**Report**

Large Matrix data for iPhone and Samsung Galaxy were used to predict the sentiment for both the smart phones.

The steps that were followed for both the smart phones were as below

1. Feature selection (by eliminating the correlated features)
2. Preprocessing the data (discretization of the data)
3. Collecting the sentiment counts by splitting the data into 7 levels
4. Creating training and test sets
5. Training the algorithms
6. Choosing the best performing model to make predictions with test test

**iPhone**

A total of 60 attributes and 12877 instances were used

Random Forest model was selected as it gave the highest accuracy for the trained model

The accuracy achieved was 96%

The **15** attributes that were used to train the classifier were

iphone

samsunggalaxy

htcphone

ios

googleandroid.

iphonedispos

iphonedisneg.

iphonedisunc

iphoneperpos

iphonecampos

iphonecamneg

iphonecamunc

iphoneperneg

iphoneperunc

iphoneSentiment

The rationale used to eliminate other attributes

* These variables have zero variances: nokiacampos, nokiacamneg, nokiacamunc, nokiadispos, nokiadisneg, nokiadisunc, nokiaperpos, nokiaperneg, nokiaperunc

All the above columns have same values per column except for one or two rows so they can be removed.

* Highly correlated attributes were removed. This was achieved by using the findCorrelation function from the caret package.

Comparative performance of all the classifiers

|  |  |  |
| --- | --- | --- |
|  | 15 Attributes |  |
|  | Accuracy | Kappa |
| KNN | 0.8980785 | 0.852365 |
| RF | 0.9606218 | 0.8838985 |
| SVM | 0.9549223 | 0.8677274 |
| C5.0 | 0.9639896 | 0.8939760 |

**Samsung Galaxy**

A total of 60 attributes and 12877 instances were used

Random Forest model was selected as it gave the highest accuracy for the trained model

The accuracy achieved was 98.9%

The **19** attributes that were used to train the classifier were

iphone

samsunggalaxy

htcphone.

ios

googleandroid

iphonedispos

iphonedisneg

iphonedisunc

iphoneperpos

samsungcampos

samsungcamneg

samsungcamunc

samsungdispos

samsungdisneg

samsungdisunc

samsungperpos

samsungperneg

samsungperunc

samsunggalaxySentiment

The rationale used to eliminate other attributes

* These variables have zero variances: nokiacampos, nokiacamneg, nokiacamunc, nokiadispos, nokiadisneg, nokiadisunc, nokiaperpos, nokiaperneg, nokiaperunc

All the above columns have same values per column except for one or two rows so they can be removed.

* Highly correlated attributes were removed. This was achieved by using the findCorrelation function from the caret package.

Comparative performance of all the classifiers

|  |  |  |
| --- | --- | --- |
|  | 19 Attributes |  |
|  | Accuracy | Kappa |
| KNN | 0.9577720 | 0.7836743 |
| RF | 0.9891710 | 0.8464041 |
| SVM | 0.9715026 | 0.8572868 |
| C5.0 | 0.9704663 | 0.8524378 |

By reducing the dimensionality of the data using the attribute selection process and reducing the number of attributes which were highly correlated the performance was much better for all the models as seen in the figure below

**What worked well?**

It was easy to analyze the results using different machine learning methods and predict the sentiment towards both the smart phones.

**What was difficult?**

It was initially difficult to come up with right number of attributes to be used for the prediction. It took several iterations and few trials and errors before I was able to confidently select the number of attributes both for iPhone and Samsung Galaxy

**What did not work well?**

As I am still not familiar (a lot to learn) with different ways to plot the graphs, I was not able to capture all my observations in the graph format using R