Paper SAS282-2017

Applying Text Analytics and Machine Learning to Assess Consumer Financial Complaints

Tom Sabo, SAS Institute Inc., Arlington, VA

ABSTRACT

The Consumer Financial Protection Bureau (CFPB) collects tens of thousands of complaints against companies each year, many of which result in the companies in question taking action, including making payouts to the individuals who filed the complaints. Given the volume of the complaints, how can an overseeing organization quantitatively assess the data for various trends, including the areas of greatest concern for consumers?

In this paper, we apply a repeatable model of text analytics techniques to the publicly available CFPB data. Specifically, we use SAS® Contextual Analysis to explore sentiment and machine learning techniques to model the natural language available in each free-form complaint against a disposition code for the complaint, primarily focusing on whether a company paid out money. This process generates a taxonomy in an automated manner. We also explore methods to structure and visualize the results, showcasing how areas of concern are made available to analysts using SAS® Visual Analytics and SAS® Visual Statistics. Finally, we discuss the applications of this methodology for overseeing government agencies and financial institutions alike.

INTRODUCTION

The CFPB is one of a number of overseeing institutions that ensure that the consumer is treated fairly by corporations and financial institutions. It stands alongside other international organizations such as the United Kingdom Financial Conduct Authority, the Australian Competition and Consumer Commission, and EEC-Net, which assists European Union consumers in resolving international purchase complaints. The CFPB was established in 2011. It was created after the financial crisis of 2008 to help consumers resolve problems at the transactional level, and to address larger macro-level issues before they become unmanageable. The CFPB is responsible for more than 11.7 billion dollars of relief for consumers due to enforcement actions[1].

The CFPB has handled more than one million complaints since its inception, and this number is increasing annually[1]. As more complaints are filed, is the solution to handling the increasing workload adding more readers to manually address the complaints and identify trends? Generally speaking, scaling up manual analysis of textual data has three challenges. First, unless very specific standards (bordering on definitive rules) are adopted, the method that one reader uses to address and tag a complaint can be quite different from the method a second reader uses. Scale this difference up to many readers, and you have many different, qualitative interpretations of the textual data. Second, reader fatigue ensures that the way a reader will address the first 10 complaints of the day will not necessarily be the same as the way they address the last 10 complaints. Vital information might be missed or skipped. Finally, suppose a trend is uncovered, and the directive arises to go back and retag all the data from the past year with this new trend. This is a case where manual analysis doesn't scale, and often enough, a simple search operation for a trend pattern will not be sufficient.

The benefit for potential analysis of the CFPB data is that each of the records is tagged with a disposition code, denoting the action taken by the organization against which the complaint was filed. With this information (and to a lesser extent, independent of it), we can uncover trends surrounding the actions taken (for example, what were the defining characteristics of complaints where the organization in question paid out monetary compensation to the individuals filing the complaints vs. complaints with a lower disposition such as those closed simply with an explanation?) In the sections that follow in this paper, we will explore a short end-to-end implementation that showcases how an analyst can use SAS technology to quantitatively assess the complaint data for various trends. This includes the consumers' areas of greatest concern, as well as areas of complaint that are in need of legislative correction. We show how to apply a sentiment model to the text as well as machine learning methods through SAS

Contextual Analytics to accomplish this. Finally, we will assess the results using visualization capabilities to highlight actionable information.

Specifically, we will apply a process built upon three previously presented papers at SAS Global Forum: one in 2014 to define a framework for research analytics[2], a second in 2015 to extend this framework for government spending[3], and a third in 2016 to apply the framework to auto-categorization of event data in conflict affected regions[4]. We encourage the reader to refer back to these papers to gain a sense for the wide applicability of these capabilities across several public sector use cases.

The five-step process for generating and using the framework is as follows:

1. **Data acquisition and preparation for text analytics**: Data is acquired for our example use case through web interfaces and is converted into a SAS data set using SAS® Enterprise Guide®.
2. **Text analytics**: We use SAS Contextual Analysis for sentiment analysis, as well as for modeling and rule-building techniques to generate hierarchical categorical data. This newly generated sentiment and categorical data serves as additional structured information for subsequent analysis and visualization against the CFPB data set.
3. **Data preparation for visual analysis**: We first use categorical scoring code outlined in the 2016 paper mentioned above[4]. To add a layer of sentiment information, we leverage the code provided by SAS Contextual Analysis and merge with the categorically scored table. This enables hierarchical exploration of the data in the subsequent visualization steps.
4. **Ad hoc exploration and modeling**: This is accomplished with SAS Visual Analytics and SAS Visual Statistics.
5. **Interactive report generation and use**: This is also accomplished with SAS Visual Analytics.

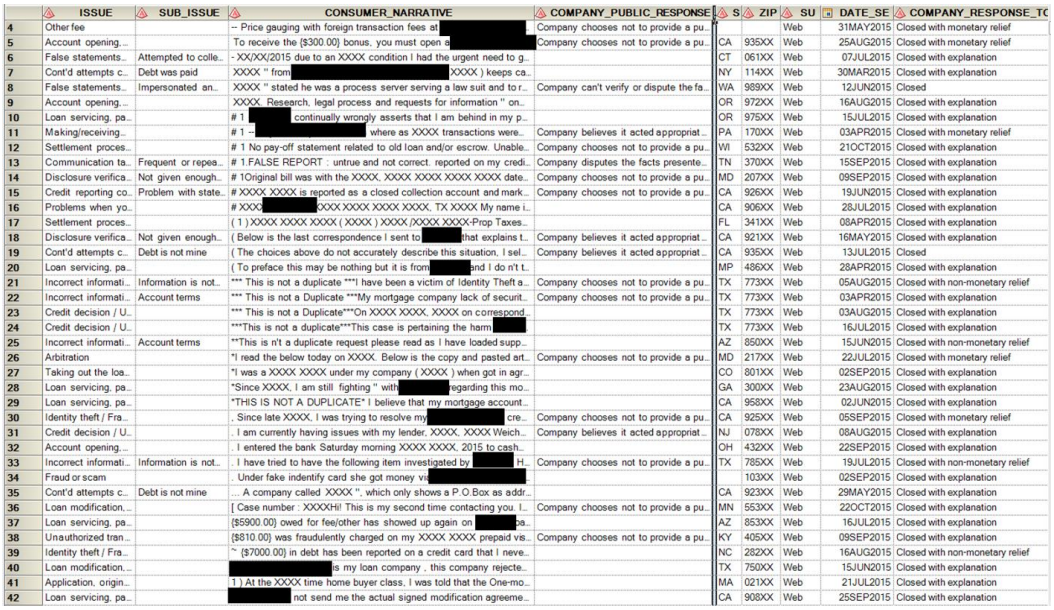
DATA ACQUISITION AND PREPARATION FOR TEXT ANALYTICS

We obtained the CFPB data from the interface available on its publicly facing website[5] and specified that only data with a narrative should be pulled. For the project, we used data from March 19, 2015 to October 30, 2015. This amounted to 37,619 complaints with a narrative. We imported the data into a SAS data set using SAS Enterprise Guide. In the process, we retained the original SAS data set of 37,619 complaints, but we also generated a new representative data set of 15K complaints for the interactive and modeling-based text analytics work.

TEXT ANALYTICS – SAS CONTEXTUAL ANALYTICS

After registering the SAS data set in metadata using SAS® Management Console, we select and load the data set within the SAS Contextual Analysis interface. For this project we selected the option within the interface to run a document-level (complaint level) sentiment model. In our case, we chose to run the default sentiment model. However, we could also use a specialized sentiment model developed using SAS® Sentiment Analysis Studio® in its place.

When selecting the data set, we specify the CONSUMER_NARRATIVE column as the freeform text field to perform text analytics against, as well as specifying the COMPANY_RESPONSE_TO_CONSUMER as a categorical target variable. See Figure 1 below for an example snippet of the data, including additional structured data columns. The names of the various financial and retail organizations called out in the narrative complaints have been obscured. Also note that the CFPB has also already obfuscated all personally identifiable information for each consumer using XXXX notation.



	ISSUE	SUB_ISSUE	CONSUMER_NARRATIVE	COMPANY_PUBLIC_RESPONSE	S	ZIP	SU	DATE_SE	COMPANY_RESPONSE_TO
4	Other fee		-- Price gauging with foreign transaction fees at [redacted]	Company chooses not to provide a pu...			Web	31MAY2015	Closed with monetary relief
5	Account opening, ...		To receive the [$300.00] bonus, you must open a [redacted]	Company chooses not to provide a pu...	CA	935XX	Web	25AUG2015	Closed with monetary relief
6	False statements...	Attempted to colle...	- XX/XX/2015 due to an XXXX condition I had the urgent need to g...		CT	061XX	Web	07JUL2015	Closed with explanation
7	Cont'd attempts c...	Debt was paid	XXXX " from [redacted] XXXX) keeps ca...		NY	114XX	Web	30MAR2015	Closed with explanation
8	False statements...	Impersonated an...	XXXX " stated he was a process server serving a law suit and to r...	Company can't verify or dispute the fa...	WA	989XX	Web	12JUN2015	Closed
9	Account opening, ...		XXXX Research, legal process and requests for information " on...		OR	972XX	Web	16AUG2015	Closed with explanation
10	Loan servicing, pa...		# 1 [redacted] continually wrongly asserts that I am behind in my p...		OR	975XX	Web	15JUL2015	Closed with explanation
11	Making/receiving...		# 1 - [redacted] where as XXXX transactions were...	Company believes it acted appropriat...	PA	170XX	Web	03APR2015	Closed with monetary relief
12	Settlement proces...		# 1 No pay-off statement related to old loan and/or escrow. Unable...	Company chooses not to provide a pu...	WI	532XX	Web	21OCT2015	Closed with explanation
13	Communication ta...	Frequent or repea...	# 1 FALSE REPORT : untrue and not correct reported on my credi...	Company disputes the facts presente...	TN	370XX	Web	15SEP2015	Closed with explanation
14	Disclosure verifica...	Not given enough...	# 1Original bill was with the XXXX, XXXX XXXX XXXX XXXX date...	Company chooses not to provide a pu...	MD	207XX	Web	09SEP2015	Closed with explanation
15	Credit reporting co...	Problem with state...	# XXXX XXXX is reported as a closed collection account and mark...	Company chooses not to provide a pu...	CA	926XX	Web	19JUN2015	Closed with explanation
16	Problems when yo...		# XXX[redacted]XXX XXXX XXXX XXXX, TX XXXX My name i...		CA	906XX	Web	28JUL2015	Closed with explanation
17	Settlement proces...		(1) XXXX XXXX XXXX (XXXX) XXXX /XXXX XXXX-Prop Taxes...		FL	341XX	Web	08APR2015	Closed with explanation
18	Disclosure verifica...	Not given enough...	(Below is the last correspondence I sent to [redacted] that explains t...	Company believes it acted appropriat...	CA	921XX	Web	16MAY2015	Closed with explanation
19	Cont'd attempts c...	Debt is not mine	(The choices above do not accurately describe this situation. I sel...	Company believes it acted appropriat...	CA	935XX	Web	13JUL2015	Closed
20	Loan servicing, pa...		(To preface this may be nothing but it is from [redacted] and I do n't t...		MP	486XX	Web	28APR2015	Closed with explanation
21	Incorrect informati...	Information is not...	*** This is not a duplicate ***I have been a victim of Identity Theft a...	Company chooses not to provide a pu...	TX	773XX	Web	05AUG2015	Closed with non-monetary relief
22	Incorrect informati...	Account terms	*** This is not a Duplicate ***My mortgage company lack of securit...	Company chooses not to provide a pu...	TX	773XX	Web	03APR2015	Closed with explanation
23	Credit decision / U...		*** This is not a Duplicate***On XXXX XXXX, XXXX on correspond...		TX	773XX	Web	03AUG2015	Closed with explanation
24	Credit decision / U...		***This is not a duplicate***This case is pertaining the harm [redacted]		TX	773XX	Web	16JUL2015	Closed with explanation
25	Incorrect informati...	Account terms	**This is n't a duplicate request please read as I have loaded supp...		AZ	850XX	Web	15JUN2015	Closed with non-monetary relief
26	Arbitration		*I read the below today on XXXX. Below is the copy and pasted art...	Company chooses not to provide a pu...	MD	217XX	Web	22JUL2015	Closed with monetary relief
27	Taking out the loa...		*I was a XXXX XXXX under my company (XXXX) when got in agr...		CO	801XX	Web	02SEP2015	Closed with explanation
28	Loan servicing, pa...		*Since XXXX, I am still fighting " with [redacted] regarding this mo...		GA	300XX	Web	23AUG2015	Closed with explanation
29	Loan servicing, pa...		*THIS IS NOT A DUPLICATE* I believe that my mortgage account...		CA	958XX	Web	02JUN2015	Closed with explanation
30	Identity theft / Fra...		. Since late XXXX, I was trying to resolve my [redacted] cre...	Company chooses not to provide a pu...	CA	925XX	Web	05SEP2015	Closed with monetary relief
31	Credit decision / U...		. I am currently having issues with my lender, XXXX. XXXX Weich...	Company believes it acted appropriat...	NJ	078XX	Web	08AUG2015	Closed with explanation
32	Account opening, ...		. I entered the bank Saturday morning XXXX XXXX, 2015 to cash...		OH	432XX	Web	22SEP2015	Closed with explanation
33	Incorrect informati...	Information is not...	. I have tried to have the following item investigated by [redacted] H...	Company chooses not to provide a pu...	TX	785XX	Web	19JUL2015	Closed with non-monetary relief
34	Fraud or scam		. Under fake indentify card she got money via [redacted]			103XX	Web	02SEP2015	Closed with explanation
35	Cont'd attempts c...	Debt is not mine	... A company called XXXX ", which only shows a P.O.Box as addr...		CA	923XX	Web	29MAY2015	Closed with explanation
36	Loan modification, ...		[Case number : XXXXHi! This is my second time contacting you. I...	Company chooses not to provide a pu...	MN	553XX	Web	22OCT2015	Closed with explanation
37	Loan servicing, pa...		{$5900.00} owed for fee/other has showed up again on [redacted] ce...		AZ	853XX	Web	16JUL2015	Closed with explanation
38	Unauthorized tran...		{$810.00} was fraudulently charged on my XXXX XXXX prepaid vis...	Company chooses not to provide a pu...	KY	405XX	Web	09SEP2015	Closed with explanation
39	Identity theft / Fra...		~ {$7000.00} in debt has been reported on a credit card that I neve...		NC	282XX	Web	16AUG2015	Closed with non-monetary relief
40	Loan modification, ...		[redacted] is my loan company , this company rejecte...		TX	750XX	Web	15JUN2015	Closed with explanation
41	Application, origin...		1.) At the XXXX time home buyer class, I was told that the One-mo...		MA	021XX	Web	21JUL2015	Closed with explanation
42	Loan servicing, pa...		[redacted] not send me the actual signed modification agreeme...		CA	908XX	Web	25SEP2015	Closed with explanation

Figure 1: Sampling of Complaint Data Including Consumer Narrative and Company Response to Consumer

We specify a target category variable of COMPANY_RESPONSE_TO_CONSUMER as part of the SAS Contextual Analysis project definition to tell SAS to model Boolean textual rules against the CONSUMER_NARRATIVE. These textual rules differentiate term and phrase combinations that appear in each category value from other term and phrase combinations that appear in the other category values. In the context of this data set, one of the category values for COMPANY_RESPONSE_TO_CONSUMER is “Closed with Monetary Relief”. SAS Contextual Analysis can tell us what common terms, phrases, and term-phrase combinations are most often associated with monetary relief, but not typically not associated with the other category values, such as “Closed with Explanation”. Two examples of these phrases or terms are mentions of specific retail organizations not mentioned anywhere in the structured data, or the term GFE (Good Faith Estimates). This automated rule-building technology helps the analyst by characterizing the complaints that result in monetary relief, and backs up the analyst with quantitative analysis. A researcher could manually generate a similar categorical taxonomy to capture these instances, but creating this taxonomy from scratch is highly time consuming, compared to the method just described, which produces results in minutes. In addition, the researcher is likely unaware of all the patterns in the textual data. The power of the approach presented here is that it automatically generates a taxonomy that fits and describes each data set, and that taxonomy can subsequently be refined using subject matter expertise. Refining a taxonomy makes much better use of the subject matter expert’s time and resources than creating a taxonomy from scratch. Figure 2 illustrates the process of generating a project against the data using the COMPANY_REPONSE_TO_CONSUMER as a categorical variable and the CONSUMER_NARRATIVE as the text variable.

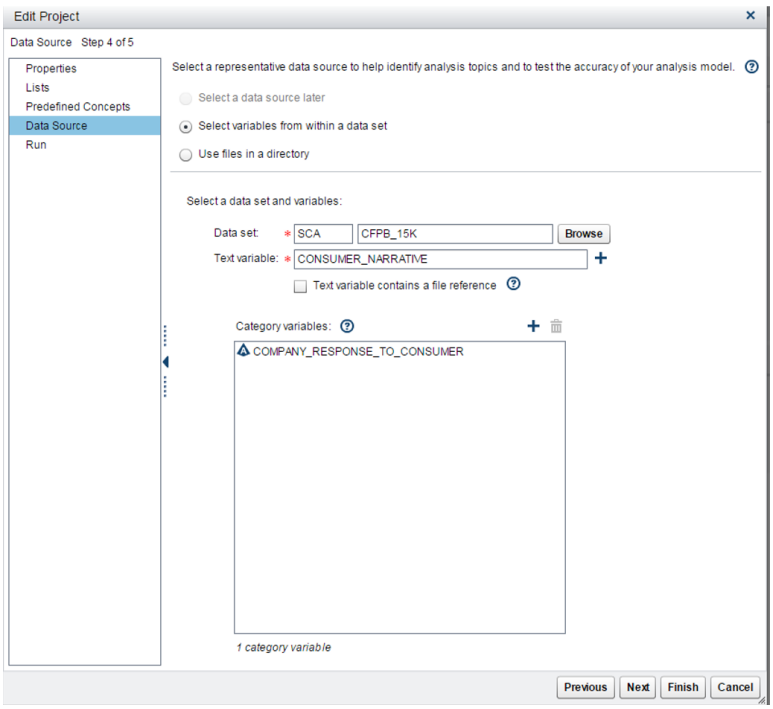


Figure 2: Defining a New Project in SAS Contextual Analytics, Including a Category Variable

SAS Contextual Analysis includes a number of exploratory capabilities, including term and topic exploration. In this paper, we focus on its capability to generate textual rules against categorical data, and the subsequent scoring and augmentation of the original data set using these rules overlaid by a sentiment model. For a further study of terms and topic exploration in the context of a research-oriented data set, please see SAS Global Forum paper 061-2014, "Uncovering Trends in Research using Text Analytics with Examples from Nanotechnology and Aerospace Engineering." [2]

Figure 3 illustrates the textual rules that SAS Contextual Analysis auto-generates against the CONSUMER_NARRATIVE textual column for the COMPANY_RESPONSE_TO_CONSUMER value of "Closed with Monetary Relief". These rules are combinations of certain terms and phrases that appear in the narrative for complaints that result in monetary relief but that don't tend to occur for the other complaint disposition codes. These terms and phrases can be used to auto-classify new narratives as being ones that are likely to result in monetary relief, which enables analysts to prioritize the complaints that they receive. The terms and phrases are also used in subsequent steps to characterize and sub-divide the various terminology that surrounds financial compensation, which enables the analyst to explore these divisions separately and identify trends and patterns in monetary relief.

Figure 3: Rules Generated by SAS Contextual Analysis Related to Monetary Relief

Each Boolean rule consists of terms and phrases joined by "&," indicating "AND," as well as modified by a "~," indicating "NOT." In addition, each of the terms presented include all stemmed versions of those terms. For example, the term "steal" is representative of "stole" and "stolen" as well. Putting it all together in an example, the rule "fee & ~modification & bank & steal" indicates a complaint containing all of the terms "fee", "bank", and "steal", or different stemmed versions of these, so "fees", "banks", and "stolen" would suffice. Also, the term "modification" or any of its stemmed versions must not be present, indicating that this rule primarily applies outside of the mortgage process, where loan modifications are common in the complaint data.

The Boolean rules are represented by a colored bar, which includes blue, yellow, and red components. The blue component of the bar represents cases where a rule correctly matches the given event type. These are true positives. The yellow component of the bar represents cases where the rule also matches for a different event type. These are false positives. The red component of the bar is primarily applicable to the "Closed with Monetary Relief" level, rather than the individual rule level. At the "Closed with Monetary Relief" level, the red component of the bars represents cases where SAS Contextual Analysis is unable to define a consistent rule to differentiate these complaint dispositions from the other disposition types such as "Closed with Explanation" or "Closed with Non-Monetary Relief".

Rules that are generated against the narrative take a variety of forms when interpreted, all of which are indicative of some trend related to monetary compensation following a consumer complaint. They can be indicative of retail organizations that are supported by financial organizations via a company debit card, for example. This can be powerful information, because no retail organization is named anywhere in the structured complaint data. However, if the customer service for the card in question is very poor for say, ACME retail organization, individuals are apt to complain about ACME in their free-form complaints, and this connection can be made only through text analysis of the free-form narrative.

Rules are also indicative of problems with the supporting financial organizations themselves, such as account trouble that is related to a particular bank that supports a variety of retail debit cards. This might be due to widespread poor customer service involving this bank. Finally, these rules can be indicative of bank or lending practices, such good faith estimates, which often result in some type of monetary compensation when they are mentioned in a complaint. It is possible that lending organizations are taking advantage of the complication of good faith estimate statements. They could be giving confusing or inaccurate statements in order to hide fees. This last example in particular is valuable. Text Analytics quantitatively depicts that there is a practice by lending institutions that is likely being abused or misused, and provides the opportunity for an overseeing organization to take action.

DATA PREPARATION FOR VISUAL ANALYSIS

As mentioned, we are interested in both the categorical scoring and sentiment scoring of the data set. This involves a few steps, and two segments of SAS code.

1. First, we use a SAS code segment to invoke the categorical taxonomy model generated with the COMPANY_RESPONSE_TO_CONSUMER structured data field. We score against the full data set of CFPB complaints, as opposed to the 15K sample that we used to generate the model. This scores each document for, among other things, rules indicative of complaints that lead to monetary relief. To generate this code and subsequently score our complaints data set, we used methods similar to those described in the DATA PREPARATION FOR VISUAL ANALYSIS section of the SAS Global Forum paper "Extending the Armed Conflict Location and Event Data Project with SAS® Text Analytics"[4]. This method extends the out-of-the-box score code to produce a categorical hierarchy suitable for visual exploration. Please refer to the paper for information on augmenting the out-of-the-box code provided by SAS Contextual Analysis for this purpose.
2. Second, we obtain a code snippet in SAS Contextual Analysis by selecting the Sentiment Code option from the View drop-down menu. As discussed previously, this option uses a generalized sentiment model, but it could also leverage a model built from SAS® Sentiment Analysis Studio. The out-of-the-box DS2 code needs to be modified only slightly from its original format to designate input and output SAS data sets. Figure 4 shows how to access the sentiment scoring code in SAS Contextual Analysis, which is subsequently modified in a SAS programming environment.

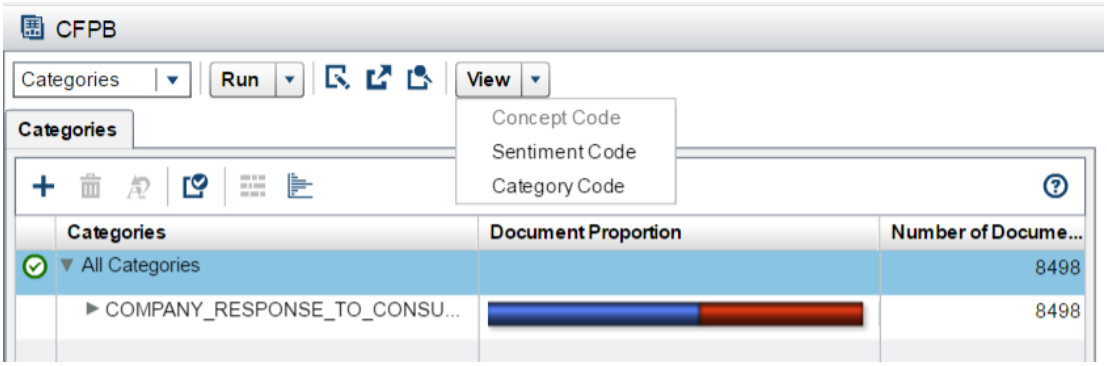


Figure 4: Option to Depict Sentiment Code in SAS Contextual Analysis

In order to define input file locations for the environment used in this project, we modify the early lines of the out-of-the-box sentiment code to look like the following:

```
/*************************************************************
* SAS Contextual Analysis
* Sentiment Score Code
*
* Modify the following macro variables to match your needs.
*************************************************************/
/* check if the variables were defined elsewhere - this is used for
 embedding code into SAS Text Miner */
%sysfunc(ifc(%symexist(tm_defined_vars),, %nrstr(
/* the path to the directory containing the data set you would like to
 score */
%let lib_path= D:\data\sca;
/* the data set you would like to score */
%let input_ds = _my_lib.cfpb_full;
/* the column in the data set that contains the text data to score */
%let document_column = CONSUMER_NARRATIVE;
)));
```

In order to save the output sentiment data for the environment used in this project, we add the following lines to the end of the sentiment code:

```
libname outputlib 'D:\data\sca\out';
data outputlib.cfpb_sentiment_scored; set &output_ds;
run;
```

3. Finally, using SAS Enterprise Guide, we simply join all the fields of the categorical results table from step 1 above with the _sentiment_probability_ field from step 2. We join for every row from table 1 against the ID column, which is present in both tables.

AD HOC EXPLORATION AND MODELING

We load the SAS data set of complaints, which includes newly generated hierarchical category scoring and document level sentiment, into SAS Visual Analytics for exploration, modeling, and reporting. SAS Visual Statistics, a set of predictive capabilities within SAS Visual Analytics, provides interactive decision trees that illuminate differentiating trends in the data. In this example, we use a decision tree to highlight textual rule combinations that are indicative of various disposition codes, such as monetary relief. To do this, we set the COMPANY_RESPONSE_TO_CONSUMER as a target, and use only the textual rules generated from the text analytics exercise associated with each event as input to the model. This will highlight branches of the decision tree where various phrases present in the narrative correlate with the different disposition codes of COMPANY_RESPONSE_TO_CONSUMER. If you re-create this example, you should also consider using the pre-existing structured data that is associated with each complaint in conjunction with the newly generated structured rules. These combinations also yield illuminating results. See Figure 5 for a high-level depiction of the generated tree, whose resulting bins characterize the data in meaningful ways. In particular, it is interesting to note branches of the tree that result in predominant dispositions other than “Closed with Explanation”, which is the overall predominant disposition. Figure 6 zooms in on the top of the decision tree to visually depict how, in general, complaints that result in monetary relief, denoted by a red bar, are significantly less frequent than ones with a “Closed with Explanation” disposition, denoted by a green bar..

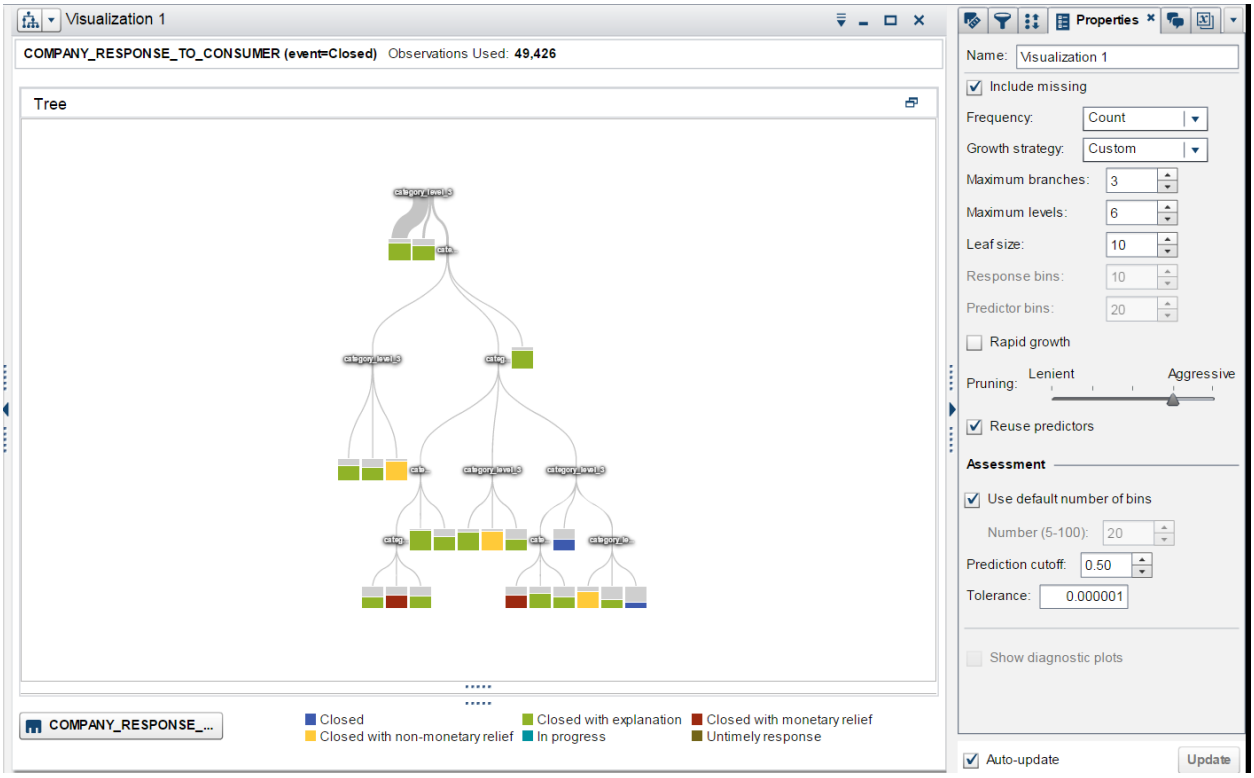


Figure 5: High-Level Depiction of COMPANY_RESPONSE_TO_CONSUMER Decision Tree

Figure 6: Top of the COMPANY_RESPONSE_TO_CONSUMER Decision Tree

Figure 7 shows how the proportion of responses that are closed with monetary relief dramatically increases as we traverse certain branches of the tree. There are three rules depicted in the highlighted node, including one that mentions the terms “fee”, “bank”, and “steal”, where “modification” is not mentioned. On the right hand side of the tree, note how one of the nodes is associated with “gfe” or good faith estimates, and that this warrants its own node that is strongly correlated with monetary relief. Insight garnered from this step is useful in explorations using the interactive reports.

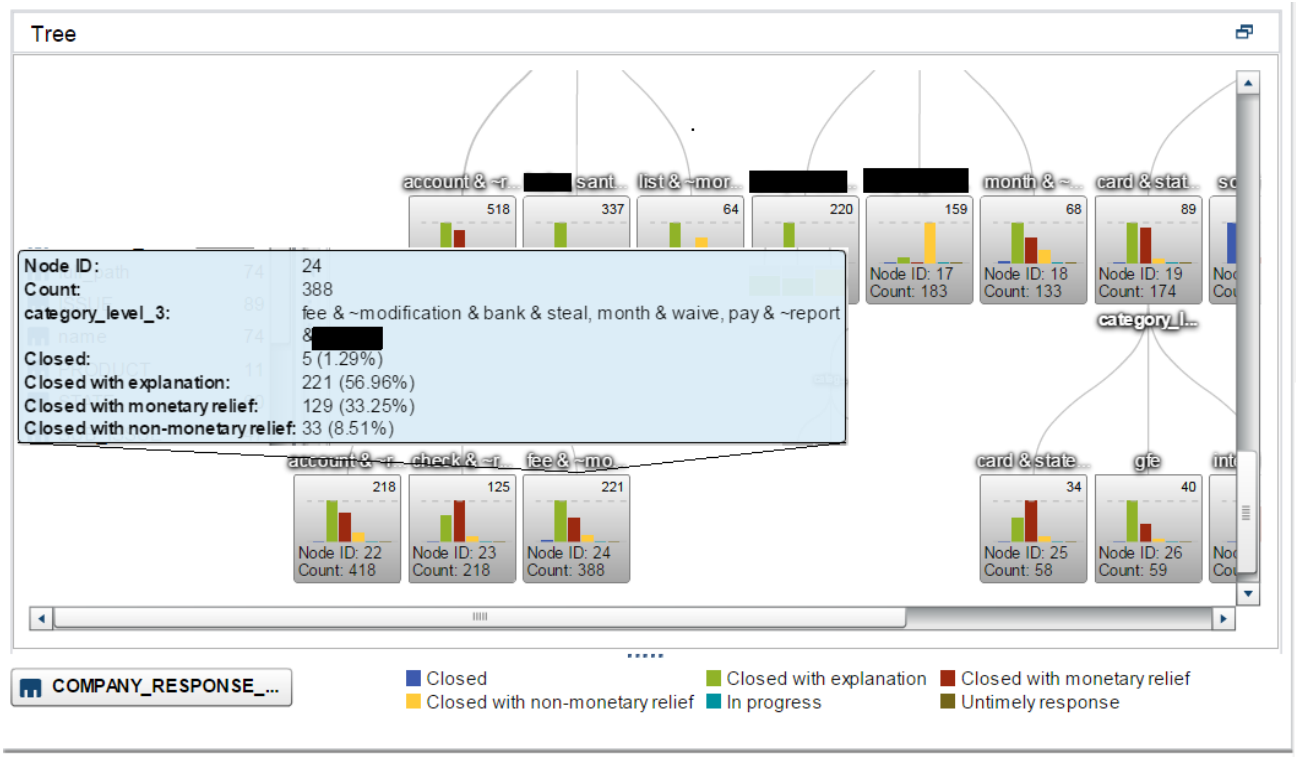


Figure 7: View of Decision Tree Highlighting Cases Strongly Associated with Monetary Relief

INTERACTIVE REPORT GENERATION AND USE

Interactive reporting enables the end-user analyst to explore the pre-existing data for complaints enhanced with the sentiment and rules that are generated from text analytics. This allows the analyst to sub-divide and prioritize exploration avenues according to the auto-categorization, while being guided by the relative levels of sentiment toward each of the categories. The analyst uses a dashboard, which depicts the rules and sentiment information in a tree map, the geospatial information and sentiment in a geospatial map, and information surrounding structured data issues and products in a pie chart. Links are provided from the tree and tile maps to drill down into the textual complaint data in a separate drilldown report. This drilldown report also includes a time series line chart so that analysts can observe trends over time.

Figure 8 depicts the use of a dashboard to explore one of the rules, “gfe” (good faith estimate). We have already determined in both SAS Contextual Analysis and SAS Visual Statistics that complaints containing this term are strongly correlated with monetary relief actions. From the dashboard, we can assess positive and negative sentiment, particularly at the geospatial level, so that we can begin to evaluate US states that might have been more particularly affected by misuse of the good faith estimate. We can also determine that this rule, as would be expected, entirely relates to mortgage products, and is associated with various issues surrounding the mortgage process.

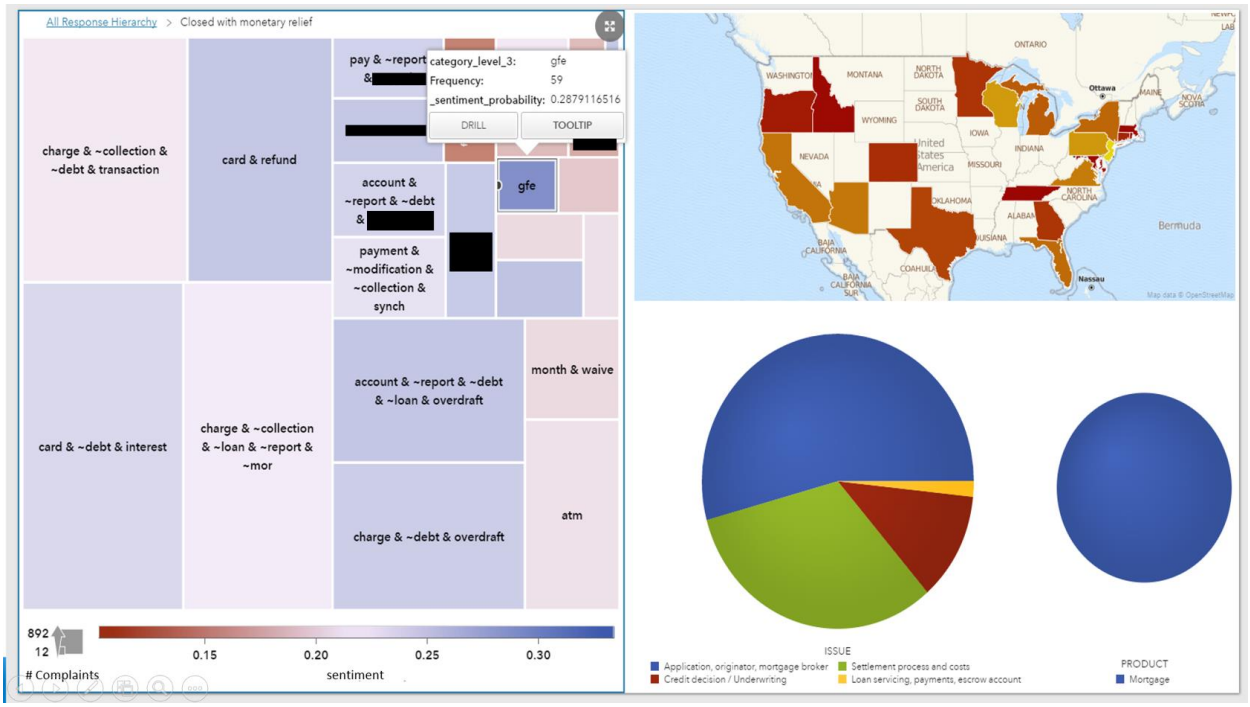


Figure 8: Dashboard Depicting Information Surrounding Good Faith Estimates in Complaint Data

Drilling down into the complaint data, as shown in Figure 9, we can see from the timeline that there are no new complaints related to good faith estimates after August 2015. This is because Congress directed the CFPB to combine two forms, the “Initial TIL Disclosure” and the “Good Faith Estimate” into the new “Loan Estimate and Closing Disclosure” form. This change took effect on October 3, 2015[6]. This new form is intended to be more transparent to the consumer, and therefore more difficult for financial organizations to misuse. The question remains, if analytics had been part of the assessment, could this ultimate decision to protect the consumer have been reached sooner?

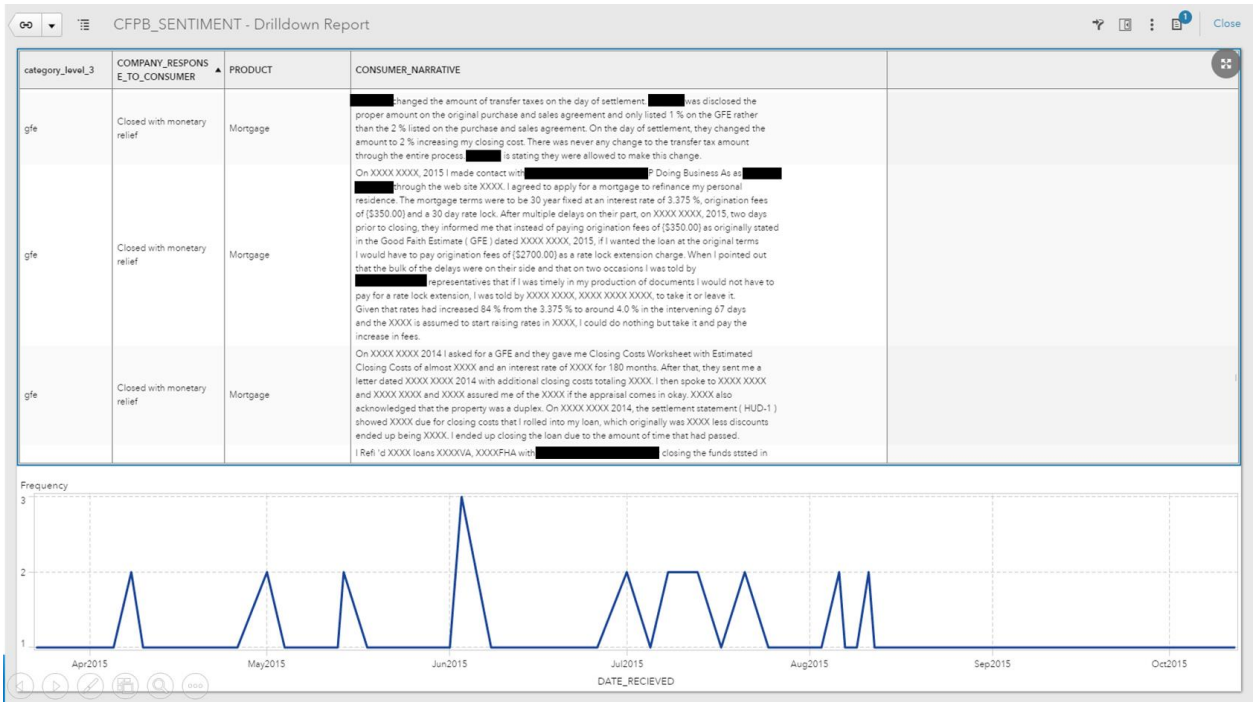


Figure 9: Drilldown Report on Good Faith Estimate Complaint Narratives Including Timeline

One rule mentioned previously was indicative of “fees,” “stealing,” and “banks,” where the term modification is not mentioned. A dashboard depicting this rule is shown in Figure 10. By using this dashboard to explore the related complaints, we discovered that these complaints are split into two groups. One group includes cases where individuals claim that their banks are stealing from them based on the various fees these banks assess. The other group covers cases of identity theft and associated fees. This provides excellent feedback into the auto-generated taxonomy. SAS Contextual Analysis users can take this information and refine the rules for this particular sub-category to distinguish the identity theft cases from the excessive fees cases. After rescoring, they can better explore trends associated with the two new sub-categories. This example illustrates how statistical analysis goes hand-in-hand with capabilities to leverage subject matter expertise, refine the rules-based taxonomy, and better enable search and discovery.

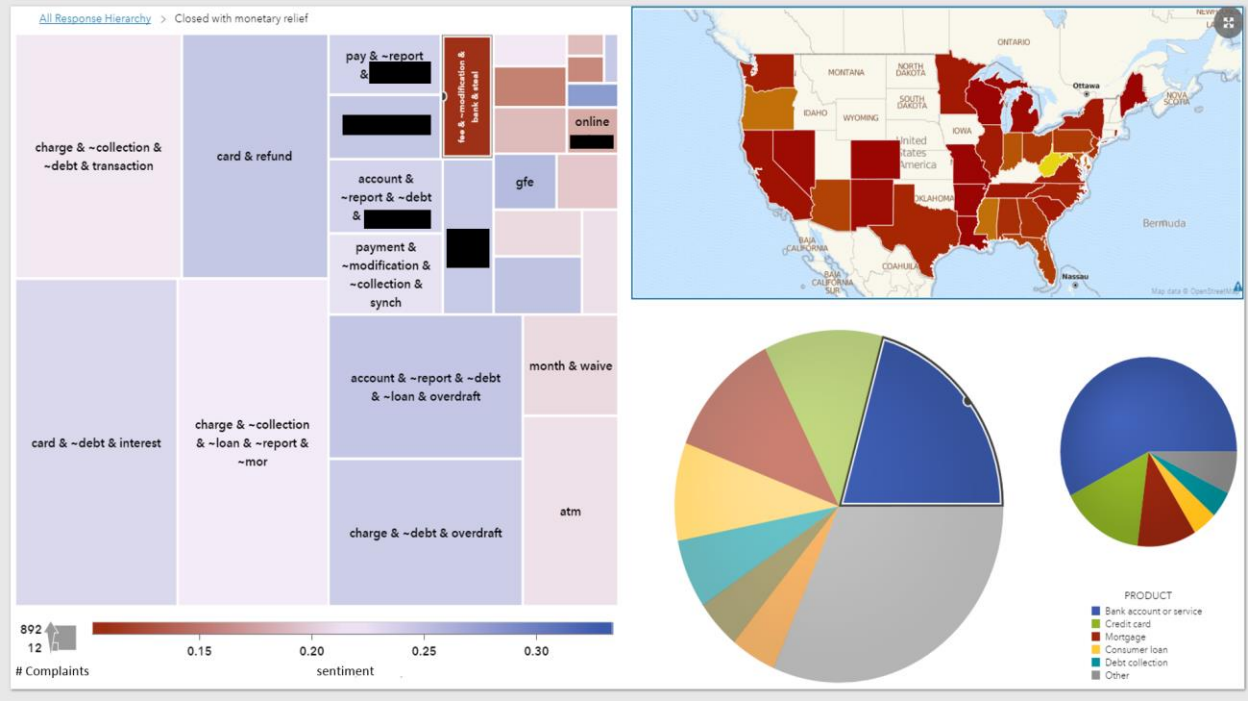


Figure 10: Dashboard Depicting Information Surrounding Excessive Fees and Identify Theft

CONCLUSION

In summary, we showcased a repeatable process that combined the benefits of both statistical and classification-based text analytics against the Consumer Financial Protection Bureau complaint data in order to assess these complaints for areas that trended toward monetary relief. In doing so, we identified several patterns, including one pattern that highlighted flaws related to good faith estimates, which is a part of the mortgage loan process on which CFPB has taken action. The quantitative analysis presented in this paper serves to validate the actions of the CFPB.

Overseeing organizations can use the methodology presented in this paper to improve time to value and quality of analysis when assessing complaint data. Financial organizations who support retail organizations through, for example, a debit card should use this methodology to help assess the quality of their customer care and their organizational satisfaction. Retail organizations should pay attention to assess whether the financial organization that is supporting their company is negatively impacting their brand.

The methodology depicted here is widely applicable. It relies on having substantial rows of data, generally 500 or more, in the context of a target variable of concern related to the text. The length of the text should typically be between a single line and several pages in order for this methodology to produce actionable information. An additional area where we can apply these capabilities is generating taxonomy around stand-up clinics after natural disasters. By analyzing a subset of the medical issues for which individuals are seen at these clinics, and for which we have a diagnosis code that differentiates between issues such as respiratory issues and bodily injuries, we can generate a taxonomy that characterizes these various issues in more detail. We can identify actionable information such as the type and quantity of materials that are needed at these clinics in order to ensure that medical needs are met for disaster survivors. For more areas of text analytics and subsequent visualization application, please see the SAS Global Forum paper, “Text Analytics in Government: Using Automated Analysis to Unlock the Hidden Secrets of Unstructured Data[7].

The methodology also stands up to manual coding of textual data. Organizations that leverage machine learning capabilities can label social media content, for example, with tags that differentiate anything that needs to be analyzed. For example, an analyst could tag 1,000 Twitter entries related to food poisoning

with a flag that differentiates actual instances of food poisoning. This can help build a model that more accurately identifies these instances from more generic talk. In an implementation of this example, the analyst might discover that certain terminology tends to surround the actual instances of food poisoning, such as the mention of a time-related term such as “hour(s)” or “day(s)”. Because this model also characterizes the tangible instances of food poisoning, an analyst might be interested in exploring all the cases in which the term “hour(s)” is mentioned because these might be more immediate. This is important if the analyst is looking for indications of a suspected epidemic.

A semi-automated feedback loop would enhance a machine-learning solution. In the context of this paper, this feedback loop is self-contained in SAS Contextual Analytics, enabling the user to modify the auto-generated categorical rules or to provide new complaint information in the context of the existing auto-generated rules. Feedback occurs when the user subsequently re-runs the models. Extension of this capability is something that should be considered, and could assist in determining whether documents end up fitting well in particular categorical buckets. Hence, a user of a visual exploration system would be able to dynamically re-assign primary reasons for monetary relief, and these re-assignments will be taken into account by the modeling software the next time the models are run.

The taxonomy, which was auto-generated and modified with subject matter expertise, could be used for auto-coding of new complaint data. It might be helpful for a reviewer to see, for example, that a new complaint matched a historical pattern such as issues with card refunds. It might also be helpful to see recommendations and contextual information surrounding the complaint. An example is “Here are a number of additional recent complaints matching the general pattern for the current one, related to card refunds, and here also are the general disposition for these complaints, such as how often they resulted in some form of monetary relief.” All of this information would assist in both the speed and quality of processing new complaints. This would not be difficult to implement using the SAS capabilities presented in this paper.

Finally, many data sources are not as structured as the data we obtained from the CFPB. For a demonstration of tokenization on a data set of large documents and subsequent analysis, see the SAS Global Forum papers “Getting More from the Singular Value Decomposition (SVD): Enhance Your Models with Document, Sentence, and Term Representations”[8] and “Star Wars and the Art of Data Science: An Analytical Approach to Understanding Large Amounts of Unstructured Data”[9].

REFERENCES

1. Website of the Consumer Financial Protection Bureau. Available http://www.consumerfinance.gov/. Accessed on February 1, 2017.
2. Sabo, Tom. 2014. “Uncovering Trends in Research Using Text Analytics with Examples from Nanotechnology and Aerospace Engineering.” *Proceedings of the SAS Global Forum 2014 Conference*. Cary NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings14/SAS061-2014.pdf.
3. Sabo, Tom. 2015. “Show Me the Money! Text Analytics for Decision-Making in Government Spending.” *Proceedings of the SAS Global Forum 2015 Conference*. Cary NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings15/SAS1661-2015.pdf.
4. Sabo, Tom. 2016. “Extending the Armed Conflict Location and Event Data Project with SAS® Text Analytics.” *Proceedings of the SAS Global Forum 2016 Conference*. Cary NC: SAS Institute Inc. Available https://support.sas.com/resources/papers/proceedings16/SAS6380-2016.pdf.
5. “Consumer Complaint Database" Consumer Financial Protection Bureau. Available http://www.consumerfinance.gov/data-research/consumer-complaints/#download-the-data. Accessed on February 1, 2017.
6. “How we improved the disclosures” Consumer Financial Protection Bureau. Available http://www.consumerfinance.gov/know-before-you-owe/compare/. Accessed on February 1, 2017.

7. Sabo, Tom. 2014. SAS Institute white paper. “Text Analytics in Government: Using Automated Analysis to Unlock the Hidden Secrets of Unstructured Data.” Available http://www.sas.com/en_us/whitepapers/text-analytics-in-government-106931.html.
8. Albright, Russ. Cox, James. Jin, Ning. 2016. “Getting More from the Singular Value Decomposition (SVD): Enhance Your Models with Document, Sentence, and Term Representations” *Proceedings of the SAS Global Forum 2016 Conference*. Cary NC: SAS Institute Inc. Available https://support.sas.com/resources/papers/proceedings16/SAS6241-2016.pdf.
9. Osborne, Mary. Maness, Adam. 2014. “Star Wars and the Art of Data Science: An Analytical Approach to Understanding Large Amounts of Unstructured Data.” Proceedings of the SAS Global Forum 2014 Conference. Cary NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings14/SAS286-2014.pdf.

ACKNOWLEDGMENTS

Thanks to Sophia Victor for providing insights into the visualizations that would best convey the text analytics results of this project to a wider audience. Thanks to Emily McRae, Sophia Victor, and Mary Beth Simmons for reviewing and refining this paper.

RECOMMENDED READING

- Chakraborty, G., M. Pagolu, S. Garla. 2013. *Text Mining and Analysis; Practical Methods, Examples, and Case Studies Using SAS®. SAS Institute Inc.*
- *Reamy, Tom. 2016. Deep Text; Using Text Analytics to Conquer Information Overload, Get Real Value from Social Media, and Add Big(ger) Text to Big Data*. Medford NJ: Information Today, Inc.

CONTACT INFORMATION

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

Tom Sabo, Principal Solutions Architect
1530 Wilson Blvd.
Arlington, VA 22209
SAS Federal LLC
+1 (703) 310-5717
tom.sabo@sas.com
@mrTomSab
http://www.sas.com