GBDT and Random Forests w2vec tfidfw2vec[M]

June 2, 2018

In [2]:
```
!pip install gensim
!pip install imblearn
!pip install xgboost
```

```
Requirement already satisfied: gensim in /usr/local/lib/python3.6/site-packages
Requirement already satisfied: numpy>=1.11.3 in /usr/local/lib/python3.6/site-packages (from gensim)
Requirement already satisfied: scipy>=0.18.1 in /usr/local/lib/python3.6/site-packages (from gensim)
Requirement already satisfied: six>=1.5.0 in /usr/local/lib/python3.6/site-packages (from gensim)
Requirement already satisfied: smart-open>=1.2.1 in /usr/local/lib/python3.6/site-packages (from gensim)
Requirement already satisfied: bz2file in /usr/local/lib/python3.6/site-packages (from smart-open>=1.2.1->gensi
Requirement already satisfied: boto3 in /usr/local/lib/python3.6/site-packages (from smart-open>=1.2.1->gensim
Requirement already satisfied: boto>=2.32 in /usr/local/lib/python3.6/site-packages (from smart-open>=1.2.1->g
Requirement already satisfied: requests in /usr/local/lib/python3.6/site-packages (from smart-open>=1.2.1->gen
Requirement already satisfied: s3transfer<0.2.0,>=0.1.10 in /usr/local/lib/python3.6/site-packages (from boto3->
Requirement already satisfied: jmespath<1.0.0,>=0.7.1 in /usr/local/lib/python3.6/site-packages (from boto3->sm
Requirement already satisfied: botocore<1.11.0,>=1.10.31 in /usr/local/lib/python3.6/site-packages (from boto3-
Requirement already satisfied: python-dateutil<3.0.0,>=2.1; python_version >= "2.7" in /usr/local/lib/python3.6/si
Requirement already satisfied: docutils>=0.10 in /usr/local/lib/python3.6/site-packages (from botocore<1.11.0,>=
You are using pip version 9.0.1, however version 10.0.1 is available.You should consider upgrading via the 'pip i
Requirement already satisfied: imblearn in /usr/local/lib/python3.6/site-packages
Requirement already satisfied: imbalanced-learn in /usr/local/lib/python3.6/site-packages (from imblearn)
Requirement already satisfied: numpy in /usr/local/lib/python3.6/site-packages (from imbalanced-learn->imblear
Requirement already satisfied: scipy in /usr/local/lib/python3.6/site-packages (from imbalanced-learn->imblearn)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.6/site-packages (from imbalanced-learn->imb
You are using pip version 9.0.1, however version 10.0.1 is available.You should consider upgrading via the 'pip i
Requirement already satisfied: xgboost in /usr/local/lib/python3.6/site-packages/xgboost-0.7-py3.6.egg
Requirement already satisfied: numpy in /usr/local/lib/python3.6/site-packages (from xgboost)
Requirement already satisfied: scipy in /usr/local/lib/python3.6/site-packages (from xgboost)
You are using pip version 9.0.1, however version 10.0.1 is available.You should consider upgrading via the 'pip i
```

In [3]:
```
from sklearn.model_selection import train_test_split

from sklearn.grid_search import GridSearchCV
from sklearn.grid_search import RandomizedSearchCV
from scipy.stats import randint

from imblearn.over_sampling import SMOTE
```

```
import sqlite3

import pandas as pd
import numpy as np

from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
import gensim

from sklearn.metrics import classification_report, accuracy_score, confusion_matrix

from sklearn.ensemble import RandomForestClassifier
from xgboost.sklearn import XGBClassifier
```

In [20]:
```
con = sqlite3.connect('final.sqlite') # this is cleaned dataset
final = pd.read_sql_query("""
SELECT Score, Text_not_included
FROM reviews
""", con)[:2000]

for i, seq in enumerate(final['Text_not_included']):
    final['Text_not_included'][i]=final['Text_not_included'][i].decode('UTF-8')
X_train, X_test, y_train , y_test = train_test_split(final['Text_not_included'], final['Score'], test_size=0.2
```

In [21]:
```
# Generate count BoW
count_vect = CountVectorizer(ngram_range=(1,2) )
count_vect.fit(X_train)
bow_train=count_vect.transform(X_train)
bow_test=count_vect.transform(X_test)

# Generate tf idf
tf_idf_vect=TfidfVectorizer(ngram_range=(1,2), min_df=10, dtype=float)
tf_idf_vect.fit(X_train)
tf_idf_train=tf_idf_vect.transform(X_train)
tf_idf_test=tf_idf_vect.transform(X_test)

# Generate average word2vec
sentences=[]
for review in X_train:
    sentence=[]
    for word in review.split():
        sentence.append(word)
    sentences.append(sentence)

w2vec_model=gensim.models.word2vec.Word2Vec(sentences, min_count=10)

avg_w2vec_train=np.zeros(shape=(len(X_train), 100), dtype=float)
```

```
for i, sentence in enumerate(sentences):
    for word in sentence:
        try:
            avg_w2vec_train[i]+=w2vec_model.wv[word]

        except KeyError:
            pass

    avg_w2vec_train[i]/=len(sentence)

sentences=[]
for review in X_test:
    sentence=[]
    for word in review.split():
        sentence.append(word)
    sentences.append(sentence)

avg_w2vec_test=np.zeros(shape=(len(X_test), 100), dtype=float)

for i, sentence in enumerate(sentences):
    for word in sentence:
        try:
            avg_w2vec_test[i]+=w2vec_model.wv[word]

        except KeyError:
            pass

    avg_w2vec_test[i]/=len(sentence)

# Generate tf idf weighted word2vec
sentences=[]
for review in X_train:
    sentence=[]
    for word in review.split():
        sentence.append(word)
    sentences.append(sentence)

tf_idf_w2vec_train=np.zeros((len(X_train), 100), dtype=float)
feat=tf_idf_vect.get_feature_names()
for i, sentence in enumerate(sentences):
    tf_idf_sum=0
    for word in sentence:
        try:
            tf_idf_w2vec_train[i]+=w2vec_model.wv[word]*tf_idf_train[i, feat.index(word)]
            tf_idf_sum+=tf_idf_train[i, feat.index(word)]
        except KeyError:
            pass
```

```
            except ValueError:
                pass
        tf_idf_w2vec_train[i]/=tf_idf_sum

    sentences=[]
    for review in X_test:
        sentence=[]
        for word in review.split():
            sentence.append(word)
        sentences.append(sentence)

    tf_idf_w2vec_test=np.zeros((len(X_test), 100), dtype=float)

    for i, sentence in enumerate(sentences):
        tf_idf_sum=0
        for word in sentence:
            try:
                tf_idf_w2vec_test[i]+=w2vec_model.wv[word]*tf_idf_test[i, feat.index(word)]
                tf_idf_sum+=tf_idf_test[i, feat.index(word)]
            except KeyError:
                pass
            except ValueError:
                pass
        tf_idf_w2vec_test[i]/=tf_idf_sum
```

```
/usr/local/lib/python3.6/site-packages/sklearn/feature_extraction/text.py:1089: FutureWarning: Conversion of the
  if hasattr(X, 'dtype') and np.issubdtype(X.dtype, np.float):
```

0.1 Upsampling - Decision trees affected by imbalanced dataset

```
In [22]: # Upsampling minority class
    over_sampler = SMOTE(ratio='minority')
    bow_train_resampled, y_train_resampled = over_sampler.fit_sample(bow_train, y_train)
    tf_idf_train_resampled, y_train_resampled = over_sampler.fit_sample(tf_idf_train, y_train)
    avg_w2vec_train_resampled, y_train_resampled = over_sampler.fit_sample(avg_w2vec_train, y_train
    tf_idf_w2vec_train_resampled, y_train_resampled = over_sampler.fit_sample(tf_idf_w2vec_train, y_tr
```

1 Classification using RandomForest

```
In [25]: tuned_parameters = {'n_estimators': np.arange(1,100,1)}
    gscv = GridSearchCV(RandomForestClassifier(n_jobs=-1), tuned_parameters, scoring = 'accuracy', c

    tuned_parameters = {'n_estimators' : randint(low=1, high=101)}
    rscv = RandomizedSearchCV(RandomForestClassifier(n_jobs=-1), tuned_parameters, scoring = 'acc
```

1.0.1 Word2Vec

In [26]:
```
gscv.fit(avg_w2vec_train_resampled, y_train_resampled)

        predictions=gscv.best_estimator_.predict(avg_w2vec_test)
        print(classification_report(y_test, predictions))
        print(confusion_matrix(y_test, predictions).T)
        tn, fp, fn, tp = confusion_matrix(y_test, predictions).ravel()

        print("TPR = {}\n TNR = {}\n FPR = {}\n FNR = {}".format(tp/(fn+tp), tn/(tn+fp), fp/(tn+fp), fn/(fn+tp)))
```

```
             precision    recall  f1-score   support

   negative       0.38      0.17      0.24        87
   positive       0.80      0.92      0.86       313

avg / total       0.71      0.76      0.72       400

[[ 15  24]
 [ 72 289]]
TPR = 0.9233226837060703
 TNR = 0.1724137931034483
 FPR = 0.8275862068965517
 FNR = 0.07667731629392971
```

In [27]:
```
rscv.fit(avg_w2vec_train_resampled, y_train_resampled)

        predictions=rscv.best_estimator_.predict(avg_w2vec_test)
        print(classification_report(y_test, predictions))
        print(confusion_matrix(y_test, predictions).T)
        tn, fp, fn, tp = confusion_matrix(y_test, predictions).ravel()

        print("TPR = {}\n TNR = {}\n FPR = {}\n FNR = {}".format(tp/(fn+tp), tn/(tn+fp), fp/(tn+fp), fn/(fn+tp)))
```

```
             precision    recall  f1-score   support

   negative       0.41      0.22      0.29        87
   positive       0.81      0.91      0.86       313

avg / total       0.72      0.76      0.73       400

[[ 19  27]
 [ 68 286]]
TPR = 0.9137380191693291
 TNR = 0.21839080459770116
 FPR = 0.7816091954022989
 FNR = 0.08626198083067092
```

In [28]: gscv.best_estimator_

Out[28]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
max_depth=None, max_features='auto', max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, n_estimators=99, n_jobs=-1,
oob_score=False, random_state=None, verbose=0,
warm_start=False)

In [29]: rscv.best_estimator_

Out[29]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
max_depth=None, max_features='auto', max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, n_estimators=93, n_jobs=-1,
oob_score=False, random_state=None, verbose=0,
warm_start=False)

1.0.2 TF IDF weighted word2Vec

In [30]: gscv.fit(tf_idf_w2vec_train_resampled, y_train_resampled)

```
predictions=gscv.best_estimator_.predict(tf_idf_w2vec_test)
print(classification_report(y_test, predictions))
print(confusion_matrix(y_test, predictions).T)
tn, fp, fn, tp = confusion_matrix(y_test, predictions).ravel()

print("TPR = {}\n TNR = {}\n FPR = {}\n FNR = {}".format(tp/(fn+tp), tn/(tn+fp), fp/(tn+fp), fn/(fn+tp)))
```

	precision	recall	f1-score	support
negative	0.23	0.07	0.11	87
positive	0.78	0.94	0.85	313
avg / total	0.66	0.75	0.69	400

```
[[  6  20]
 [ 81 293]]
TPR = 0.9361022364217252
 TNR = 0.06896551724137931
 FPR = 0.9310344827586207
 FNR = 0.06389776357827476
```

In [31]: rscv.fit(tf_idf_w2vec_train_resampled, y_train_resampled)

```
predictions=rscv.best_estimator_.predict(tf_idf_w2vec_test)
```

```
print(classification_report(y_test, predictions))
print(confusion_matrix(y_test, predictions).T)
tn, fp, fn, tp = confusion_matrix(y_test, predictions).ravel()

print("TPR = {}\n TNR = {}\n FPR = {}\n FNR = {}".format(tp/(fn+tp), tn/(tn+fp), fp/(tn+fp), fn/(fn+tp)))
```

```
             precision    recall  f1-score   support

   negative       0.25      0.08      0.12        87
   positive       0.78      0.93      0.85       313

avg / total       0.67      0.75      0.69       400

[[  7  21]
 [ 80 292]]
TPR = 0.9329073482428115
 TNR = 0.08045977011494253
 FPR = 0.9195402298850575
 FNR = 0.0670926517571885
```

In [32]:
```
gscv.best_estimator_
```

Out[32]:
```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=73, n_jobs=-1,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False)
```

In [33]:
```
rscv.best_estimator_
```

Out[33]:
```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=72, n_jobs=-1,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False)
```

2 Classification using GBDT

In [41]:
```
from scipy.stats import uniform

tuned_parameters = {'n_estimators': np.arange(10,60,10),'max_depth' : np.arange(1,5,1), 'learning_ra
gscv = GridSearchCV(XGBClassifier(), tuned_parameters, scoring = 'accuracy', cv=5)

tuned_parameters = {'n_estimators': randint(low=10, high=61), 'max_depth' : randint(low=1, high=6), '
rscv = RandomizedSearchCV(XGBClassifier(), tuned_parameters, scoring = 'accuracy', cv=5, n_iter=
```

2.0.1 Word2Vec

In [42]:
```
gscv.fit(avg_w2vec_train_resampled, y_train_resampled)

predictions=gscv.best_estimator_.predict(avg_w2vec_test)
print(classification_report(y_test, predictions))
print(confusion_matrix(y_test, predictions).T)
tn, fp, fn, tp = confusion_matrix(y_test, predictions).ravel()

print("TPR = {}\n TNR = {}\n FPR = {}\n FNR = {}".format(tp/(fn+tp), tn/(tn+fp), fp/(tn+fp), fn/(fn+tp)))
```

```
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
```

```
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:


             precision    recall  f1-score   support

    negative       0.36      0.34      0.35        87
    positive       0.82      0.83      0.83       313

avg / total        0.72      0.72      0.72       400

[[ 30  53]
 [ 57 260]]
TPR = 0.8306709265175719
 TNR = 0.3448275862068966
 FPR = 0.6551724137931034
 FNR = 0.16932907348242812


/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
```

In [37]:
```
rscv.fit(avg_w2vec_train_resampled, y_train_resampled)

        predictions=rscv.best_estimator_.predict(avg_w2vec_test)
        print(classification_report(y_test, predictions))
        print(confusion_matrix(y_test, predictions).T)
        tn, fp, fn, tp = confusion_matrix(y_test, predictions).ravel()

        print("TPR = {}\n TNR = {}\n FPR = {}\n FNR = {}".format(tp/(fn+tp), tn/(tn+fp), fp/(tn+fp), fn/(fn+tp)))
```

```
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
```

```
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:

             precision    recall  f1-score   support

   negative       0.37      0.31      0.34        87
   positive       0.82      0.85      0.83       313
```

```
avg / total               0.72          0.73          0.73          400

[[ 27      46]
 [ 60 267]]
TPR = 0.853035143769968
 TNR = 0.3103448275862069
 FPR = 0.6896551724137931
 FNR = 0.14696485623003194


/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
   if diff:


In [43]: gscv.best_estimator_

Out[43]: XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
                       colsample_bytree=1, gamma=0, learning_rate=0.5, max_delta_step=0,
                       max_depth=4, min_child_weight=1, missing=None, n_estimators=50,
                       n_jobs=1, nthread=None, objective='binary:logistic', random_state=0,
                       reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
                       silent=True, subsample=1)

In [39]: rscv.best_estimator_

Out[39]: XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
                       colsample_bytree=1, gamma=0, learning_rate=0.30371216290769487,
                       max_delta_step=0, max_depth=5, min_child_weight=1, missing=None,
                       n_estimators=53, n_jobs=1, nthread=None,
                       objective='binary:logistic', random_state=0, reg_alpha=0,
                       reg_lambda=1, scale_pos_weight=1, seed=None, silent=True,
                       subsample=1)
```

2.0.2 TF IDF weighted word2Vec

```
In [44]: gscv.fit(tf_idf_w2vec_train_resampled, y_train_resampled)

            predictions=gscv.best_estimator_.predict(tf_idf_w2vec_test)
            print(classification_report(y_test, predictions))
            print(confusion_matrix(y_test, predictions).T)
            tn, fp, fn, tp = confusion_matrix(y_test, predictions).ravel()

            print("TPR = {}\n TNR = {}\n FPR = {}\n FNR = {}".format(tp/(fn+tp), tn/(tn+fp), fp/(tn+fp), fn/(fn+tp)))

/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
   if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
   if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
```

```
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value
  if diff:


             precision    recall  f1-score   support

   negative       0.33      0.23      0.27        87
   positive       0.80      0.87      0.84       313

avg / total       0.70      0.73      0.71       400

[[ 20  40]
 [ 67 273]]
TPR = 0.8722044728434505
 TNR = 0.22988505747126436
```

```
FPR = 0.7701149425287356
FNR = 0.12779552715654952
```

```
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
  if diff:
```

In [45]:
```
rscv.fit(tf_idf_w2vec_train_resampled, y_train_resampled)

predictions=rscv.best_estimator_.predict(tf_idf_w2vec_test)
print(classification_report(y_test, predictions))
print(confusion_matrix(y_test, predictions).T)
tn, fp, fn, tp = confusion_matrix(y_test, predictions).ravel()

print("TPR = {}\n TNR = {}\n FPR = {}\n FNR = {}".format(tp/(fn+tp), tn/(tn+fp), fp/(tn+fp), fn/(fn+tp)))
```

```
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
  if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
```

```
    if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
    if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
    if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
    if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
    if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
    if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
    if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
    if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
    if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
    if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
    if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
    if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
    if diff:
/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
    if diff:


             precision    recall  f1-score   support

   negative       0.29      0.23      0.26        87
   positive       0.80      0.85      0.82       313

avg / total       0.69      0.71      0.70       400

[[ 20  48]
 [ 67 265]]
TPR = 0.8466453674121406
 TNR = 0.22988505747126436
 FPR = 0.7701149425287356
 FNR = 0.15335463258785942


/usr/local/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: DeprecationWarning: The truth value 
    if diff:
```

In [46]: gscv.best_estimator_

Out[46]: XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
colsample_bytree=1, gamma=0, learning_rate=0.5, max_delta_step=0,
max_depth=4, min_child_weight=1, missing=None, n_estimators=50,
n_jobs=1, nthread=None, objective='binary:logistic', random_state=0,
reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
silent=True, subsample=1)

In [47]: rscv.best_estimator_

Out[47]: XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
colsample_bytree=1, gamma=0, learning_rate=0.28470023723952187,
max_delta_step=0, max_depth=5, min_child_weight=1, missing=None,
n_estimators=46, n_jobs=1, nthread=None,
objective='binary:logistic', random_state=0, reg_alpha=0,
reg_lambda=1, scale_pos_weight=1, seed=None, silent=True,
subsample=1)

3 Conclusions

Random forest performance
Using W2vec TPR = 0.91 TNR = 0.22
best n_estimators = 93
Using TF IDF W2vec TPR = 0.93 TNR = 0.08
best n_estimators = 72
GBDT performance
Using W2vec TPR = 0.83 TNR = 0.34
best n_estimators = 93 best max_depth = 4 best eta = 0.5
Using TF IDF W2vec TPR = 0.87 TNR = 0.22
best n_estimators = 50 best max_depth = 4 best eta = 0.5
W2Vec representations provide better results with both classifiers. GBDT provides better TNR.