Paper SAS2225-2018

Harvesting Unstructured Data to Reduce Anti-Money Laundering (AML) Compliance Risk

Austin Cook and Beth Herron, SAS Institute Inc.

ABSTRACT

As an anti-money laundering (AML) analyst, you face a never-ending job of staying one step ahead of nefarious actors (for example, terrorist organizations, drug cartels, and other money launderers). The financial services industry has called into question whether traditional methods of combating money laundering and terrorism financing are effective and sustainable. Heightened regulatory expectations, emphasis on 100% coverage, identification of emerging risks, and rising staffing costs are driving institutions to modernize their systems. One area gaining traction in the industry is to leverage the vast amounts of unstructured data to gain deeper insights. From suspicious activity reports (SARs) to case notes and wire messages, most financial institutions have yet to apply analytics to this data to uncover new patterns and trends that might not surface themselves in traditional structured data. This paper explores the potential use cases for text analytics in AML and provides examples of entity and fact extraction and document categorization of unstructured data using SAS® Visual Text Analytics.

INTRODUCTION

Financial Institutions dedicate substantial resources in support of government's efforts to curb money laundering and terrorism financing. Money laundering is the process of making funds that were gained through illegal channels appear legitimate, typically through a process of placement, layering, and integration. Terrorism financing is often more challenging to identify, as the funding can be raised through legitimate means, but later used to fund an act of terror or support a terrorist organization. Detecting these patterns can often feel like a game of "whack-a-mole;" by the time a new control is implemented to identify a known risk, the criminals have already changed their behavior to elude your efforts. The stakes are high, as the amount of money laundered per year is estimated to be 2 to 5% of global GDP. That's 2 trillion in USD according to the United Nations Office on Drugs and Crime (UNODC). In today's big-data environment, using modern technology to quickly identify financial crimes is critical.

A lot has changed globally since the early AML regimes of the 1970s. A growing regulatory landscape has led to higher penalties for program deficiencies. Banking has fundamentally changed with the creation of digital channels, faster payments, and new financial instruments. Data storage has become cheaper, opening the opportunity to process big data rapidly. Financial institutions have mostly adapted to these changes through enhancements to their rule-based detection programs and, as a result, have seen their headcount and costs soar. There's an appetite to overhaul the system to reduce false positive rates, increase the detection of money laundering, and automate many of the tedious tasks required in the investigations process. With the help of SAS® Visual Text Analytics, we can leverage artificial intelligence techniques to scale the human act of reading, organizing, and quantifying free-form text in meaningful ways, uncovering a rich source of underused risk data.

UNSTRUCTURED DATA SOURCES

While structured data such as transaction, account, and demographic information has been used in combating money laundering for years, financial institutions are just now beginning to see the value in harvesting unstructured data sources. These data sources are both vast and rich with valuable information that provides new data points, creates linkages, and identifies trends. Here is a list of the more notable sources of unstructured data that can be used for AML:

- **Wire Data** - Wire transfers between financial institutions contain much more valuable information than just the amount of money being sent. Along with origination, intermediary, and beneficiary data, wires

often include free-form text including payment instructions and other messaging.

- **Transaction Review Memos** - The branch employees and client managers are the first line of defense when it comes to protecting the bank from money laundering. Typically, these individuals report valuable insight to the AML group through a transaction review memo. The details included in these memos are at the branch attendee's discretion, but often they have supporting detail on why the transaction was deemed suspicious that might not be apparent in the transaction alone.
- **Case Data** - Anti-money laundering case data contains information enriched by the investigator during the life of the investigation. Cases generally contain several free-form text fields including notes, comments, and email correspondence as well as a narrative report explaining the final disposition. If suspicious activity is identified, a suspicious activity report (SAR) will be filed.
- **Suspicious Activity Report Data** - SARs are documents that financial institutions must file with their in-country government agency following the identification of potential unusual behavior related to money laundering or fraud. These documents are typically free-form text and generally contain several pieces of information about the person, company, and entity or entities of interest; the general findings from the investigator as to what the suspicious activity was; as well as any supporting evidence for the suspicious activity.
- **Negative News** - Beyond unstructured data your financial institution generates, there is a vast amount of publicly generated data from news and media organizations. Public news data can be used to identify supporting information about your customers including relationships to businesses or risky behaviors and criminal activity.
- **Email/Phone/Chat** - In addition to transactional data, risk factors might be identified in the non-transactional data stored by the bank in the form of email, phone, or chat conversations between the customer and employee.
- **Law Enforcement Requests** - Financial institutions have an obligation to identify subjects of law enforcement requests and file SARs where appropriate. Grand jury subpoenas, national security letters, and other requests are received in electronic format and contain text regarding persons of interest and requests for information.
- **Trade Documents** - The global trade system remains primarily a paper-based system. The trade documents (letters of credit, bills of lading, commercial invoices, other shipping documents) contain critical risk information in free-form text such as boycott language, dual use goods, inconsistent unit pricing, and other trade-based, money-laundering vulnerabilities.

USE CASES IN AML

Mining your unstructured data can be valuable in uncovering new insights to help combat money laundering in your financial institutions. Processing techniques such as theme detection, categorization, and entity or fact extraction are all ways to provide structure to free-form text. Once text is structured, there are several use cases to apply this data to ensure compliance:

- **Negative News Monitoring** - As an industry standard, financial institutions typically look for negative news related to high-risk customers and customers who have an open AML case. With the wide array of digital news made available daily, the identification of credible news can be challenging. Negative news not relevant to compliance can bias an investigator's decision process, while missed news can leave an institution open to reputational risk. Coupled with bank policy and risk tolerance, an automated process to identify negative news and successfully link this information to customers provides both cost and time savings through automation.
- **Network Analytics** - Perhaps one of the best pieces of information for investigating AML is to understand relationships among your customers, as well as non-customers. Most institutions have structured data for known relationships among their customers, but often there are gaps with unknown relationships and those relationships with non-customers. Relationships and networks often surface through normal investigative procedures and are documented in case notes and SAR data. Storing this valuable information and displaying it for future use along with geographic tagging

provides deeper insights to the investigations process.

- **SAR Attribution Detection** - The detection of money laundering is an exercise in correctly identifying rare events in vast amounts of data. As the AML compliance industry starts to explore the application of artificial intelligence and machine learning to replace Boolean rules, the need for reliably labeled data (target variables) for training becomes even more important. Often, SARs are filed based on external information, but are attributed to the success of one or more rule-based scenarios. Text mining can help determine the correlation. This is critical to not only tune existing models, but also to allow banks to predict newly identified patterns in the future.
- **Trade Finance Document Categorization** - Deciphering trade documents is a tedious, manual process. We've been testing cognitive computing capabilities that are used for character recognition and natural language processing for document categorization. In a pilot with a tier 1 bank, our models read trade finance documents with ~99% accuracy and reduced the time to manually process the documents from several weeks to 26 seconds in an automated process.

EXAMPLE FRAMEWORK USING SAS® VISUAL TEXT ANALYTICS

This paper explores the process of processing unstructured data to support any of the use cases listed above. To demonstrate the potential applications, we will follow the framework below, primarily using SAS Visual Text Analytics as the enabling technology.

- **Data Acquisition –** Data is acquired for the example use case utilizing web scraping tools and is imported into SAS Visual Text Analytics.
- **Concept Extraction** – Predefined and customized concepts are generated to extract key facts from the unstructured data.
- **Text Parsing** – The individual records are parsed to enumerate the terms contained in the documents and apply filtering with start and stop lists.
- **Topic Generation** – Individual records are grouped into a collection of related themes containing similar subject matter automatically based on a bottom-up approach using the underlying terms.
- **Categorization** – Documents are classified into predetermined categories based on a top-down approach of the areas of interest using linguistic rules.
- **Post-Processing** – Output from SAS Visual Text Analytics is processed and prepared for use in modeling or investigative tools.

DATA ACQUISITION

While SAR information is not publicly available, we wanted to conduct our analysis on text data with similar content and format. The Internal Revenue Service (IRS) publishes summaries of significant money laundering cases each fiscal year, dating back to 2015. This data is rich with information, including people, organizations, risk typologies, locations, and other interesting data related to financial crimes. Below is an example of an IRS case from our data set:

"**Former Owners of Money Transmitter Business Sentenced for Conspiring to Structure Financial Transactions**
On October 25, 2016, in Scranton, Pennsylvania, German Ossa-Rocha was sentenced to 27 months in prison and two years of supervised release. On October 26, 2016, Mirela Desouza was sentenced to 18 months in prison and two years of supervised release. Ossa-Rocha and Desouza were the former owners of Tropical Express, a money transmitter service business located in Stroudsburg. Beginning in approximately January of 2008 and continuing through December 2011, Ossa-Rocha and Desouza structured financial transactions that represented the proceeds of drug trafficking in a manner intended to avoid reporting and recording requirements. The amount of funds involved in the structuring was approximately $340,000. The funds were transmitted by Ossa-Rocha and Desouza via wire transfers to the Dominican Republic." (IRS)

Web scraping tools were used to extract the various money laundering examples and write to a CSV file with four columns: observation number, year, title, and text narrative. The CSV file was then imported into SAS Visual Text Analytics for analysis.

CONCEPT EXTRACTION

After initializing a project and loading the data, the first step in the process was focused on concept and fact extraction. With our data being rich in entities and facts, we wanted to extract these from the text for potential use in further analysis and research by investigators. In our model pipeline, this was done by dragging a Concept node and placing it on top of the Data node. SAS Visual Text Analytics comes with predefined concepts out of the box, as well as the ability to write your own custom concepts using LITI (language interpretation and text interpretation) syntax. For our analysis, we enabled the predefined concepts and wrote several custom concepts that are highlighted below.

The predefined concepts are common points of interest in which the rules come out of the box to immediately apply to your data, saving you time and helping you gain instant insights. Here are the predefined concepts of interest for our analysis:

- **nlpDate –** Identifies and extracts all dates and date ranges in your data in several formats (for example, May 2003, 05/15/2007, between 2007 and 2009, and so on).
- **nlpMeasure –** Identifies and extracts measures of time and quantities (for example, 30 years, 500 kilograms, and so on).
- **nlpMoney –** Identifies and extracts all references to currencies (for example, $272,000, more than $3 million, and so on).
- **nlpOrganizations –** Identifies and extracts all organization names (for example, U.S. Treasury, Department of Agriculture, and so on).
- **nlpPerson –** Identifies and extracts all names (for example, Joyce Allen, Robert L. Keys, and so on).
- **nlpPlace –** Identifies and extracts all places (for example, Asheville, North Carolina, Newport Beach, California, and so on).

Error! Reference source not found. below shows a set of matched concepts for the predefined concept nlpMoney.

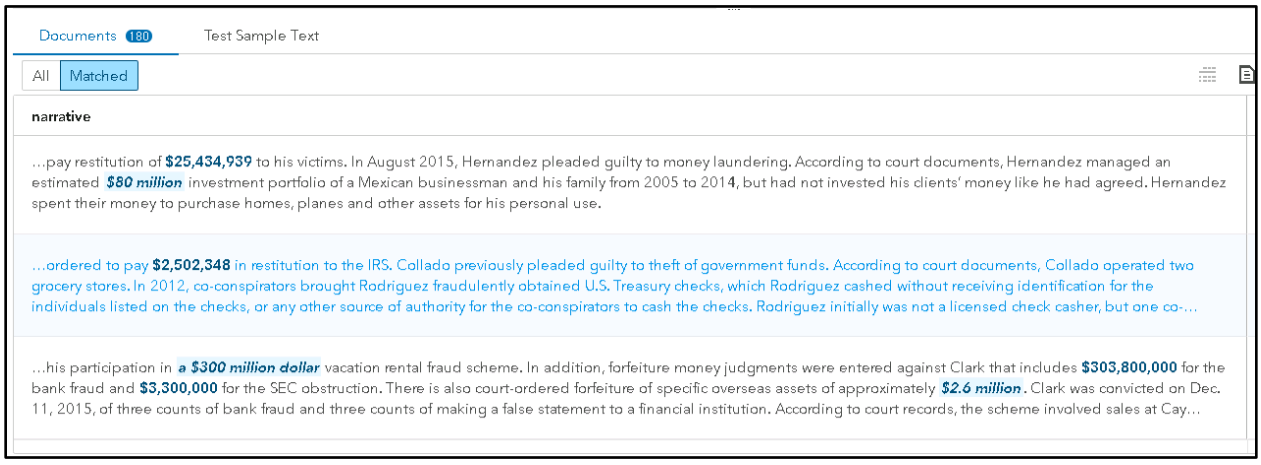

Figure 1. Matched Concepts for Predefined Concept nlpMoney

While the predefined concepts are valuable in and of themselves, they are also useful for referencing in your custom concepts. An example of this can be seen with our custom concept Fine_Amount. The predefined concept nlpMoney will extract out all references to money, but suppose we want to exclusively extract out the fines associated with each record for further analysis. Instead of filtering through all references to money, we can define a custom concept to pull out only currencies associated with a fine.

Figure 2 below shows the LITI syntax to generate this rule:

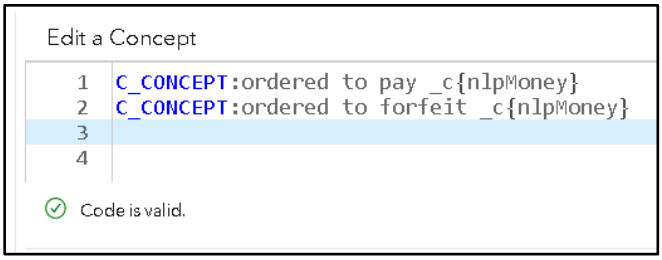

Figure 2. Custom Concept Fine_Amount LITI Syntax

The Fine_Amount custom concept uses the C_CONCEPT rule, which enables you to return matches that occur only in the context that we desire. In our case, we want to return the currency found by the nlpMoney predefined concept, but only in the context of a fine as in “ordered to pay” or “ordered to forfeit”.

A set of custom concepts was built on top of the predefined concepts to extract additional useful facts that could be helpful for indexing and searching, as well as additional analysis. Table 1 below summarizes the custom concepts that were developed, the type of concept used, and an example of the output.

Custom Concept	Concept Type	Example Output
Drug_Names	CLASSIFIER	Marijuana
Prison_Sentence	C_CONCEPT	60 months
Drug_Amount	CONCEPT_RULE	15 kilograms
Investment_Fraud_Amount	CONCEPT_RULE	$200 million
Investment_Fraud_Victims	CONCEPT_RULE	70 victims
Case_Charges	CLASSIFIER	Identity theft
Sentence_Location	CONCEPT_RULE	Providence, Rhode Island

Table 1. Custom Concept Definitions

TEXT PARSING

The next step in our analysis was to parse the text and create our term document matrix. In our model studio pipeline, this is done by dragging the Text Parsing node and placing it on top of the Concept node. SAS Visual Text Analytics allows you to customize how terms are parsed by configuring the minimum number of documents the term must be found in to be included for analysis, as well as using custom start, stop, and synonym lists. For the purposes of our example, we used the Text Parsing node to further explore some terms of interest for additional context and understanding. Figure 3 is an example of a term map used for exploration purposes.

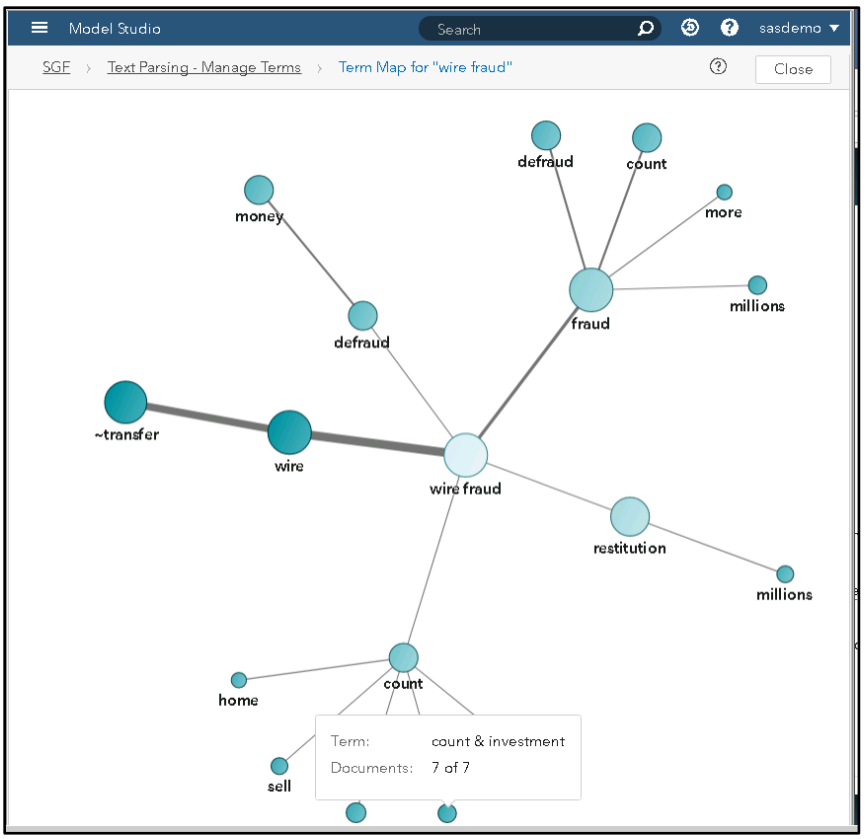

Figure 3. Term Map for "wire fraud"

TEXT TOPICS

Continuing with our analysis, we wanted to understand any relevant themes found in the data with the underlying terms that were parsed. For this, we dragged a Topic node and placed it on top of the Text Parsing node. SAS Visual Text Analytics allows you to automatically generate topics or choose the number of topics to generate, as well as set several other configurations including the term and document density. With a few iterations, we found the most informative results by setting the number of topics generated at 20, as well as term and document density of 2 and 1, respectively. Here is the output of the text topics.

Topic	Documents
+investor, +investment, +invest, capital, +return	34
cocaine, cocaine, +possess, +residence, +kilogram	28
marijuana, california, +sale, +drug, marijuana	28
+victim, costa, costa rica, rica, +co-conspirator	28
+church, +client, plan, boston, +asset	26
lee, +victim, portland, +live, oregon	25
+loan, +false statement, +statement, bank fraud, false	24
+check, +cash, +refund, +tax, +check	23
+report, +avoid, +structure, +casino, cash	23
+request, information, +order, +purchase, +supply	22
+stock, shell, u.s., arrest, +trade	21
equipment, +steal, carolina, north carolina, unlawful	21
fictitious, +employee, +client, +company, +create	20
+prescription, +patient, oxycodone, +physician, +substance	17
jr., +dollar, diego, united, san	17
+buyer, +mortgage, straw, +straw buyer, +application	16
construction, +bond, +bond, +contract, +project	16
law firm, firm, +law, +client, marijuana	16
silk road, silk, road, +user, +website	12
reserve, liberty, liberty, reserve, +user	9

Figure 4. Text Topics and Associated Document Count

Upon inspecting the topics, we were interested in two themes that were promoted to categories for ongoing analysis. The topics that were automatically generated provided a new lens on the data that we would like to track further and categorize new documents moving forward.

Topic Terms	Topic Theme	Percent of Documents
+buyer, +mortgage, straw, +straw buyer, +application	Real Estate Investment Fraud	9.4%
silk road, silk, road, +user, +website	Dark Web Drug Trade	7.0%

Table 2. Text Topics Promoted to Categories

TEXT CATEGORIES

Previously, we discussed text topics and the bottom-up approach of using the underlying terms to generate topics of interest. Our next step in our analysis was to take a top-down approach and define categories of interest using linguistic rules available in SAS Visual Text Analytics. In our model pipeline, this is done by dragging a Category node and placing it on top of the Topic node.

Categorizing your documents can be valuable for several reasons, such as creating tags for searching or for assigning similar documents for workflow purposes. Previously, we identified two categories of interest that we converted from the topics that were generated using the Topic node. In addition to these, we created a custom hierarchy of categorization that will help with future analysis. The table below shows the hierarchy of categories we were interested in.

Level 1	Level 2	Percentage of Matches
Drug Activity	Pharma Drugs	3%

	Illegal Drugs	15%
High Risk Customer Groups	Casino	3%
	Real Estate	23%
	Shell Company	3%
Financial Crime Charges	Bank Fraud	14%
	Bulk Cash Smuggling	4%
	Check Fraud	1%
	Identity Theft	6%
	Investment Fraud	8%
	Mail Fraud	16%
	Structuring	3%
	Tax Fraud	5%
	Wire Fraud	28%

Table 3. Custom Category Matches

Each category uses Boolean and proximity operators, arguments, and modifiers to effectively provide matches to only desired documents. Through the authors' domain expertise and the capabilities of SAS Visual Text Analytics, we were able to provide relevant matches on several categories of interest. An example of this concept is outlined below using the text category for the custom category "Identify Theft":

Figure 5. Text Category for "Identity Theft" with Matched Output

The "Identity Theft" rule can be broken up into two main components using the OR operator. The first component is simply looking for a direct match for the two sequential terms "identity theft", which provides several simple matches in the output found in the bottom of Figure 5. The second component uses the

SENT operator and will trigger a match if two sub-components exist in the same sentence somewhere within the document. The first sub-component is looking for some form of the word “identity” or a close combination of “personal” and “information”. The second sub-component is looking for the action of theft including terms such as “split”, “dual”, “stole”, “fabricate”, or “obtain”. The fourth and fifth matches in Figure 5 highlight the types of matches this will create in the form of “stolen identities” and “obtained identities” in the fourth and fifth match, respectively.

POST-PROCESSING

Once your project is set up in SAS Visual Text Analytics, you can produce score code and apply this to new data for ongoing tracking and monitoring. There are several types of post-processing that can happen depending on your use case and what the type of output you are working with. The most common types of post-processing can be found below:

- **Categorical Flags** – Typically, the presence or match for a category is used as a binary indicator for each document and can be used in filtering or searching, or as inputs to machine learning algorithms.
- **Network Analysis** – Extracted concepts such as locations, people, and organizations can be post-processed to show linkages and used as input to network diagrams for analysis.
- **Numerical Analysis** – Extracted concepts such as duration, fine amounts, or other numerical fields extracted from the documents can be post-processed to derive summarizations and averages of areas of interest.

CONCLUSION

There is a lot of excitement in the financial crime and compliance industry around the application of artificial intelligence and automation techniques. We see many opportunities available today to apply these methods to improve the effectiveness of detection programs and automate the manual tasks being performed by investigators. Text analytics is one area that has enormous potential, given that compliance departments have vast amounts of untapped, unstructured data sources. These sources contain rich information including who, where, what, when, and how that can be used as an input to many financial crimes use cases such as Negative News Monitoring, Trade Finance Monitoring, and SAR/STR Quality Assurance. With SAS Visual Text Analytics, banks can extract and derive meaning from text and organize it in a way that helps them perform these complex tasks that were previously accessible only through manual human review.

REFERENCES

UNODC (United Nations Office on Drugs and Crime). *n.d.* “Money-Laundering and Globalization.” Accessed February 20, 2018. Available https://www.unodc.org/unodc/en/money-laundering/globalization.html.

IRS (Internal Revenue Service). 2017. “Examples of Money Laundering Investigations - Fiscal Year 2017.” Accessed February 20, 2018. Available https://www.irs.gov/compliance/criminal-investigation/examples-of-money-laundering-investigations-for-fiscal-year-2017.

ACKNOWLEDGMENTS

The authors would like to thank David Stewart for his guidance and thought leadership regarding AML compliance. In addition, we would like to thank Adam Pilz for his guidance and thought leadership regarding text analytics.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors:

Austin Cook
100 SAS Campus Drive
Cary, NC 27513

SAS Institute Inc.
Austin.Cook@sas.com
http://www.sas.com

Beth Herron
100 SAS Campus Drive
Cary, NC 27513
SAS Institute Inc.
Beth.Herron@sas.com
http://www.sas.com