Modeling and Text Analysis to Empower FAERS Adverse Event Assessment

Qais Hatim[1], Lilliam Rosario[1], Eileen Navarro Almario[1], Kendra Worthy[1], Tom Sabo[2], Emily McRae[2], Monica Munoz[1], Marc Stone[1], Sonja Brajovic[1], Allen Brinker[1], Soundar Kumara[3]

ABSTRACT

Timely adverse event pattern detection and resolution, related to drugs, vaccines, and medical devices, can save lives and further quality of life. However, determining conclusions from the data can prove too time consuming for manual analysis, and can produce qualitative results at best.

This paper applies a combination of exploratory and predictive text analytic techniques against FAERS data to signal primary textual cues from narratives associated with serious adverse drug events. Specifically, we will run models to generate Boolean combinations of terms and phrases for serious versus non-serious events, including the absence of terms. Additionally, we will assess two characteristic scores tied to drug associations with adverse events via association and market basket analysis. This project provides global observations about the relationship between drugs and adverse events, depicts trends and characterizes the details of serious events, and highlight events misclassified as non-serious for re-investigation. This project serves as proof of concept that modeling and text analysis is useful in adverse event review.

INTRODUCTION

Many industries have realized the benefits of deriving valuable insight from converting unstructured corpus of documents into structured data and applying data mining on the structured data. Firms with efficient algorithms for text mining have a competitive advantage over those who do not. Many researchers have published work related to applications of text mining in various domains. These applications mainly fall under the general categories of text categorization, information retrieval and measurement (Miller, 2005). In recent years, text mining has been applied in health care systems to discover adverse event patterns and trends in textual data. Natural Language Processing (NLP) is a widely-used method for knowledge extraction that, given a set of medical reports, can be used to structure narrative clinical data and classify content into different topics that exist in the text. NLP is especially powerful in extracting predefined patterns (or existing knowledge) that can help in mining observations and descriptive modifiers from free-text reports, for future usages by machine learning algorithms.

Typically, several pipelined NLP subtasks are used by text mining process to format the text in preparation for the statistical analysis or pattern discovery phase. The subtasks include a set of foundational low-level syntactic tasks, and a set of high-level tasks that build on the low-level tasks and involve semantic processing (Table 1) (Harpaz, et al., 2014).

1 U.S. Food and Drug Administration, 2 SAS Institute Inc., 3 Pennsylvania State University/University Park

Task	Description
Segmentation	Splitting a document along sentence and section boundaries
Tokenization	Splitting sentences up into their parts—individual words and punctuation
Part of speech (POS) tagging	Assigning grammatical parts of speech to individual tokens e.g. 'drug' is a noun, 'administers' is a verb, 'quickly' is an adjective, 'the' or 'a' are determiners
Parsing	Determining the grammatical structure of sentences and the relationship between groups of words that together form noun phrases, verb phrases, clauses, etc. Shallow parsing, often used instead of deep parsing, only identifies the constituents (e.g. noun phrases) but not the internal structure of the sentence
Named entity recognition (NER)	Identifying terms or phrases of interest ('entities') in the text. NER may go beyond just recognizing terms to also categorizing, normalizing, and mapping them to standardized vocabularies, e.g. identifying 'rofecoxib' as a drug, and 'myocardial infarction' as a medical condition
Negation detection	Determining whether a named entity is present or absent, e.g. 'patient does not exhibit symptoms of …', 'patient was ruled out for myocardial infarction'
Word sense disambiguation (WSD)	Determining which sense of a homograph (words with identical spellings but different meanings) is appropriate in the context of the sentence
Temporal inference	Establishing temporal order of events from text, e.g. 'adverse event occurred after prescription of drug'
Relation detection	Determining whether two or more named entities recognized in the text form specific relationships, e.g. 'drug A treats disease B', 'drug A induces disease B'

Table 1: Natural Language Processing Subtasks

Various statistical classification and machine learning techniques can be applied to text categorization, including regression models, nearest neighbor classifiers, decision trees, Bayesian classifiers, Support Vector Machines, rule learning algorithms, relevance feedback, voted classification and neural networks (Aas & Eikvil, 1999). *Aas et al.* enumerated and analyzed the main methods and algorithms used within text mining process. Text classification generally involves a multi-step process (see Figure 1, reproduced with permission from Ikonomakis, Kotsiantis, & Tampakas, 2005).

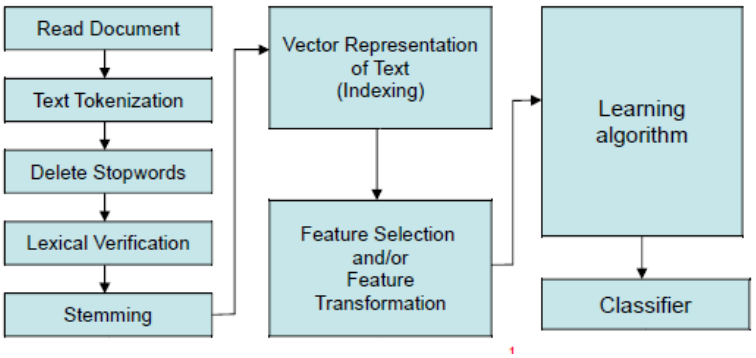


Figure 1: Text Classification Process

During the post-marketing period when drugs are used on larger populations and for more varied periods of time, unexpected adverse drug events may occur, which alter the risk–benefit ratio enough to require regulatory action. Post-marketing adverse events impose a significant burden for healthcare and result in unnecessary, often fatal, harm to patients (Harpaz, et al., 2013). Therefore, a critical component in pharmacovigilance is the early discovery and analysis of emerging adverse event signals in the post-marketing period.

The Food and Drug Administration (FDA) uses the FDA adverse event reporting system (FAERS) to compute signal scores (statistical reporting associations) for the millions of drug–event combinations in its adverse event reporting system. Nevertheless, these signals by themselves do not establish a causal adverse drug reactions relationship, but are rather considered initial warnings that require further assessment by domain experts to establish causation. This further assessment normally consists of a complicated process whereby drug safety evaluators look for supporting information such as temporal relationships, published case reports, biological and clinical plausibility, clinical trials data, and epidemiological studies in several large healthcare databases (Schneeweiss, 2010).

Much of the supporting information is textual data (i.e. unstructured data) that requires manual review by safety experts. Specifically, the experts review the unstructured free text fields to identify the clinical entities in any given case report, decide upon the acquisition of additional information (e.g., request a copy of the medical records), and consider whether any regulatory action is warranted. FAERS case reports should be distinguished from any other type of medical documentation (e.g., discharge summaries), since both experts (physicians) and non-experts (patients and relatives) can reporter these adverse events. Consequently, special processing is needed to handle the frequent non-medical syntax and semantics.

A FAERS report contains both structured (e.g., drug dosage and case event date) and unstructured data (e.g., product reported verbatim). To meet the challenges posed by unstructured text, text mining, which employs a wide range of statistical, machine learning, and linguistic techniques for natural language processing (NLP), is utilized in this paper. Moreover, while some of the FAERS information is structured, a significant portion of it remains in narrative formats. Much of the information that is critical to risk assessments, such as signs and symptoms, disease status and severity, and medical history, are typically imbedded in narrative text. Accordingly, we exploited both the structured and unstructured FAERS data items and developed different analytical models such as text rule builder, decision tree and neural network to extract meaningful information to understand the themes related to drug adverse events. Quantitatively processing this extracted text will helped us understand which adverse events are most common within a cluster of FDA-approved drugs. In particular, we addressed adverse events related to hepatic failure and its range of consequences when used in broader patient populations and with other concomitant medications. By applying computational approaches in the analysis of FAERS data, we will advance the understanding of the feasibility of applying thee techniques to uncover relationships between these drugs and hepatic failure, offer information on the drug combinations related to hepatic failure, clarify factors that can predict a greater degree of the hepatic failure from serious (death) to minor events (treatable), rate the serious impact of using these drugs on patients.

DATA DESCRIPTION and PREPARATION

In any analytical project, data preprocessing is the most tedious and time-consuming task. Data analysts state the 80/20 rule that 80% of the project timeline is dedicated for data preparation and only 20% of the timeline is utilized for analysis and model building. The following section describes this process, starting from downloading FAERS data, through data pre-processing to gain acquaintance with data content and ending with an exploratory analysis.

We utilized the Drug Safety Analytics Dashboards (MERCADO) to retrieve FAERS reports of hepatic failure adverse events received from November 1997 up to March 2018. However, some of the reports received prior to November 1997 did not contain narrative text in FAERS. Since drug-induced liver injury (DILI) is increasingly being recognized as a cause of clinically significant acute and chronic liver disease (Fontana, et al., 2010), we customized our event definition using Standard MedDRA Query (SMQ) in order to select drug related hepatic disorders-severe events only with narrow scope searches. Using such custom search enabled us to group terms from one or more MedDRA System Organ Classes (SOCs) terms related to the defined medical condition or area of interest inclusive of terms of signs, symptoms, diagnoses, syndromes, physical findings, laboratory and other physiologic test data, etc., related to the medical condition or area of interest, in this case, DILI, with greater specificity. Data was downloaded in seven time intervals since MERCADO allows only around 40,000 cases to be retrieved in one search. Some obsevational analysis was exploited at each time interval in order to understand the characteristics of the data at each time interval. For instance, for a time interval from January 01, 2014 till December 31, 2016 the following analysis, not all inclusive, was performed (Tables 2,3,4,5,6) .

Patient Sex	Total Cases	% of Cases
Female	19,567	45.7%
Male	18,250	42.6%
Not Reported	4,990	11.7%
Unknown	12	0.0%
Total	**42,819**	**100.0%**

Table 2: Case Count by Patient Sex

Reported Outcomes	Total Cases
Death	9,324
Hospitalized	20,190
Life Threatening	3,027
Disabled	965
Congenital Anomaly	58
Required Intervention	115
Other Outcome	27,090
Total (Distinct Cases)	**42,819**

Table 3: Case Count by Reported Outcomes

		Report Type			Seriousness		Reported Outcomes						
Age Group	Total Cases	Direct	Expedited	Non-Expedited	Non Serious	Serious	DE	HO	LT	DS	CA	RI	OT
<1 year	134	3	125	6	1	133	50	58	21	4	5	0	69
1 - <3 years	174	16	153	5	3	171	67	84	15	2	2	0	76
3 - <7 years	225	23	198	4	7	218	50	110	31	0	0	1	119
7 - <17 years	770	49	685	36	16	754	131	397	71	14	1	3	469
17 - <65 years	19,456	901	17,385	1,170	753	18,703	3,849	10,106	1,530	482	8	70	12,100
>=65 years	10,444	363	9,559	522	320	10,124	2,811	5,819	934	242	2	36	6,139
NOT REPORTED	11,616	112	10,157	1,347	1,057	10,559	2,366	3,616	425	221	40	5	8,118
Total	**42,819**	**1,467**	**38,262**	**3,090**	**2,157**	**40,662**	**9,324**	**20,190**	**3,027**	**965**	**58**	**115**	**27,090**

Table 4: Displays total case count by age group, report type, seriousness and outcome.

		Report Type			Seriousness		Reported Outcomes						
Country	Total Cases	Direct	Expedited	Non-Expedited	Non Serious	Serious	DE	HO	LT	DS	CA	RI	OT
Foreign	27,364	34	26,912	418	187	27,177	6,420	13,362	2,207	525	31	26	17,840
USA	15,429	1,412	11,347	2,670	1,964	13,465	2,897	6,815	814	440	27	88	9,241
Not Reported	26	21	3	2	6	20	7	13	6	0	0	1	9
Total	**42,819**	**1,467**	**38,262**	**3,090**	**2,157**	**40,662**	**9,324**	**20,190**	**3,027**	**965**	**58**	**115**	**27,090**

Table 5: Displays total case count by country, report type, seriousness and outcome.

	Report Type		
Initial FDA Received Year	Direct	Expedited	Non-Expedited
2016	445	14,392	998
2015	547	12,950	1,303
2014	475	10,920	789

Table 6: Displays Case Count by Initial FDA Received Year or Event Year

After a high level understanding of our data corpus of 304,000 cases, we prepared the data for both unsupervised and supervised learning. In unsupervised learning, for instance, it is important to reject variables which are unnecessary or irrelevant to the stated objective(s), in our case, a binary objective serious vs. non-serious event of liver failure. For example, the basis variables used in the unsupervised learning, clustering algorithms, should be meaningful to the analysis objective; have low correlation between input variables; and have low kurtosis and skewness in order to reduce the possibility of producing small outlier clusters for DILI cases. Likely basis variables include case demographic, products information, patient history, report type, and reporter information. Moreover, intervals variables have a propensity to take over a cluster information, when assessed as categorical variables. Since our data consisted largely of class variables and text, we utilized text mining techniques such as text clustering, text rule builder, and text profile to transform these data into interval variables. More than 241 variables were

available for modeling but these were reduced to a smaller set, 121 variables, which had the possible to be analytically beneficial.

Since the data is dominated by cases with serious outcome value of Yes (Y=1), building any model with such dominant outcome will be biased towards predicting serious adverse event. To compensate for the rare proportion of No (No=0) in the raw data, over-sampling of the data was done to produce a balanced data and to enable to the patterns that appear in the whole data to be traceable in the sample. Over-sampling rare classes often leads to more accurate predictions (ref?). To illustrate the data over-sampling, we used the FAERS data collected till December 31, 2000. Figure 2 shows that 88% of the target level was Yes (Y=1) while only 12% was No (N=0).

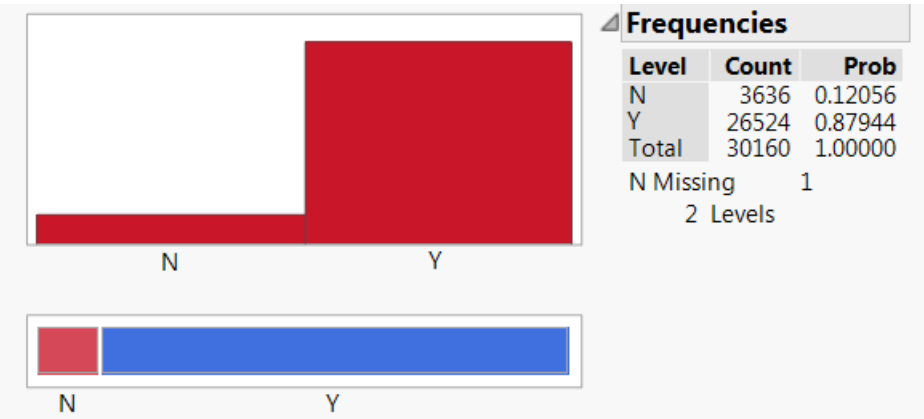


Figure 2: Summary Statistic for Serious Outcome

Oversampling was employed to create a frequency variable with sampling weights, and our final final No (N=0) proportion was increased to 34%.
Before building any predictive models and in order to get the correct decision consequences, we specify the inverse priors based on the original proportion of rare events (12%) to correctly adjust model predictions regardless of what the proportions in the training set are. If no adjusted prior probabilities are used, the estimated posterior probability for the No event class will be over-estimated. SAS Enterprise Miner uses profit matrix with elements equal to the inverse of the prior distribution for each outcome instead of a traditional profit matrix. The reason for such modification (i.e. inverse prior distribution) is to get accurate specification from model-based decision since it is a difficult, if not impossible, task to tune and assess predictive models based on the profit or loss consequence of model-based decision (SAS Course Note, 2016).

$$Let\ \pi_i = prior\ distribution\ for\ serious\ outcome\ at\ level\ i$$

Therefore, the inverse prior profit matrix for serious outcome, binary target, will be

		Decision	
		1	0
Serious Outcome	1	$\frac{1}{\pi_1}$	0
	0	0	$\frac{1}{\pi_0}$

Using the above modification, cases predicted more likely than average to have serious outcome with level=1 (primary outcome) will receive primary decision (decision=1). This instinctively related to the fact that the inverse prior profit matrix for binary response variable allocates decision=1 to each case with a posterior probability more than prior distribution probability for serious outcome at level 1(i.e. π_1).

Data set was partitioned in a proportion of 60/30/10 for the training, validation, and test, consecutively. The 60% training data portion was used for preliminary model fitting. The 30% validation data portion was used for assess the adequacy of the model in models comparison as well as to prevent model overfitting. For obtaining a final, unbiased estimate of the generalization error of the model (i.e. how well the model generalizes on true hold out data), the 10% test data portion was used.

METHODOLOGY: MATERIALS and METHODS

1. TEXT MINING

Text Mining starts with text parsing which identifies unique terms in the text variable and identifies parts of speech, entities, synonyms and punctuation (Rajman & Besancon, 1997). The terms identified from text parsing are used to create a term-by-document matrix with terms as rows and documents as variables. A typical text mining problem has more terms than documents resulting in a sparse rectangular terms-by-document matrix. Stop lists help in reducing the number of rows in the matrix by dropping some of the terms (SAS Enterprise Miner, 2018). Stop list is a dictionary of terms that are ignored in the analysis. A standard stop list removes words such as "a, about, again, and, after, etc." However, a custom stop lists can be designed by analyst to obtain more informative text mining results. Based on the preliminary analysis for our aggregated data as well as by communicating with experts at the FDA, we created a custom stop list that includes terms appearing in fewer than 5 FAERS cases as well as terms with highest frequencies (i.e. drop term with frequency more than 6000). These terms are deemed as to not add any value to the analysis. Examples of such terms in the custom stop lists are patient, drug,

liver, FDA, and so on. We also created a custom synonym data set using the terms extracted from the four data sets. For instance, terms hepatic, hepaticopsida, leafy liverwort, and liver failure are considered as synonyms for this research.

Even after using customized stop lists, in a corpus of several thousands of documents, the term-by-document matrix can contain hundreds and thousands of terms. It becomes computationally very difficult to analyze a matrix with high dimensional sparse data. Singular Value Decomposition (SVD) creates orthogonal columns that characterize the terms data set in fewer dimensions than the document by term matrix. Therefore, SVD can be used to reduce the dimensionality by transforming the matrix into a lower dimensional and more compact form. A high number of SVD dimensions usually summarizes the data better but requires a lot of computing resources. In addition, the higher the number, the higher the risk of fitting to noise (Sanders & DeVault, 2004). However, a careful decision needs to be made on how many SVD high dimensions to use. A high number for SVDs can give better results, but high computing resources are required. It is recommended to try low, medium, and high different values for number of dimensions and compare the results. In this paper, we selected 25 SVD dimensions. Each term identified in text parsing is given a weight based on different criteria. The log frequency weighting (local weights) is selected to assign weights to term/document matrix to control the effect of high-frequency terms in a document. Moreover, mutual information is selected for term weights (global weights) to help in identifying significant terms in separating cases from other cases in the corpus by distinguishing terms that occur in only few documents, but occur many times in those few documents.

2. TEXT CATEGORIZATION-CLUSTERING

Clustering technique is used for text categorization. In this paper, the text cluster uses a Expectation-Maximization clustering algorithm to describe the contents clusters. Since we described to display m=25 descriptive terms for each cluster, then the top 2*25 most frequently occurring terms in each cluster are used to compute the descriptive terms. For each of the 2*25 terms, a binomial probability for each cluster is computed. The probability of assigning a term to cluster j is prob=F(k|N, p) where F is the binomial cumulative distribution function, k is the number of times that the term appears in cluster j, N is the number of documents in cluster j, p is equal to (sum-k)/(total-N), sum is the total number of times that the term appears in all the clusters, and total is the total number of documents. The m=25 descriptive terms are those that have the highest binomial probabilities.

The primary purpose of the text cluster is to derive clusters. This is achieved by deriving a numeric representation for each document. Latent Semantic Analysis (LSA), implemented through Singular Value Decomposition (SVD), produces the numeric representation and the derived numeric representation produces the clusters. The SVD is applied to approximate the term/document matrix, and documents are projected into a reduced dimensional space. The

generated SVD dimensions are those that fit the subspace the best in terms of least-square best fit.

3. TEXT CATEGORIZATION- TEXT RULE BUILDER & MEMORY-BASED REASONING

Both text rule builder and memory-based reasoning (MBR) are implemented in this paper. The text rule builder provides a stand-alone predictive modeling solution for supervised text categorization. In the text rule builder, Boolean rules are created from small subsets of terms to predict a categorical target variable, in our case the binary serious outcome. To prevent overtraining, control the complexity of rules, and determine the exhaustiveness of the rule search process, different settings for these specifications were tried and several analyses were compared to set the optimal text rule builder setting.
Memory-based reasoning is a process that identifies similar cases and applies the information that is obtained from these cases to a new record. The memory-based reasoning (MBR) uses a k-nearest neighbor algorithm to categorize observations. The k-nearest neighbor algorithm takes a data set and a probe, where each observation in the data set is composed of a set of variables and the probe has one value for each variable. The distance between an observation and the probe is calculated. The k observations that have the smallest distances to the probe are the k-nearest neighbor to that probe. The k-nearest neighbors are determined by the Euclidean distance between an observation and the probe. Based on the target values of the k-nearest neighbors, each of the k-nearest neighbors votes on the target value for a probe. The votes are the posterior probabilities for the binary or nominal target variable (SAS Reference Help, 2018).

4. TEXT TOPIC

To explore the document collection, the text topic technique is used to automatically associate terms and documents according to both discovered and user-defined topics. Topics are collections of terms that describe and characterize a main theme or idea. However, the approach is different from clustering because clustering assigns each document to a unique group while the Text Topic node assigns a score for each document and term to each topic. Then thresholds are used to determine if the association is strong enough to consider that the document or term belongs to the topic. Thus, documents and terms may belong to more than one topic or to none. Since the number of topics are directly related to the size of the document collection, we created user-defined topics in addition to the discovered topics to add expert knowledge in the analysis.

5. NEURAL NETWORK-SUPERVISED LEARNING

Neural networks attempt to mimic key aspects of a brain, in particular its ability to learn from experience. Although inspired by cognitive science, neurophysiology, neural networks largely draw their methods from statistical physics (Hertz, Krogh, & Palmer, 1991). There are dozens, if not hundreds, of neural network algorithms. But despite this plethora of models, all neural networks fall into one of two broad classes supervised learning and unsupervised learning.
In this paper, we used supervised learning, in particular, on the multilayer perceptron and radial basis function algorithms.
In some ways, neural networks are similar to regressions. The most prevalent problem for neural networks is missing values. Like regressions, neural networks require a complete record for estimation and scoring. Neural networks resolve this complication in the same way that regression does by imputation. Also, extreme or unusual values also present a problem for neural networks. The problem is mitigated somewhat by the hyperbolic tangent activation functions in the hidden units. These functions compress extreme input values to between *-1* and *+1*.
In other ways, neural networks are different from regressions. Nonnumeric inputs pose less of a complication to a properly tuned neural network than they do to regressions. This is mainly due to the complexity optimization process. Unlike standard regression models, neural networks easily accommodate nonlinear and non-additive associations between inputs and target. In fact, the main challenge is over-accommodation—that is, falsely discovering nonlinearities and interactions.

6. DECISION TREE

The goal of data mining is to create a good predictive model, which provides us with knowledge and the ability to identify key attributes of business processes that target opportunities (for example, target customers, control risks, or identify fraud). Decision tree models represent one of the most popular types of predictive modeling. Decision trees partition large amounts of data into smaller segments by applying a series of rules. These rules split the data into pieces until no further splits can occur on those pieces. The goal of these rules is to create subgroups of cases that have a lower diversity than the overall sample of population. The purpose of partitioning the data is to isolate concentrations of cases with identical target values. Decision trees are visually represented as upside-down trees with the root at the top and branches emanating from the root. Branches terminate with the final splits (or leaves) of the tree.
In this paper, we utilized decision tree to perform the three essential tasks that predictive models perform which are predict new liver failure cases, select useful inputs, and optimize complexity. Each of these essential tasks applies to a general principle as shown in (Table 7) below. Decision trees, similar to other modeling methods, address each of the modeling essentials. Cases are scored using prediction rules. A split-search algorithm facilitates input selection. Model complexity is addressed by pruning.

Predictive Modeling Task	General Principle	Decision Trees
Predict new cases	Decide, rank, or estimate	Prediction Rules
Select useful inputs	Eradicate redundancies and irrelevancies	Split Search
Optimize complexity	Tune models with validation data	Pruning

Table 7: Decision Tree Essential Tasks

Moreover, we utilized these three methods for constructing decision tree models which are interactive method or by hand, the automatic method, and the autonomous method. Many parameter settings for building decision tree have been adjusted in this work. These parameters can be divided in to five groups 1) the number of splits to create at each partitioning opportunity, 2) the metric used to compare different splits, 3) the rules used to stop the autonomous tree growing process, 4) the method used to prune the tree model, and 5) the method used to treat missing values.

RESULTS and DISCUSSION

Several analytical models have been developed in this paper, although we limited our result and discussion section for some of these analyses. In the following section, we sequentially discuss the text filtering and important concept links for selected terms, text rule builder, and decision tree. Moreover, we illustrate a model comparison for selecting a champion model based on the specified fit statistic. Finally, we used the winning model for scoring new data.

1. TEXT FILTER and CONCEPT LINKS

Text filtering is used to reduce the total number of parsed terms or documents that are analyzed. Therefore, we eliminated unnecessary information so that only the most valuable and relevant information is considered. Experimental analysis and subject matter expert input were applied to remove unwanted terms and to keep only documents that discuss a liver injury, to develop a smaller corpus of useful distinct terms.

Zipf's Law identifies important terms for purposes such as describing concepts and topics. The number of meanings of a word is inversely proportional to its rank (Konchady, 2006). Figure 3 exhibits the exponential decay for the Zipf's Law which is typical for the English language and indicates that our data does not deviate from this law.

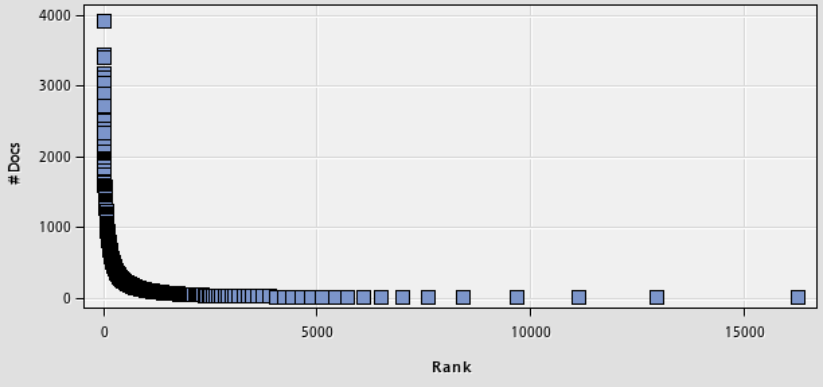


Figure 3: Zipf Plot

The number of documents by frequency plot (Figure 4) exhibits a monotonic behavior indicating the usefulness of our data preprocessing methods. Frequency counts that deviate substantially from an approximate linear relationship are considered suspicious and usually indicate data quality problem. Therefore, our results indicate that we can proceed with modeling with a reasonable likelihood of obtaining informative conclusions.

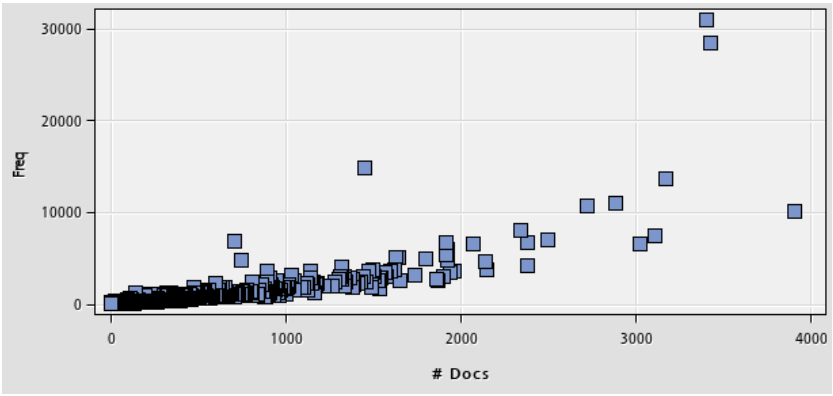


Figure 4: Number of Documents by Frequency

Domain knowledge was utilized to suggest expected frequency distribution for the Role by Freq table. Based on the guidelines from the FDA medical officers as well as some other health care experts, verbs, nouns, adjectives, noun group, and miscellaneous proper nouns should be the expected frequency distribution in the FAERS data (Figure 5).

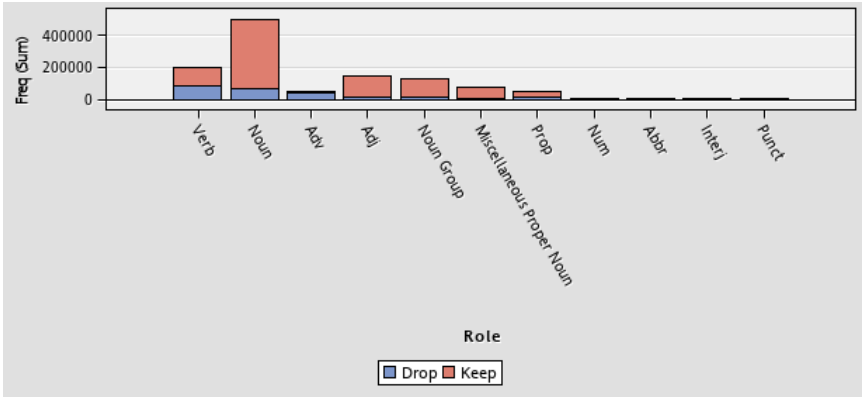


Figure 5: Role by Frequency Distribution

To understand the association between words identified in the corpus, concept linking is an interactive view that illustrates for a given pair of terms their strength of association with one another which is computed using the binomial distribution (Cerrito, 2006). Concept linking is graphical representation where the width of the line between the centered term and a concept link represents how closely the terms are associated. A thicker line indicates a closer association. As an example, Figure 6 shows below the concept linking for the noun group "hepatic failure". decades.

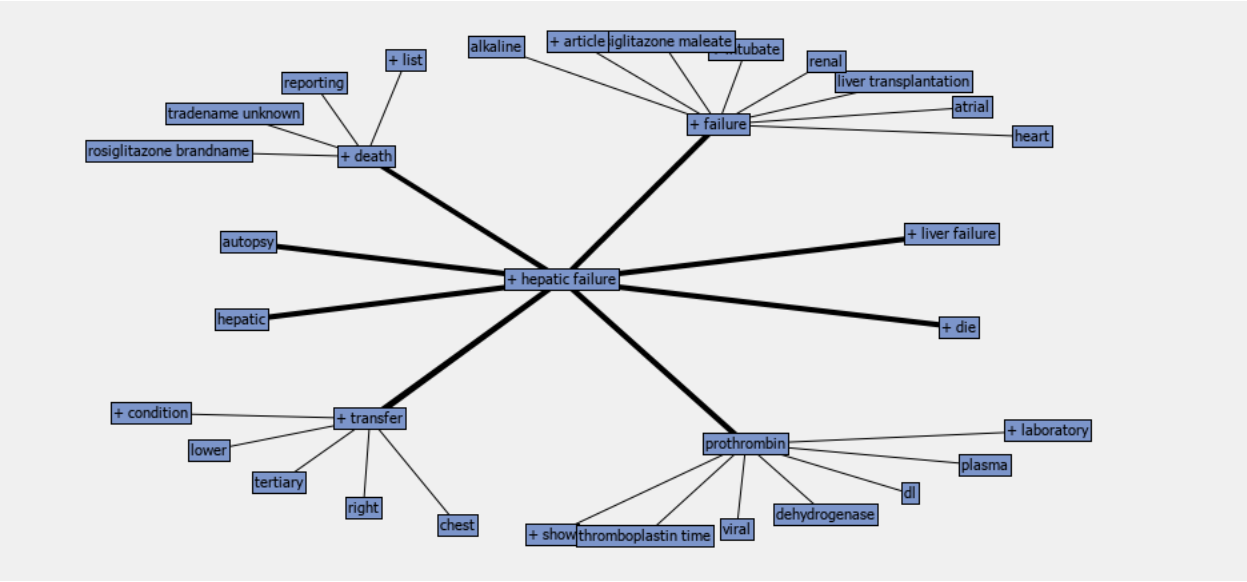


Figure 6: Hepatic Failure Concept Linking

The concept hepatic failure is mainly associated with terms such as autopsy, transfer, prothrombin, die, failure, etc. which indicates that FAERS cases with hepatic failure mentioned in the case narrative might be serious adverse events or need careful investigation for identifying the causality of death. The association between the hepatic failure and death can be obtained as the following. Let n be is the number of documents that contain term death and let k be the number of documents containing both term hepatic failure and term death. Then, p = k/n is the probability that term death occurs when term hepatic failure occurs, with the assumption that they are independent of each other. Death appears in 412 cases and out of these cases, 81 cases death appears with hepatic failure. Then the actual metric that is used to judge strength of association between the hepatic failure and death, for a given r cases, is as follows:

$$Strength = \log_e\left(\frac{1}{Prob_k}\right) \text{ Where } Prob_k = \sum_{r=k}^{r=n} prob(r) \text{ and } prob(r) = \left[\frac{n!}{[r!(n-r)!]}\right] p^r (1-p)^{(n-r)}$$

Therefore, the strength between death and hepatic failure is around 65% which is also showed in the graph with thicker line, closer association. This information can be made available to domain

experts and intuitively understand the association between the terms of interest and care such association and knowledge for further analysis.

2. TEXT-BASED RULE BUILDER

As we mentioned above, that text rule builder is a powerful tool for text categorization. We developed several text rule builder models so that we can detected the settings that are most likely to wrongly predict seriousness of liver injury based on the generated model. Moreover, we utilized active learning (i.e., dynamically interact with the algorithm to iteratively build a predictive model) by altering the target that is assigned to some of these cases and rerun the analysis. The results for the aggressive rule builder scenario will only be discussed in this section. The generalization error which determines predicted probability for rules on untrained data is set very high to exhaustively control the number of terms tested and to decide to add to or replace a rule conjunction. Whereas purity of rules is set very low so that more rules will be generated to cover more cases. The effectiveness of the rules in classifying the FAERS cases are graphically displayed in Figure 7. We can examine whether a particular rule tends to favor a particular outcome by visualizing the rules that dominated by one color. With 53 rules are generated, most of our rules are accurate in classifying the serious outcome (i.e., Y vs. N).

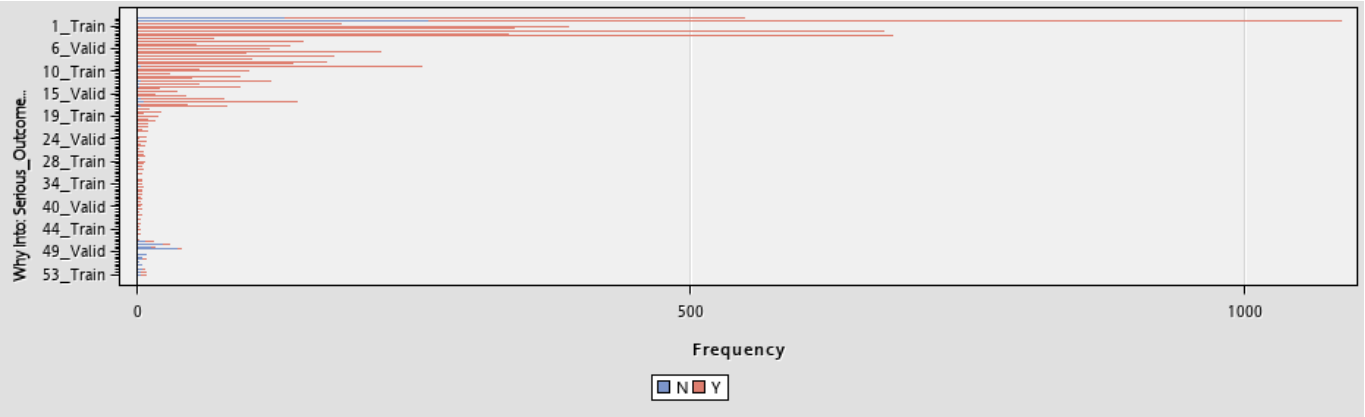


Figure 7: Rule Success

The first 13 rules obtained are shown below (Figure 8). The first rule is ***clinical outcome***. This rule probably indicates that serious hepatic outcomes can be realized from the patient clinical outcome. However, clinical outcome can be wide-ranging based on the type of the test, patient health status and history, etc. In general, detecting clinical outcome in FAERS cases narratives will predict seriousness with specific precision and recall. For the FAERS training data set, 390 cases exhibited the term clinical outcome, and 390 these documents were flagged as a serious outcome equal Y. This produces a precision rule for this rule of 390/390=100%. On the other hand, the training data has 4589 training FAERS cases with serious outcome=Y, so recall for this rule is 390/4589= 8.5%. When the term (i.e., clinical outcome) exists in a FAERS case report, it is almost certainly a serious adverse event, but there are many serious adverse events that do not use this term. The harmonic mean (F1 score) measures the trade-off between precision and recall.

Target Value	Rule #	Rule	Precision	Recall	F1 score	Valid Precision	Valid Recall	Valid F1 score	True Positive/ Total	Valid True Positive/ Total
Y	1	clinical outcome	100.0%	8.50%	15.67%	100.0%	8.06%	14.92%	390/390	185/185
Y	2	concomitant medication & drug	100.0%	23.19%	37.64%	99.62%	22.83%	37.15%	722/722	370/372
Y	3	medical & ~drug induced liver injury & ~auto-antibody & ~opioid dependence & ~epstein-barr & hepatitis	100.0%	38.05%	55.12%	99.77%	37.47%	54.48%	1,093/1...	571/571
Y	4	mg	100.0%	41.32%	58.47%	99.79%	40.48%	57.59%	334/334	156/156
Y	5	article	100.0%	44.35%	61.44%	99.80%	42.83%	59.94%	243/243	88/88
Y	6	information & ~treatment information & ~auto-antibody & ~normal liver & ~fall & initial	100.0%	49.16%	65.92%	99.73%	48.02%	64.82%	580/580	309/310
Y	7	rezulin & ~lfts	100.0%	53.04%	69.32%	99.75%	52.29%	68.61%	582/582	292/292
Y	8	iu & ~healthy	100.0%	56.77%	72.42%	99.77%	56.82%	72.40%	697/697	356/356
Y	9	failure & ~normal	99.83%	62.28%	76.70%	99.65%	62.88%	77.10%	698/703	375/377
Y	10	suspect	99.76%	64.44%	78.30%	99.53%	65.23%	78.81%	623/625	317/319
Y	11	induce & ~ana	99.77%	66.49%	79.80%	99.41%	66.45%	79.66%	460/462	213/215
Y	12	significant & ~specific	99.65%	69.03%	81.57%	99.37%	68.58%	81.15%	900/906	465/467
Y	13	daily	99.57%	71.02%	82.91%	99.21%	70.89%	82.69%	1,163/1...	607/613

Figure 8: Rules Obtained

$$F1 = \frac{1}{\left(0.5*\left(\frac{1}{Precision}\right)+\left(0.5*\left(\frac{1}{Recall}\right)\right)\right)};$$

The F1 gets larger as precision and recall get closer to each other in value. Because recall is so small, the F1 value for the first rule is small: F1=15.67%.

The statistics in Figure 8 are cumulative. The second rule is ***concomitant medication & drug*** which is interpreted to apply when FAERS case contains the word concomitant medication and does contain another drug that taken by this case. After removing the 390 cases that contain the term clinical outcome, there are 722 training cases that satisfy the rule, and 722 have serious adverse event (i.e., serious outcome=Y). The overall precision for the first two rules is

Precision= (390+722)/ (390+722) =100%.

To score a case, the rules are applied in order. If a rule is satisfied, then the target value associated with the rule is assigned, and the variable w-serious outcome (Why the observation was assigned the INTO value that it was. Why Into: Serious Outcome) is assigned the rule number that triggered the classification. If no rule is satisfied, the secondary event target value is assigned (i.e., for a binary target, the primary event value is 1 and the secondary event value is 0). If no rule applies to a case, then w-serious outcome is assigned a missing value, and I_serious outcome (formatted predicted target value with classification role) is assigned a value of 0. As shown in Figure 9 below that the dominant rule is rule 2. A total of 27 FAERS cases do not satisfy the first rule, but do satisfy the rule number 2. The frequency 27 is for the random sample of FAERS cases selected for exploration. However, the total of cases in the training data set that satisfy rule 2 are 218. This number has been obtained after running the SAS code node running PROC FREQ.

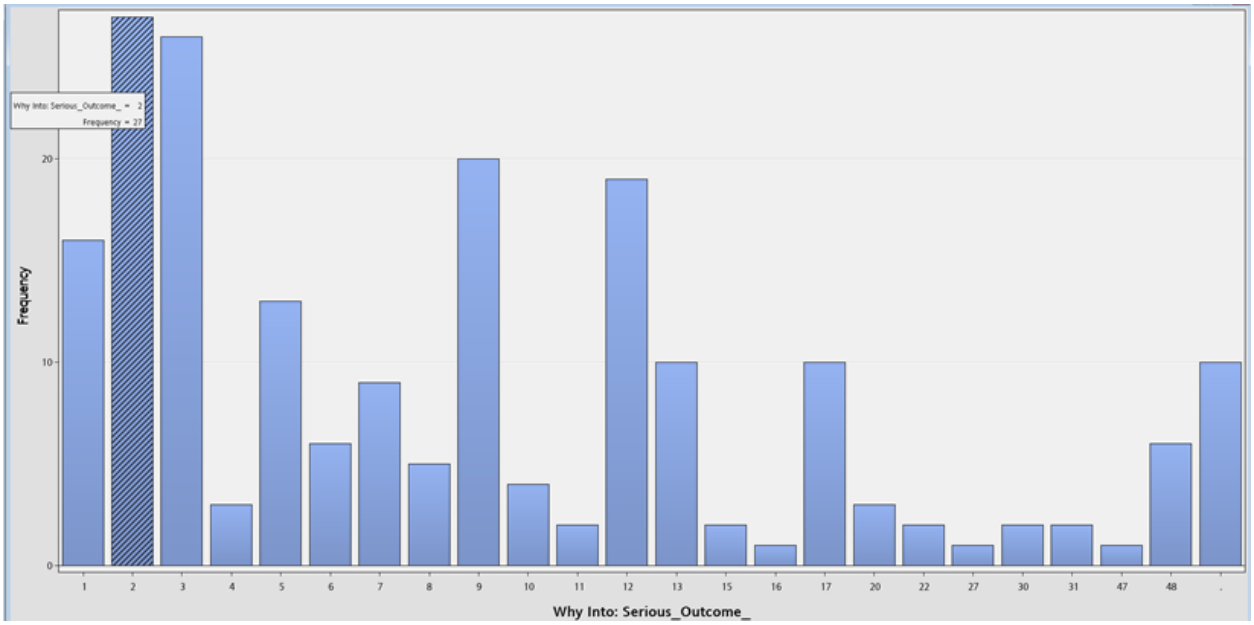


Figure 9: Rules by their frequency

3. DECISION TREE

Decision tree models are constructed using a recursive algorithm that attempts to partition the input space into regions with mostly primary outcome cases and regions with mostly secondary outcome cases. Model predictions are based on the percentage of primary outcome cases found in each partition. The models can easily accommodate missing values and therefore do not require imputed data. Decision tree models make few assumptions regarding the nature of the association between input and target, making them extremely flexible predictive modeling tools.

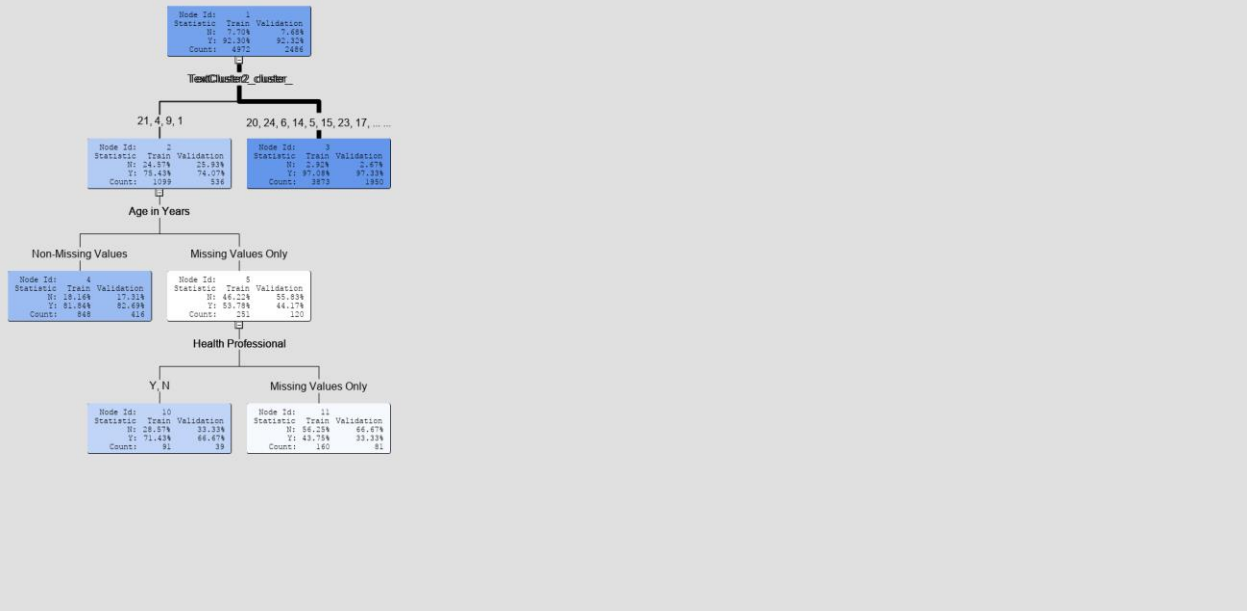
To utilized unstructured data in building the decision tree, text cluster is built prior to decision tree. The aim of text cluster is to create clusters that will help with identifying the desired value of the target variable (serious outcome). FAERS cases are assigned to mutually exclusive clusters so each document can belong to only one cluster which is described by a set of terms (Figure 10). This is achieved by deriving a numeric representation for each document. Producing the numeric representation for each cluster is implemented through Singular Value Decomposition (SVD) to organize terms and documents into a common semantic space based upon term co-occurrence. When cases are parsed, a frequency matrix is generated. Depending on the application, the user can define the number of dimensions. For the purpose of text segmentation, a recommended number of dimensions' ranges from 2 to 50, but for prediction and classification higher values from 30 to 200 are used (Berry & Kogan, 2010). Therefore, we

selected the number of cluster to be 25 (Figure 10).

Cluster ID	Descriptive Terms
1	mi-god betaseron-god
2	ap dronedarone ggt +value causal alat +relationship +increase further present +comment +recover written +bilirubin +phosphatase first alt +antibody +study alkaline +day substantial sclerosis +admit +purpose
3	ast alt troglitazone alanine aminotransferase aspartate +'additional information' 'aspartate aminotransferase' 'alanine aminotransferase' +'total bilirubin' +request rezulin alkaline additional +'alkaline phosphatase' total +bpe +'clinical outcome' +phosphatase +laboratory follow-up daily +value diabetes +bilirubin
4	tysabri unknown iv 'monitored therapy' +onset 'tysabri therapy' +program monitored qm biogen therapy +enrol idec sclerosis 'biogen idec' +update 'multiple sclerosis' +causality +assess female 'onset unknown' +confirm +mar +nurse written
5	+disease +perform +biopsy +hospital +admit +alcohol +week first +condition +body +cell jaundice +initiate present acute +follow +severe +grade literature +injury +confirm +relationship +pain beta +failure
6	+program betaseron sclerosis +interferon +refer 'multiple sclerosis' beta +consumer +state patient +reporter +dose unspecified +condition previously +causality +request +number +enzyme +drug +event +follow-up +confirm +time +outcome
7	troglitazone 'cardiovascular system' +'liver damage' +'liver problem' +'medical condition' +'potential injury' +'unspecified medical condition' +body +complaint +condition +damage +develop +experience +form +injury +pain +system +threaten attorney cardiovascular experienced life medical mental patient
8	+antibody +disorder 'diabetes mellitus' mellitus iu dl sgot hepatic daily sgpt +purpose +function +laboratory +value diabetes mg +body +test +reference +condition +'total bilirubin' total +comment +follow +mar
9	+'tracking number' +assign +purpose +track naltrexone +'no further detail' +reference +initiate +number +detail spontaneous male oral unknown serious +case +time further +'unknown date' +resolve alanine +status registered +event aminotransferase
10	tysabri unknown sclerosis iv 'multiple sclerosis' +assess +causality +onset 'tysabri therapy' +update qm +'unknown date' +reference 'onset unknown' +mar apr written 'old female patient' female +confirm significant +antibody +nurse medically +diagnose
11	+onset iu mellitus health 'diabetes mellitus' +transaminase sgot sgpt diabetes +damage +physician concomitant daily +case +hospitalize +history mg +'medical history' medical +increase causal total +'concomitant medication' +number hepatic
12	alk phos 'alk phos' dronedarone sgpt sgot written abdominal normal +'liver biopsy' +'alkaline phosphatase' er +ultrasound alkaline +normal +phosphatase alt 'monitored therapy' +'total bilirubin' total ast +biopsy monitored +study +threaten
13	+nurse registered +request troglitazone +'additional information' +bpe rezulin +'total bilirubin' ggt 'diabetes mellitus' 'old female patient' diabetes total mellitus additional further +old aspartate 'aspartate aminotransferase' +'clinical outcome' ast follow-up +discontinue +'troglitazone therapy' +begin
14	+study +enrol +grade +result literature dl +update acute +dose ap +'medical condition' 'alanine aminotransferase' +disorder +assign spontaneous oral aspartate biogen causal concomitant +ultrasound +detail +drug +relate +include
15	iv +grade +cell +study +relationship +enrol +death acute er +condition +assess +die +dose +treat +initiate causal +male +system +failure +severe +blood +reference clinical +comment +perform
16	iv spontaneous +infection +assign +detail +comment +relationship +initiate serious +purpose +admit +male acute +day +condition sgot +reporter potential oral +antibody +hospital +treat causal sgpt +causality
17	beta betaseron +comment +female literature +severe +increase pt +cause 'alk phos' phos sclerosis +enzyme previously alk +cell apr +interferon +medication first +confirm +condition +begin +biopsy +follow-up
18	er pt +begin +admit +pain +ultrasound jaundice +study +'no further detail' +detail +lab +discharge ap +enzyme +system alk elevated +female +day +time normal +complaint phos first +week
19	+ultrasound pt +disease +show +infection +status +'female patient' hepatic negative +decrease +level betaseron alat +cause +normal +resolve acute first mental +test normal abdominal literature +reveal +day
20	alat +male first acute hepatic +infection oral dronedarone +dose +comment causal +drug +resolve +decrease +relationship +event spontaneous daily 'aspartate aminotransferase' aminotransferase +treat +grade +'female patient' 'alanine aminotransferase' +disorder
21	liver literature +alcohol +comment er +disorder ap +injury +biopsy written +'liver biopsy' potential normal +assign +form +week +condition +'liver problem' +enzyme medical serious +month +drug +begin +cell
22	march alat +treatment +severe literature +male oral +injury +'medical condition' +death +dose +die +time +'unknown date' male +treat hepatic +follow dronedarone +day +condition +case +failure jaundice +cause
23	sgot +state +consumer +admit +cause +death +hospital sgpt +treat +blood +follow-up +discharge +infection +physician +reveal +cell +day +form +biopsy pt +die elevated follow-up +enzyme +perform
24	liver troglitazone +damage unspecified sgpt diabetes sgot rezulin mellitus +injury 'diabetes mellitus' +'liver damage' +'troglitazone therapy' +'clinical outcome' attorney anguish +patient +pain jaundice daily +physician medically summons +year +discontinue
25	phos alk 'alk phos' +'liver biopsy' pt +biopsy limited literature +bpe +test rezulin +month +perform 'diabetes mellitus' +'troglitazone therapy' sgot +improve hepatitis abdominal troglitazone +hospital +relationship mellitus +antibody negative

Figure 10: Cluster Description

The output from the cluster analysis is the input to the decision tree modeling. Two decision tree models have been developed. In the first one, the numeric values for the 25 SVDs have been assigned rejected role so that only the nominal values of cluster numbers (TextCluster_cluster_) will input the decision tree modeling with other FAERS input variables. While on the second model, the SVDs assigned new role as input to the decision tree with other FAERS variables and cluster number variable has been rejected. Figure 11 demonstrates the tree construction for the first model as well as the variables that were important in growing this decision tree.



Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
TextCluster2_cluster_		1	1.0000	1.0000	1.0000
Age_in_Years	Age in Years	1	0.6163	0.7513	1.2189
Health_Professional	Health Professional	1	0.3328	0.3681	1.1060

Figure 11: Decision Tree where the role of cluster numbers set as input role

The classification chart for assigning the serious outcome for this model (Figure 12) shows that 91 % of the serious outcome with level =Y was classified correct (Serious Outcome=Y) while 4.9% of the serious outcome with level=N was misclassified as (Serious Outcome=Y).

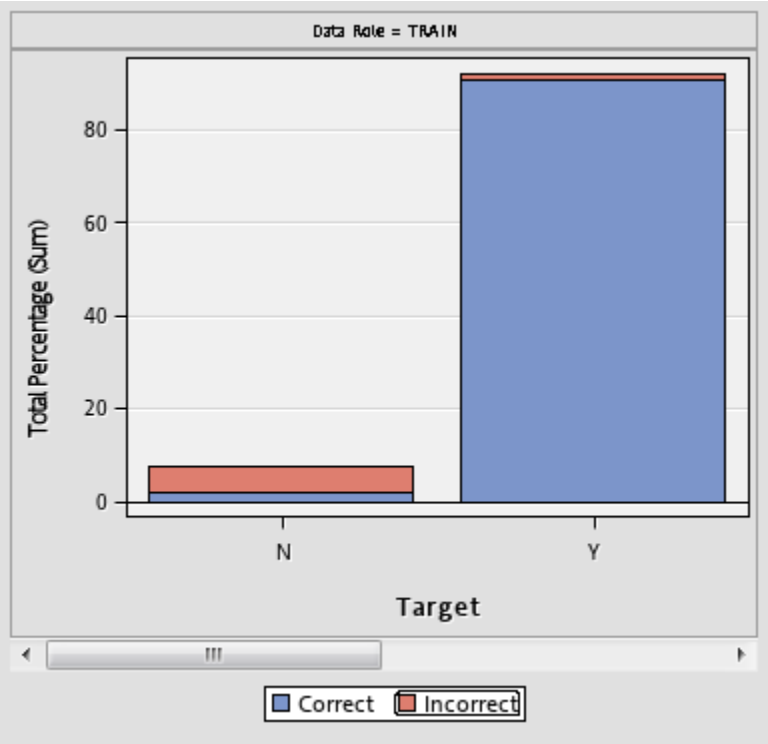
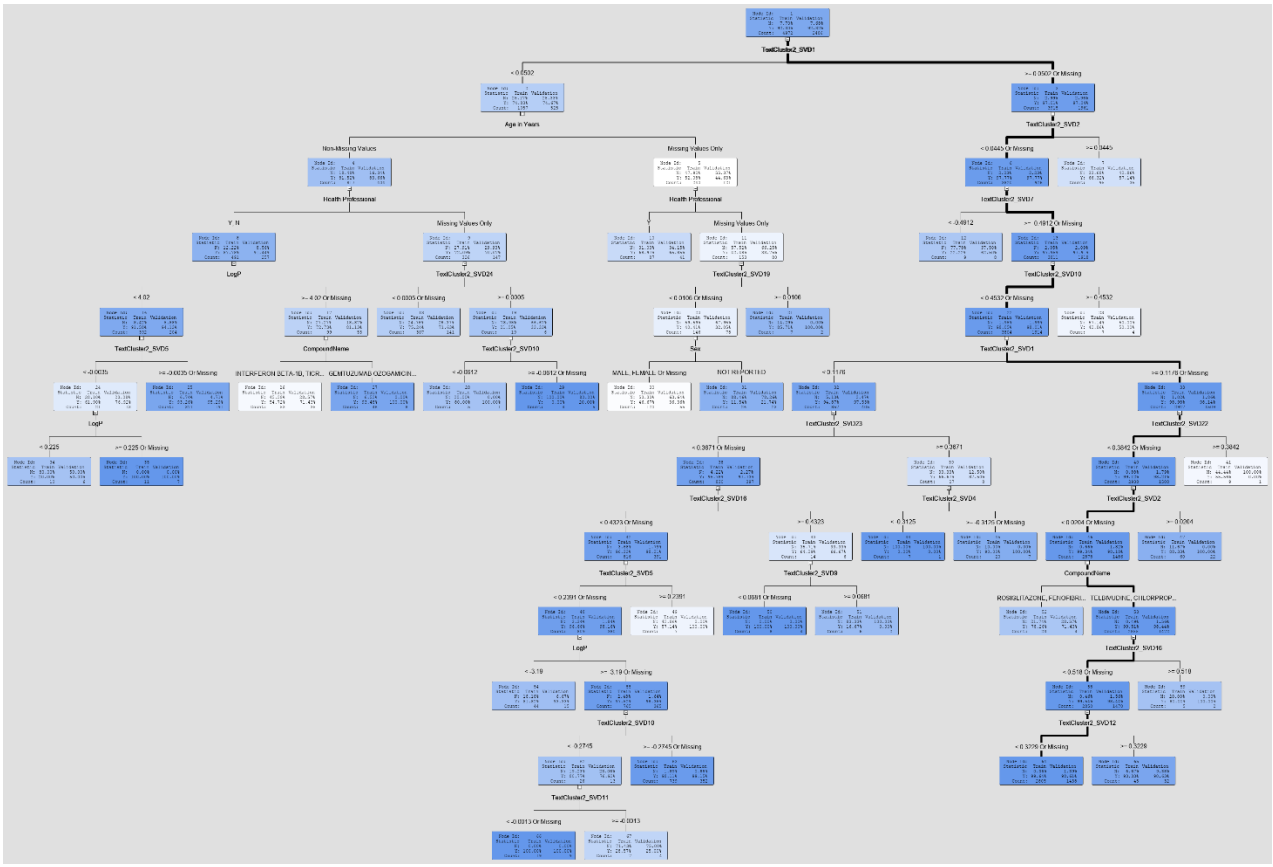


Figure 12: Classification Chart: Serious Outcome? where the role of cluster numbers set as input role

On the other hand, when SVDs assigned a role of inputs in the metadata while rejecting the cluster number, Figure 13 shows more variables have been contributed to construct this tree.



Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
TextCluster2_SVD1		2	1.0000	1.0000	1.0000
Age_in_Years	Age in Years	1	0.6182	0.8019	1.2972
TextCluster2_SVD2		2	0.4848	0.5093	1.0506
Health_Professional	Health Professional	2	0.4552	0.5756	1.2645
LogP	LogP	3	0.4134	0.2313	0.5595
TextCluster2_SVD24		1	0.3535	0.1742	0.4929
TextCluster2_SVD10		3	0.3530	0.3004	0.8510
TextCluster2_SVD7		1	0.3499	0.0000	0.0000
CompoundName	CompoundName	2	0.3354	0.2472	0.7370
TextCluster2_SVD4		1	0.3160	0.2357	0.7460
TextCluster2_SVD5		2	0.2688	0.0906	0.3370
Sex	Sex	1	0.2504	0.0576	0.2301
TextCluster2_SVD11		1	0.2491	0.2753	1.1053
TextCluster2_SVD9		1	0.2379	0.2486	1.0448
TextCluster2_SVD23		1	0.2296	0.0000	0.0000
TextCluster2_SVD22		1	0.2012	0.1753	0.8715
TextCluster2_SVD16		2	0.1953	0.1696	0.8686
TextCluster2_SVD19		1	0.1805	0.2112	1.1697
TextCluster2_SVD12		1	0.0648	0.0989	1.5276

Figure 13: Decision Tree where the role of SVDs set as input role

The classification chart for the second model (Figure 13) does not differ much than the first model. However, based on the fit statistics of the first model, the overall misclassification rate on the validation data set is 69.8 %around while the overall misclassification rate on the validation data set for the second mode 61.154%. Therefore, using SVDs as input variable improved the model by reducing the misclassification rate for the FAERS serious outcome around 9%.

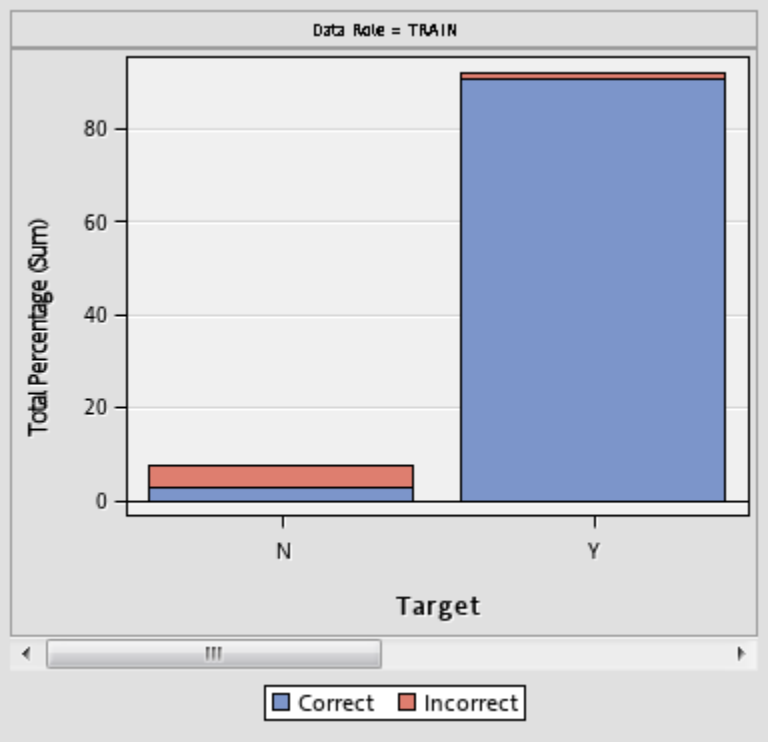


Figure 14: Classification Chart: Serious Outcome? where the role of SVDs set as input role

4. MODEL COMPARISON

The model comparison tool facilitates selection between model tools using the summary statistics. Summary statistics transform model performance to numerical scores. The model comparison tool collates statistics from different modeling nodes for easy comparison. Out of the 8 models that have developed in this research, the winning model is the decision tree_probability (Figure 15) which is indicated with letter Y as winning model.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable
Y	Meta10	Tree	Decision Tree_Probability	Serious_Outcome_
	Meta13	TextRule2	Text Rule Builder_Aggressive	Serious_Outcome_
	Meta8	Tree2	Decision Tree_DefaultDecisionT...	Serious_Outcome_
	Meta12	TextRule7	Text Rule Builder_VeryAggressive	Serious_Outcome_
	Meta11	TextRule3	Text Rule Builder_Default	Serious_Outcome_
	Meta9	MBR	MBR	Serious_Outcome_
	Meta7	Neural	Neural Network	Serious_Outcome_

Figure 15: Model Comparison

The gains chart is an excellent way to show the performance of a model. The gains chart (Figure 16) shows the percent difference between the overall proportion of events and the observed proportion of events in all groups up to the current group. Therefore, gains are calculated relative to the baseline (overall average) rate. Champion model is the one with the greatest gain.

$$Gain = \left[\left[\frac{\%\ of\ events\ in\ decil}{random\ \%\ events\ in\ decile}\right] - 1\right]$$

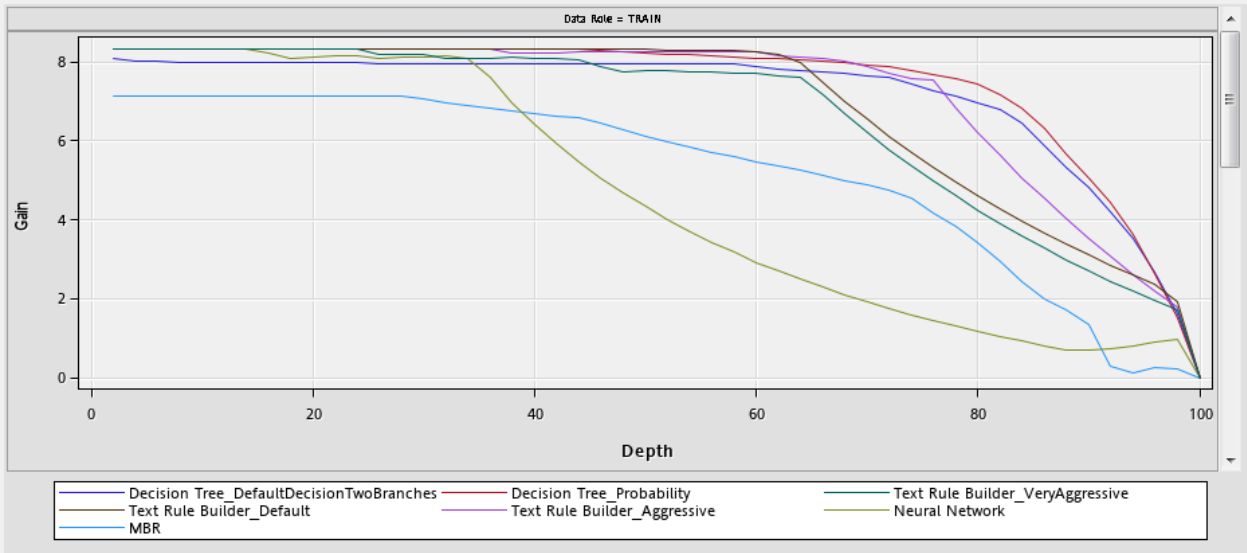


Figure 16: Gain Chart

Using the champion model, the gain value of 8.1 at depth of 20% indicated by using the predictive model, we are able to get a gain of 8.1 in predicting a serious outcome in FAERS cases if we only retrieved the top 20% of the FAERS cases compared to overall averages (i.e., without the predictive model as randomly chosen FAERS cases).

5. SCORING

To see how accurately the champion model classifies new data that do not contain a target value (i.e., serious outcome has not been assigned yet to the FAERS cases), 4000 cases have been selected randomly and assigned the role of score data set. The score table (Figure 17) illustrate the model's predicted FAERS serious outcome category and the model probability for this category (probability of classification).

Narrative	Probability for level Y of Serious_Outcome_	Probability of Classification	Prediction for Serious_Outcome_	Target Variable: Serious_Outcome_ ▲
This spontaneous report was received from a MRI technician refers to a female patient of unknown ...	0.315789	0.684211	N	N
Report of liver damage, splenomegaly, portal hypertension, ascites, listlessness, and withdrawal sy...	0.166667	0.833333	N	N
A 48 year old female patient on TYSABRI (300 mg, IV, QM) since 01 Oct 2006 reported experiencing l...	0.076923	0.923077	N	N
A 43 year old female patient on TYSABRI (300 mg, IV, QM) for Multiple Sclerosis since 17 Dec 2012 ...	0.076923	0.923077	N	N
A currently 64 year old female patient on TYSABRI (300 mg, IV, QM) for Multiple Sclerosis since 14 M...	0.884615	0.884615	Y	N
A 51 year old female patient on TYSABRI (300 mg, IV, QM), for Multiple Sclerosis, from 18 Jan 2012 t...	0.076923	0.923077	N	N
Case reference number 2016BI00283601 is a Spontaneous post marketing case report by a nurse....	0.941176	0.941176	Y	N
Case number# PHEH2017US025482, is an initial spontaneous report received from a consumer (p...	0.666667	0.666667	Y	N
This is a spontaneous report from a contactable physician. A patient of unspecified age, ethnicity, a...	0.614035	0.614035	Y	N
** Abstracted per FDA rep **Drug Induced Liver InjuryA 34 year old man with history of intravenous dr...	0	1	N	N
GI BLEEDESOPHAGEAL VARICESCIRRHOSIS WITH MICRONADULAR CHANGESCHRONIC ACTIV...	0.904762	0.904762	Y	N
This is a spontaneous report received from a consumer (via medical information) which described ...	0.97619	0.97619	Y	N
Serious; related; unlistedThis case report from BRAZIL was derived from medical literature on 24-J...	0.990991	0.990991	Y	Y
Initial report received from a patient counselor executive on 29 Jan 2010: This patient was prescribe...	1	1	Y	Y
Case number PHHY2012KR060672 is an initial PMS (CLDT600AKR01) report received from an inv...	1	1	Y	Y
This serious post-marketing case report from France received on 10-MAR-2014 by an investigator d...	0.991131	0.991131	Y	Y
THIS IS A SECOND FOLLOW UP TO A REPORT INITIALLY SUBMITTED ON 08APR99. THE FIRST F...	1	1	Y	Y
THIS IS A FOLLOW-UP BASED ON INFORMATION REPORTED TO PFIZER ON 04MAY00. THE INITI...	0.723684	0.723684	Y	Y
This spontaneous case, reported by a cardiologist via a company sales representative, concerns a f...	0.941176	0.941176	Y	Y
Initial consumer report was received from the patient's wife on 28 Sep 2009: This polymedicated pat...	1	1	Y	Y
Septic shockDrug ineffectiveRenal failure acuteOesophageal varices haemorrhage	0.972222	0.972222	Y	Y
Literature report: Boyle-Vavra S, Jones M, Gourley BL, Holmes M, Ruf R, Balsam A, et al. Comparativ...	0.991131	0.991131	Y	Y
Initial report received on 24 Jun 2009 : Patient 17 from centre 1486 was enrolled in study CSPP10...	1	1	Y	Y
THIS IS A FOLLOW-UP REPORT BASED ON INFORMATION REPORTED TO PFIZER ON 02AUG99. ...	1	1	Y	Y
Information was received from a healthcare professional describing a 68 year-old female who recei...	1	1	Y	Y

Figure 17: Score Table

CONCLUSION and FUTURE WORK

We applied different analytical models with text mining to better understanding FAERS data by using SAS Enterprise Miner. Both structured and unstructured data were utilized to increase our predictive power and provide an informative analysis. Our analysis can be easily extended to other database such as the vaccine adverse event reporting system (VAERS) which contains information on unverified reports of adverse events (illnesses, health problems and/or symptoms) following immunization with US-licensed vaccines. Our work illustrates a proof of concept of modeling FAERS database and the feasibility of utilizing the unstructured data in such modeling. This work is in progress and more improvement will be adapted for refining the analysis and utilizing more powerful techniques.

Bibliography

Aas, K., & Eikvil, L. (1999). *Text Categorisation: A Survey.*

Berry, M., & Kogan, J. (2010). *Text Mining-Application and Theory.* John Wiley & Sons, Ltd. .

Cerrito, B. (2006). *Introduction to Data Mining using SAS Enterprise Miner.* SAS Publishing .

Fontana, R., Seeff, L., Andrade, R., Bjornsson, E., Day, C., Serrano, J., & Hoofnagle, J. (2010). Standardization of Nomenclature and Causality Assessment in Drug-Induced Liver Injury: Summary of a Clinical Research Workshop. *Hepatology, 52*(2), 730-742.

Harpaz, R., Callahan, A., Tamang, S., Low, Y., Odgers, D., Finlayson, S., . . . Shah, N. H. (2014). Text Mining for Adverse Drug Events: the Promise, Challenges, and State of the Art. *Drug Saf, 37*, 777-790.

Harpaz, R., Vilar, S., DuMouchel, W., Salmasian, H., Haerian, K., Shah, N. H., . . . Friedman, C. (2013). Combing Signals from Spontaneous Reports and Electronic Health Records for Detection of Adverse Drug Reactions. *Journal of the American Medical Informatics Association, 20*, 413-419.

Hertz, J., Krogh, A., & Palmer, R. (1991). *Introduction to the Theory of Neural Computation.* Redwood City, CA: Addison-Wesley Publishing Co.

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text Classification Using Machine Learning Techniques. *WSEAS Transactions on Computers 4.8*, 966-974.

Konchady, M. (2006). *Text Mining Application Programming.* Boston: Charles River Media.

MERCADO. (n.d.). *MERCADO Integrated Marketplace for Regulatory Data*. Retrieved January 2018, from http://mercado.fda.gov/analytics/saw.dll?dashboard&PortalPath=%2Fshared%2FDrug%20Safety%20Adverse%20Events%2F_portal%2FQuick%20Search

Miller, T. W. (2005). *Data and Text Mining: A Business Applications Approach.* Pearson Pentice Hall.

Rajman, M., & Besancon, R. (1997). *Text Mining: Natural Language Techniques and Text Mining Applications.* Lausanne, Switzerland: Chapman & Hall.

Sanders, A., & DeVault, C. (2004). Using SAS at SAS: The Mining of SAS Technical Support. *SUGI 29 Analytics.* Cary, NC.

SAS Course Note, E. (2016). *Advanced Predictive Modeling Using SAS Enterprise Miner.* Cary, NC: SAS Institute Inc.

SAS Enterprise Miner. (2018). Intriduction to Text Miner. Cary, NC.: SAS Institue Inc.

SAS Reference Help. (2018). SAS Enterprise Miner 14.2 . Cary, NC: SAS Institute Inc.

Schneeweiss, S. (2010). A Basic Study Design for Expedited Safety Signal Evaluation Based on Electronic Healthcare Data. *Pharmacoepidemiol Drug Safety, 19*, 858-868.