GALAXY ATTRIBUTES SELECTION/ELIMINATION

The attribute elimination criteria was based on a threshold imposed on Chi-Squared merit scores equal or less than 150.

The threshold value was chosen because it eliminated no more than 1/3 of the original attributes. This criteria allowed us to keep a reliable number of a significant attributes that reflect real web pages content.

We used the merit score rather than the ranking to select attributes that could be eliminated. By doing that, we assure that the correlation to galaxysentiment was weak and will not impact the general overall prediction outcome.

List of Galaxy attributes eliminated: 26, 20, 57, 31, 41, 11, 4, 46, 51, 30, 45, 36, 35, 21,16, 40, 15, 50.

GALAXY CLASSIFIER SELECTION

1. Four classifiers were applied on the reduce data set: J43, Forest Random, IBK and SMO.
2. The classifiers summaries were added to a comparison table (See below). SMO was eliminated because the time it takes to apply this algorithm to the data exceeded in more 100% the time required by the other classifiers. Besides, we experienced memory saturation problems.
3. Forest Random Test showed the best performance, followed by J48.
4. J48 showed the quickest process time, and the most comprehensive visual output.
5. Since performance differences between J48 and Forest Random were in around 0.008% but processing time for Forest Random was almost 9 times higher than J48's, J48 was chosen as the best prediction model because for very large data sets such as our Big Matrix, it will provide mostly the same accuracy in less time and cost.
6. J48 was tuned by testing how different levels of confidence impact on performance. Levels of confidence ranged from 0.25, 0.3, 0.4. The best performance for J48 on the Galaxy data set was C=0.3 M=2 Cross-Validation 10 folds. This was the selected model.

GALAXY SENTIMENT	J48	Forest Random	IBK
Correctly Classified Instances	10135 /78.1238%	10218 / 78.7636%	10083 / 77.723%
Incorrectly Classified Instances	2838 / 21.8762%	2755 / 21.2364%	2890 / 22.277%
Kappa statistic	0.5627	0.5764	0.5656
Mean absolute error	0.1177	0.1161	0.1134
Root mean squared error	0.2473	0.2438	0.2509
Relative absolute error	58.986%	58.180%	56.851%
Root relative squared error	78.299%	77.182%	79.429%
Total Number of Instances	12973	12973	12973
Time taken to build model	0.29 s	8.97s	0.01s

7. J48 provides a comprehensive, strong and concise set of rules to classify the current sentiment towards galaxy. However, the data collection procedures do not consider context and in that sense are limited. This bias has been overcome by assigning sentiment scores manually that counteract the lack of context analysis.