Text Analytics using High Performance SAS® Text Miner

Edward R. Jones, Ph.D.
Exec. Vice Pres.; Texas A&M Statistical Services

Abstract: The latest release of SAS® Enterprise Miner™, version 13.1, contains high performance modules, including new modules for text mining. This paper compares the new High Performance Text Mining modules to those found in SAS Text Miner. The advantages and disadvantages of HP Text Miner are discussed. This is illustrated using data from the National Highway and Transportation Safety Administration for the GMC recall of 258,000 SUVs for a potential fire hazard .

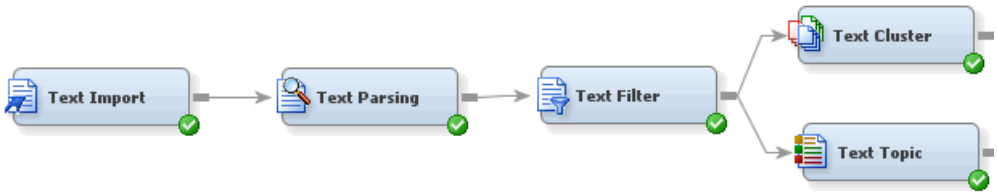
Introduction

In August, 2014, General Motors announced the third recall of a family of 258,000 SUVs to correct a potential fire hazard. The SUVs included the Chevrolet TrailBlazer, GMC Envoy, Saab 9-7x, Buick Rainier and Isuzu Ascender after reports of 28 fires in the power windows and locks of these vehicles from 2005-2007. The NHTSA, National Highway Traffic Safety Administration, maintains a public database of consumer complaints for these and other vehicles. The question is whether these complaints could have been used to identify this hazard earlier than 2014.

Over 1 million complaints have been registered and are available for download from the NTHSA. To examine this question, both SAS Text Miner and SAS HP Text Miner, version 13.1, were used to analyze 2,122 complaints recorded by the NTHSA for the recalled SUVs from 2005-2007. These data include not only written complaints but also information on the vehicle, the driver and whether or not the vehicle was involved in a fire, a crash, and the number of injuries and deaths. This paper describes how to use SAS HP Text Miner to incorporate the written complaints with the structured data to identify potential vehicle problems.

SAS® Text Miner

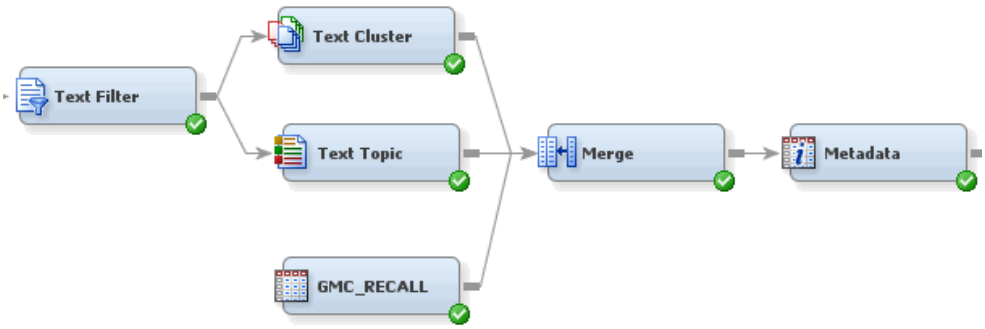
SAS® Text Miner is incorporated within SAS® Enterprise Miner™. The text import node requires all documents to be placed in individual files. Unfortunately, the NTHSA data file contained all 2,122 complaints in one file. These had to be extracted and placed into 2,122 files. The analytics process of for text mining these complaints is described in this figure.



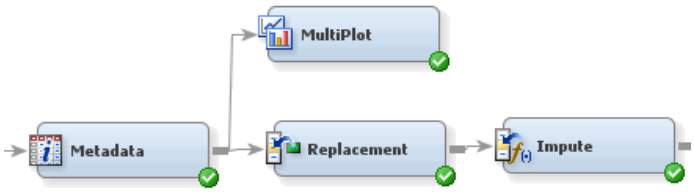
The default document size in text import is 100 characters. This had to be increased to 2100 characters. Text parsing was done using the default settings. Text filtering was done using inverse document frequency term weights, and the filter viewer was used to review terms to identify those that were not useful in this analysis. This consisted mainly of the initials of the NTHSA recorder.

Using the default settings, the text cluster node found only 3 clusters and 45 SVD vectors. The text topic node was used iteratively to identify ten user word clusters including one word cluster for a door switch fire and 9 others for various mechanical problems such as air bag and fuel system problems.

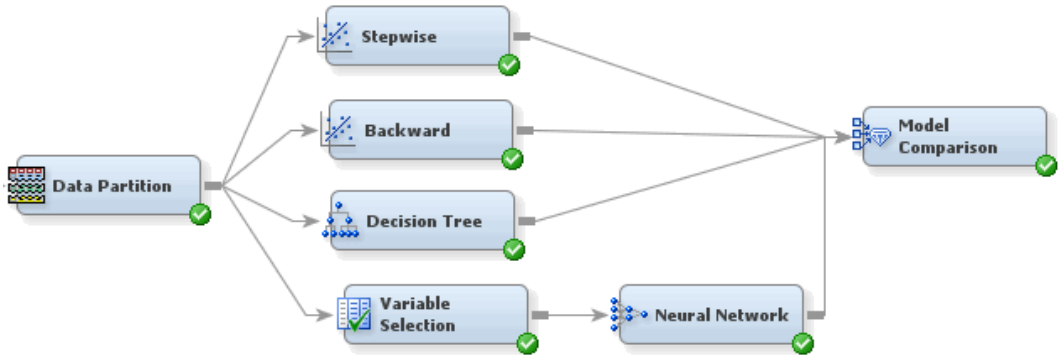
The output from text mining process consisted of 10 binary attributes associated with the ten user developed topics and 3 interval attributes consisting of the estimated probabilities for the three text clusters. These were merged into the original structured data to produce a single file for modeling whether fires in these vehicles might be related to some of these complaint topic groups or text clusters.



After the merge, the roles of all attributes were modified as either rejected, input or target. The text topic groups and text cluster estimated probabilities were all marked as inputs. The SEMMA process in Enterprise Miner was followed. After preparing this integrated file, outliers were set to missing using the replacement node, and missing values were imputed.



The model step was implemented using the logistic regression node, the decision tree node and neural network node. Experience with problems involving many input attributes has found it's prudent to reduce the number of input attributes using variable selection.



The logistic regression and decision tree models all had the same misclassification error rate for predicting whether the vehicle was involved in a fire based upon the topic of the complaint and the driver and vehicle characteristics. The logistic regression model developed using stepwise selection had the smallest Validation Average Squared Error (VASE). It incorporated only one of the input attributes: text topic 2, which was the topic related to a door switch problem. The GMC recall reported that the fires were related to shorts in the door switch.

The complaints to point to fire problems related to the door switch. There were 222 SUVs out of 2,122, or over 10% that reported this problem. The GMC recall only mentioned 28 fires out of 258,000 vehicles.

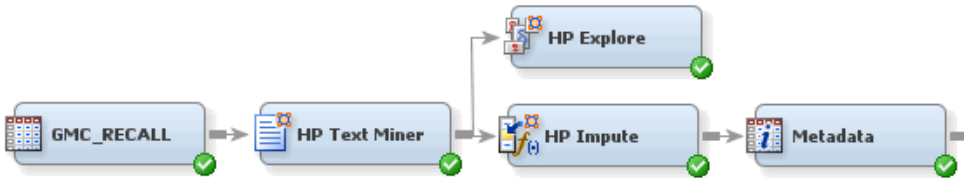
SAS® High Performance Text Miner

SAS® High Performance Text Miner is also incorporated within SAS® Enterprise Miner™ since version 12.1. The latest version, 13.1 is an improvement and expansion over 12.1. There are two primary features that differentiate High Performance Text Miner from the original version of SAS Text Miner. First all High Performance nodes are designed to run faster by using high-speed algorithms and multi-core and cluster capabilities.

The second, and most important, difference in text mining is that unlike SAS Text Miner, High Performance Text Miner does not require all documents to be stored in individual files inside a single director. Instead, High Performance Text Miner will accept any SAS file that contains two attributes, one with the role ***key*** and a second with the role ***text***. The key field should be a unique interval value associated with each of the text fields. This type of input is ideal for the NTHSA data, surveys and other data containing written comments.

In this case, the key field was the unique NTHSA ID assigned to each complaint and the text field was the written comment. In these data it was restricted to a maximum of 2100 characters, but High Performance Text Miner can accept written comments up to 32,000 characters.

The Sample-Explore-Modify process in SEMMA using High Performance Text Miner is much simpler than what was described for SAS Text Miner. First, there is no need to separate the text mining from the original data. In addition, at this time there is no equivalent to the Text Filter and Text Topic nodes. Instead, High Performance Text Miner by default exports topic groups defined by the SVD analysis. These will be orthogonal to one another.



Using the default setting for the High Performance Text Miner node produced 20 topic groups. These were marked automatically as ***interval inputs***. Version 13.1 of High Performance Enterprise Miner also contains a new ***Impute node*** which was used to estimate missing values, such as vehicle mileage.

Notice that the topic attributes are named COL1…COL20, but High Performance Text Miner puts the names of the words associated with each attribute in its label. Notice that COL4 is associated with a comment about a vehicle fire. COL8 is talking about a burning odor, and COL1 is talking about a problem with the driver side door.

Properties - EMWS2.HPTM_TRAIN

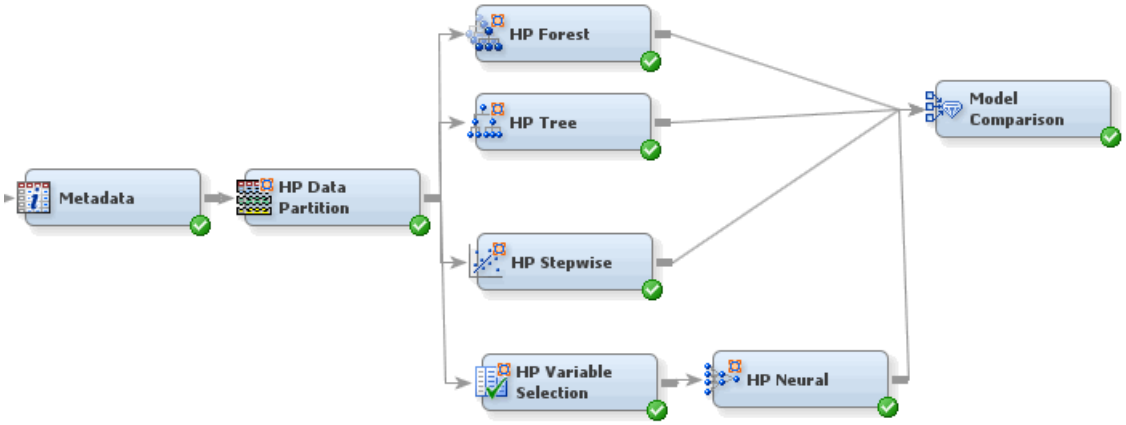
Table | Variables

Columns: Label | Mining | Basic

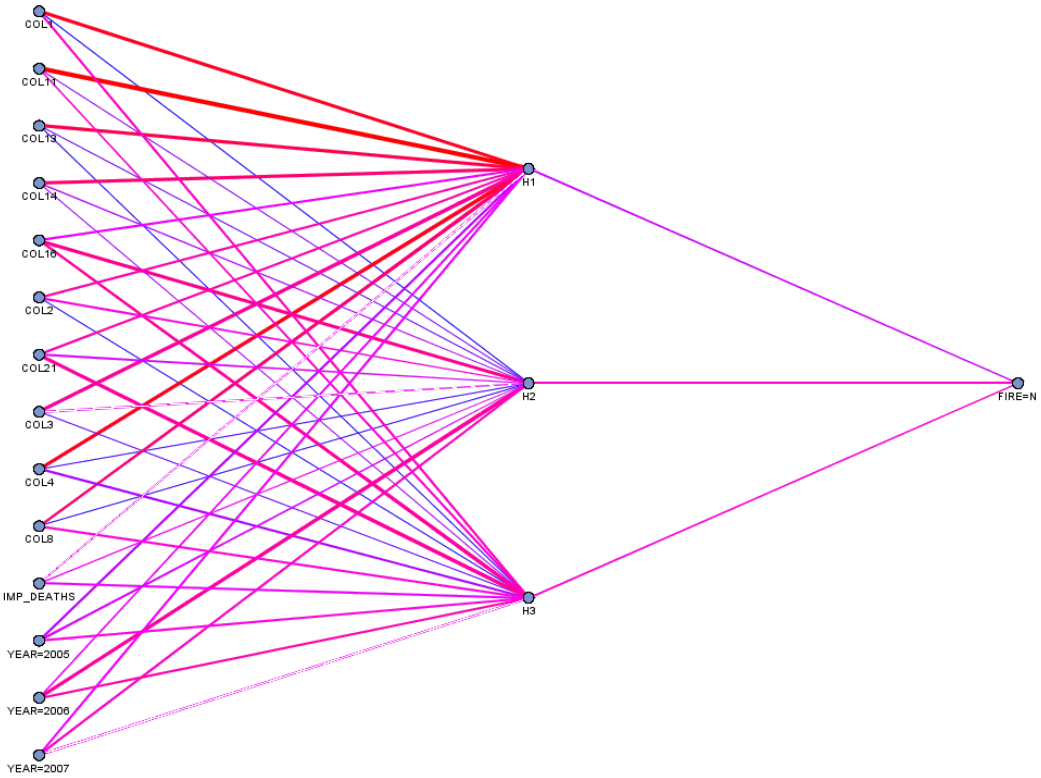
Name	Label	Role	Level
ABS	ABS	Input	Binary
CLYS	CLYS	Rejected	Nominal
COL1	+driver,+door,+window,side,side	Input	Interval
COL10	+module,+control,+state,+repair,+vehicle	Input	Interval
COL11	+melt,+blower,+fire,+motor,resistor	Input	Interval
COL12	approximate,approximate failure mileage,+manufacturer,+notify,+repair	Input	Interval
COL13	smoke,+smell,+emit,+door,box	Input	Interval
COL14	+burn,+wire,+smell,+window,+flame	Input	Interval
COL15	+switch,power,master,+failure,+vehicle	Input	Interval
COL16	+window,+vehicle,+manufacturer,+park,+notify	Input	Interval
COL17	side,front,+passenger,+driver side window,driver side door	Input	Interval
COL18	+update,js,+notice,bf,lj	Input	Interval
COL19	power,+lock,side,+catch,+smell	Input	Interval
COL2	+contact,+mileage,+failure,tl,+current mileage	Input	Interval
COL20	+manufacturer,panel,+vehicle,+file,+report	Input	Interval
COL21	+control,panel,+smell,+window,+lock	Input	Interval
COL3	campaign,+number,nhtsa,12v406000,id	Input	Interval
COL4	+fire,+department,+fire department,+catch,+burn	Input	Interval
COL5	+door,+lock,+unlock,+passenger,outside	Input	Interval
COL6	electrical,+electrical system,+system,nhtsa campaign number,+state	Input	Interval
COL7	+side,front,+passenger,+vehicle,+park	Input	Interval
COL8	burning,odor,burning odor,+smell,+smell	Input	Interval
COL9	+vehicle,+notice,information,+park,+recall	Input	Interval
COMPLAINT	COMPLAINT	Text	Nominal
CRASH	CRASH	Input	Binary
CRUISE	CRUISE	Input	Binary
DEATHS	DEATHS	Input	Interval
DIED	DIED	Input	Binary
FIRE	FIRE	Target	Binary
INJURED	INJURED	Input	Binary
INJURIES	INJURIES	Input	Interval

The modeling approach is almost idential to that used with SAS Text Miner. Stepwise logistic regression is used, along with decision tree and neural network.

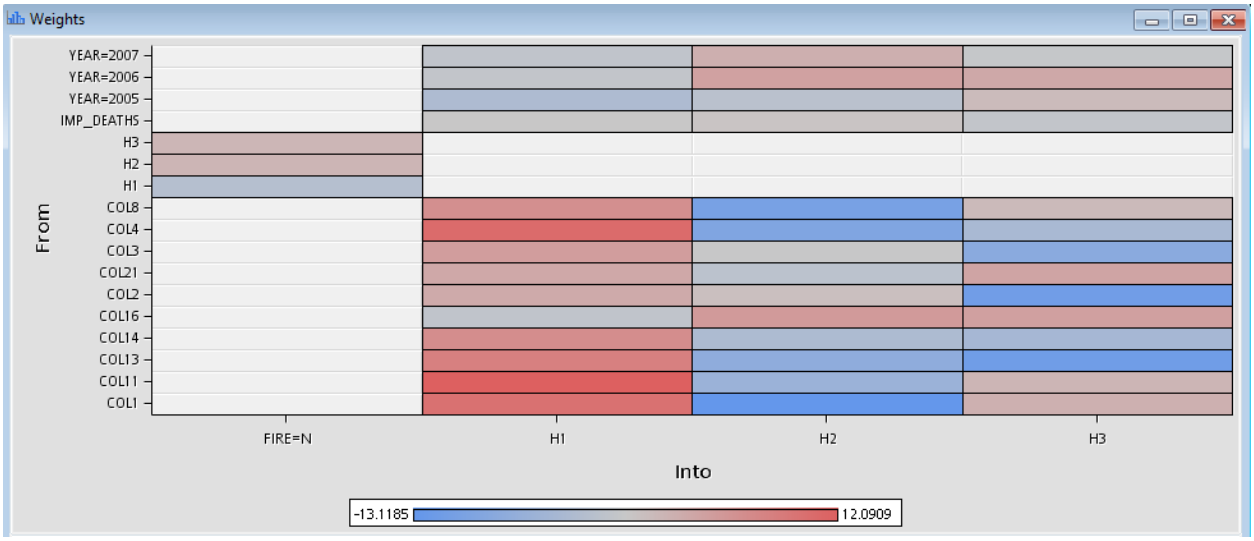
However there is one additional modeling node that was used this is only available in the High Performance nodes: High Performance Random Forests.



In this case, the model that produced the smallest misclassification error is the Neural Network Model. It was 2.8%, lower than that seen with the non-High Performance nodes. The new High Performance Neural Network provides clues about the relative importance of attributes in the model. First the network is displayed with colorful lines. The darker the line, the larger the weight for that attribute.



In this case the attributes with the highest weights were text topics. Another plot that describes the relative size of the network weights is the weights plot. From this it's apparent the the topics with the larger weights are COL1, COL11, COL4 and COL8.



These correspond to the following toics:

- COL1: the driver-side door
- COL11: melting and fire
- COL 4: fire
- COL 8: a burning odor.

This is also seen in the importance statistics from the random forest node:

Variable Importance

Variable Name	Label	Number of Splitting Rules	Gini Reduction
COL4	+fire,+department,+fire department,+catch,+b...	59	0.010705
COL8	burning,odor,burning odor,+smell,+smell	40	0.003145
COL1	+driver,+door,+window,side,side	37	0.003395
COL11	+melt,+blower,+fire,+motor,resistor	37	0.003691
YEAR	YEAR	33	0.001412
COL13	smoke,+smell,+emit,+door,box	30	0.002243
COL9	+vehicle,+notice,information,+park,+recall	23	0.001852
COL7	+side,front,+passenger,+vehicle,+park	19	0.001248
COL14	+burn,+wire,+smell,+window,+flame	18	0.000441
COL3	campaign,+number,nhtsa,12v406000,id	18	0.000736
COL10	+module,+control,+state,+repair,+vehicle	13	0.000318
IMP_DEATHS	Imputed DEATHS	11	0.000145
COL19	power,+lock,side,+catch,+smell	9	0.000354
COL20	+manufacturer,panel,+vehicle,+file,+report	9	0.000441
IMP_MILEAGE	Imputed MILEAGE	9	0.000505
COL16	+window,+vehicle,+manufacturer,+park,+notify	8	0.000323
COL5	+door,+lock,+unlock,+passenger,outside	5	0.000294
COL15	+switch,power,master,+failure,+vehicle	4	0.000221
COL6	electrical,+electrical system,+system,nhtsa c...	4	0.000260
CRASH	CRASH	4	0.000060
COL12	approximate,approximate failure mileage,+m...	3	0.000146
COL18	+update,js,+notice,bf,lj	3	0.000164
IMP_INJURI...	Imputed INJURIES	3	0.000184
IMP_MPH	Imputed MPH	3	0.000015
MODEL	MODEL	3	0.000005
COL2	+contact,+mileage,+failure,tl,+current mileage	2	0.000016
COL21	+control,panel,+smell,+window,+lock	2	0.000112

Conclusion

Both SAS Text Miner and High Performance Text Miner were able to uncover about 222 complaints (10%) that were associated with fires caused by a problem with the electric switches in the vehicle door. The latest version of High Performance Text Miner, version 13.1, is easier to use since it processes a single file containing multiple comments. In addition, the modeling output from High Performance Neural Network models and Random Forest Models makes it easier to identify key complaint topics associated with the target.