IMDB

Project First Milestone - IMDB Movie Sentiment

Prepared for: Annie Lee, PHD, MMATH

Prepared by: Jairo Melo, Ing. MSC

April 23, 2019

5. Project First Milestone

I chose three algorithms for Text Classification: Model Random Forest, Gradient Boosting and Flair NLP State of the Art. We evaluated performance, accuracy, precision, recall, and score. All three are great algorithms but only one showed the best performance and accuracy.

Training Random Forest Model

Random forest shows a great performance fitting the model as well as predicting. The features importance start with the length, punctuation and tokenized text.

```
Fit time: 375.802 / Predict time: 11.189 ---- Precision: 0.865 / Recall: 0.86 / Accuracy: 0.864 / F1 Score: 0.863
```

```
1 #Feature Importance
2 sorted(zip(rf_model.feature_importances_, X_train.columns), reverse=True)[0:10]
```

```
[(0.0032239412208359213, 'len'),
 (0.002586011500927889, 'punct%'),
 (0.00257477283747453, 'text_nostop'),
 (0.0025358750164764608, 'word count')]
```

```
1 print(confusion_matrix(y_test,y_pred_rf))
2 print(classification_report(y_test,y_pred_rf))
3 print(accuracy_score(y_test, y_pred_rf))
```

```
[[4394  661]
 [ 694 4251]]
              precision    recall  f1-score   support

           0       0.86      0.87      0.87      5055
           1       0.87      0.86      0.86      4945

   micro avg       0.86      0.86      0.86     10000
   macro avg       0.86      0.86      0.86     10000
weighted avg       0.86      0.86      0.86     10000

0.8645
```

Training Gradient Boosting Model

Fitting and Predicting is very expensive. The feature considered are only Punctuation and Tokenized words. Word count and Length are disregarded.

```
Fit time: 18183.007 / Predict time: 44.431 ---- Precision: 0.779 / Recall: 0.869 / Accuracy: 0.813 / F1 Score: 0.821
```

```
1 #FTheeature Importance
2 sorted(zip(gb_model.feature_importances_, X_train.columns), reverse=True)[0:10]
```

```
[(0.0002817854324905849, 'punct%'),
 (4.5316218563653224e-05, 'text_nostop'),
 (0.0, 'word count'),
 (0.0, 'len')]
```

```
1 print(confusion_matrix(y_test,y_pred_gb))
2 print(classification_report(y_test,y_pred_gb))
3 print(accuracy_score(y_test, y_pred_gb))
```

```
[[3836 1219]
 [ 649 4296]]
              precision    recall  f1-score   support

           0       0.86      0.76      0.80      5055
           1       0.78      0.87      0.82      4945

   micro avg       0.81      0.81      0.81     10000
   macro avg       0.82      0.81      0.81     10000
weighted avg       0.82      0.81      0.81     10000

0.8132
```

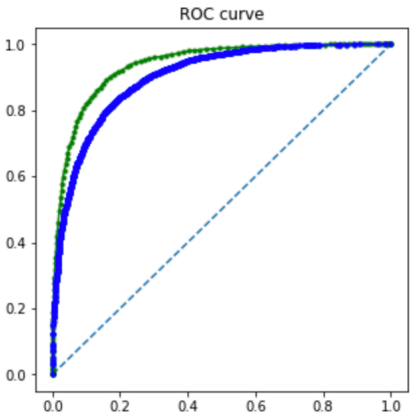
IMDB

Random forest and Gradient Boosting Model Evaluation

As shown in the charts, both models have great skill of learning and good precision curve which both have great predicting power. However, based on performance, accuracy and precision we conclude Random Forest is the selected model from these two.

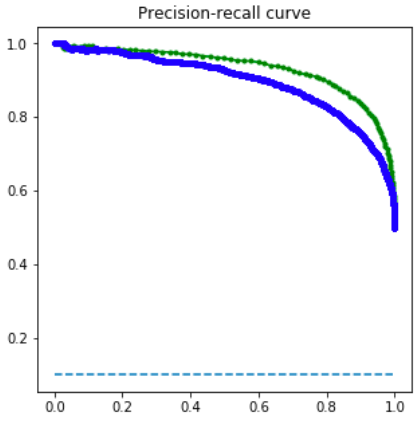
```
Randomforest AUC: 0.939

GradientBoosting AUC: 0.902
```




```
Randomforest: f1=0.863 auc=0.933

GradientBoosting: f1=0.821 auc=0.896
```



Flair Text Classifier

We trained a Flair Text classifier Model 150 epochs during two days; Accuracy of 47% and F1 Score of 65%.

```
-------
2019-04-22 19:02:38,666 EPOCH 150 done: loss 0.0001 - lr 0.0500 - bad epochs 1
2019-04-22 19:03:09,868 DEV  : loss 0.10816629 - f-score 0.6000 - acc 0.4286
2019-04-22 19:08:28,407 TEST : loss 0.10048141 - f-score 0.6532 - acc 0.4850
2019-04-22 19:08:47,224 ----------------------------------------------------------------------------------------------------
-------
2019-04-22 19:08:47,233 Testing using best model ...
2019-04-22 19:08:47,254 loading file resources/taggers/sentiment/best-model.pt
2019-04-22 19:12:52,708 MICRO_AVG: acc 0.4736 - f1-score 0.6428
2019-04-22 19:12:52,734 MACRO_AVG: acc 0.4736 - f1-score 0.64275
2019-04-22 19:12:52,736 neg        tp: 825 - fp: 468 - fn: 425 - tn: 782 - precision: 0.6381 - recall: 0.6600 - accur
acy: 0.4802 - f1-score: 0.6489
2019-04-22 19:12:52,737 pos        tp: 782 - fp: 425 - fn: 468 - tn: 825 - precision: 0.6479 - recall: 0.6256 - accur
acy: 0.4669 - f1-score: 0.6366
2019-04-22 19:12:52,739 ----------------------------------------------------------------------------------------------------
-------
```

Running a prediction we identify that Flair model accurately identify that the review is positive.

```
2019-04-22 23:43:47,197 loading file resources/taggers/sentiment/final-model.pt
[pos (1.0)]
[Sentence: "I liked French Kiss, Kevin Kline and Meg Ryan delivered an amazing performance; love the story and the co
nnection to grapes and wine in France" - 25 Tokens]
```

Conclusion: Random Forest is our best predicting model based on our initial criteria:

Model	Training perf	Predicting Perf	Accuracy	Precision	Recall	F1 Score
Random Forest	6mins	18 secs	86%	87%	86%	86%
Gradient Boosting	5 hrs	1.4 mins	81%	78%	87%	82%
Flair	+2Days	~30 secs	47%	65%	62%	64%