ATTRIBUTES SELECTION/ELIMINATION

The attribute elimination criteria was based on a threshold for 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 iphonesentiment was weak and will not impact the general overall prediction outcome.

List of Iphone attributes eliminated:

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 Iphone data set was C=0.25.

|  |  |  |  |
| --- | --- | --- | --- |
| IPHONE SENTIMENT | J48 | Forest Random | IBK |
| Correctly Classified Instances | 9871 / 76.0888 % | 9953/76.7209% | 9832 / 75.7882% |
| Incorrectly Classified Instances | 3102 / 23.9112 % | 3020/23.2791% | 3141 / 24.2118% |
| Kappa statistic | 0.5351 | 0.5488 | 0.5412 |
| Mean absolute error | 0.1269 | 0.1251 | 0.1223 |
| Root mean squared error | 0.2571 | 0.2531 | 0.26 |
| Relative absolute error | 61.75% | 60.85% | 56.51% |
| Root relative squared error | 80.19% | 78.94% | 81.11% |
| Total Number of Instances | 12973 | 12973 | 12973 |
| Time taken to build model | 0.91s | 8.32s | 0.01s |

1. J48 provides a comprehensive sets of rules to classify
2. **Summarize the model you developed to predict sentiment toward iPhone in a 2-3 page Word document, include the following explanations:**
   1. The**c**lassifier you selected, the final input parameters for the model, and the features (attributes) you used to train the classifier.
   2. Your rationale for selecting the classifier that you did.
   3. The features you eliminated from the data matrix and your rationale for doing so.
   4. Your level of confidence that the model will accurately predict the sentiment for the instances in the Large Matrix. Include any scenarios where you think sentiment will not be predicted correctly by this analysis. Include any types of documents you think will be missed because of the words being used to identify pages with relevant sentiment.
   5. What classifiers you tried and their comparative performance (you can explain in text or with a chart).