Find Entities	None
Custom Entities	
⊟Ignore	
Ignore Parts of Speech	'Abbr' 'Aux' 'Conj' 'Det' 'I
Ignore Types of Entities	
Ignore Types of Attributes	'Num' 'Punct'
⊟Synonyms	
Stem Terms	Yes
Synonyms	SASHELP.ENGSYNMS
⊟Filter	
Start List	
Stop List	SASHELP.ENGSTOP

Figure 2: Property panel of Text Parsing Node.

In the Text Filter node, we have changed the default option of 'Check Spelling' from "No" to "Yes". All other parameters are set to default options. Check spelling option specifies whether to check and correct the spelling of terms in the input data set. Topics extracted from the topic node are groups of terms that define a compact representation of the document collection. In this research, after various trials and errors, we selected ten as the number of topics to be extracted as shown in Table 1. For example, Topic ID 1 shows as "+good, job, +sale, +rep, +personnel" which seems very relevant for this analysis. We can identify the important grouped terms to analyze the customer satisfaction by using these text topics.


Topics

Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs	Category
1	0.978	0.465	+good,always,salesperson,job,+sale	4	14	Mult
2	0.857	0.421	+sale,+rep,+good,job,+help	3	19	Mult
3	0.672	0.373	+great,always,salesman,+service,job	2	9	Mult
4	0.596	0.406	+delivery,+stock,availability,+product,+problem	6	30	Mult
5	0.566	0.368	+help,+time,+product line,+stock,+rep	4	8	Mult
6	0.505	0.391	always,+call center,+center,+time,+center	8	19	Mult
7	0.572	0.362	tech,support,+sale,+system,+supplier	4	11	Mult
8	0.531	0.385	+supplier,+stock,poor,+inventory,+product	7	28	Mult
9	0.578	0.375	+long,+time,+delivery,+freight,+inventory	8	27	Mult
10	0.501	0.347	+company,salesperson,+product line,business,+freight	5	15	Mult

Table 1: Topics identified from Topic Node

The topic node associates terms and documents using the discovered topics. Topics are terms grouped together to describe the main theme. From Table 1, consider the Topic ID 3. The terms "great, always, salesman, +service, job" are grouped together to characterize the theme Salesman and Good Service. Text Topic node also assigns a score for each document and term cutoff for each topic. Then thresholds are used to determine if the association is strong enough to consider that a document or a term belongs to the topic. In this analysis we used only multi-term topics.

For a better understanding of the valence of each topic we calculated the percentage of negative vs non-negative documents for the topic. Table 2 shows the percentage values for negative vs. non-negative for all the ten topics. In addition, we have appended the mean of the numeric value of satisfaction (from the numeric variables in the survey) for generating additional insights.

Topic ID	Rating=0 (%) - Non Negative	Rating=1 (%) - Negative	Mean of Satisfaction	No. of Documents
Topic1	78.57	21.43	8.64	14
Topic2	57.89	42.11	8.58	19
Topic3	77.78	22.22	8.55	9
Topic4	23.33	76.67	6.86	30
Topic5	75.00	25.00	9.5	8
Topic6	36.84	63.16	8.1	19
Topic7	63.64	36.36	7.72	11
Topic8	7.14	92.86	6.25	28
Topic9	7.41	92.59	6.76	17
Topic10	73.33	26.67	9.28	15

Table 2: Association between identified topics and customer satisfaction

For example, the terms that define Topic 8 are "supplier, stock, poor, inventory, and product". While it is very clear from the terms that this is a negative topic, we can also validate the same using the percentage of negative documents (92.86%) for this topic. The survey included a question about overall satisfaction with the current supplier, with an 11-point measurement scale (0-0%satisfied, 11 –100% Satisfied). The average rating for this variable across the topics can reveal the relationship between text comments and overall satisfaction measured on a numerical scale. In Table 2, the mean of satisfaction column shows the average rating on the overall satisfaction variable for each topic. It is quite clear that those customers who were unhappy due to poor inventory and stocks-outs rated the company very poorly on the overall satisfaction measure.

PREDICTIVE MODELING

The topics extracted can also be used as inputs in a predictive model. Each topic may represent an input variable. In our case, we have 10 input variables because the topic node extracted 10 topics. We used 80:20 split in creating training and validation data sets for building predictive models. The target variable used for modeling is binary

(Negative or Non-Negative comments) based on experts' ratings of comments as described earlier.

We tried various modeling algorithms as shown in Figure 4. In addition to the traditional models such as Decision Trees, Logistic Regression and Neural Networks, we also used Memory Based Reasoning (MBR) algorithm which have been suggested by prior researchers as better for textual data. MBR uses a k-nearest neighbor algorithm to categorize or predict observations. For MBR, in the property panel, we have used Method as 'Rd-Tree' which is used to store the training set observations and then retrieve the nearest neighbors.

In all the models we used Average Square Error in validation data as the model selection criteria. We used default options for all other properties in the decision tree model. For Logistic Regression, we used stepwise selection method and default options for all other properties. We have also used an ensemble model to predict the target. The ensemble node builds a new model by combining the posterior probabilities or the predicted values from multiple predecessor models. In the property panel of Ensemble node, 'Voting' is chosen as a function to combine models for class targets. All other parameters values are set to default values. This new model may then be used to score new data.

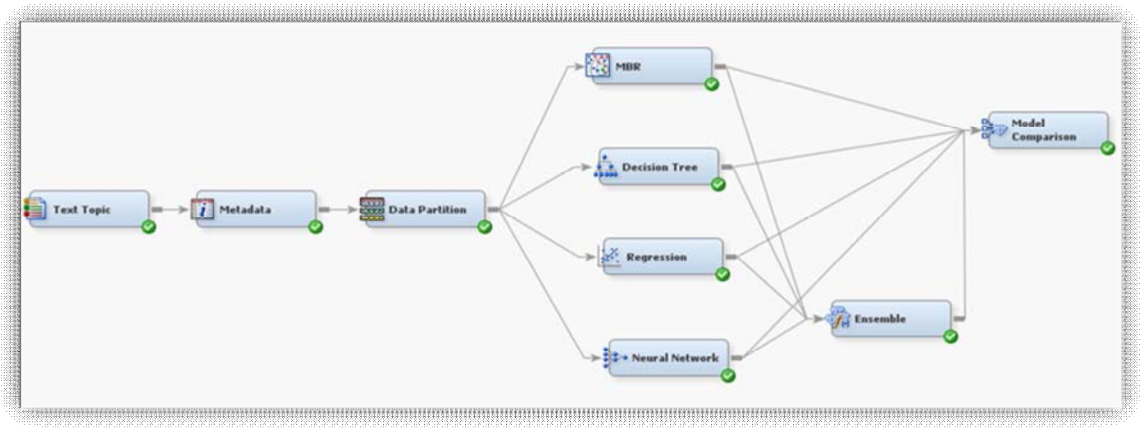


Figure 4: Process Flow diagram with predictive modeling nodes and model comparison node

MODEL COMPARISION

Model comparison node was used to compare and evaluate all the five models based on Average Squared Error. Neural network is selected as the best model for this data with Average Square Error as 0.09 in the validation data set. Table 2 shows Average Square Error, Sensitivity and Misclassification rates of the all models on Validation data. The best model can be deployed on future surveys to classify customer's comments automatically.

Model	Misclassification rate	Sensitivity	Avg Squared Error
Regression	0.19	0.93	0.12
Ensemble	0.19	0.93	0.12
Neural (Y)	**0.12**	**0.93**	**0.09**
Tree	0.21	0.89	0.17
MBR	0.24	0.71	0.15

Table 2: Model Selection statistics for all five models

While the Neural net turns out to be the best model, it is difficult to explain. Therefore, as an example, we will explain the findings form the Logistic Regression model which has the same sensitivity but a slightly higher misclassification rate than the winning neural net model. The regression identified the important and significant Topic IDs as 1, 10, 3, 8 and 9. These contain positive comments such as "+good, always, salesperson, job, +sale" and negative comments as "+long, +time, +delivery, +freight, +inventory".

Using the topics identified from this method we can also perform a qualitative study in understanding the sentiments expressed by the users. We can identify several features of service that are considered unsatisfactory by the customer. From the topic node results, interpreting Topic ID 1 and Topic Id 3 show some of the good supplier characteristics as "Good Service and Good Sales Representatives". The negative aspects can be interpreted from Topic IDs 7, 8 and 9 for services "Poor inventory and lack of technical knowledge with support staff". Table 3 summarizes the good and bad services as perceived by the customers of this company.

Good Service	Bad Service
Excellent Salesman	Poor Inventory
Service is good	Bad technical support
	System needs to be improved.

Table 3: Customer Opinions

CONCLUSION

Topic mining is an excellent way to summarize the theme of a collection of documents using topics. If a set of documents have already been classified into groups by experts, then one can also build predictive models using the topics to automatically classify new customer responses into those groups. As more data becomes available, the predictive models can be fine-tuned to improve its performance. This will save considerable time and money for companies who presently rely on experts only in classifying customer's comments. This type of analysis can perhaps be better performed using a sophisticated sentiment analysis package that is available from SAS®. However we did not have access to sentiment analysis package at the time of this research. It is however gratifying to see that with some trial-and-error, one can build a reasonable good model using the topics mined via SAS® Text Miner. In future we would like to extend this research by using a nominal target variable that will have more than two levels to include Positive, Negative, Neutral and Mixed levels. That type of research would require a much larger sample (i.e., a larger document collection). We will also like to use SAS's sentiment analysis to explore if the model performance in classifying comments can be improved by using a specialized package.

REFERENCES

1. Sullivan, Dan and Ellingsworth, Marty. 2003 "Text Mining Improves Business Intelligence and Predictive Modeling in Insurance." DM Review Magazine.
2. K Markellos, P Markellou, G Mayritsakis. 2009. "Text Mining for Business Intelligence". Available at: http://www.igi-global.com/viewtitlesample.aspx?id=11086
3. Nareddy and Chakraborty. 2011. "Improving Customer Loyalty Program through Text Mining of Customers' Comments". Proceedings SAS® Global Forum 2011.
4. Dobson. 2010. David Dobson, Dobson Analytics Inc., Segmenting Textual Data for Automobile Insurance Claims. Proceedings SAS® Global Forum 2010.
5. Kleij and Musters. 2003. Text analysis of open-ended survey responses: a complementary method to preference mapping. *Food Quality and Preference* 14 1, pp. 43–52
6. Miller, W.T. 2005. Data and Text Mining-A Business Applications Approach. Pearson Prentice Hall.

ACKNOWLEDGMENTS

The authors would like to thank Satish Garla, Karthik Nakkeeran and Hari Harasudan for their valuable comments and suggestions. They are Masters students and Research assistants at Oklahoma State University.

CONTACT INFORMATION

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

Name: Anil Kumar Pantangi
Enterprise: Oklahoma State University
Work Phone: (203)917-2199
E-mail: anil.pantangi@okstate.edu
City, State ZIP: Stillwater, OK – 74078

Anil Kumar Pantangi is a Master's student in Management Information Systems at Oklahoma State University with specialization in Data Mining and Business Intelligence Tools. Has three years of professional experience as a Remedy/ITSM Consultant with IBM and Wipro Technologies.

Name: Dr. Goutam Chakraborty
Enterprise: Oklahoma State University
Work Phone: (405)744-7644
E-mail: goutam.chakraborty@okstate.edu
City, State ZIP: Stillwater, OK - 74078

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 chaired the national conference for direct marketing educators in 2004 and 2005 and co-chaired the M2007 data mining conference. He is also a Business Knowledge Series instructor for SAS.

TRADEMARKS