DATA MINGING FINAL PROJECT:

Debate tweets and political leanings on 2016 Election

Team7: dtpl

- Team members:
- Liu, Disheng (email: dil36@pitt.edu)
- Zhu, Zixuan (email: ziz57@pitt.edu)
- Lei, Chen (email: chl276@pitt.edu)

Purpose of this project

Predicting which candidate (Hillary Clinton or Donald Trump) that a person will follow and support after based on their tweets. We try to figure out factors that can distinguish people among two groups the most and we believe a good model of dichotomizing twitter users will help us make less mistake in the next election.

The importance of the project?

Social media has changed the rules of the political game, allowing candidates to communicate directly with voters on everything from macro policies to what’s for dinner.

Example: the 2016 US Presidential Election Result is Dramatic

Our data source:

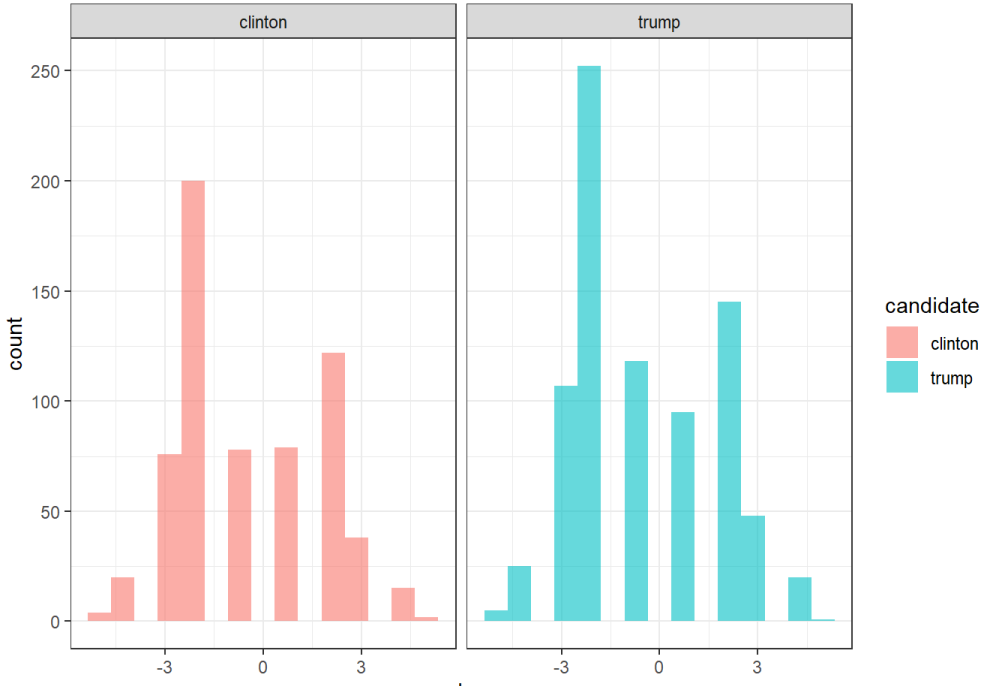
SET B:TWITTER 3

What does the data set look like?

Clinton

Trump

Sentiment analysis



For Clinton the positive rate is 0.47855 and for trump is 0.378676503

Clinton has less negative and positive words in tweets related to her. But this is only from the 10k data set we extracted from our data set1.

Data cleaning

- Remove candidate name
- Remove RT retweet tag
- Remove special signs and non-english characts
- Romove useless but commonly seen words such as debates, support and etc.

Model Establishment 1:

Pre-Processing data

method:TF-IDF

MODELS:

1.K-Means

2.SVM

3.Classification Tree

Pre-Processing data

method:TF-IDF+PCA

MODELS:

1.K-Means

2.SVM

3.Classification Tree

4.Logistic Regression

5.Neural Network

6. Random Forest

1. K-Means

TF_IDF

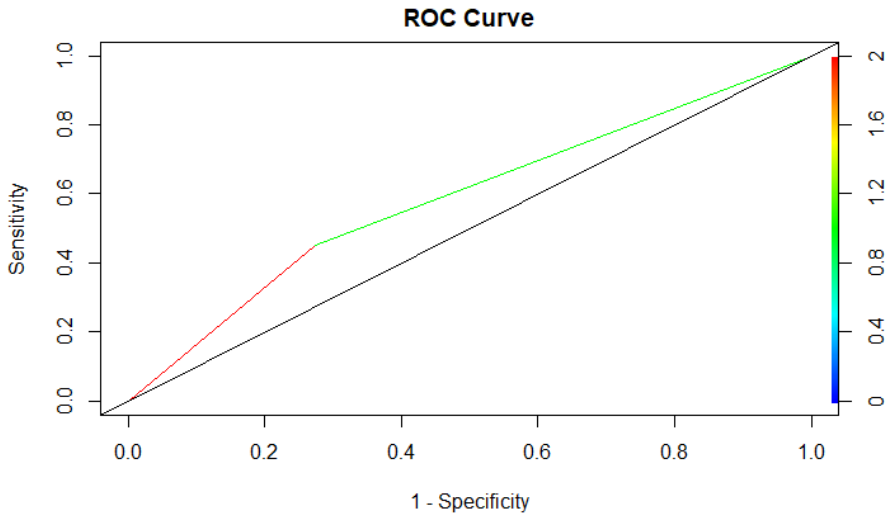
```
Accuracy: 0.5005061
Sensitivity: 2/(2+2) = 0.5
Specificity: 987/(987+985) = 0.5005071
F1 score: 2*(0.5 * 0.5005071)/(0.5 + 0.5005071) = 0.5002534
```

PCA+TF_IDF

```
Accuracy:0.5
Sensitivity: 0
Specificity: 988/(988+987) = 0.5002532
F1 score: 0
```

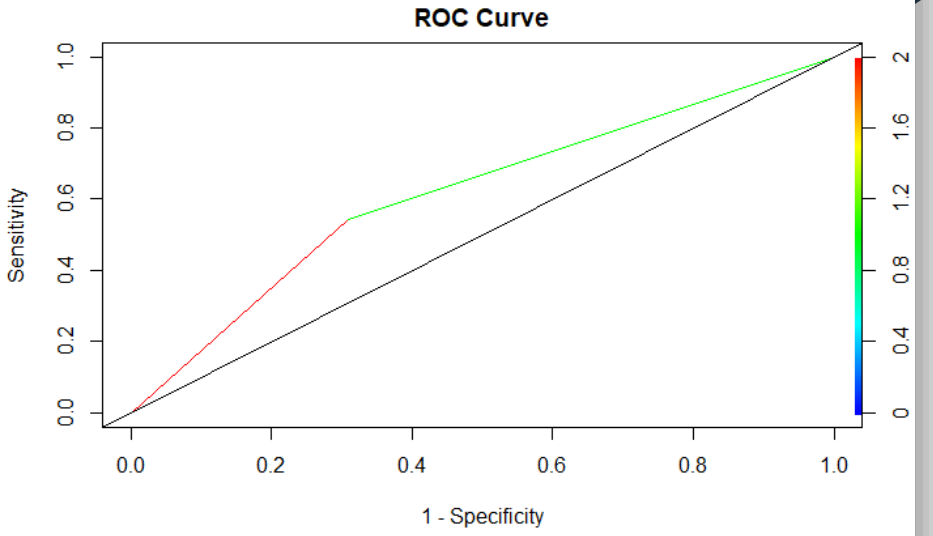
2.SVM

TF_IDF




```
Accuracy: 0.6167513
Sensitivity: 107/(107+61) = 0.6369048
Specificity: 136/(136+90) = 0.6017699
F1 score: 2*(0.6369048 * 0.6017699)/(0.6369048 + 0.6017699) = 0.6188391
```

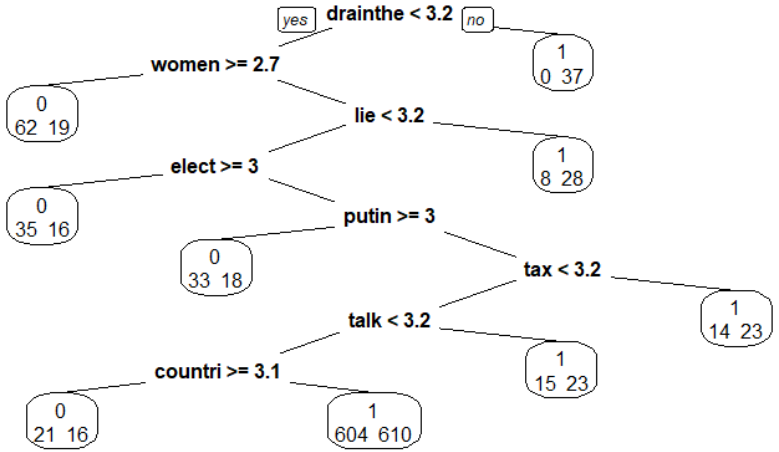
PCA+TF_IDF




```
Accuracy: 0.5888325
Sensitivity: 89/(89+54) = 0.6223776
Specificity: 143/(143+108) = 0.5697211
F1 score: 2*(0.6223776 * 0.5697211)/(0.6223776 + 0.5697211) = 0.5948864
```

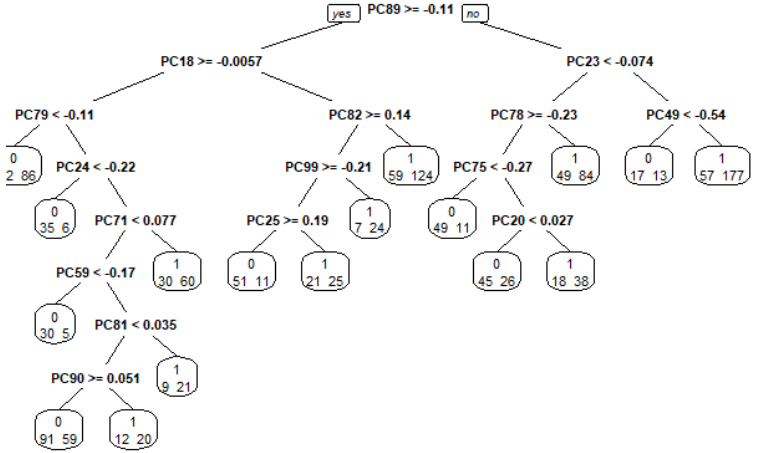
3. Classification Tree

TF-IDF



Accuracy:0.5253807
Sensitivity: 183/(183+14) = 0.928934
Specificity: 24/(24+173) = 0.1218274
F1 score: 2*(0.928934 * 0.1218274)/(0.928934 + 0.1218274) = 0.215405

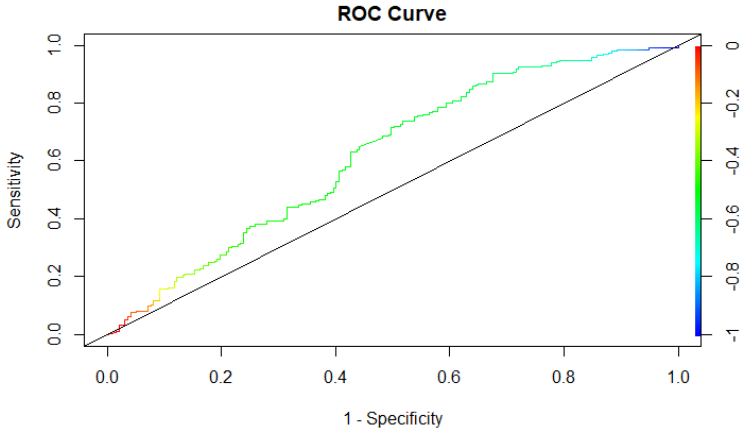
PCA+TF_IDF



Accuracy:0.5736041
Sensitivity: 120/(120+77) = 0.6091371
Specificity: 106/(106+91) = 0.5380711
F1 score: 2*(0.6091371 * 0.5380711)/(0.6091371 + 0.5380711) = 0.5714029

4. Logistic Regression

PCA+TF_IDF

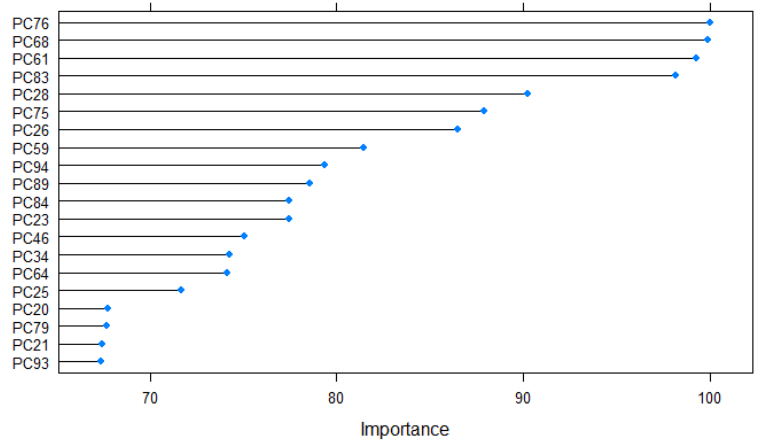



```
Accuracy: 0.5558376
Sensitivity: 88/(88+109) = 0.4467005
Specificity: 131/(131+66) = 0.6649746
F1 score: 2*(0.4467005 * 0.6649746)/(0.4467005 + 0.6649746) = 0.5344088
```

```
AUC value: 0.6206421
```

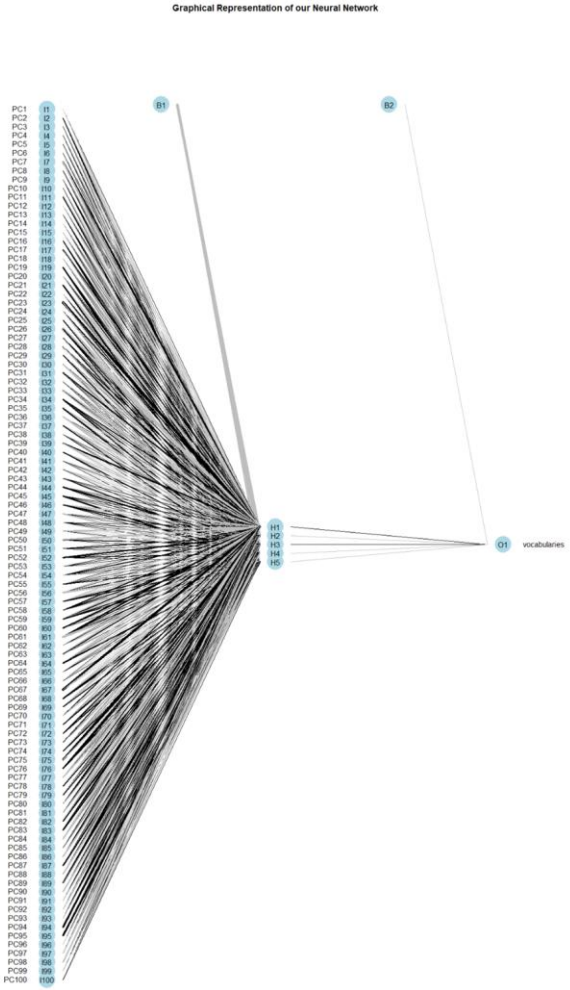
5.Neural Network

PCA+TF_IDF



size	decay	ROC	Sens	Spec
1	0e+00	0.6237088	0.6651131	0.5631441
1	1e-04	0.6152711	0.5743168	0.6170538
1	1e-01	0.6285382	0.5591396	0.6423627
3	0e+00	0.6235344	0.5612367	0.6270164
3	1e-04	0.6409704	0.5722658	0.6534995
3	1e-01	0.6334486	0.5692970	0.6281803
5	0e+00	0.6263246	0.5883556	0.5774240
5	1e-04	0.6520488	0.6654310	0.5532277
5	1e-01	0.6192705	0.5490130	0.6120084

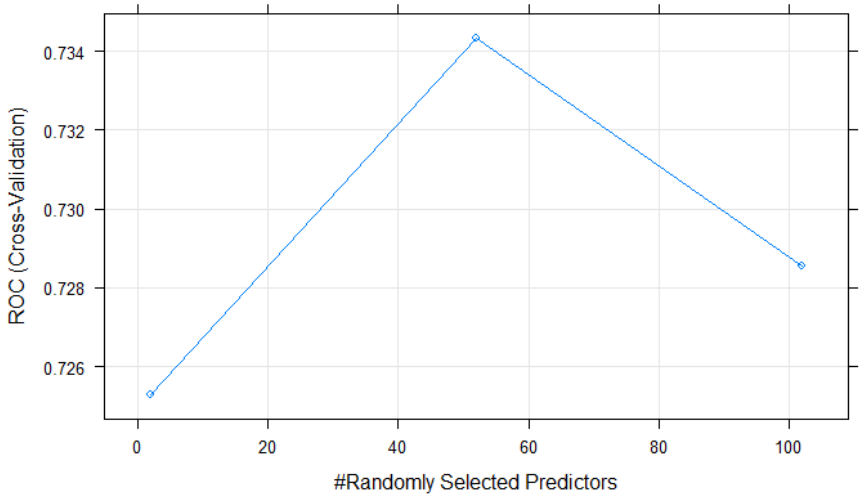
ROC was used to select the optimal model using the largest value.
The final values used for the model were size = 5 and decay = 1e-04.



6. Random Forest

PCA+TF_IDF

```
Random Forest

1976 samples
 100 predictor
   2 classes: 'clinton', 'trump'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 1581, 1581, 1581, 1581, 1580
Resampling results across tuning parameters:

  mtry  ROC        Sens       Spec
    2   0.7127792  0.6805158  0.6169410
   51   0.7184402  0.6865816  0.6199764
  100   0.7118406  0.6956878  0.6088448

ROC was used to select the optimal model using the largest value.
The final value used for the model was mtry = 51.
```



Model Establishment Plus:

The second feature—sentiment

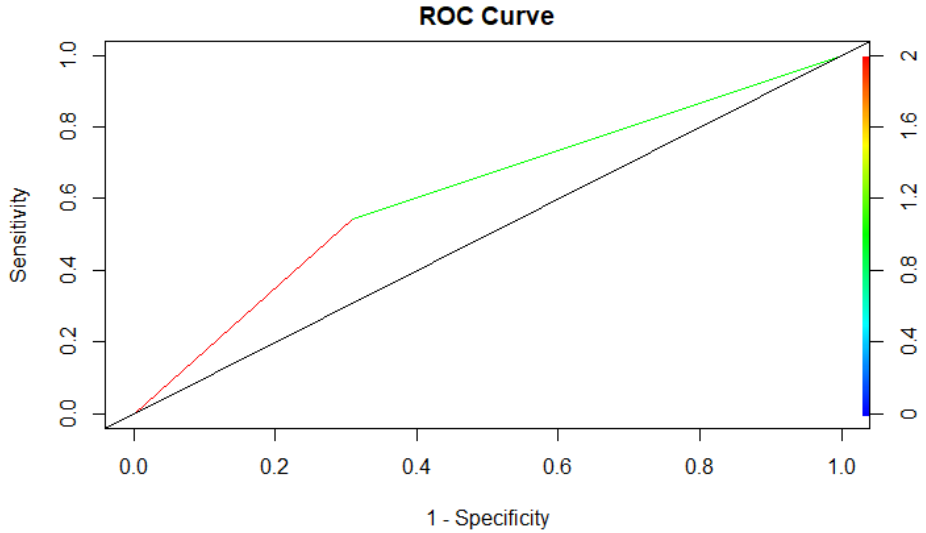
Pre-Processing data method: PCA+TF_IDF+Sentiment

1.SVM

2.Random Forest

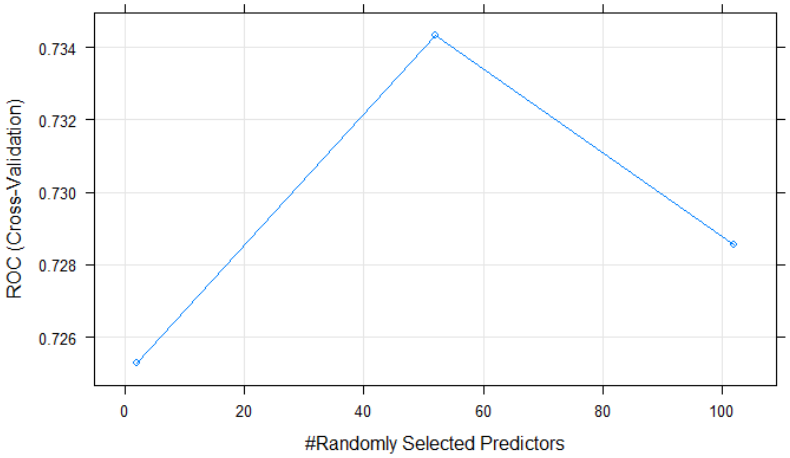
3.Neural Network

SVM



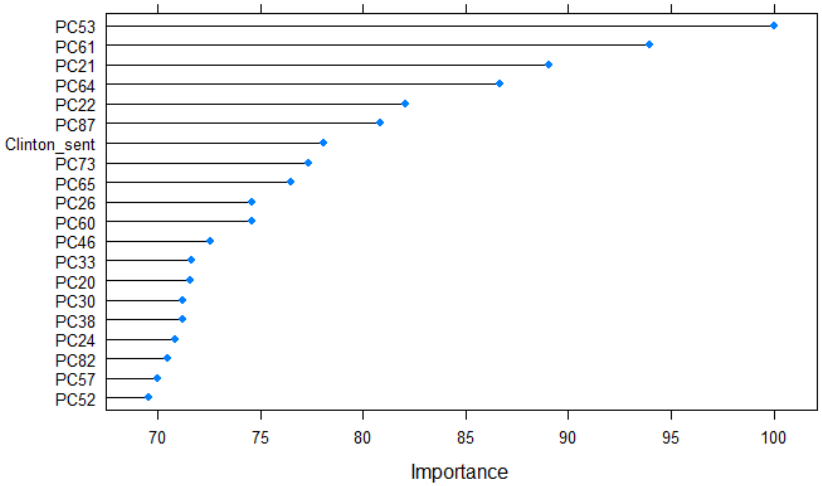

```
 [1] "Accuracy"
 [1] 0.5761421

Sensitivity: 77/(77+47) = 0.6209677
Specificity: 150/(150+120) = 0.5555556
F1 score: 2*(0.6209677 * 0.5555556)/(0.6209677 + 0.5555556) = 0.5864433
```

Random Forest



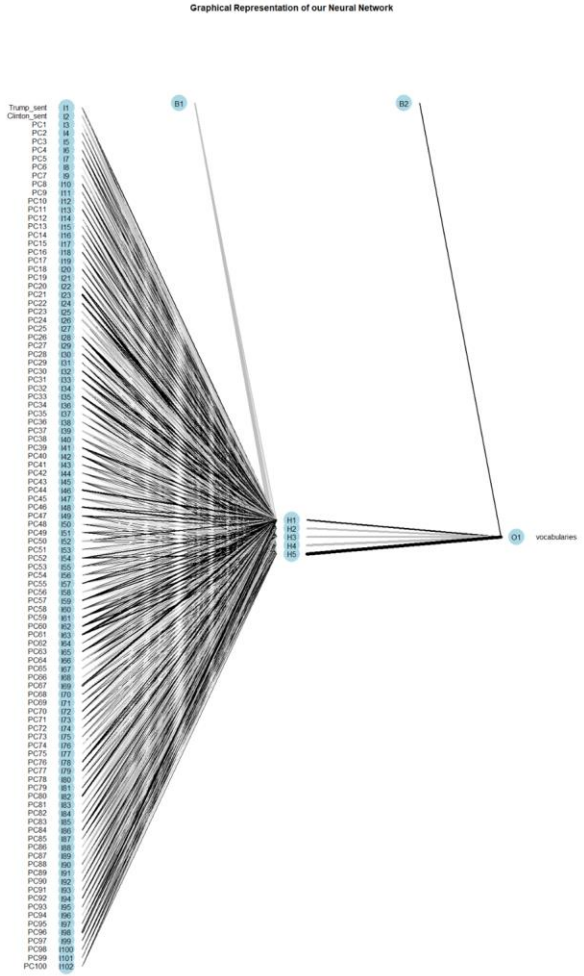

```
Random Forest

1976 samples
 102 predictor
   2 classes: 'clinton', 'trump'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 1581, 1581, 1581, 1581, 1580
Resampling results across tuning parameters:

  mtry  ROC        Sens       Spec
    2   0.7252978  0.6886376  0.6362098
   52   0.7343155  0.7037840  0.6240783
  102   0.7285477  0.7088192  0.6129313

ROC was used to select the optimal model using the largest value.
The final value used for the model was mtry = 52.
```

Neural Network



size	decay	ROC	Sens	Spec
1	0e+00	0.6171415	0.6369738	0.5804851
1	1e-04	0.6083246	0.6814746	0.4993591
1	1e-01	0.6377688	0.5975388	0.6220787
3	0e+00	0.6255285	0.6138594	0.5806389
3	1e-04	0.6406426	0.5823668	0.6291545
3	1e-01	0.6377149	0.6217813	0.5713172
5	0e+00	0.6321193	0.5925242	0.6129519
5	1e-04	0.6230166	0.6218530	0.5502282
5	1e-01	0.6471131	0.6097165	0.6220992

ROC was used to select the optimal model using the largest value.
The final values used for the model were size = 5 and decay = 0.1.



Overall model fit

	Accuracy	Sensitivity	Specificity	F1 score	ROC
kmeans(TDM)	0.5	0.5	0.5005	0.5002	NA
kmeans(TDM+PCA)	0.5	0	0.5002	0	NA
logistic(TDM+PCA)	0.555	0.446	0.664	0.534	NA
classification tree(TDM)	0.525	0.928	0.121	0.215	NA
classification tree(TDM+PCA)	0.573	0.609	0.538	0.571	NA
svm(TDM)	0.616	0.636	0.601	0.618	NA
svm(TDM+PCA)	0.588	0.622	0.569	0.594	NA
svm(TDM+PCA+sentiment)	0.576	0.62	0.555	0.586	NA
random forest(TDM+PCA)	NA	0.686	0.619	NA	0.718
random forest(TDM+PCA+sentiment)	NA	0.703	0.624	NA	0.734
neural network(TDM+PCA)	NA	0.665	0.553	NA	0.652
neural network(TDM+TDM+PCA+sentiment)	NA	0.609	0.622	NA	0.647