

From Words to Actions: Using Text Analytics to Drive Business Decisions

Reid Baughman, Zencos Consulting LLC

ABSTRACT

Companies from a variety of industries collect free-form text data that can be used to identify new patterns and relationships. Because unstructured text data does not fit the standard row and column format, it can be more difficult to analyze and utilize. This growing array of data has the potential of yielding new insights for companies seeking to better understand customers and gain an edge relative to competitors. By transforming free-form text data into a structure that can be analyzed and visualized, analysts can use supervised and unsupervised data mining techniques to shed new light on old business problems or develop fresh insights. Call center transcripts, medical records, practitioner's notes, survey responses, or any free-form response fields can all hold valuable insights for any organization. This paper will outline an example of how to manage and transform free-form text data, apply advanced analytical methods to extract useful patterns, and develop actionable insights.

INTRODUCTION

In the wake of the 2007 financial crisis, the Consumer Protection Financial Bureau (CFPB) was created to "promote fairness and transparency for mortgages, credit cards, and other consumer financial products and services". As part of that mission, they established a system for consumers to log complaints regarding financial products and made the database of complaints available to the public. There are two columns of data that are of interest for this analysis: "Consumer complaint narrative" and "Company response to customer". The first column provides the opportunity to use unsupervised modeling to listen to the voice of the customer and understand common reasons for dissatisfaction. Combining those results with the second variable – Company response to customer – allows you to see which complaint topics are potentially avoidable by employing a supervised approach with relief¹ as the target. The assumption here is that if the company offered relief, then they accept wrongdoing and that the complaint should have been avoidable. To demonstrate these two types of modeling, we'll use SAS Enterprise Miner due to its robust text mining capabilities.

The data chosen for the analysis comes from the Wells Fargo complaints. They provide an interesting study since they ran afoul² of the CFPB and were required to pay a multi-million dollar fine because of a pattern of complaints uncovered from the complaint database. In the period measured (March 2015 – May 2017) there were 4,643 complaints that included a consumer narrative and of those 756 resulted in the bank providing some sort of relief. In the period measured (March 2015 – May 2017) there were 4,643 complaints that included a consumer narrative and of those 756 resulted in the bank providing some sort of relief.

UNSUPERVISED MODELING: TEXT TOPICS GENERATION

DATA PREPARATION

The first step in any text mining project is to parse and filter the data with the Text Parse and Text Filter nodes. These nodes parse (or split), stem (combine all variations of the same word), drop unimportant filler words, and create the term-frequency matrix. The term-frequency matrix is then used as the input for both the unsupervised and supervised methods.

¹ Relief = 1 when Company response = "Closed with monetary relief", "Closed with non-monetary relief" or "Closed with relief". Relief = 0 when Company response = "Closed with explanation", "Closed without relief" or "Closed".

² <https://www.consumerfinance.gov/about-us/newsroom/consumer-financial-protection-bureau-fines-wells-fargo-100-million-widespread-illegal-practice-secretly-opening-unauthorized-accounts/>

TOPIC CLUSTERING

Once we've cleaned up and structured our data, we're ready for analysis. We'll first try grouping and clustering our terms using the Text Topic and Text Cluster nodes. Both are considered unsupervised methods because they lack a target they're trying to predict. We're just trying to see if any natural term clusters arise from the data. Terms cluster together if they appear together frequently. For example, if a high percentage of complaints frequently contained the words "late", "fee" and "unfair" then those words would likely appear as a topic. The Text Topic node does this topic clustering for us and generates a couple tables of interest.

Text Topic Node

Table 1 below shows an excerpt of the top terms found across all complaints. Unsurprisingly, we find "wells fargo" as the most common term and we might consider adding it to the list of stop terms since it's uninteresting for our analysis. The remaining terms shown are ordered by the "NUMDOCS" column which shows the number of documents (in this example complaints) that term is found in. The "FREQ" column shows the overall frequency of the term across all documents. The "topic" columns show how strongly weighted each term is across each of the generated topics with higher weights signifying a higher association of that term with that topic.

TERM	rolestring	WEIGHT	FREQ	NUMDOCS	topic1	topic2	topic3
wells fargo	Company	0.007021	8480	3329	0.001	0.01	0.005
+ account	Noun	0.133036	5870	2025	0.017	0.942	-0.038
+ payment	Noun	0.018661	5302	1794	-0.01	0.002	-0.006
+ pay	Verb	0.079639	3172	1651	0.014	0.009	0.041
+ loan	Noun	0.023302	4210	1645	0.002	-0.008	-0.008
+ receive	Verb	0.004322	2356	1437	0	0.002	0.001
+ time	Noun	0.004624	2101	1423	0.001	0	0.001
+ credit	Noun	0.053453	2984	1379	-0.001	0.008	0.192

Table 1. Term Table from Text Topic node

Table 2 corresponds to Table 1 except that the topics are now displayed as rows instead of columns. The benefit of this view, is that you can easily see which terms comprise which topics. The first topic identified relates to complaints regarding fees (this may not be very surprising). Topic #2 contains the terms "+account³, +close, +open, +bank account, +bank" which appears to be directly related to the rampant unauthorized opening of bank accounts of which Wells Fargo was found guilty.

_topicid	_name	_numterms	_numdocs
1	+fee,+refund,+waive,nsf,+nsf fee	48	720
2	+account,+close,+open,+bank account,+bank	50	855
3	+card,+credit card,+credit,+credit card account,+credit +modification,+loan	51	588
4	modification,+home,+foreclosure,+mortgage	193	626
5	+charge,+service,\$35.00,+account,+close	22	677
6	+overdraft,+overdraft fee,+fee,\$35.00,overdraft protection	43	284
7	+check,+deposit,+cash,+deposit,+hold	68	518

Table 2. Topic Table from Text Topic node

³ Note: the "+" indicates that several variations of the word were stemmed and combined to create that term (e.g. "Account", "Accounts", etc.)

Text Cluster Node

The Text Cluster node is another approach to topic generation that works similar to the Text Topic node but which employs a methodology that allows for additional customization and flexibility. The resulting clusters are shown below in Table 3 (Note: for brevity, this table does not include all clusters calculated). Similar to the Topic ID #2 from the Topic results above, we notice that Cluster ID #7 above appears to contain the complaints related to the widespread unrequested account opening performed by some Wells Fargo employees. Since both methods arrive at similar yet different results, it's typically a good idea to try both methods and see which results are more interpretable. The trends from these topics and clusters are extremely valuable for any company trying to better understand their customers and to make changes that will have the largest possible impact.

CLUSTER	clus_desc	freq	percent
1	services dealer 'dealer services' +vehicle +car +'credit report' +report +loan +purchase +late +number +payment 'wells fargo' +pay +dispute	102	0.02
2	+loan +mortgage +home +modification home +foreclosure +deny +house +property +'loan modification' +sale +keep +help +document +review	753	0.16
3	+transaction +debit +bank +card +claim +account +money +fraud +transfer +dispute +issue +check +deposit +balance +branch	162	0.03
4	+post +transaction +date +show +balance +day +overdraft +debit +bank +issue +deposit +payment +fee +apply +check	85	0.02
5	+day +sale +review +process +document +date +home +modification +request +foreclosure +file +receive +property +house +work	295	0.06
6	+balance +interest +amount +owe monthly +pay +payment +loan +rate +home +year +show +month +mortgage +purchase	165	0.04
7	+card +credit +'credit card' +debit +open +close +account +customer +branch +'credit report' +number +report +transfer +balance +apply	376	0.08

Table 3. Term Clusters from Text Cluster node

SUPERVISED MODELING: PREDICTING AVOIDABLE COMPLAINTS

Now that we have a better understanding of some of the dominant complaint topics, let's look now to see which topics are potentially avoidable. Using the "relief" indicator explained earlier, we'll attempt to model which topics are associated with relief to uncover areas where the company can improve.

SAS Enterprise Miner contains two Text Mining nodes for supervised learning: Rule Builder and Text Profile. While it is possible to use the topics and clusters identified in the Text Topic and Text Cluster nodes as inputs into classification models like Logistic Regression, Decision Trees and Neural Networks, we'll limit our discussion to just the text mining nodes.

RULE BUILDER NODE

The Text Rule Builder attempts to predict which complaints received relief by creating a variety of AND/AND NOT combinations of individual terms and using them as predictors. Table 4 below shows the rules generated from the Rule Builder node.

Relief	conj_label	t_precision	t_recall	t_f1	TP_ratio
No	asc	100.00%	1.41%	2.79%	55/55
No	modify	98.96%	4.91%	9.36%	145/147
No	modification	98.05%	18.09%	30.54%	631/645
No	short sale	97.58%	20.79%	34.27%	133/140
No	equity	97.15%	23.69%	38.10%	145/154
No	income	96.75%	27.55%	42.89%	338/355
No	mortgage & ~fee	94.56%	41.60%	57.78%	1048/1125
No	home	93.84%	46.64%	62.31%	809/871
Yes	charge & insufficient	90.91%	1.32%	2.61%	10/11
Yes	\$70.00	79.31%	3.04%	5.86%	15/21
Yes	bank rep	78.95%	3.97%	7.56%	7/9
Yes	cycle	70.59%	6.35%	11.65%	19/33
Yes	\$12.00	67.42%	7.94%	14.20%	12/22
Yes	card & late fee	65.00%	10.32%	17.81%	18/31
Yes	account & overdraft fee	53.25%	17.33%	26.15%	66/143
Yes	\$35.00	51.38%	19.71%	28.49%	50/105
Yes	car	48.20%	21.30%	29.54%	56/193
Yes	fee & ~modification & ~sale	36.89%	39.29%	38.05%	284/804
Yes	card	36.89%	39.29%	38.05%	190/716

Table 4. Rules from Rule Builder node

The Rules Obtained table from the results shows us the rules created⁴ as well as several important metrics detailing their effectiveness in predicting relief. The Target Value column in the far left shows if the rule is trying to predict a one or zero and then lists the Precision, Recall, F1 score and Total Positive/Total measures. Precision and Recall refer to true positives/total predicted positive and true positives/total actual positives respectively. The F1 score is a weighted average of those two measures where a larger value means a better balance of both and potentially a more desirable result⁵. Judging by the F1 metric the top five rules that predict relief are:

1. account & overdraft fee
2. \$35.00
3. car
4. fee & ~modification & ~sale
5. card

Most of these complaints appear to be related to fees charged to consumers. Since removing fees would likely cause a non-trivial impact on the banks profit, perhaps a better solution might be to train customer service staff to remove fees from customers' accounts when they call in to complain. A cursory investigation of such claims appears to show that many customers contacted the bank regarding the fees prior to filing the complaint with the CFPB so the problem could have been taken care of before providing a permanent blot on the company's public record and most likely being paid in the end anyway!

⁴ The tilde (~) in some of the rules is a logical NOT operator meaning NOT EQUAL to some value (i.e. "loan & ~fee" would be read as "contains the term loan but not fee").

⁵ There are instances where one might be more interested in maximizing one or the other, especially when the costs of incorrectly guessing each are asymmetric.

TEXT PROFILE NODE

The Text Profile node provides corroboration of these results using a hierarchical Bayesian model⁶ to predict which terms are the most likely to accurately distinguish whether a customer received relief (1 or 0). The terms in Figure 1 below with blue/gray in the left corresponding to red on the right are those which most accurately distinguished the two. The closer the color is to blue, the more the model predicted 0 and the closer to red the more it predicted 1. The two columns coincide with whether the customer *actually* received relief. Similar to the Text Rule Builder node above, we see that “fee”, “late fee” and “overdraft” and found to be the most related to avoidable complaints.

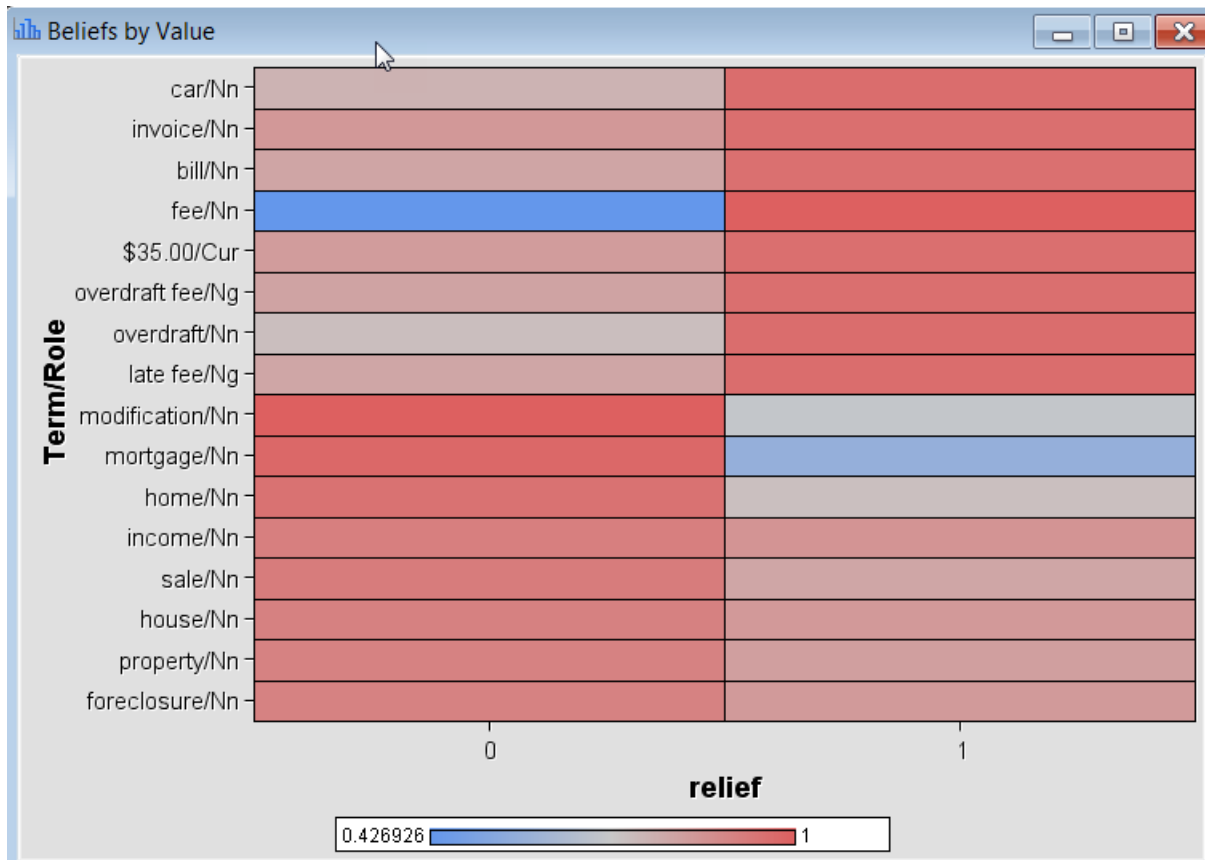


Figure 1. Term Significance Chart from Text Profile node

CONCLUSION

The text mining techniques demonstrated here uncover otherwise unreachable insights trapped in potentially thousands of customer responses. Sifting through the deluge of text, these techniques allow us to quickly filter down to the most meaningful words that describe our customer's experience. Perhaps more importantly, these techniques generalize to practically any body of text from tweets, to transcripts, to even books to name a few. Just think, what insights could be hiding in your stores of unstructured text?

⁶ In order to avoid merely selecting the most common terms, prior probabilities are used to down-weight terms that are common in more than one level of the target variable.

REFERENCES

Chakraborty, Goutam, et al. 2013. *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS®*. Cary, NC: SAS Institute Inc.

SAS Institute Inc. 2015. *SAS® Text Miner 14.1: Reference Help*. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

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

Reid Baughman
Zencos Consulting LLC
rbaughman@zencos.com