Warranty/Performance Text Exploration for Modern Reliability

Scott Wise, SAS Institute, Inc.

Key Words: Text Mining for Reliability, Warranty and Performance Analytics

ABSTRACT

With the increasing amount of warranty/performance data available to reliability professionals, we have a golden opportunity to learn much more about our designs/products and come up with better reliability/maintainability models. However, this additional info often comes in the form of unstructured text from technicians, researchers, and customers. Inability to find groupings and trends in this data can hide the real impacting issues that our traditional reliability data won't be able to uncover. This session will seek to show how modern analytic software can help provide a way to quickly and easily explore this unstructured text data to gain better insights into our true reliability. We will cover the basics of how to visualize, group, analyze, model and ultimately find trends using warranty/performance unstructured text data.

1 INTRODUCTION

In the age of bigger and better access to information, reliability practitioners now have access to even more data. By adding basic text mining methodology to their tool set, one can unlock unstructured text about project/service warranty and performance that previously went unused. In the past, these techniques were often only taught and applied by statisticians who studied and worked in the growing field of text mining. However, modern software has enabled all professionals to easily apply these basics in a way to easily explore text data, even in ways that can help build better and deeper reliability models.

2 PROBLEM DESCRIPTION

Often when we collect field warranty, there is more information available than just what and when things failed along with possibly a code for the type of failure. Included with the data is often text notes and descriptions about failure that can add a lot more to our understanding of reliability. However, in the past this unstructured text data was largely overlooked and only used when one was looking for more information about an individual failure entry. However, modern analytic software can now help analyze this unstructured text data by identifying patterns and trends on failure causes that can be directly incorporated into our reliability models. This can help in two ways. First is by uncovering new failure trends to help better focus our reliability activities. The second is by allowing for the creation of more accurate reliability practices by incorporating these findings into our models.

3 TECHNICAL DESCRIPTION

To incorporate text exploration skills into your tool kit, reliability professionals will need to learn some new methodology that may have not been part of their past education or daily practices. This involves incorporating methods to summarize, prepare, analyze, visualize and model text data. By learning these basics, we can explore our warranty/performance unstructured text to improve our reliability.

3.1 Text Exploration Definitions

As with any new analytic area, there are a few new helpful definitions that are utilized to better describe text analytics. The document is used to describe the individual body of text you are analyzing (basically each row of text in our data). A corpus is then a set of these documents (all rows in a text column of our data). Lastly a term describes the unit of analysis (A single or a multiword phrase). So, to put this all together, if we have a *corpus* (or column of unstructured text data), we can analyze the *documents* (or rows of text words) to find *terms* (common words and phrases) that are of interest. See Figure 3.1 below for an example of where these definitions fit within an example of unstructured text warranty/performance field data.

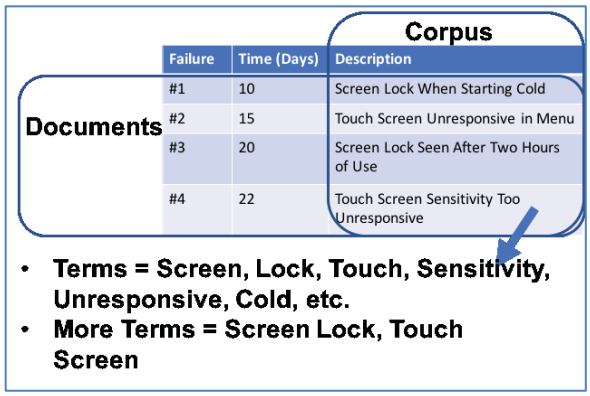

Figure 3.1 Text Exploration Definitions Example

3.2 Business Case Intro

To show how Text Exploration can add value to reliability, we will utilize typical computer warranty/performance failure data. This data provides 197

978-1-5090-5284-4/17/$31.00 ©2017 IEEE

rows of failure data from different stages of product testing. The following columns are available for analysis: Model, Issue ID, Phase Found, Severity, Submit Date/Time, Close Date/Time, Time Delta, and Text. While there are several graphs and tables we can make using the Model, Severity, Phase, Time Delta, etc. we currently cannot make use of the Text column (unstructured text comments about each failure) which contains valuable information to understand product reliability. See Figure 3.2 to see a snapshot of the data and graphing on several areas of interest. Note that we will be using a modern analytic software package called JMP Pro Statistical Analytic Software (Version 13.2) from the SAS Institute, Inc. for this analysis.

Model	Phase Found	Severity	Close Date/Time	Time Delta	Text
Mango B	Develo...	1	05/08/2003 4:46 PM	:9:00:44	MP Sync fails in Mango
Mango B	Develo...	1	05/16/2003 1:37 PM	:14:04:40	Bringup: FSMI_USB_POST_INIT causes system loc...
Mango B	Develo...	1	05/08/2003 4:47 PM	:6:07:42	Bringup: Incorrect chipset initialization values
Mango B	Develo...	1	05/08/2003 4:47 PM	:6:07:11	Bringup: Failed to identify SATA HD and CDROM drive
Mango B	Develo...	1	05/08/2003 4:47 PM	:6:07:02	Failed to boot W2K on SATA HD
Mango B	Develo...	1	05/08/2003 4:48 PM	:6:07:02	W2K does not recognize logical processor with Hyp...
Mango B	Develo...	1	05/08/2003 4:48 PM	:2:23:31	Windows2000 blue-screens saying BIOS is not ACP...
Mango B	Develo...	1	05/29/2003 10:08 AM	:4:17:32	W2K hangs with blue screen during the installation ...

Figure 3.2 Case Data Snapshot & Graph

3.3 Summarizing Warranty/Performance Text

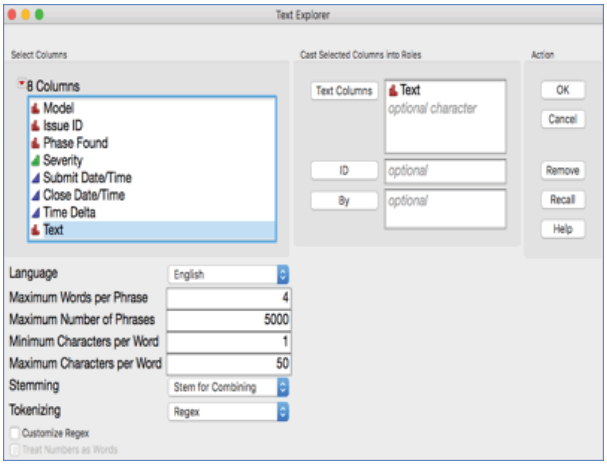

Figure 3.3 Common Text Exploration Setup

The first step is to *Summarize* our unstructured text data into a list of the most frequently occurring words. Modern analytic software allows you to go beyond just a simple count among all the words available. *Tokenizing* algorithms parse the text terms into the most common terms that fit common patterns. These patterns can be things like domain names, money words, html links, time, numbers, etc. Often done at the same time is *stemming* which combines terms that share the same root word. So, for the word "fail," the terms fail, fails, failing, failed, etc. would be put under the root word. See Figure 3.3 for a view of a common text exploration setup screen. Note that among the Phrase and Word controls, the software allows for Tokenizing (Regex selected to access the built-in regular expression library) and Stemming (Stem for Combining to merge needed words with endings into common root words).

3.4 Preparing Warranty/Performance Text

The results of this text analytics lead us to *Preparing* the text data. This involves looking over the list of generated common terms and phrases from the summarize step. It is at this point at which we will use our domain expertise about the failure/performance data to *build on terms and phrases*. Modern analytic software allows you to easily combine those phrases and terms that should belong together into one. Therefore, if you have several phrases or terms that all point to a known failure, you can combine these into one for better categorization. For example, if we saw phrases such as "no data connection, data streaming failure, and data connection broke" you would want to combine these under one common phrase like "data streaming" to get a larger term count for these common failure types. Additionally, software allows for the adding of *stop words* where you can remove those terms and phrases that are noisy and would just confuse the analysis. Very common stop words would be pronouns, like "and, the, or, etc." that don't add much value to exploring your text.

When we ran initial Text Exploration on our business case data, we were able to look at the list of terms and phrase. The software automatically removed common stop words, so we don't see any pronouns like "is, and, but, etc." in the list. Next, we used our engineering domain knowledge and realized that the phrases "ac power recovery, ac power, and power recovery" all are referring to the same failure type. So, we asked the software to combine these phrases when doing further analysis. See Figure 3.4 for a look at the term and phrase list after combining "power" related phrases.

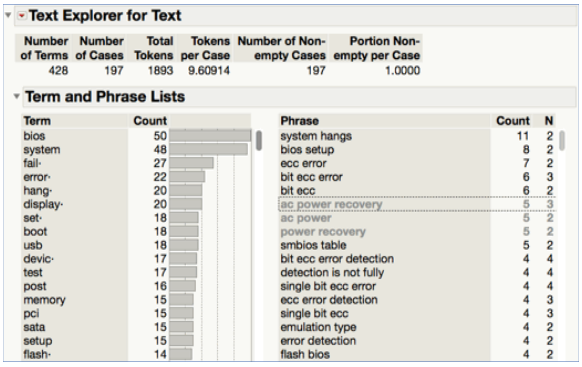

Figure 3.4 Terms & Phrases List After Combining "Power" Related Phrases

3.5 *Visualizing Warranty/Performance Text*

Visualizing the data is the next step in our text exploration. While this can take many graphical forms, the most popular one is the Word Cloud. This creates a graph where the largest terms in the data are plotted and arranged in a cloud-like visual. The terms are then sized by frequency, making the largest occurring terms really stand out on the graph. A color gradient can be added for another factor to further make important terms stand out in the graph. Visualizations like Word Clouds can help us make last minute preparations to our terms. For our case data, we asked for a word cloud colored by the severity rating. Now we can more easily see with the large font the high-count failure terms and how they compare in terms of the severity color gradient. In this case, "Bios" looked like a very interesting failure type for future reliability analysis.

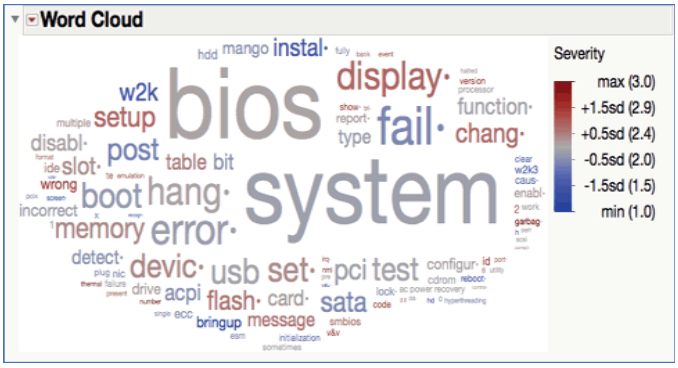

Figure 3.5 Word Cloud

3.6 *Analyzing Warranty/Performance Text*

We are now ready for *Analyzing* the text data. To do so modern analytic software will make use of the terms we found in the preceding steps to make an analysis ready table of our summarized and prepared text terms. This can be as simple as making an indicator column of 0s and 1s on one term of interest, where a 1 indicates that that row contains the term. Or we can prepare for a bigger analysis by making indicator columns for all terms in our text analysis, creating what is called a Document Term Matrix (DTM). In the case of a DTM, we can perform further dimension reduction to reduce down to a subset of terms that have the strongest weights in our analysis.

A multivariate algorithm Latent Semantic Analysis with SVD does a sparse singular value decomposition on the DTM. Then we can save the reduced DTM for further analysis. This approach is similar to the popular PCA (Principle Components Analysis). See Figure 3.6.1 for a snapshot of the reduced DTM. The resulting SVD Scatterplot Matrix allows us to graphically see documents (rows) and terms that share the same dimensional direction and possibly can be grouped together. Further Topic Analysis also allows us to group common terms into a list of common themes and is similar to Factor Analysis. Other multivariate methods such as clustering on terms and documents are also available. These methods allow us to further explore and discover trends in our warrant/performance data. With our business case data, we can see that "Bios" related failures not only stand out dimensionally on the SVD Matrix, but are associated with several related descriptive terms such as "IPMI & DRAC3/XT" in Topic 1 of the Topic Analysis. See Figure 3.6.2 SVD Matrix and Topic Analysis.

86/0 Cols

	Text	bios Binary	system Binary	fail Binary
1	MP Sync fails in Mango	0	0	1
2	Bringup: FSMI_USB_POST_INIT causes system loc...	0	1	0
3	Bringup: Incorrect chipset initialization values	0	0	0
4	Bringup: Failed to identify SATA HD and CDROM drive	0	0	1
5	Failed to boot W2K on SATA HD	0	0	1
6	W2K does not recognize logical processor with Hyp...	0	0	0
7	Windows2000 blue-screens saying BIOS is not ACP...	1	0	0
8	W2K hangs with blue screen during the installation ...	0	0	0
9	Bringup: Bus parking enable/disable not functional	0	0	0
10	BIOS tries to do ESM initialization even though there...	1	0	0

Figure 3.6.1 Reduced DTM Snapshot

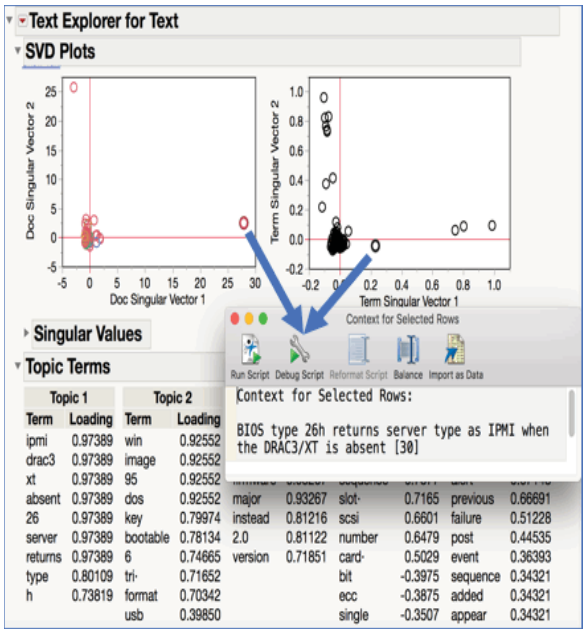

Figure 3.6.2 SVD Matrix & Topic Analysis

3.7 *Modeling Warranty/Performance Text*

Our last step is to incorporate what we have learned through our text exploration of the warrant/performance data and integrate it into our Reliability Modeling. Term indicator columns from the saved DTM can be brought directly into models for reliability modeling of failure causes. In a typical Fit Life Reliability model, the indicator column can also serve as the censor data in the model. As well we can build additional modeling using indicator columns or even topics to further predict expected failure/performance on a dependent variable.

Looking at our business case data, we decided to first create a Reliability Life Distribution model around the "Bios"

text term. Since we have time to failure ("Close Date/Time"), but no censored data, we can treat the "Bios" indicator column from the DTM where we have a 1 in the column indicating those rows where Bios shows up in the failure and a 0 in the rows where "Bios" does not show up. After running a Life Distribution for Reliability, our software indicated that picking a Frechet underlying distribution would provide the most accurate reliability fit. Lastly, we are ready to use the graphical profilers, especially the Quantile Profiler, to interactively answer reliability questions like finding the time at which we expect to see 50% of the Bios issues occur (the B50). See figure 3.7.1 for the Reliability Model Setup and Analysis.

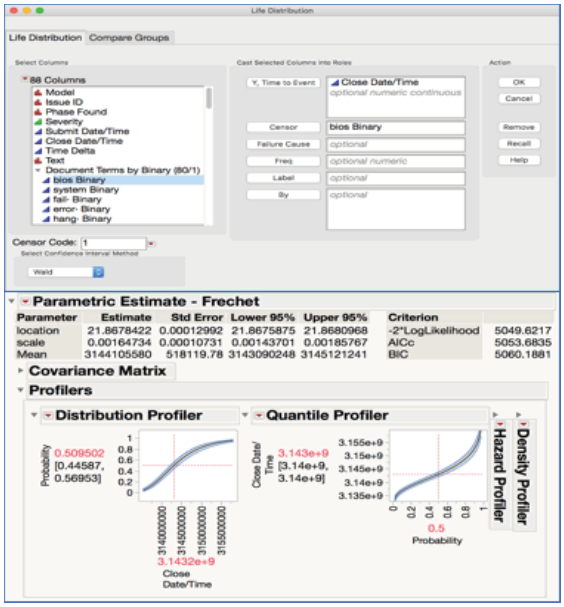

Figure 3.7.1 Reliability Model Setup & Results

Another advantage of doing the Text Exploration is the ability to also make Predictive Models based on a dependent output. Looking at our business case, we can utilize all the failure terms in our DTM and see if we can build a model to predict severity of the failure. By using an Ordinal Logistic Regression, we can reduce down from our original 80 to 11 important term factors in the model (those with p values below .01). See Figure 3.7.2 for the Predictive Model Setup and Analysis. Lastly, we can view results with a Profiler to see what the severity would be if we activate failure terms such as the word "Setup." See Figurer 3.7.3 for a view of the Predictive Model Results Profiler.

4 CONCLUSION

In conclusion, learning and utilizing text exploration techniques unlocks the power of unstructured text that often accompanies our warranty/performance data. This gives us the ability to learn much more about our designs/products and come up with better reliability/maintainability models. Modern analytic software makes it easy for reliability engineers to use these methods, without having to be a statistician. By following Basic Text Exploration steps of Summarizing, Preparing, Visualizing, Analyzing and Modeling, one can find groupings and trends in this data that our traditional reliability model won't uncover. Incorporated into our tool kit, this additional info unlocks ways to truly meet the challenges of modern reliability.

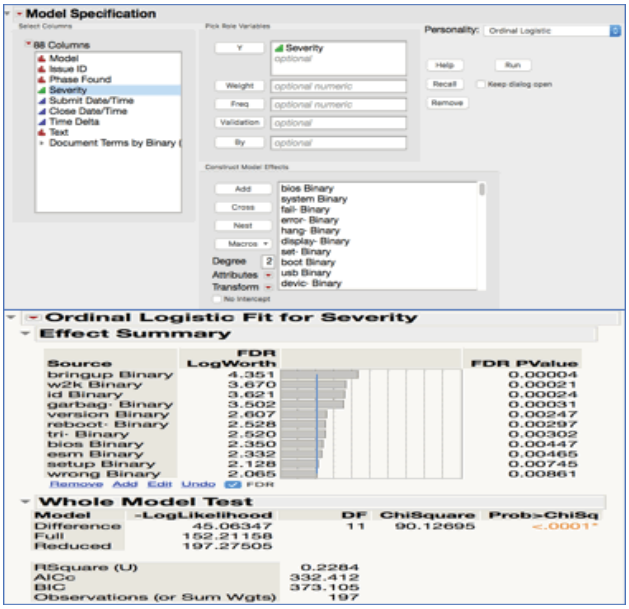

Figure 3.7.2 Predictive Model Setup & Analysis

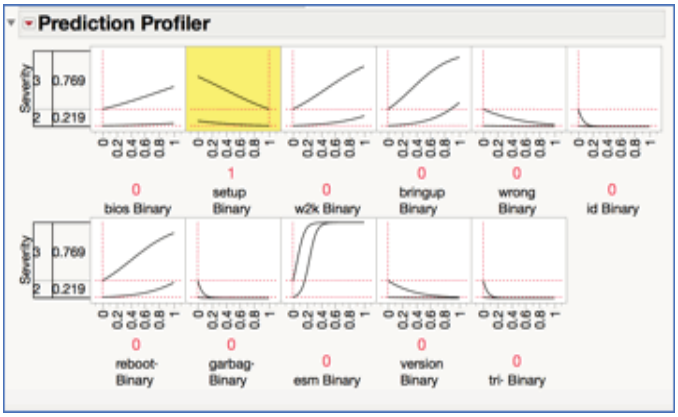

Figure 3.7.3 Predictive Model Results Profiler

5 FUTURE RESEARCH

There is ample research around Text Mining methods and descriptions of applied uses in fields such as customer survey and social media analysis. However, more research is definitely needed in using these methods around reliability uses, in ways that can alter and improve upon our existing practices. With access to modern analytic software, more and more reliability professionals will be utilizing Text Exploration techniques in unique ways to add value to their work.

ACKNOWLEDGEMENTS

We would like to acknowledge Dr. Chris Gotwalt, JMP Director of Research and Development at SAS Institute, and

Dr. Laura Higgins, Sr. Systems Engineer at SAS Institute, for their generous help and direction in developing Text Exploration approaches on warranty/performance data for reliability.

REFERENCES

Recommendations on where to learn more about Text Exploration:

- Fundamentals of Predictive Analytics with JMP, Second Edition, Ron Klimberg & B. D. McCullough, 2016, SAS Institute
  - Chapter 15: Text Mining
- Data Mining: The Textbook, Charu C. Aggarwal, 2015, Springer
  - Chapter 13: Mining Text Data
- JMP 13 Online Documentation: Text Explorer – Unstructured Text in Your Data, JMP Business Division, SAS Institute, Inc.
  - https://www.jmp.com/support/help/13/Text_Explorer_Overview.shtml

BIOGRAPHY

Scott Wise
Technical Manager
The SAS Institute, Inc., JMP Business Division
11920 Wilson Parke Ave
Austin, TX 78726 USA

e-mail: scott.wise@jmp.com

Scott Wise is a Technical Manager for the JMP Business Division at the SAS Institute, Inc. since 2008. Before joining SAS, Scott worked in a variety of industrial Quality Management and Engineering positions. While at Dell, Inc. he received the company's first-ever Black Belt and Master Black Belt certifications. Scott also holds ASQ Certified Quality Engineer, PMI Project Management Professional, and ASA Professional Statistician certifications.