Paper 125-2010

Segmenting Textual Data for Automobile Insurance Claims

David Dobson, Dobson Analytics Inc.

ABSTRACT

Text-based data contains a wealth of information. They can provide insight into information that may not be possible with structured quantitative data. In this paper, actual textual data is analyzed using SAS® Text Miner. Customers posted auto insurance claim related questions online, hoping to gain expert advice. These questions were analyzed and classified based on the nature of the claims. This text mining exercise has numerous business improvement possibilities, such as enabling online consultants to answer a customer's question more efficiently, based on the consultant's expertise.

INTRODUCTION

A large portion of today's business information is stored as text (Herschel 2005). Widespread use of the Internet, has led text data to be stored more so in "Question and Answer" services on the Internet. These online services answer individual customer questions well. However, they do not analyze their collected large data sets to look for patterns or to gain further insight about their customers. In this paper, questions are analyzed that were asked in an online auto insurance claims forum, where customers sought expert advice regarding their claims after an auto accident. Using SAS® Text Miner, the text clustering mining technique is applied to group the textual claim data. The text mined data was further analyzed with structured data. Cluster analysis helps us understand the nature of the questions and the frequency of key terms that are mentioned in the questions.

OBJECTIVES

The primary objective is to apply pattern discovery to perform data reduction and profiling of the text data. The secondary objective is to calculate the weight for each key term, a measure of their relative importance, which auto insurance claimants mention in their questions.

DATA

For the purpose of text analysis, 225 questions were picked that customers asked regarding their auto insurance claims from the **AllExperts** website - http://en.allexperts.com/q/Auto-Insurance-Claims-2055. AllExperts is a free question and answer service on the Internet. Three of the website's expert volunteers, Bennie, Richard, and Sheldon, answered 75 questions each. The data preprocessing for creating SAS data files was conducted in SAS 9.1. The prepared data file had the following five variables:

`Question`: Actual question customer asked (i.e. textual data)

`Q_Length`: The length of the question (number of characters in a question)

`Gender`: Gender of who submitted the question

`Day`: Day of the week when the question was asked

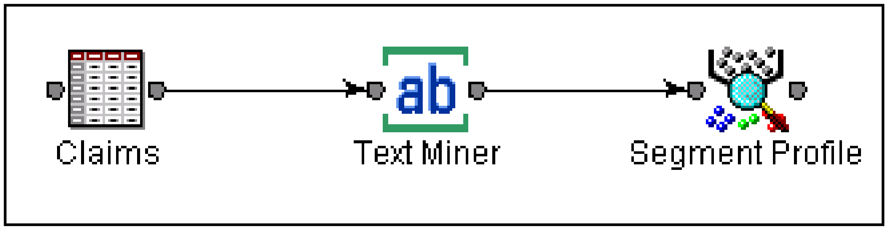
`Expert`: Volunteer who answered the question

`Case No`: Question identifier

METHODOLOGY

Using SAS Text Miner tool of SAS® Enterprise Miner, the text clustering mining technique to cluster the textual claim data was applied. Once the clusters were identified, the cluster data was examined and the factors that differentiated the clusters using the Segment Profile tool was identified. Figure 1 below illustrates the process flow chart using SAS Enterprise Miner's graphical Text Miner tool.

Figure 1. Auto Insurance Claims Process Flow Diagram



ANALYSIS RESULTS

Cluster Analysis

Numerous cluster solutions were produced, but the 3-cluster solution was selected, as the clusters were relatively large and distinct. The clusters were uniquely characterized by a single term, as the Classification column in figure 2 below shows.

Figure 2: Clusters with Relative Frequency and Descriptive Terms

Cluster	Percent	Descriptive Terms (20)	Classification
1	21%	medical, + medical bill, + limit, + bill, + injury, neck, + policy, + law, + chiropractor, + lawyer, + doctor, car accident, + release, + pain, attorney, + settlement, + settle, liability, + feel, + record	Bodily Injury
2	39%	police, information, + lane, + officer, + deny, + report, police report, + file, parking, + side, front, + park, + right, + hit, + road, + lot, + license, + stop, truck, + turn	Accident Description and Incident Reporting
3	39%	+ shop, body shop, + estimate, + total, + loss, + cost, + value, body, rental, + fix, + repair, + check, rental car, + run, + accept, + damage, + insurance company, money, + cause, + bumper	Vehicle Damages and Repairs

Figure 2 shows that the most frequently asked questions about insurance claims included the subject of Accident Description and Incident Report (Cluster 2) and Vehicle Damages and Repairs (Cluster 3), with 39% of all the questions asked, respectively. Cluster 1, questions related to Bodily Injury claims, is the smallest cluster with 21% of all questions.

Term Extraction

Figure 3 below shows the output of key terms in the documents.

Figure 3: Selected List of Terms and Weights

Term	Freq	# Documents	Weight
+ doctor	18	9	0.6200
body shop	13	11	0.5658
+ chiropractor	11	11	0.5573
Attorney	18	13	0.5517
Neck	24	15	0.5391
parking lot	17	14	0.5277
+ medical bill	15	14	0.5171
+ bumper	30	19	0.4926
+ settlement	26	20	0.4674
+ value	31	24	0.4387
+ turn	36	27	0.4077
+ stop	45	28	0.4063
+ file	36	27	0.4050
+ estimate	42	29	0.4006
+ repair	49	33	0.3816
+ injury	37	31	0.3774
Police	84	46	0.3575
+ total	66	53	0.2798

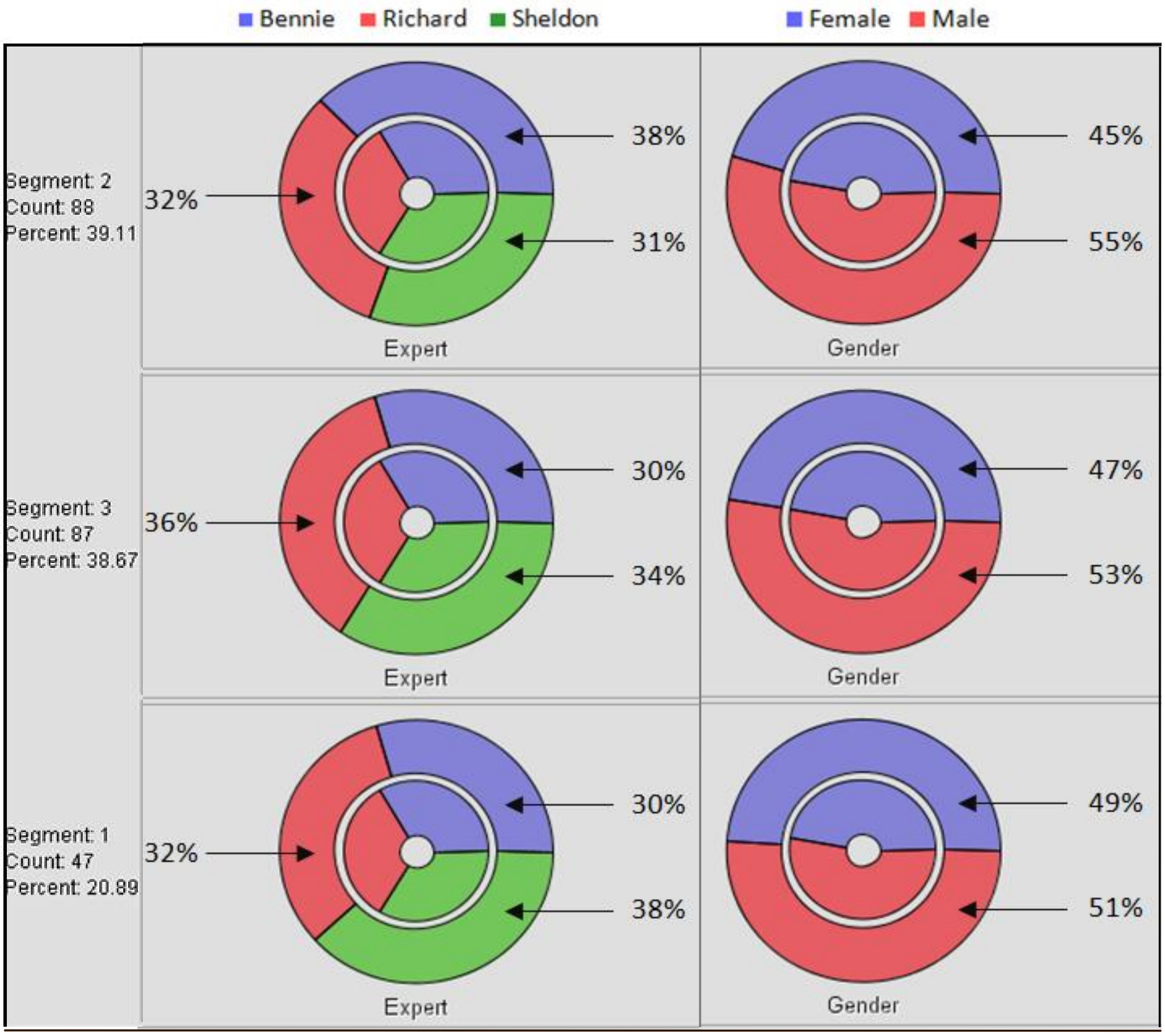
The Term column, in the above table represents the distinct word mentioned in the question. The plus sign (+) in the beginning of the term means that they are “Rollup Terms”. The Freq column represents the frequency of term occurrence. The # of Document column represents the number of documents the term occurred in. The Weight column is the weight of the term; the higher the weight, the more distinguish the term is between the documents.

PROFILING SEGMENTS

To describe and interpret the composition of three cluster solution, the cluster with structured data, variables such as *expert* who answer the questions, *gender* of the inquirers who asked the questions, *question length*, and *day* of the week when the question was asked were analyzed.

Figure 4, using the Segment Profile tool in SAS Enterprise Miner, shows the proportion of expert and gender by cluster.

Figure 4: Expert and Gender, showing % within Cluster



PROFILING SEGMENTS (Cont'd)

Experts

For Cluster 1 (Bodily Injury), Sheldon answered proportionally more questions than Richard and Bennie. For Cluster 2 (Accident Description and Incident Reporting), Bennie answered more questions than Richard and Sheldon; and, for Cluster 3 (Vehicle Damage and Repair), Richard answered slightly more questions than the others.

Based on this information, though each expert is capable of answering a variety of questions, we could generalize volunteer expertise based on the frequency of questions answered for each cluster:

- **Sheldon:** Bodily injury questions (Cluster 1)
- **Bennie:** Accident description and Incident reporting questions (Cluster 2)
- **Richard:** Vehicle damage and repair questions (Cluster 3)

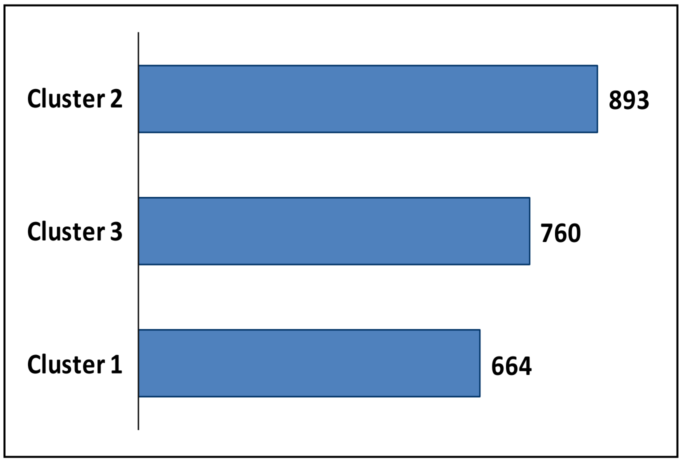
Gender

There were slightly more men asking questions than women. Of the 225 questions, 52.4% were from men and 47.6% from women. There was a noticeable difference in Cluster 2, where 55% of men, compared to 45% of females, asked questions about accident descriptions and incident reporting. Cluster 1, bodily injury related questions, did not show much difference in proportion for men and women. Comparing women in all the clusters, a higher proportion of them asked bodily injury related questions (Cluster 1).

Question Length

Figure 5 shows Cluster 1 (Bodily injury) had the shortest questions, with an average of 664 characters counts per question. Cluster 2 (Accident/ incident description) had the longest questions, with an average of 893 characters per question. Cluster 3 (Vehicle Damage and Repair) has an average 760 characters per question.

Figure 5. Question Length - Characters Counts

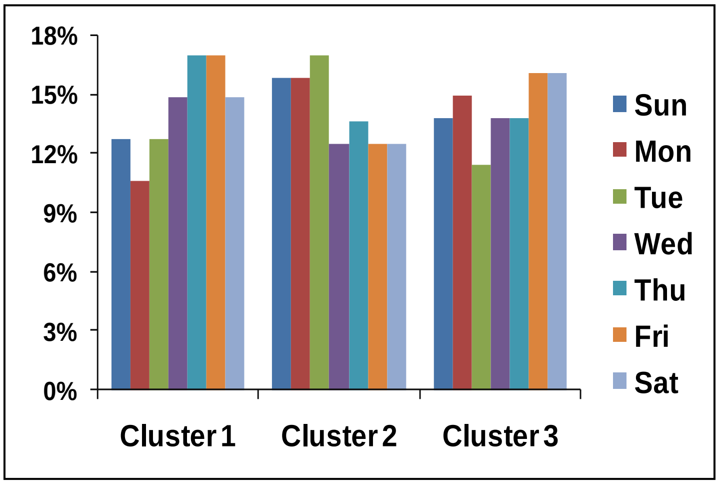


PROFILING SEGMENTS (Cont'd)

Day of the Week

Figure 6 shows that Thursday and Friday were busy days for Cluster 1 (Bodily injury), with more than one-third of the total questions being asked. For Cluster 2 (Accident/ incident description), Sunday through Tuesday were busy, with almost half of all the questions being asked. For Cluster 3 (Vehicle Damage and Repair), Friday and Saturday were busy, with almost one-third of all questions asked on these two days.

Figure 6. Day of the Week when Question was asked



CONCLUSION AND LESSONS LEARNED

- **Text Reduction:** The original textual data was reduced from 225 documents (questions) consisting of 34,942 words to 3 meaningful clusters consisting of 20 descriptive terms per cluster.

- **Types of Question:** About 8 out of 10 questions asked were related to Accident Description and Incident Reporting (Cluster 2) and Vehicle Damages and Repairs (Cluster 3). 2 out of 10 questions were related to Bodily Injury claims (Cluster 1).

- **Expert Specialization:** Expert specialization was generalized, for example Sheldon answered mostly Bodily Injury questions (Cluster 1). This information is helpful in directing the question to the appropriate expert and to prioritize their workload.

- **Day of the Week:** Day of the week has a slight impact on the frequency of the questions asked by cluster. More questions are asked on Thursday and Friday for Cluster 1; Sunday, Monday and Tuesday are busy for Cluster 2; and Friday and Saturday are busy for Cluster 3. This information is useful for volunteer allocation by cluster and by the day of the week.

- **Document Order vs. Cluster Structure:** Notably, if the document order changed in the cluster analysis, the structure of the cluster changed slightly. This is due to a small sample size.

- **# of Clusters vs. # of Documents:** If the number of clusters is reduce from a higher number to a lower number, not all the documents can be classified within the assigned clusters. The text data was characterized by three natural clusters, but three documents were not clustered. This suggests document outliers, or the possibility of adding another cluster.

REFERENCES

- Dobson, David. "Customer Segmentation: The Application of Data Mining for the Automobile Insurance Industry". Paper presented at the Joint Statistical Meetings in Denver, Colorado, August 3-7, 2008.
- Herschel, Richard T. and Jones, Nory E. "Knowledge management and business intelligence: the importance of integration" *Journal of Knowledge Management*, vol. 9 (4), pp. 45-55, 2005.
- Sullivan, Dan and Ellingsworth, Marty "Text Mining Improves Business Intelligence and Predictive Modeling in Insurance" DM Review Magazine, July 2003.

ACKNOWLEDGEMENT

The author would like to thank Dr. Tao Xie, Professor of Computer Science at NC State University for his valuable feedback on this paper. The author would also like to thank former classmates Sean Harris, Petr Hustak and Dong Wang, MSA Class of 2008 at NC State University, who worked as team members in a similar Text Mining project.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. You may contact the author via email at: david.dobson1@gmail.com