Paper 1288-2014

Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining

Dr. Goutam Chakraborty, Professor, Department of Marketing, Spears School of Business, Oklahoma State University
Murali Krishna Pagolu, Analytical Consultant, SAS® Institute Inc., Cary, NC

ABSTRACT

The proliferation of textual data in business is overwhelming. Unstructured textual data is being constantly generated via call center logs, emails, documents on the web, blogs, tweets, customer comments, customer reviews, and so on. While the amount of textual data is increasing rapidly, businesses' ability to summarize, understand, and make sense of such data for making better business decisions remain challenging. This paper takes a quick look at how to organize and analyze textual data for extracting insightful customer intelligence from a large collection of documents and for using such information to improve business operations and performance. Multiple business applications of case studies using real data that demonstrate applications of text analytics and sentiment mining using SAS® Text Miner and SAS® Sentiment Analysis Studio are presented. While SAS® products are used as tools for demonstration only, the topics and theories covered are generic (not tool specific).

INTRODUCTION

The opportunity cost of any business to ignore unstructured data is paramount in today's fierce competitive world. According to an IDC survey, unstructured data takes a lion's share in digital space and approximately occupies 80% by volume compared to only 20 for structured data. While the unstructured data is available in abundance, the number of software products and solutions that can accurately analyze the text, present insights in an understandable manner along with the ability to integrate such insights readily into other extant models that use numerical only data are rare. A lot of the challenges in this space arise from the fact that natural language provides the flexibility to convey exactly the same meaning in umpteen different ways, or worse, exactly the same statement in a different context may convey completely different meaning. Machine learning algorithms designed to analyze numerical data exactly know the structure of numbers and they are pre-programmed to process the data with precision. In case of natural language, it gets very problematic. Dialects, jargon, misspellings, short forms, acronyms, colloquialism, grammatical complexities, mixing one or more languages in the same text are just some of the fundamental problems unstructured data poses. It makes extremely difficult to precisely analyze unstructured data in the same way we process structured data. SAS® Text Analytics tools provide text analysis capabilities which are at par or better with the other software available in the industry. Perhaps most importantly, SAS® Text Miner being part of the SAS® Enterprise Miner Suite, offers the ability to seamlessly integrate and use text analysis results along with analysis of numeric data only.

Using SAS® Text Analytics tools, we can collect unstructured data from wide variety of data sources and prepare it for analysis. Unstructured data can be found in databases, individual files (.txt, .xml, .doc, .xls etc.) on file systems, and websites. Chakraborty et al. (2013) in their recently published SAS® press book

discusses in details how to collect, cleanse, and standardize data using following SAS® tools and features:

- SAS® Text Import Node / TMFILTER macro
- SAS® XML Libname Engine using SAS® XML Mapper
- SAS® Crawlers

Improving Predictive Accuracy of Numeric Data Models with Textual Data

Traditional data mining methods such as predictive modeling use techniques which requires numerical data inputs to predict a response variable (binary or continuous). The underlying algorithm of a predictive model crunches the input values into a set of one or more algebraic equations (Example: Multiple Regression, Neural Networks etc.) or a set of if-then-else logical statement or rules (Example: Decision Trees) to derive the predicted probability and/or a predicted value. There are wide-variety of business applications which use predictive modeling methodology. A well-known customer relationship management application used in B2C (Business-to-Consumer) business is called the "churn modeling" where the objective is to build a predictive model and identify which customers are likely to churn in the impending future. It is important to make efforts not to lose existing customers for any business.

Recent studies indicate that along with the numerical data, the use of unstructured data specific to the individual customers may be useful in improving the predictive accuracy of the predictive models. This unstructured data can be a customer survey response to a specific service utilized or a product purchased. Some business store transcripts of complaints reported by its customers through their call center. If there are no complaints reported, one may choose to store the notes from 1-3 prior calls of the customer. Customers voices on ongoing issues, complaints, feedback, suggestions to improve and even compliments for a service may have great value and importance.

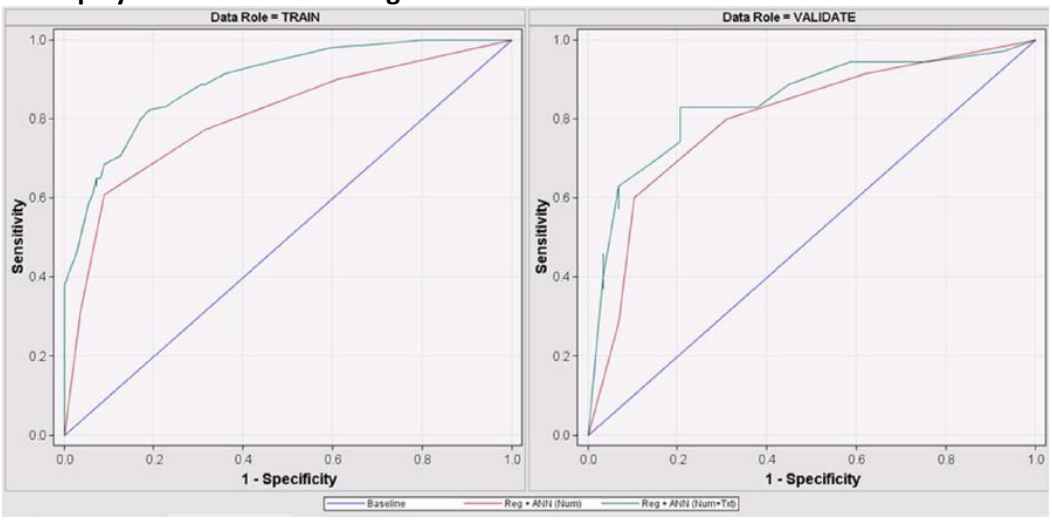
The basic premise to use text data in predictive models is that the terms contained within the text data can potentially represent the customer's experiences (bad or good) which are supposedly consistent with the customer's decision to continue with the business or churn in the nearest future. Hence the potential of mining text data in such applications cannot be undermined. Text data is first transformed into a set of numerical components called Singular Value Decomposition (SVD) units which collectively represent the text documents. These units are then used as additional inputs along with the existing structured input attributes to help improving the predictive power of the existing models. Nareddy and Chakraborty (2011) in their paper showcased how a retail customer using SAS® Text Miner achieved this. They have used the text data captured in their customer loyalty open-ended survey questionnaire to enhance a predictive model which predicts their customer's opinion of whether or not their company has the best loyalty program in the country (Binary target flag – Y/N). Display 1 shows the model fit statistics – misclassification and sensitivity for both training and validation data partitions.

Display 1: Misclassification Rates and Sensitivities of Two Best Models

Model	Role	Misclassification	Sensitivity
Reg + ANN (Num+Txt)	Train	19.52%	82.86%
Reg + ANN (Num+Txt)	Validate	**18.75%**	82.86%
Reg + ANN (Num)	Train	25.90%	60.71%
Reg + ANN (Num)	Validate	**26.56%**	60.00%

It clearly shows how combining the text data with numerical data gives better accuracy in predicting the target attribute. A regression model used for variable selection followed by an artificial neural network model (ANN) to predict the target variable were found to be the best models among a number of models that were built. Display 2 shows the Receiver Operating Characteristics (ROC) curves for both training and validation data partitions indicating a superior performance of the model which includes both numerical attributes and the text components as compared to the model which uses just the numerical attributes.

Display 2: ROC Chart Showing the Relative Performance of the Two Best Models



In their recently published SAS® press book by Chakraborty et al. (2013), the authors describe another example of how integrating textual analysis of customer call-center records improve the predictive ability of a numeric data only model. The authors report that adding just the text cluster membership to the numeric data only model performs better than the numeric model alone. But, when the SVD values from textual analysis of customer comments are added to the numeric data only model, it outperforms the model with the text cluster membership. Interestingly, authors find that addition of textual data analysis metrics not only improves the predictive ability of the numeric data only model, it also makes some of the numeric variables statistically significant which were not significant in the numeric data only model. This raises the possibility of potential synergy between information contained in numeric data and textual data.

Automatic Routing

Among all real-world applications of text analytics, automatic routing is the most useful and easiest to relate with. Well-known applications of automatic routing are e-mail forwarding and spam detection. Most e-mail exchange servers have the built-in functionality to help you setup rules that guide the incoming e-mails to your inbox into various folders or sub-folders. Those rules can be very simple or extremely complex depending on how many e-mails you receive on a daily basis and the number of folders/subfolders you have in your mailbox. Generally, the rules are setup to analyze one or more metadata attribute values of the e-mail received. Metadata attributes can be anything from the subject of the email, the name or e-mail address of the sender, the user group to which the e-mail was sent (a user can be part of many e-mail groups) and most importantly the body of the e-mail. The rules employ a set of Boolean logic operators (such as AND, OR, NOT) and check for the terms entered by the user in the respective metadata attribute fields. Accordingly, e-mails are routed to the corresponding folders or sub-folders.

In addition to this, almost every e-mail exchange server has the capability to detect spam e-mails and prevent them from entering your inbox. A majority of spam are phishing scam e-mails where the purpose is to entice you into capturing your private information to commit a financial fraud or an identity theft. Other type of spam are those which contain malicious virus if you download an attachment or click on a hyperlink within the e-mail which can potentially damage your machine. The text analytics algorithms can detect the fraudulent e-mail addresses from which phishing e-mails are pushed and subsequently route them to spam folder initiating user's attention for careful review.

An enterprise-wide application of automatic routing is to sort and route technical support request, service request and complaints reported by the customers of an organization to its appropriate team. It takes monumental manual effort if they were to be reviewed one at a time and assigned to a specific team. Usually, customers entering such requests may not always tend to update the subject line with appropriate information pertaining to the issue. Thus, an automatic routing system just relying on subject line of the request has a higher probability of assigning it to an incorrect team causing severe delays responding to the customer. The accuracy of routing increases when the body of request is included in the text analytics process.

Chakraborty et al. (2013) in their book showed a routing application to automatically identify the appropriate section to which a newly submitted SAS® Global Forum abstract should be assigned. As a first step, authors downloaded the SAS® Global Forum paper abstracts for the past 5 years from SAS® Global Forum proceedings. Authors chose top 5 sections based on the volume of submissions and their popularity among the conference attendees. Those 5 sections are Data mining and Text Analytics (DATAMINING), Business Intelligence (BUSINT), System Architecture (SYSARCH), Reports and Visualization (REPORTS) and Statistical Analysis (STATS). Using the new Text Rule Builder feature in SAS® Text Miner, authors trained a model and generated the basic content categorization code for each of those sections (Display 3).

Display 3: Automatic Content Categorization Code generated by Text Rule Builder Node

Content Categorization Code

```
F_type =DATAMINING ::
(OR
, "mining"
, "miner"
, "miner"
, (AND, (OR, "credits" , "credit" ))
, (AND, (OR, "costs" , "cost" )))
F_type =BUSINT ::
(OR
, "olap"
, "bi"
, (AND, (OR, "dashboards" , "dashboard" ))
, "intelligence" )
F_type =SYSARCH ::
(OR
, (AND, (OR, "server" , "servers" ))
, (AND, (OR, "configuration" , "configurations" ))
, "metadata"
, "storage"
, "performance"
, "operational" )
F_type =REPORTS ::
(OR
, (AND, (OR, "graphs" , "graph" ))
, (AND, (OR, "reports" , "report" ))
, (AND, (OR, "map" , "maps" ))
, (AND, (OR, "details" , "detail" )))
F_type =STATS ::
(OR
, (AND, (OR, "model" , "models" ))
, (AND, (OR, "procedure" , "procedures" ))
, (AND, (OR, "analyses" , "analysis" )))
```

While the basic codes generated by the Text Rule builder produces reasonable results, these can be improved further by copying and modifying them in SAS® Content Categorization Studio where the initial taxonomy is built for those 5 sections. The Boolean rules are further modified to improve accuracy in predicting the appropriate section. Display 4 shows the full test report generated from SAS® Enterprise Content Categorization studio once the Boolean rules have been modified further. These rules can then be used to route an abstract to an appropriate section chair without having to let the author decide which section to submit it.

Display 4: Full Test Report Using Boolean Rule Based Model with Modified Rules

Category Test Report

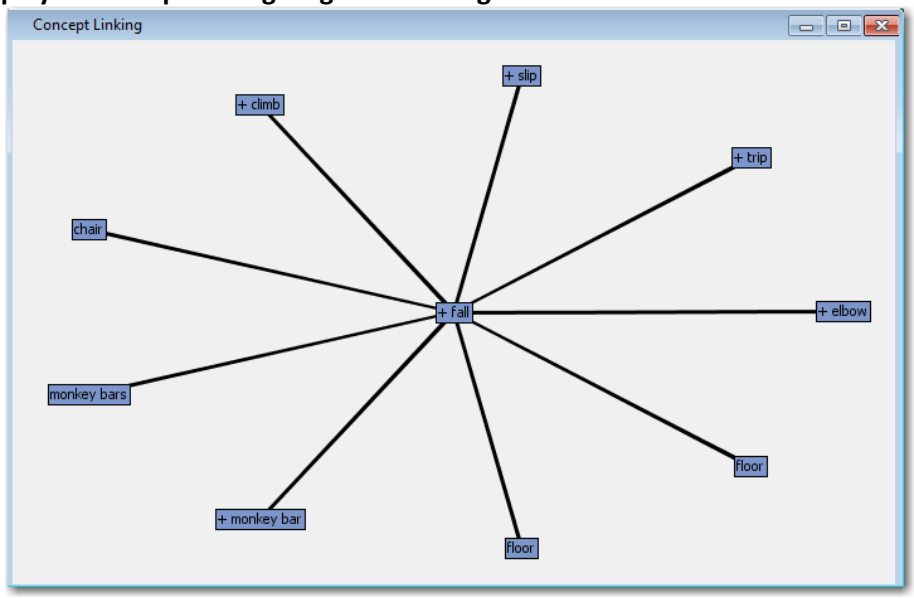
Path	All Docs	In-Cat	Total	In-Cat %	Neg	N-Tot	Neg %	Prec %	Popul...	Pop Rel
Top	0	0	0	0	0	0	0	0	0	0
Top/BI	122	74	82	90	0	0	0	60	0	0
Top/DM	147	72	79	91	0	0	0	48	0	0
Top/REPORTS	138	82	100	81	0	0	0	59	0	0
Top/STAT	279	152	174	87	0	0	0	54	0	0
Top/SYSARCH	165	84	99	84	0	0	0	50	0	0

OK View as Text

Root Cause Analysis (RCA)

Root cause analysis (RCA) is a method of problem solving that tries to identify the root causes of faults or problems. In the Chakraborty et al. (2013) book, SAS consultant Mary Osborne contributed a use case in which they have analyzed the data collected by Consumer Product Safety Commission (CPSC), an independent agency in United States which regulates sale and manufacturing of products through its extensive market research. The objective of CPSC is to issue a ban or recall of potentially faulty products which can cause serious health or physical hazards. In this use case, they used the emergency room data collected from over 100 hospitals through National Electronic Injury Surveillance System (NEISS) which contains information pertaining to several types of injuries. They considered three sets of data where injuries are associated with falls/falling, sports and swallowing. Through their analysis, they have attempted to understand the types of injuries sustained by people, actions that lead to those injuries and identify the products and their associated injuries. Using SAS® Text Miner, they were able to manage the synonyms list, correct misspellings, drop unwanted terms, retain key terms, and perform stemming to reduce the density of unique terms used in the analysis. Using the pre-built functionality of SAS® Text Miner, they developed concept link diagrams (Display 5). Concept link diagrams simply show how the terms selected in an analysis are associated with other key terms in the document collection. The selected term "fall", in this case is shown at the center of the link diagram. The terms around are those which are highly associated with the term "fall".

Display 5: Concept linking diagram showing terms associated with "fall" and its synonyms



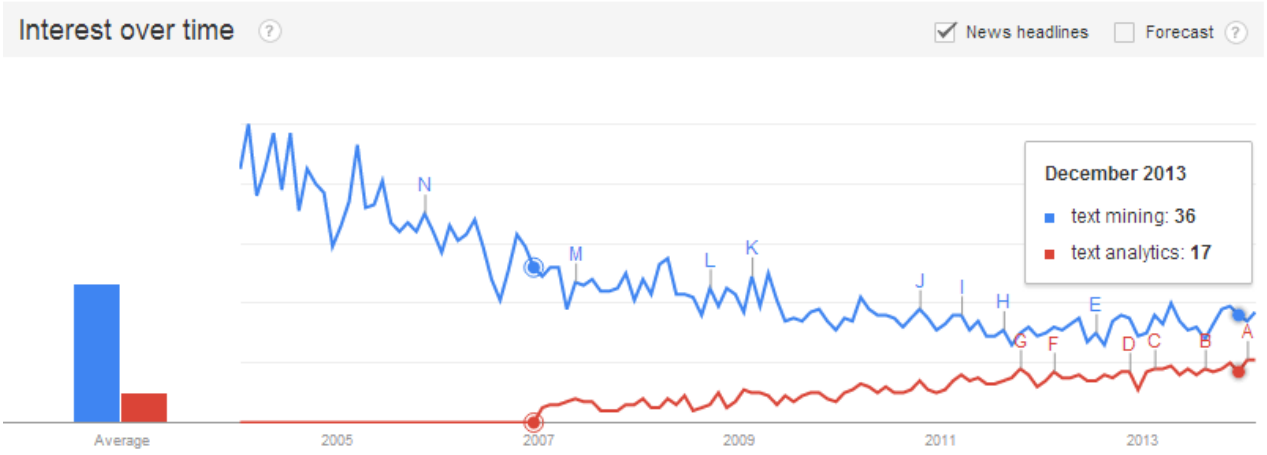
The strength of association between two terms is based on the principle of conditional probability that is, the probability that the term B exists given the term A already exists in the document. Stronger the strength of association between terms, thicker is the line connecting the terms in the display. Using this type of root causal analysis, they collected some evidences that some falls involve monkeybars and floors. Howvere, as in any association-based analysis, higher strength of association between two terms doesn't always imply a cause and effect relationship. Therefore you need to be careful while drawing causal conclusions and these must be tempered with domain expertise and method of text data cllolection.

Trend Analysis

As the name suggests, it is an analysis method to understand how certain entity under study change over a period of time. Example of such entities are part number of an appliance reported in a failure, serial number of a device being serviced, type of a customer service request, and area of technical support issue raised etc. The standard method of analyzing trends is to chart the most frequent entity surfaced in the document collection observed for a specific duration of time (a minute, an hour, a day or a month) and observed over a wider range of duration (a day, a week, a month, or an year). Trends are useful to observe how things change over a period of time. It should not be confused with a typical histogram or bar chart where you are simply considering the individual frequencies of all entities found in your document collection.

Trend analysis need not be limited to data collected internally at your company. Nowadays we keep hearing about trends in social media which are observed by means of analyzing the comments, reviews and opinions posted by the internet commuters throughout the world. Online social networking giants like Twitter, Inc. and Facebook, Inc. use text analytics algorithms and methods to determine the trends on the respective platforms where the trending entity can be the names of a person, event, place or a topic. Twitter, Inc. gives you a list of trending topics worldwide and also lets you drill down by geographic region to find trends local to your region. Adding a hash tag (#) is a common practice by twitter users when tweeting about a topic. These hashtags are useful for searching all tweets related to a particular topic (https://twitter.com/search-home). Google Inc. analyzes trends on its search engine and presents the results for public on their website http://www.google.com/trends/. Display 6 shows how the term "text analytics" is slowly gaining popularity and replacing the term "text mining" over a period of last 5 years in global searches.

Display 6: Comparison trending chart for the search term "text mining" on Google search engine

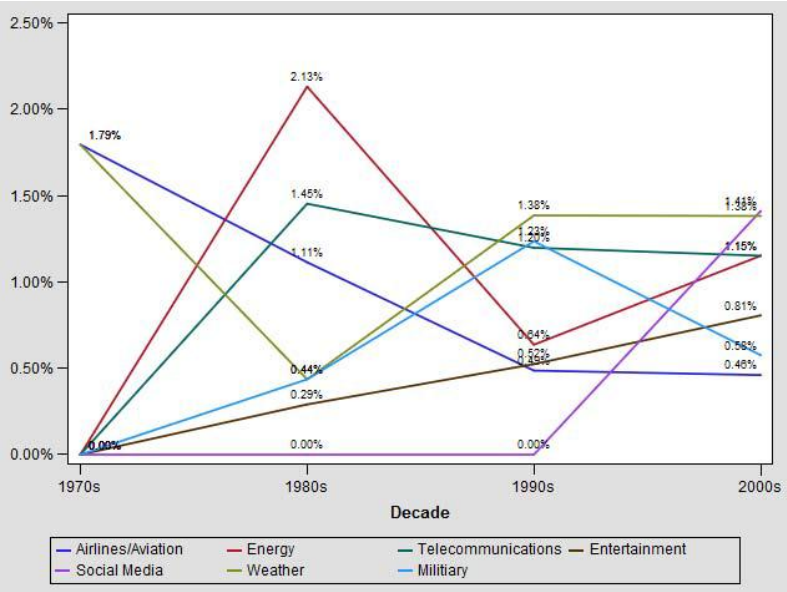


From this chart, you may infer that knowledge seekers are increasingly searching for information about "text analytics" which encapsulates much broader concepts than just "text mining" which traditionally has been used as a term for bag-of-words type of analysis involving word frequency. Facebook, Inc. recently launched www.facebookstories.com in which they publish the top trends relating to various activities Facebook users perform. Some of the interesting trends Facebook captures are for example, a geographical map showing the most frequent check-ins around the world, a pie chart displaying the top 10 global life events split by % volume and a pie chart showing the top 10 most talked about topics around the world. Analyzing trends on social media can be a powerful exercise to gauge how well your

new product launch, improved service or innovative idea has captured public interest. This type of analysis can give some insights on whether your initiative has had any impact on the public domain.

Shaik, Garla & Chakraborty (2012) in their paper discussed how they utilized the publicly available SAS® global forum proceedings between the period 1976 and 2011 and analyzed the topic trends segmenting them into 4 decades. Display 7 shows the trending chart on some traditional industry topics along with few emerging industry topics chosen for submissions.

Display 7: Trending chart shows industry topics presented in SAS® conferences in four decades



This chart signifies that while some topics are on a declining trend (aviation/airlines), others are quite stable (weather, military, entertainment, telecommunications). However, one particular industry which saw significant increase in the previous decade is the social media. It is quite obvious that this trend is going to continue for few more years.

Crime Detection and Fraud Analysis

Post 9/11, several federal agencies and governments across the world have been heavily investing in research to proactively monitor terrorist activities and foil their attempts to commit unlawful acts detrimental to civil society. Text analytics play a key role in such research areas where the objective is to intercept, translate and analyze all communication channels which the extremist organizations may be using to plot for an explosion or a serious crime and perturb peace and harmony in a community or a region. While it may be easy to track e-mail communication, it is not easy to tap telephonic conversations and convert them to text. There are several voice-to-text conversion software available in the market today but the accuracy and precision by which they perform needs improvement. Fraud analysis is playing a bigger role in major banks today to find irregular patterns and alert them when a financial fraud occurs. Certain types of frauds within organizations may be complex to detect unless appropriate monitoring systems are in place to keep a tab on its employees' key communication channels such as e-mails. SAS consultant Dan Zaratsian contributed his use case to our book in which he demonstrated how SAS® Text Analytics can be leveraged for helping agencies investigate fraudulent activities by scraping through several thousands of e-mails. The objective was to extract meaning

information from Enron e-mails data and identify trends and patterns to help investigate Enron Scandal. Using Text topic node in SAS® Text Miner multi-term text topics are generated from the Enron sample data (Display 8).

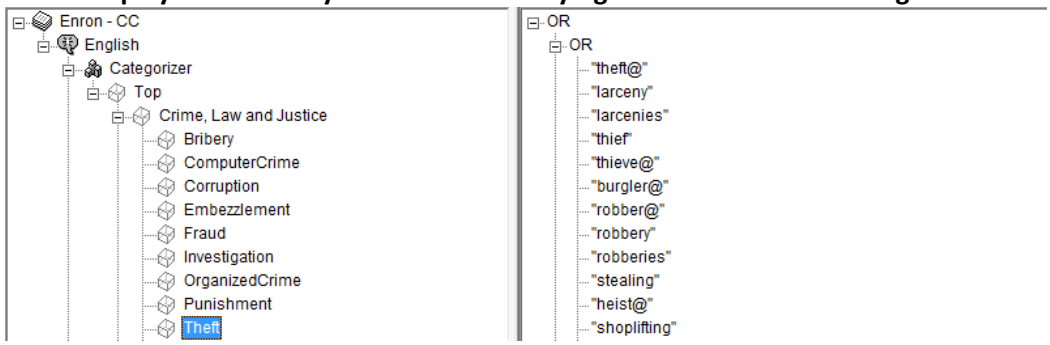
Display 8: Text Topics emerged from mining the Enron sample data

Topics

Category	Topic ID ▲	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	2.196	0.081	+business,+investor,+investment,+core,+core business	377	69
Multiple	2	1.102	0.057	+e-mail,+sender,+e-mail transmission,+message,+transmission	493	211
Multiple	3	0.730	0.048	+oil,+production,+exploration,+gas,+rig	965	195
Multiple	4	0.641	0.043	eim,+poland,newsprint,+farm,+license	418	22
Multiple	5	0.674	0.051	+thing,+time,+day,+man,+car	1694	449
Multiple	6	0.600	0.041	+eogil,enron_development,egep,+india,enron_development@enron_devel...	315	28
Multiple	7	0.658	0.045	enron_development,+hou,ect@ect,enron_development@enron_developm...	1126	238
Multiple	8	0.625	0.044	+california,+electricity,+power,+rate,+consumer	953	193
Multiple	9	0.502	0.039	+image,+search,09home,+rigzone,+map	694	180
Multiple	10	0.490	0.040	+dow,dabhol,+india,jones,+windmill	569	65
Multiple	11	0.498	0.038	+site,reregulation,+price,ap,san diego	356	48
Multiple	12	0.460	0.038	+e-mail,+karen,+jack,fielder,+carolyn	605	30
Multiple	13	0.557	0.042	+bankruptcy,+creditor,legislation,+senate,+bill	1127	342
Multiple	14	0.467	0.040	+gore,al,+election,florida,+fiction	1177	153
Multiple	15	0.532	0.040	lng,+project,+storage,+mmbtu,+facility	1076	288
Multiple	16	0.471	0.038	blockbuster,+bandwidth,+trade,+telecom,+business	488	50
Multiple	17	0.466	0.037	lng,caribbean,lng's role,+nimocks,north american	467	83
Multiple	18	0.398	0.035	+gama,trakya,+bureaucrat,+turkish,turkish	221	13
Multiple	19	0.460	0.038	+japan,+e-mail,+recipient,intended,seabron	995	187
Multiple	20	0.387	0.035	purchaser,+withholding,withholding,.ge.com,withholding tax	341	39
Multiple	21	0.365	0.035	+sa,+rua,+argentina,resignation,alvarez	393	32
Multiple	22	0.362	0.035	el paso,+paso,+el,+assembly,+california	676	84
Multiple	23	0.376	0.037	mg,+arbitration,lch,+emc,+claim	1223	157
Multiple	24	0.294	0.032	toaster,sadist,+alma,studebaker,+toast	559	12
Multiple	25	0.275	0.034	recourse,+negotiate,yr,+option,+rate	1003	37

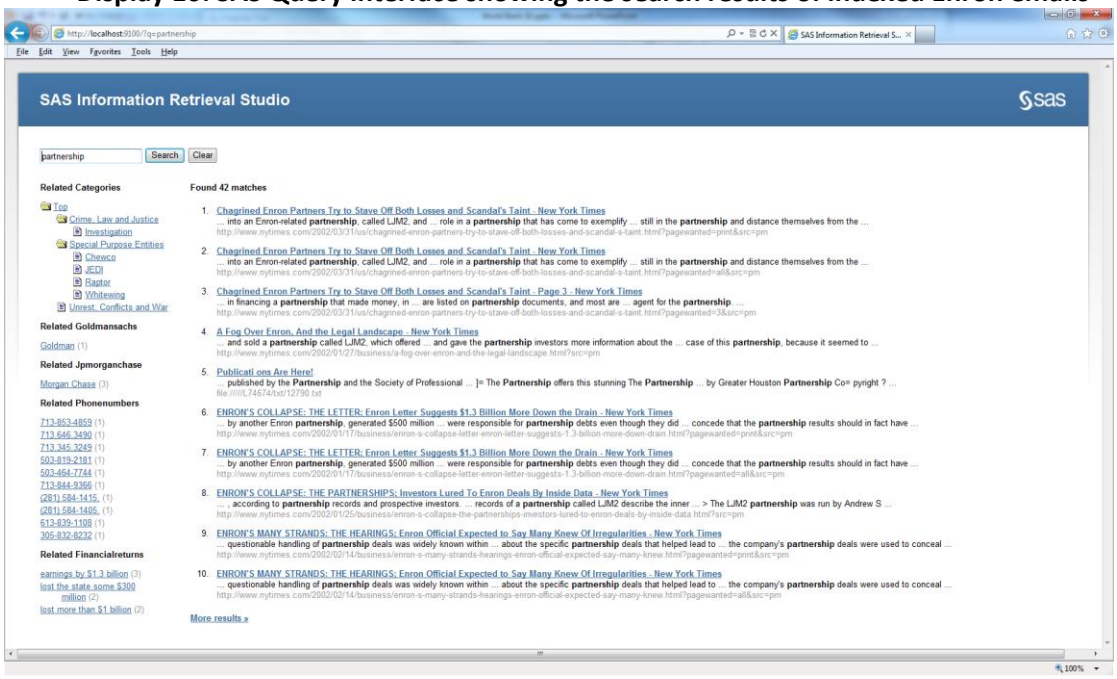
Some of these topics makes sense when we look at individual topics and the terms which describe them. Topic #3 is about oil production, #11 talks about California electricity, #14 discusses about Al Gore elections, and #16 is on trading telecommunications stocks which are all related to Enron scandal. These terms are then used to build the categories/sub-categories to categorize the e-mails data (Display 9) using SAS® Content Categorization studio.

Display 9: Taxonomy created for classifying Enron e-mails into categories



Text parsing and text topic extraction are leveraged to enhance the terms list for modifying the pre-existing categories/sub-categories from the IPTC taxonomy. The categorization project can be used in SAS® Information Retrieval Studio framework as a document processor to apply the rules on all the Enron e-mail data files. Display 10 shows the SAS® Query interface in which the categories and concepts are available to perform facetted search the Enron emails.

Display 10: SAS Query interface showing the search results of indexed Enron emails



The Enron email archive contains more than 500,000 emails from 159 personal email accounts. Reading through all these emails is intensely painstaking task and impossible to achieve. The investigator's goal to extract vital evidence of the Enron Corporation's suspicious accounting practices is made easy with powerful SAS® Text Analytics tools.

Sentiment Analysis

An interesting and important goal of analyzing unstructured data such as customer complaints, issues, opinions or comments is to get a grasp on what they perceive about an entity. An entity can be a company's brand image, product, service, person, group or an organization. Are consumers' perceptions good, bad or neutral? What attributes (features) of the product or service they feel good or bad about? What do the customers think of the various attributes of a company's product such as quality, price, durability, safety, ease of use? Typically, if customer feels good towards an entity, it is classified as a positive sentiment. If the perception towards the entity is bad, it can be considered as negative sentiment. A third kind of perception in which customer has neither good nor bad opinion implies a neutral sentiment. Social media sites such as Twitter and Facebook contains enormous volumes of customer opinions and comments on virtually all major organizations, events and products. This creates an unprecedented opportunity to mine text data in real-time to and analyze sentiment trends fluctuations over a period of time.

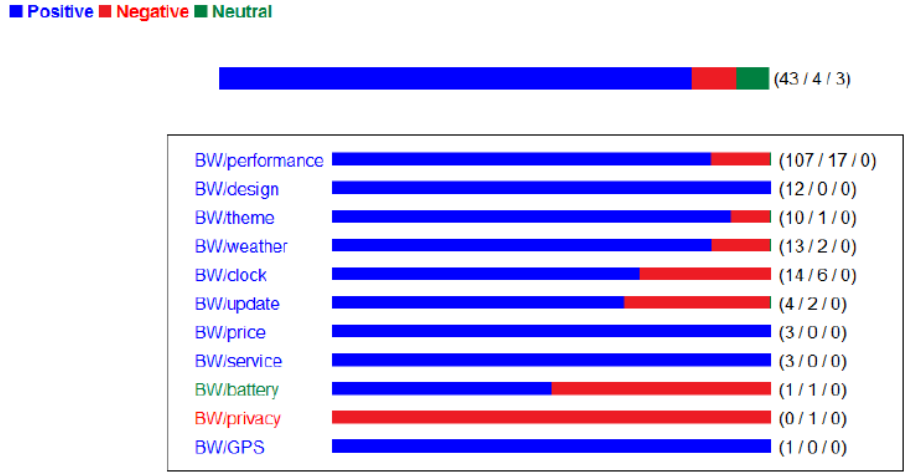
Liu et al. (2013) in their study performed feature based sentiment analysis on Android App Reviews using SAS® Sentiment analysis studio. They built sentiment analysis models which are both statistical based and rule-based. Display 11 shows the results indicating that in this study the rule-based models outperform the statistical models for both the Apps (Widget and Game). The natural language processing (NLP) capabilities of rule-based model yields higher precision.

Display 11: Comparative performance of statistical vs. rule-based sentiment analysis models

App	Statistical Model			Rule-based Model		
	Positive Precision	Negative Precision	Overall Precision	Positive Precision	Negative Precision	Overall Precision
Widget	64%	96%	80%	86%	94%	90%
Game	88%	74%	81%	94%	90%	92%

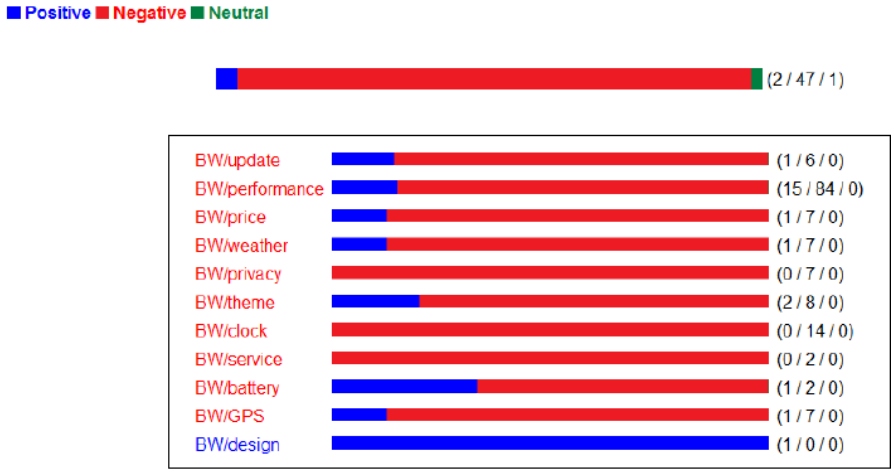
Display 12 and 13 shows feature specific sentiment scores (positive, negative and neutral) distribution when the rule-based sentiment analysis model is tested against the "Positive" and "Negative" app reviews pre-classified for the purposes of testing. This level of granularity and predictive accuracy is only possible with rule-based models where the user-written rules efficiently capture sentiment based on Adjectives, Adverbs and Verbs used in the rules which define the positive or negative sentiment.

Display 12: Widget App testing results showing feature based sentiment scores on Positive reviews



Results for selected folder:
This directory is Positive
Positive precision is 86.00%.
Number of articles:50
Number of positive articles:43
Number of negative articles:4
Number of neutral articles:1
Positive percent:86.00%.

Display 13: Widget App testing results showing feature based sentiment scores on Negative reviews

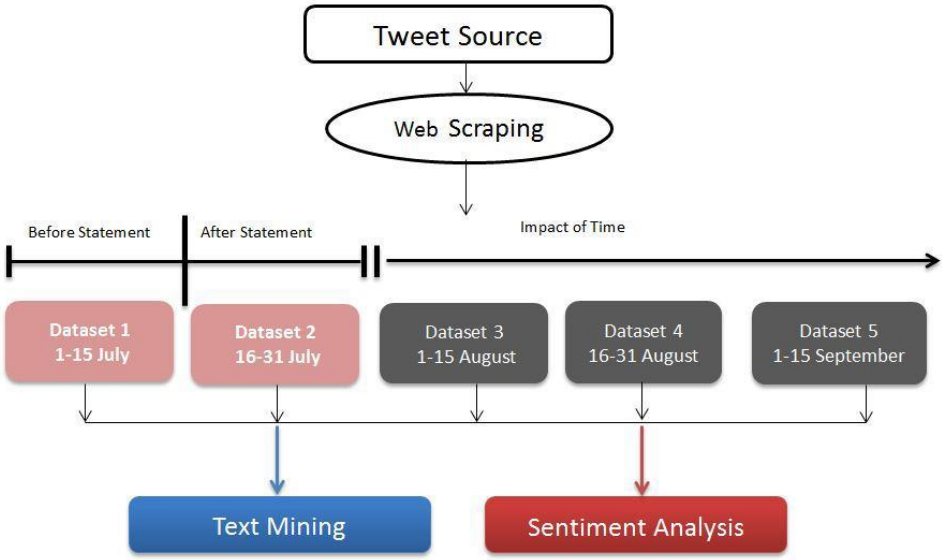


Results for selected folder:
This directory is Negative
Negative precision is 94.00%.
Number of articles:50
Number of positive articles:2
Number of negative articles:47
Number of neutral articles:0
Positive percent:4.00%.

Another way of analyzing sentiment is to assess how a significant event related to your organization might impact the way customers think about your company. Grover et al. (2013) in their paper discuss the approach they have adopted to separately analyze tweets posted on twitter by the general public

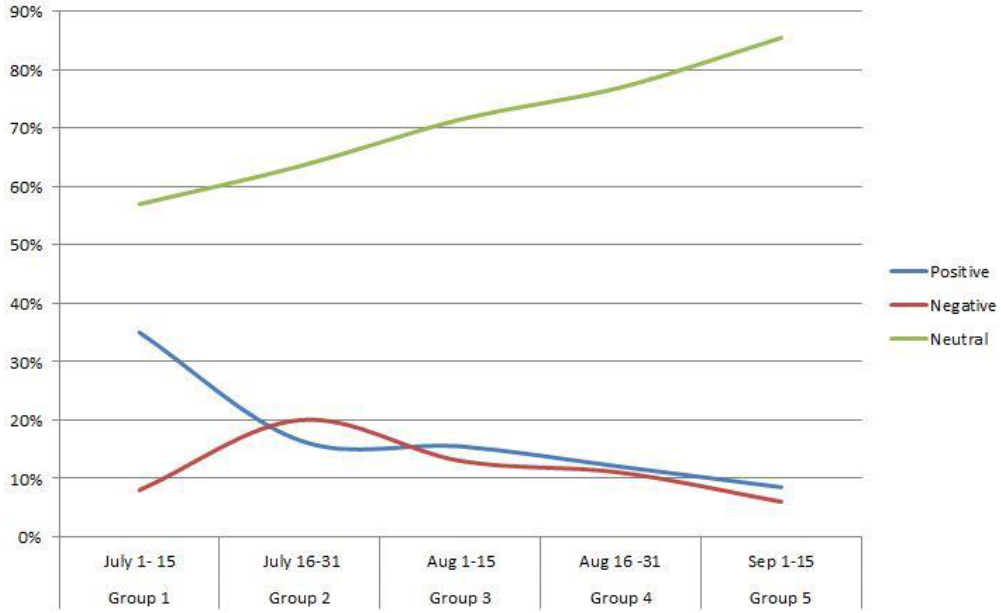
before and **after** Chik-fil-A's president Dan Cathy publicly made remarks supporting traditional marriages over same-sex marriages. His remarks invited public debate over the social media channels and the overall reaction was mostly negative. Display 14 shows how the tweets were collected before and after the event (Dan's remarks) into 5 spatial segments.

Display 14: Change in sentiment towards Chik-fil-A before and after the event



SAS® Text Miner was used to perform initial data exploration to identify clusters of tweets and understand the various aspects of twitter conversations about Chik-fil-A. SAS® Display 15 shows the sentiment analysis trend significantly changes after Dan Cathy's statement.

Display 15: Change in sentiment towards Chik-fil-A before and after the event



We can observe the positive sentiment towards Chick-fil-A takes a dip and negative sentiment increases significantly post July 15. We can also observe that the impact of his public statements is significant even after two months from the time he made those statements. From this exercise, we

can see that the % of negative comments slowly subsided two months after. We can see a tremendous increase of neutral sentiments as a result of people's opinion shifting from positive sentiment to neutral sentiment. From this study, we can infer that brand image of a company can shift overnight with an event as simple as a public statement which can trigger emotions in people that can impact their behavior.

Summary

The use of text analytics in real-world applications is growing very fast as organizations realize the untapped potential that is possible if textual data are analyzed and integrated in decision making. The field of text analytics will likely continue to grow given the exponential growth of unstructured data both within and outside the organizations. The breadth of text analytics applications also continues to expand across industries and it never ceases to amaze how many interesting applications are continuing to surface each day. We believe that text analytics applications will touch our daily lives much more frequently and impact every household that relies on internet for researching on products and services. Days are not far when every prospective consumer researching product reviews on the Internet will have his/her own thin client text analytics app to analyze the huge volumes of product/service reviews, summarize the pros and cons, display feature based sentiments and generate an overall recommendation personalized to the user's taste and preferences.

REFERENCES

Albright, R (2004). Taming Text with the SVD. SAS Institute Inc., Cary, NC.

Chakraborty, G., Pagolu, M. & Garla, S (2013). *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*. SAS Publishing.

Desai, M (2009). Applications of text analytics. Business Analytics. mybusinessanalytics.blogspot.com, 09 AUG 2009. Web. 21 FEB 2014. <http://mybusinessanalytics.blogspot.com/2009/08/applications-of-text-analytics.html>.

Grover, S., Jacob, V. J. & Chakraborty, G (2013). "Analysis of change in sentiments towards Chick-fil-A after Dan Cathy's statement about same sex marriage using SAS® Text Miner and SAS® Sentiment Analysis Studio." Proceedings of the SAS® Global 2013 Conference. Available at http://support.sas.com/resources/papers/proceedings13/251-2013.pdf

Liu, J., Sarkar, K. M. & Chakraborty, G (2013). "Feature-based Sentiment Analysis on Android App Reviews Using SAS® Text Miner and SAS® Sentiment Analysis Studio." Proceedings of the SAS® Global 2013 Conference. Available at http://support.sas.com/resources/papers/proceedings13/250-2013.pdf

Nareddy, M & Chakraborty, G (2011). "Improving Customer Loyalty Program through Text Mining of Customers' Comments." Proceedings of the SAS® Global 2011 Conference. Available at http://support.sas.com/resources/papers/proceedings11/223-2011.pdf

Shaik, Z., Garla, S & Chakraborty, G (2012). "SAS® Since 1976: An Application of Text Mining to Reveal Trends." Proceedings of the SAS® Global 2012 Conference. Available at http://support.sas.com/resources/papers/proceedings12/135-2012.pdf

TRADEMARKS

CONTACT INFORMATION

The authors welcome and encourage any questions, feedback, remarks, both on- and off-topic via email.

Goutam Chakraborty, Oklahoma State University, Stillwater, OK, goutam.chakraborty@okstate.edu

Goutam Chakraborty is a professor of marketing and founder of SAS® and OSU data mining certificate program 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 has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired M2007 data mining conference. He is also a Business Knowledge Series instructor for SAS®.

Murali Krishna Pagolu, SAS Institute Inc., Cary, NC, murali.pagolu@sas.com

Murali Pagolu is a Business Analytics Consultant at SAS Institute Inc. He has over 4 years of experience using SAS® software focused on Database Marketing, Marketing Research, Data mining, Text Mining and CRM Applications.