Configuration

In [1]:
```
# Parameters
ENABLE_COLAB = False

PROJECT_NAME = 'ML1010-Group-Project'
EXPERIMENT_NAME = 'ReviewText_Lemma_Bert2 (Random Forest)'
FILE_NAME = '01_ML1010_GP_RF_Bert2'
LOAD_FROM_EXP = False

#Root Machine Learning Directory. Projects appear underneath
GOOGLE_DRIVE_MOUNT = '/content/gdrive'
COLAB_ROOT_DIR = GOOGLE_DRIVE_MOUNT + '/MyDrive/Colab Notebooks'
COLAB_INIT_DIR = COLAB_ROOT_DIR + '/utility_files'

LOCAL_ROOT_DIR = '/home/magni//ML_Root/project_root'
LOCAL_INIT_DIR = LOCAL_ROOT_DIR + '/utility_files'
```

Bootstrap Environment

In [2]:
```
#add in support for utility file directory and importing
import sys
import os

if ENABLE_COLAB:
  #Need access to drive
  from google.colab import drive
  drive.mount(GOOGLE_DRIVE_MOUNT, force_remount=True)

  #add in utility directory to syspath to import
  INIT_DIR = COLAB_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = COLAB_ROOT_DIR

else:
  #add in utility directory to syspath to import
  INIT_DIR = LOCAL_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = LOCAL_ROOT_DIR

#Import Utility Support
from jarvis import Jarvis
jarvis = Jarvis(ROOT_DIR, PROJECT_NAME)

import mv_python_utils as mvutils
```

```
Wha...where am I?
I am awake now.
```

```
I have set your current working directory to /home/magni/ML_Root/project_root
/ML1010-Group-Project
The current time is 10:30
Hello sir. Extra caffeine may help.
```

Setup Runtime Environment

In [3]:

```
if ENABLE_COLAB:
  #!pip install scipy -q
  #!pip install scikit-learn -q
  #!pip install pycaret -q
  #!pip install matplotlib -q
  #!pip install joblib -q
  #!pip install pandasql -q
  !pip install umap_learn -q
  !pip install sentence_transformers -q
  !pip install spacytextblob -q
  !pip install flair -q
  display('Google Colab enabled')
else:
  display('Google Colab not enabled')

#Common imports
import json
import pandas as pd
import numpy as np
import matplotlib
import re
import nltk
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn import metrics
from sklearn.datasets import load_digits
from sklearn.model_selection import train_test_split as tts
#from yellowbrick.classifier import ConfusionMatrix
#from sklearn.linear_model import LogisticRegression
from yellowbrick.target import ClassBalance
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier

nltk.download('stopwords')
%matplotlib inline
```

```
'Google Colab not enabled'
[nltk_data] Downloading package stopwords to /home/magni/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [4]:
```
import importlib
import cw_df_metric_utils as cwutils
import DataPackage as dp
import DataPackageSupport as dps
import DataExperiment
import DataExperimentSupport
```

```
2022-01-15 10:30:31.038980: W tensorflow/stream_executor/platform/default/dso
_loader.cc:64] Could not load dynamic library 'libcudart.so.11.0'; dlerror: l
ibcudart.so.11.0: cannot open shared object file: No such file or directory
2022-01-15 10:30:31.039010: I tensorflow/stream_executor/cuda/cudart_stub.cc:
29] Ignore above cudart dlerror if you do not have a GPU set up on your machi
ne.
```

In [23]:
```
importlib.reload(dp)
importlib.reload(dps)
importlib.reload(DataExperiment)
importlib.reload(DataExperimentSupport)
```

Out[23]:
```
<module 'DataExperimentSupport' from '/home/magni/ML_Root/project_root/utilit
y_files/DataExperimentSupport.py'>
```

Load Data

In [5]:
```
#axis_labels=[1,2,3,4,5]
axis_labels=[0,1]
classifier = RandomForestClassifier()
ANALSYSIS_COL = 'reviewText_lemma_bert'
UNIQUE_COL = 'uuid'
TARGET_COL = 'overall_posneg'
```

In [6]:
```
if LOAD_FROM_EXP:
    #start from saved state
    myExp = jarvis.loadExperiment(FILE_NAME)
    myExp.display()

else:
    #start from source file and regenerate
    testDf = pd.read_pickle(jarvis.DATA_DIR_WORK + "/01_NL_ReviewText_All(new

    testDfBert = cwutils.getBertEncodeFrame(df=testDf,
                                            bertColumn=ANALSYSIS_COL,
                                            uniqueColumn=UNIQUE_COL,
                                            otherColumns=[TARGET_COL]
                                            )



    myExp = DataExperiment.DataExperiment(projectName=PROJECT_NAME,
                                          experimentName=EXPERIMENT_NAME,
                                          origData=testDfBert,
                                          uniqueColumn=UNIQUE_COL,
                                          targetColumn=TARGET_COL,
                                          classifier=classifier)
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Bert2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: False
---> isBaseModelPredicted: False
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: False
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: False
    ---> isTrainTestSplit: False
    Data:
    ---> isOrigDataLoaded: True
    ---> isTrainDataLoaded: False
    ---> isTestDataLoaded: False
```

In [7]:
```
myExp.processDataPackage()
```


```
Undersampling data to match min class: 0 of size: 13440
```


```
Completed train/test split (test_size = 0.2):
---> Original data size: 26880
---> Training data size: 21504
---> Testing data size: 5376
---> Stratified on column: overall_posneg
```

In [8]:

```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Bert2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: False
---> isBaseModelPredicted: False
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: False
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
```

```
---> targetColumn: overall_posneg
Process:
---> isBalanced: True
---> isTrainTestSplit: True
Data:
---> isOrigDataLoaded: False
---> isTrainDataLoaded: True
---> isTestDataLoaded: True
```

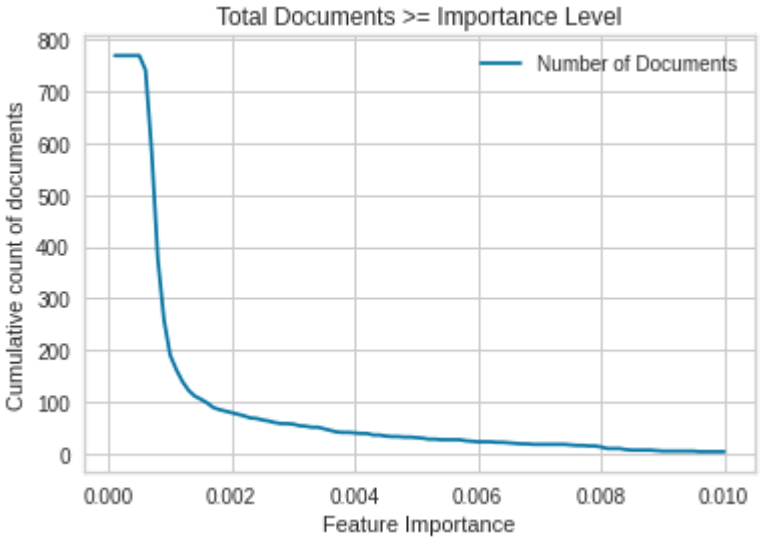
In [9]:
```
myExp.createBaseModel()
```

In [10]:
```
myExp.predictBaseModel()
```

```
Base Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.59
```

In [11]:
```
impFeatures = myExp.analyzeBaseModelFeatureImportance(returnAbove=0.002)
```

```
  0%|          | 0/101 [00:00<?, ?it/s]
Feature Importance Summary:
---> Original feature count: 768
---> Returned feature count: 80
---> Removed feature count: 688
---> Return items above (including): 0.002
```



In [12]:
```
myExp.createFinalModel(featureImportanceThreshold=0.002)
```

```
  0%|          | 0/101 [00:00<?, ?it/s]
  0%|          | 0/101 [00:00<?, ?it/s]
```

In [13]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
```

```
---> experimentName: ReviewText_Lemma_Bert2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: True
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

In [14]:

```
myExp.predictFinalModel()
myExp.display()
```

```
Final Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.59
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Bert2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

In [15]:
```
myExp.createBaseModelLearningCurve(n_jobs=10)
```

```
[learning_curve] Training set sizes: [ 1720  3440  8601 17203]
[Parallel(n_jobs=10)]: Using backend LokyBackend with 10 concurrent workers.
[Parallel(n_jobs=10)]: Done   3 out of  20 | elapsed:    4.1s remaining:   2
3.1s
[Parallel(n_jobs=10)]: Done   9 out of  20 | elapsed:   17.8s remaining:   2
1.8s
[Parallel(n_jobs=10)]: Done  15 out of  20 | elapsed:   38.8s remaining:   1
2.9s
[Parallel(n_jobs=10)]: Done  20 out of  20 | elapsed:  1.3min finished
```

In [16]:
```
myExp.createFinalModelLearningCurve(n_jobs=10)
```

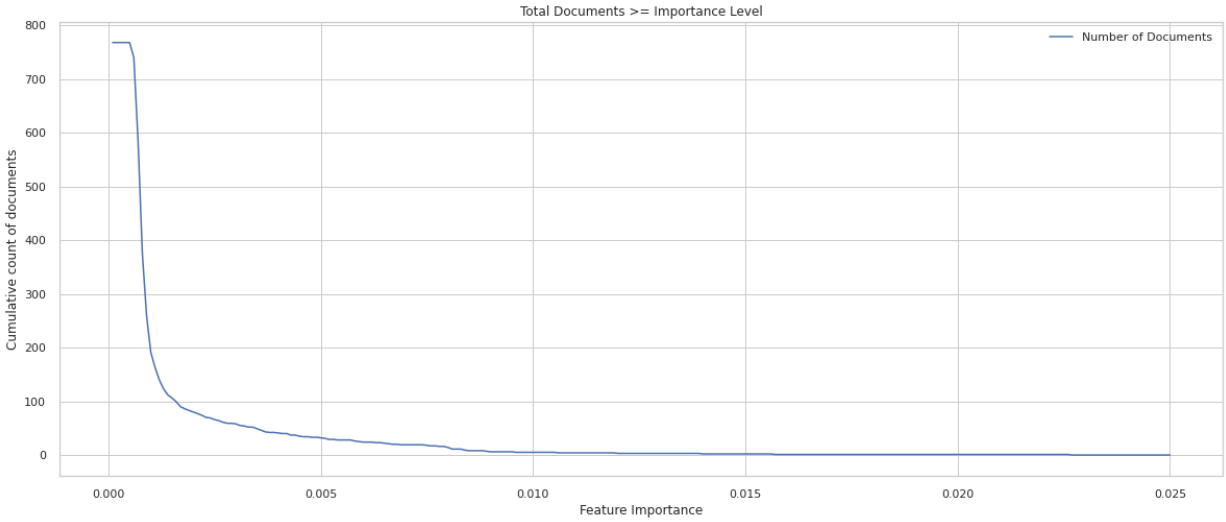
```
[Parallel(n_jobs=10)]: Using backend LokyBackend with 10 concurrent workers.
[learning_curve] Training set sizes: [ 1720  3440  8601 17203]
[Parallel(n_jobs=10)]: Done   3 out of  20 | elapsed:    1.3s remaining:
7.4s
[Parallel(n_jobs=10)]: Done   9 out of  20 | elapsed:    4.7s remaining:
5.8s
[Parallel(n_jobs=10)]: Done  15 out of  20 | elapsed:   10.3s remaining:
3.4s
[Parallel(n_jobs=10)]: Done  20 out of  20 | elapsed:   16.1s finished
```

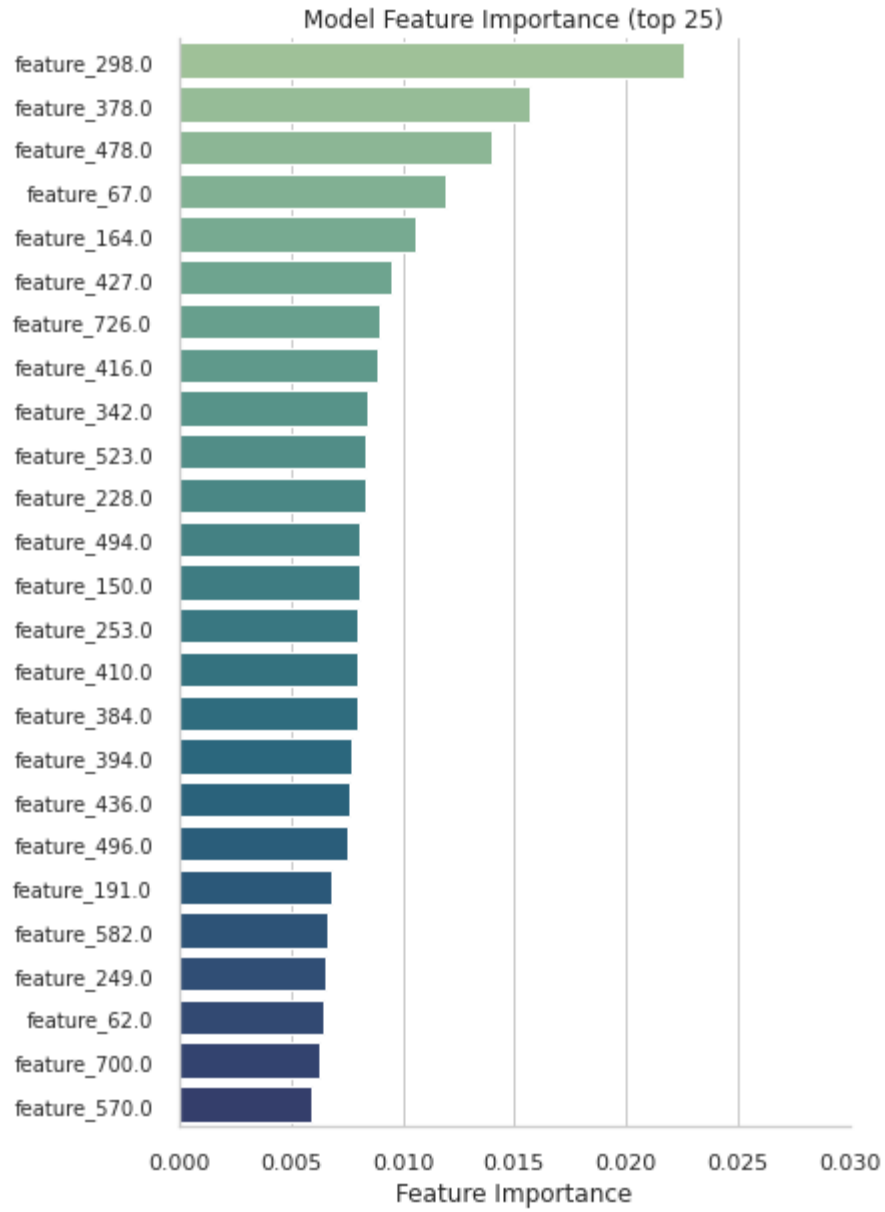
In [36]:
```
importlib.reload(dp)
importlib.reload(dps)
importlib.reload(DataExperiment)
importlib.reload(DataExperimentSupport)
```

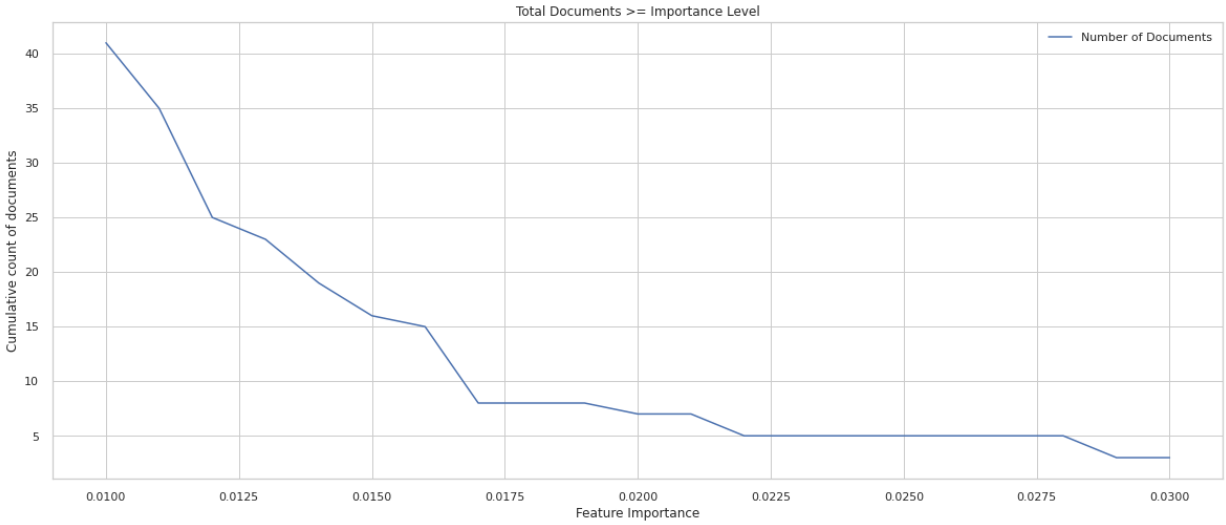
Out[36]:
```
<module 'DataExperimentSupport' from '/home/magni/ML_Root/project_root/utilit
y_files/DataExperimentSupport.py'>
```

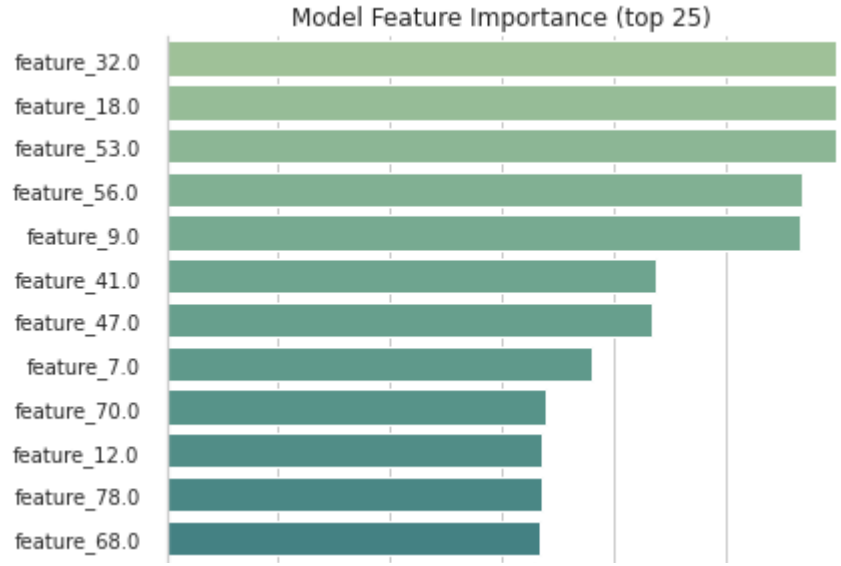
In [34]:
```
myExp.showBaseModelFeatureImportance(upperValue=0.025)
myExp.showFinalModelFeatureImportance(startValue=0.01,
                                      increment=0.001,
                                      upperValue=0.03)
```

```
  0%|          | 0/251 [00:00<?, ?it/s]
  0%|          | 0/22 [00:00<?, ?it/s]
```


Total Documents >= Importance Level
Number of Documents
Cumulative count of documents
800
700
600
500
400
300
200
100
0
0.000
0.005
0.010
0.015
0.020
0.025
Feature Importance


Model Feature Importance (top 25)
feature_298.0
feature_378.0
feature_478.0
feature_67.0
feature_164.0
feature_427.0
feature_726.0
feature_416.0
feature_342.0
feature_523.0
feature_228.0
feature_494.0
feature_150.0
feature_253.0
feature_410.0
feature_384.0
feature_394.0
feature_436.0
feature_496.0
feature_191.0
feature_582.0
feature_249.0
feature_62.0
feature_700.0
feature_570.0
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Feature Importance





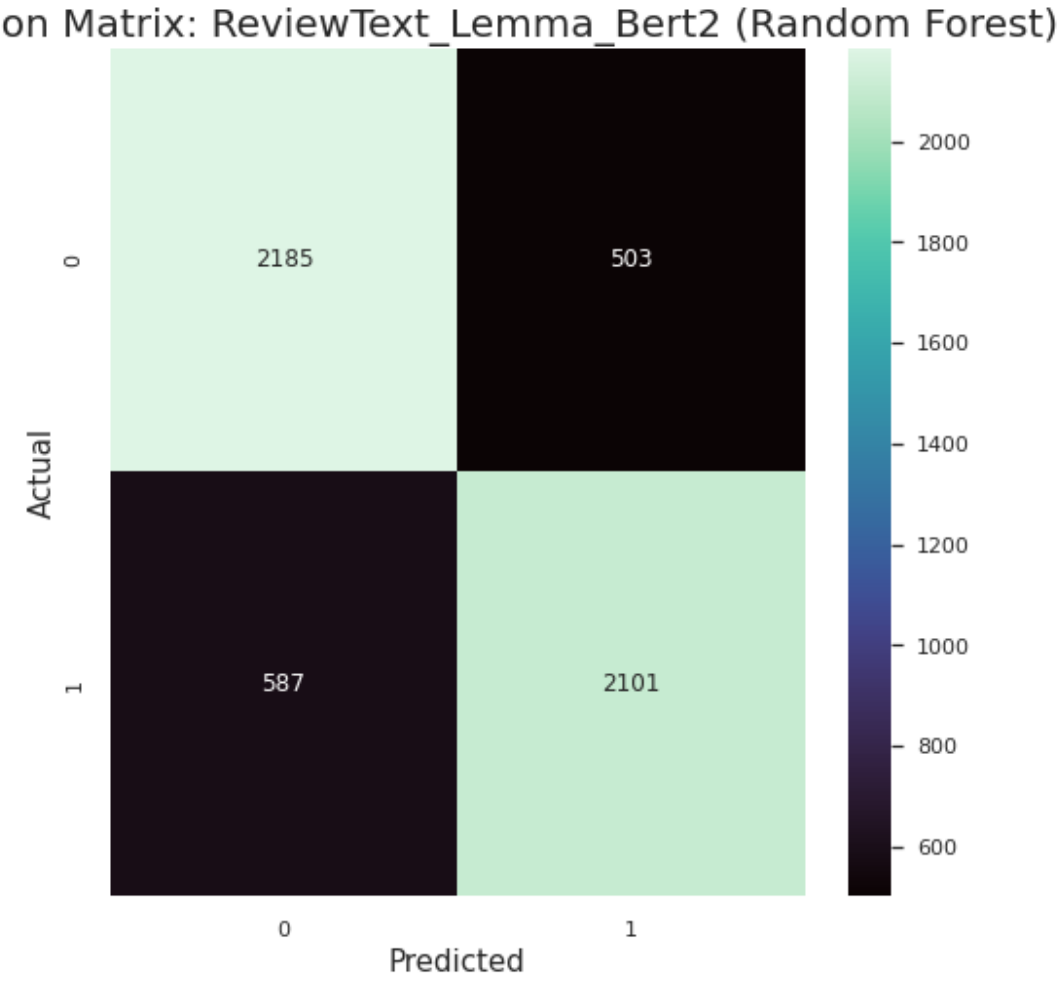
In [18]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Bert2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: True
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: True
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
```

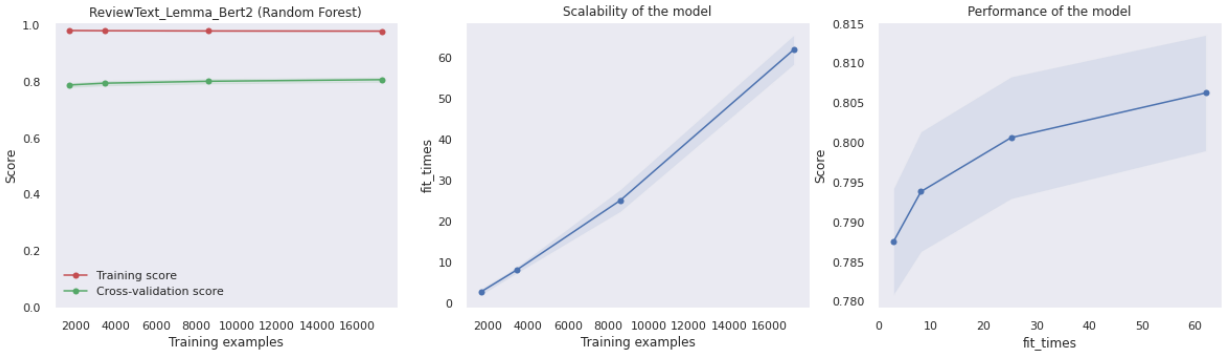
```
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

In [38]:
```
myExp.showBaseModelReport(axis_labels)
```

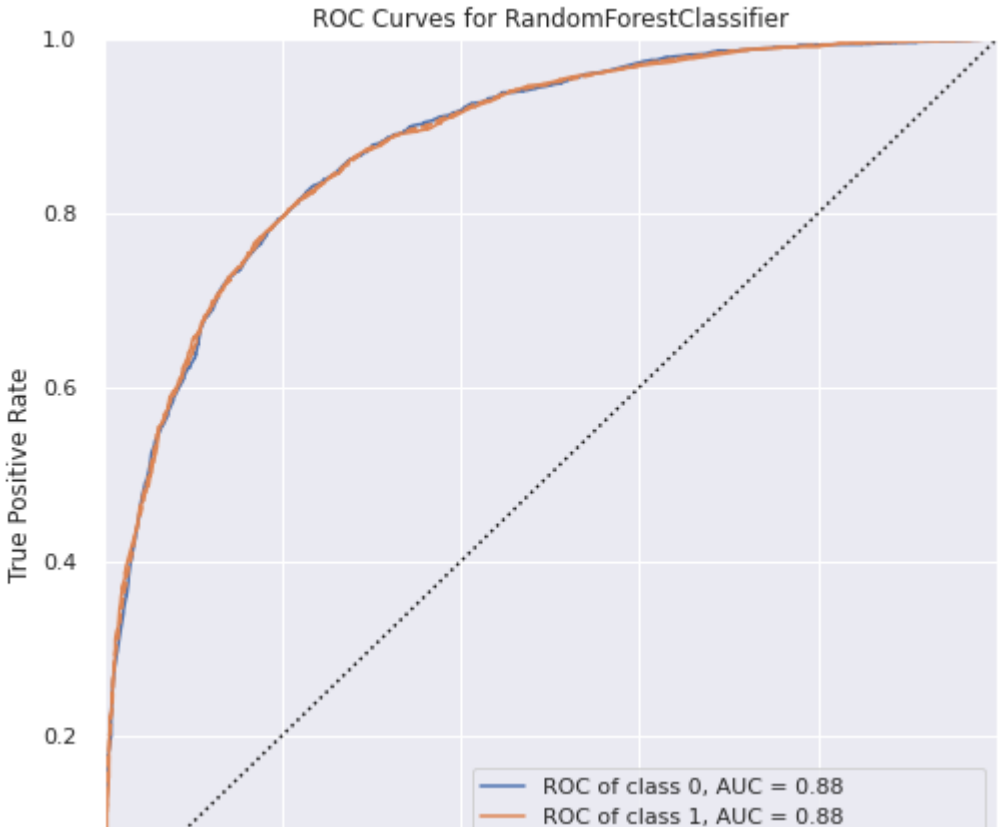
```
Base Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.59
              precision    recall  f1-score   support

           0       0.79      0.81      0.80      2688
           1       0.81      0.78      0.79      2688

    accuracy                           0.80      5376
   macro avg       0.80      0.80      0.80      5376
weighted avg       0.80      0.80      0.80      5376
```




```
<Figure size 576x576 with 0 Axes>
```

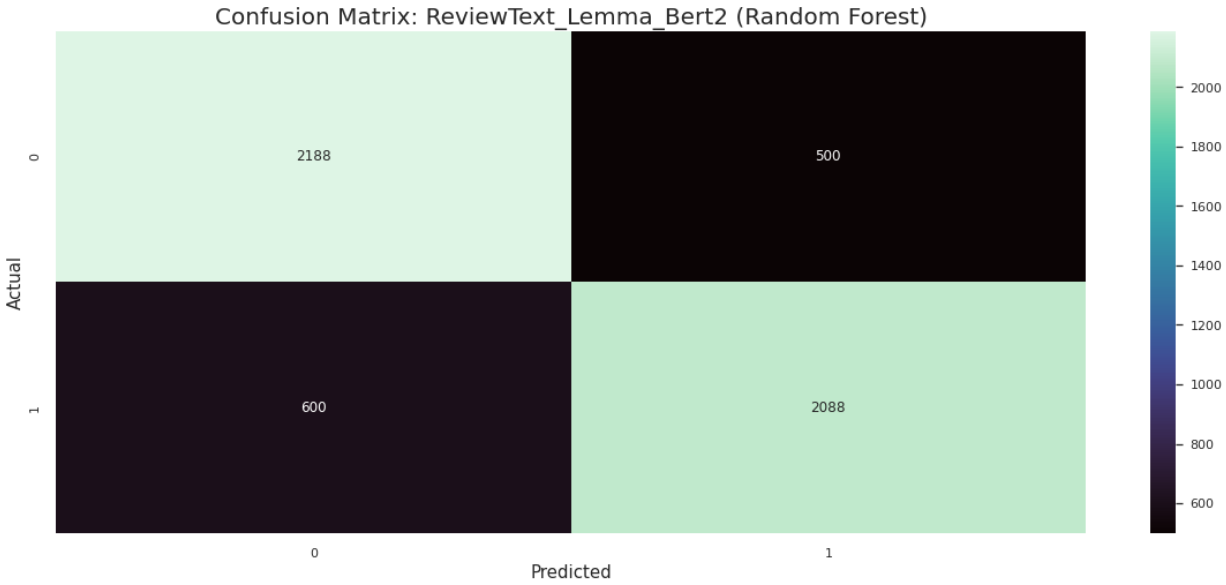


Base model ROCAUC not calculated. Starting now

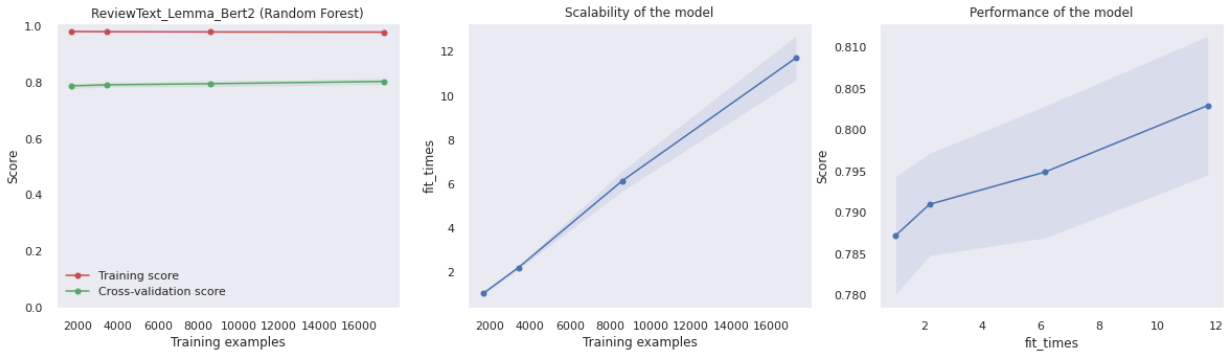


In [20]:
```
myExp.showFinalModelReport(axis_labels)
```

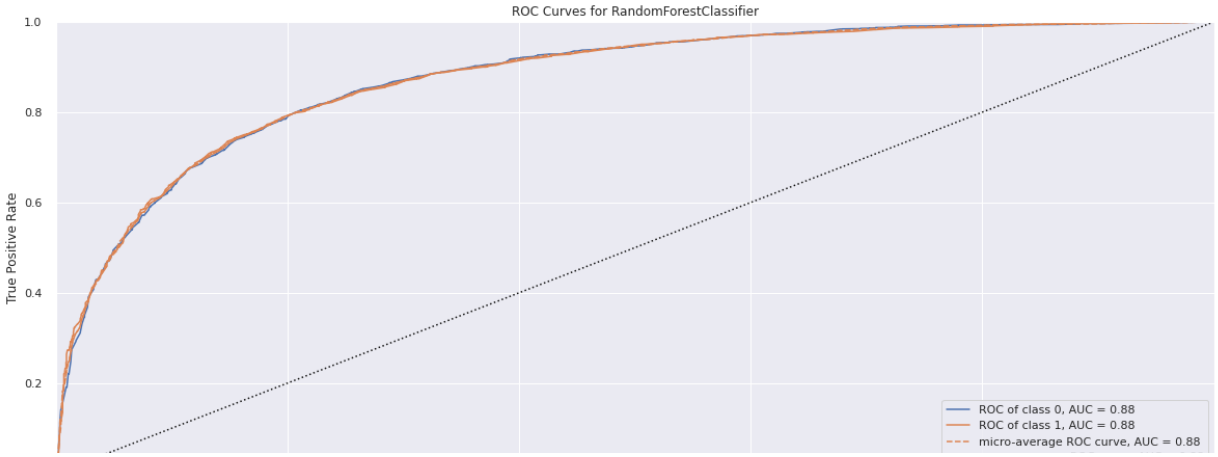
```
Final Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.59
              precision    recall  f1-score   support

           0       0.78      0.81      0.80      2688
           1       0.81      0.78      0.79      2688

    accuracy                           0.80      5376
   macro avg       0.80      0.80      0.80      5376
weighted avg       0.80      0.80      0.80      5376
```



<Figure size 1440x576 with 0 Axes>



Final model ROCAUC not calculated. Starting now



In [21]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Bert2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: True
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
```

```
---> isFinalModelLearningCurveCreated: True
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
```

Save Experiment

In [22]:
```
jarvis.saveExperiment(myExp, FILE_NAME)
```

```
[CV] END ..................., score=(train=0.978, test=0.806) total time=  2
9.3s
[CV] END ..................., score=(train=0.978, test=0.799) total time=
6.5s
[CV] END ..................., score=(train=0.979, test=0.807) total time= 1.1
min
[CV] END ..................., score=(train=0.979, test=0.793) total time=
6.7s
[CV] END ..................., score=(train=0.980, test=0.791) total time=
7.1s
[CV] END ..................., score=(train=0.980, test=0.787) total time=
9.5s
[CV] END ..................., score=(train=0.979, test=0.812) total time=  2
0.8s
[CV] END ..................., score=(train=0.979, test=0.789) total time=
1.1s
[CV] END ..................., score=(train=0.980, test=0.790) total time=
6.3s
[CV] END ..................., score=(train=0.980, test=0.790) total time=
2.4s
[CV] END ..................., score=(train=0.980, test=0.796) total time=  2
5.6s
[CV] END ..................., score=(train=0.979, test=0.792) total time=
2.3s
[CV] END ..................., score=(train=0.979, test=0.785) total time=
6.3s
[CV] END ..................., score=(train=0.977, test=0.810) total time= 1.0
min
[CV] END ..................., score=(train=0.980, test=0.793) total time=
2.3s
[CV] END ..................., score=(train=0.980, test=0.782) total time=
2.3s
[CV] END ..................., score=(train=0.979, test=0.808) total time=
5.3s
[CV] END ..................., score=(train=0.979, test=0.789) total time=
3.1s
[CV] END ..................., score=(train=0.982, test=0.776) total time=
```

```
4.3s
[CV] END ....................., score=(train=0.979, test=0.791) total time=  2
5.5s
[CV] END ....................., score=(train=0.977, test=0.810) total time=  1
2.4s
[CV] END ....................., score=(train=0.982, test=0.786) total time=
2.3s
[CV] END ....................., score=(train=0.978, test=0.805) total time= 1.1
min
[CV] END ....................., score=(train=0.979, test=0.800) total time=  1
2.7s
[CV] END ....................., score=(train=0.979, test=0.798) total time=  2
5.5s
[CV] END ....................., score=(train=0.980, test=0.794) total time=
1.2s
[CV] END ....................., score=(train=0.978, test=0.798) total time=  1
1.7s
[CV] END ....................., score=(train=0.979, test=0.797) total time=
8.1s
[CV] END ....................., score=(train=0.979, test=0.793) total time= 1.1
min
[CV] END ....................., score=(train=0.982, test=0.783) total time=
1.2s
[CV] END ....................., score=(train=0.982, test=0.775) total time=
1.0s
[CV] END ....................., score=(train=0.979, test=0.792) total time=  1
2.2s
[CV] END ....................., score=(train=0.980, test=0.787) total time=
8.1s
[CV] END ....................., score=(train=0.980, test=0.797) total time=
3.2s
[CV] END ....................., score=(train=0.980, test=0.807) total time=
8.9s
[CV] END ....................., score=(train=0.978, test=0.815) total time=  5
5.8s
[CV] END ....................., score=(train=0.980, test=0.787) total time=
2.4s
[CV] END ....................., score=(train=0.980, test=0.795) total time=
1.1s
[CV] END ....................., score=(train=0.980, test=0.800) total time=
2.2s
[CV] END ....................., score=(train=0.978, test=0.815) total time=
0.9s
```

Scratchpad

In [ ]:

Configuration

In [1]:
```
# Parameters
ENABLE_COLAB = False

PROJECT_NAME = 'ML1010-Group-Project'
EXPERIMENT_NAME = 'ReviewText_Lemma_Glove2 (Random Forest)'
FILE_NAME = '01_ML1010_GP_RF_Glove2'
LOAD_FROM_EXP = False

#Root Machine Learning Directory. Projects appear underneath
GOOGLE_DRIVE_MOUNT = '/content/gdrive'
COLAB_ROOT_DIR = GOOGLE_DRIVE_MOUNT + '/MyDrive/Colab Notebooks'
COLAB_INIT_DIR = COLAB_ROOT_DIR + '/utility_files'

LOCAL_ROOT_DIR = '/home/magni//ML_Root/project_root'
LOCAL_INIT_DIR = LOCAL_ROOT_DIR + '/utility_files'
```

Bootstrap Environment

In [2]:
```
#add in support for utility file directory and importing
import sys
import os

if ENABLE_COLAB:
  #Need access to drive
  from google.colab import drive
  drive.mount(GOOGLE_DRIVE_MOUNT, force_remount=True)

  #add in utility directory to syspath to import
  INIT_DIR = COLAB_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = COLAB_ROOT_DIR

else:
  #add in utility directory to syspath to import
  INIT_DIR = LOCAL_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = LOCAL_ROOT_DIR

#Import Utility Support
from jarvis import Jarvis
jarvis = Jarvis(ROOT_DIR, PROJECT_NAME)

import mv_python_utils as mvutils
```

```
Wha...where am I?
I am awake now.
```

```
I have set your current working directory to /home/magni/ML_Root/project_root
/ML1010-Group-Project
The current time is 10:30
Hello sir. Extra caffeine may help.
```

Setup Runtime Environment

In [3]:

```
if ENABLE_COLAB:
  #!pip install scipy -q
  #!pip install scikit-learn -q
  #!pip install pycaret -q
  #!pip install matplotlib -q
  #!pip install joblib -q
  #!pip install pandasql -q
  !pip install umap_learn -q
  !pip install sentence_transformers -q
  !pip install spacytextblob -q
  !pip install flair -q
  display('Google Colab enabled')
else:
  display('Google Colab not enabled')

#Common imports
import json
import pandas as pd
import numpy as np
import matplotlib
import re
import nltk
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn import metrics
from sklearn.datasets import load_digits
from sklearn.model_selection import train_test_split as tts
#from yellowbrick.classifier import ConfusionMatrix
#from sklearn.linear_model import LogisticRegression
from yellowbrick.target import ClassBalance
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier

nltk.download('stopwords')
%matplotlib inline
```

```
'Google Colab not enabled'
```

```
[nltk_data] Downloading package stopwords to /home/magni/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [4]:
```
import importlib
import cw_df_metric_utils as cwutils
import DataPackage as dp
import DataPackageSupport as dps
import DataExperiment
import DataExperimentSupport
```

```
2022-01-15 10:30:51.239984: W tensorflow/stream_executor/platform/default/dso
_loader.cc:64] Could not load dynamic library 'libcudart.so.11.0'; dlerror: l
ibcudart.so.11.0: cannot open shared object file: No such file or directory
2022-01-15 10:30:51.240010: I tensorflow/stream_executor/cuda/cudart_stub.cc:
29] Ignore above cudart dlerror if you do not have a GPU set up on your machi
ne.
```

In [24]:
```
importlib.reload(dp)
importlib.reload(dps)
importlib.reload(DataExperiment)
importlib.reload(DataExperimentSupport)
```

Out[24]:
```
<module 'DataExperimentSupport' from '/home/magni/ML_Root/project_root/utilit
y_files/DataExperimentSupport.py'>
```

Load Data

In [5]:
```
#axis_labels=[1,2,3,4,5]
axis_labels=[0,1]
classifier = RandomForestClassifier()
ANALSYSIS_COL = 'reviewText_lemma_glove'
UNIQUE_COL = 'uuid'
TARGET_COL = 'overall_posneg'
```

In [6]:
```
if LOAD_FROM_EXP:
    #start from saved state
    myExp = jarvis.loadExperiment(FILE_NAME)
    myExp.display()

else:
    #start from source file and regenerate
    testDf = pd.read_pickle(jarvis.DATA_DIR_WORK + "/01_NL_ReviewText_All(new

    testDfBert = cwutils.getBertEncodeFrame(df=testDf,
                                            bertColumn=ANALSYSIS_COL,
                                            uniqueColumn=UNIQUE_COL,
                                            otherColumns=[TARGET_COL]
                                            )



    myExp = DataExperiment.DataExperiment(projectName=PROJECT_NAME,
                                          experimentName=EXPERIMENT_NAME,
                                          origData=testDfBert,
                                          uniqueColumn=UNIQUE_COL,
                                          targetColumn=TARGET_COL,
                                          classifier=classifier)
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Glove2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: False
---> isBaseModelPredicted: False
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: False
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: False
    ---> isTrainTestSplit: False
    Data:
    ---> isOrigDataLoaded: True
    ---> isTrainDataLoaded: False
    ---> isTestDataLoaded: False
```

In [7]:
```
myExp.processDataPackage()
```


```
Undersampling data to match min class: 0 of size: 13440
```


```
Completed train/test split (test_size = 0.2):
---> Original data size: 26880
---> Training data size: 21504
---> Testing data size: 5376
---> Stratified on column: overall_posneg
```

In [8]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Glove2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: False
---> isBaseModelPredicted: False
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: False
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
```

```
---> targetColumn: overall_posneg
Process:
---> isBalanced: True
---> isTrainTestSplit: True
Data:
---> isOrigDataLoaded: False
---> isTrainDataLoaded: True
---> isTestDataLoaded: True
```

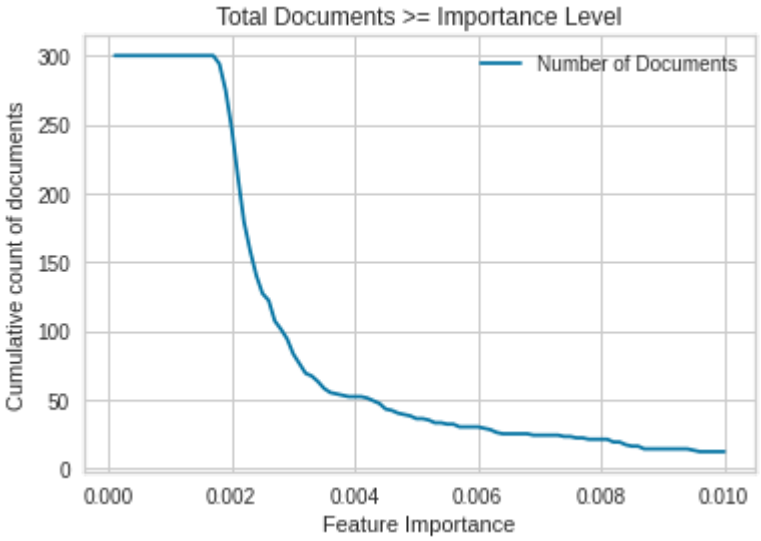
In [9]:
```
myExp.createBaseModel()
```

In [10]:
```
myExp.predictBaseModel()
```

```
Base Model Stats:
Accuracy: 0.8
Precision: 0.81
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.61
```

In [11]:
```
impFeatures = myExp.analyzeBaseModelFeatureImportance(returnAbove=0.002)
```

```
  0%|          | 0/101 [00:00<?, ?it/s]
Feature Importance Summary:
---> Original feature count: 300
---> Returned feature count: 247
---> Removed feature count: 53
---> Return items above (including): 0.002
```



In [12]:
```
myExp.createFinalModel(featureImportanceThreshold=0.002)
```

```
  0%|          | 0/101 [00:00<?, ?it/s]
  0%|          | 0/101 [00:00<?, ?it/s]
```

In [13]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
```

```
---> experimentName: ReviewText_Lemma_Glove2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: True
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

In [14]:

```
myExp.predictFinalModel()
myExp.display()
```

```
Final Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.6
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Glove2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

In [15]:
```
myExp.createBaseModelLearningCurve(n_jobs=10)
```

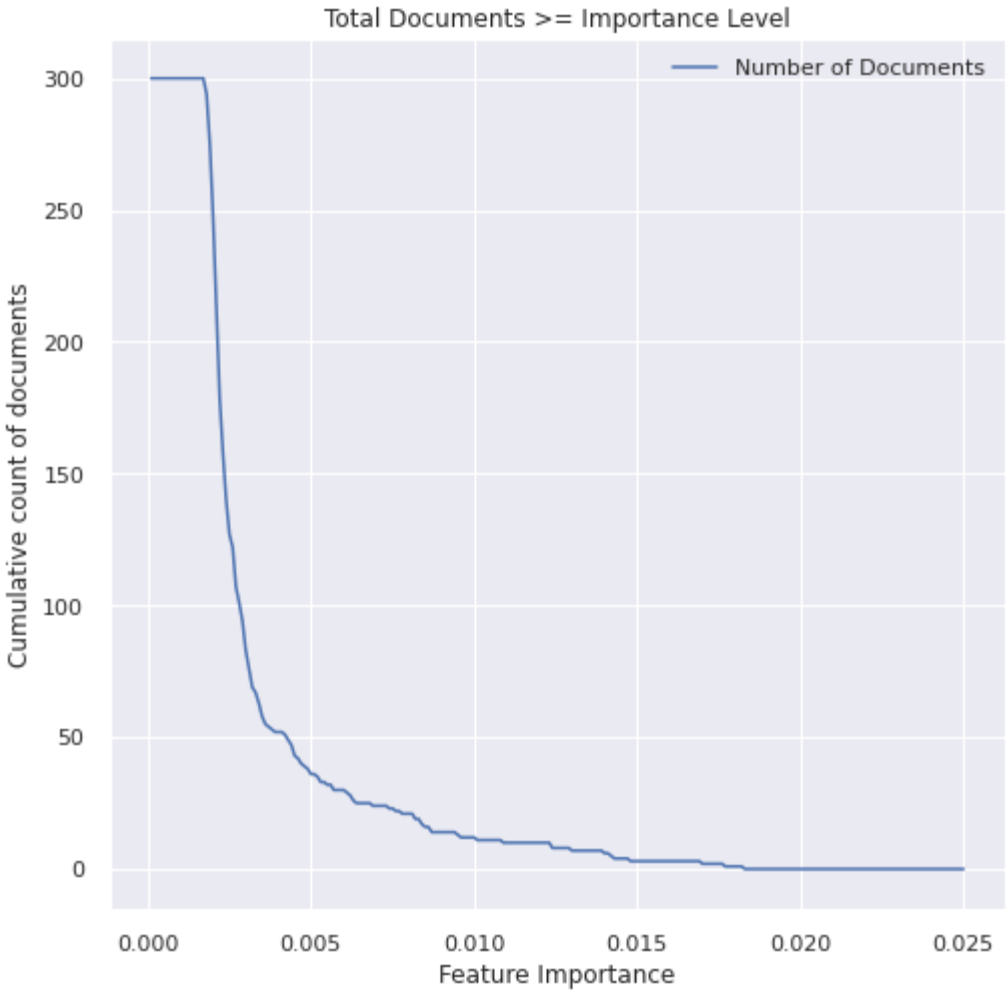
```
[learning_curve] Training set sizes: [ 1720  3440  8601 17203]
[Parallel(n_jobs=10)]: Using backend LokyBackend with 10 concurrent workers.
[Parallel(n_jobs=10)]: Done   3 out of  20 | elapsed:    4.2s remaining:   2
3.9s
[Parallel(n_jobs=10)]: Done   9 out of  20 | elapsed:   10.5s remaining:   1
2.8s
[Parallel(n_jobs=10)]: Done  15 out of  20 | elapsed:   23.6s remaining:
7.9s
[Parallel(n_jobs=10)]: Done  20 out of  20 | elapsed:   38.6s finished
```

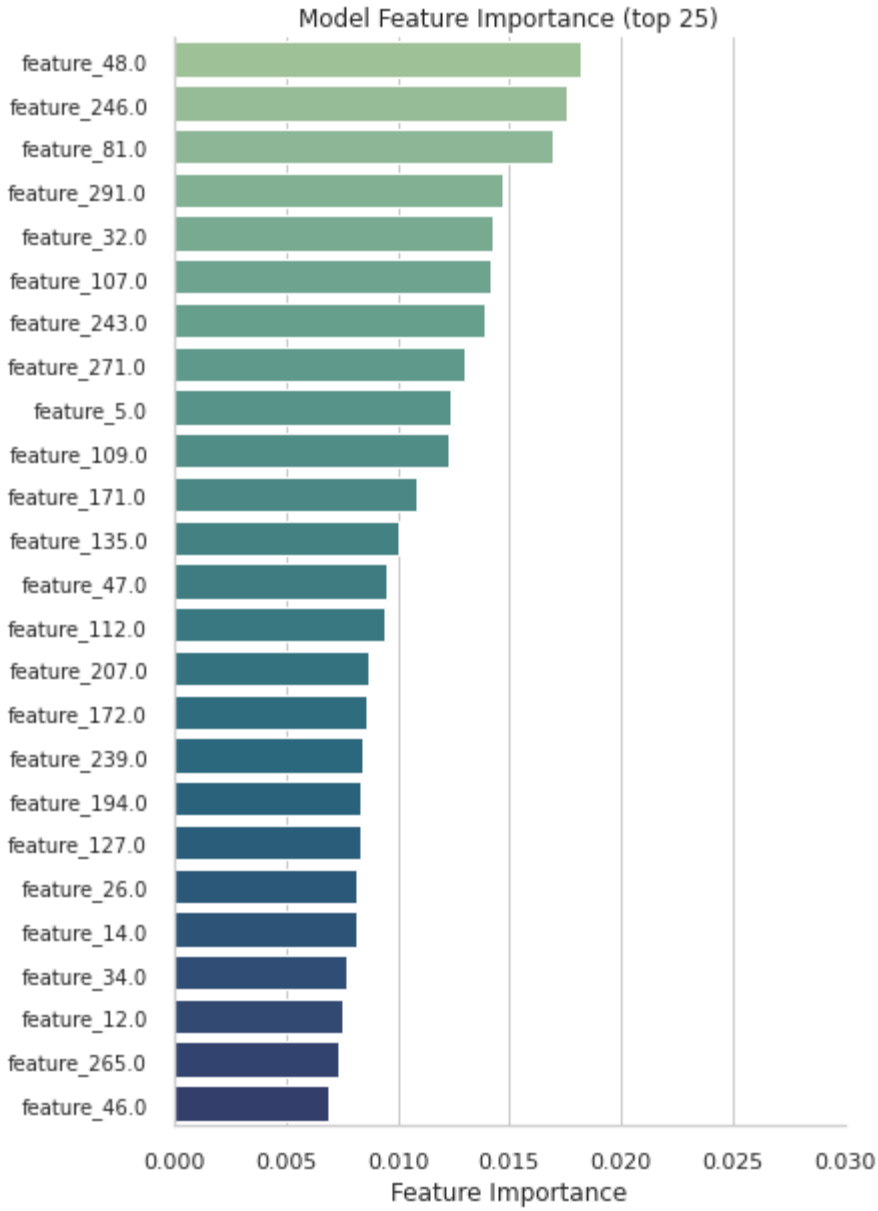
In [16]:
```
myExp.createFinalModelLearningCurve(n_jobs=10)
```

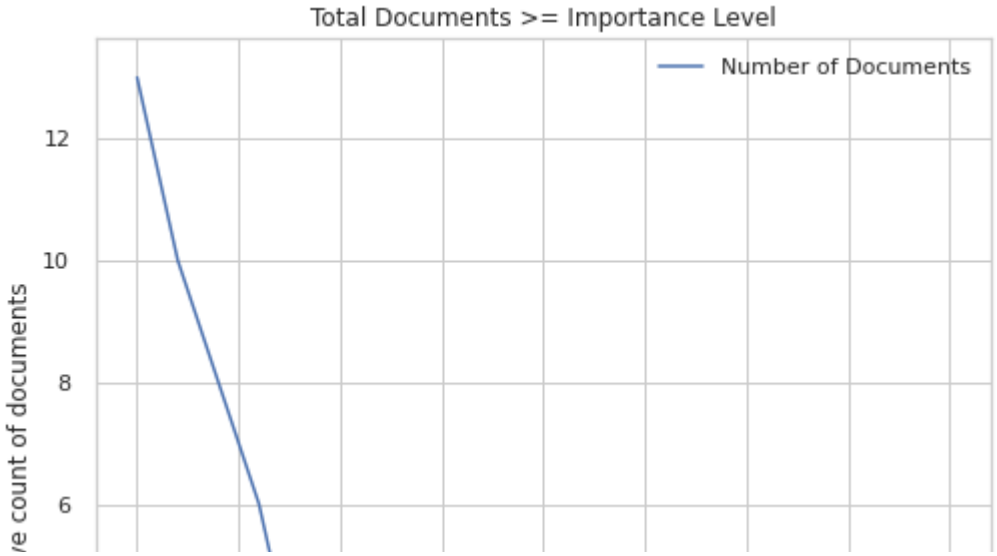
```
[Parallel(n_jobs=10)]: Using backend LokyBackend with 10 concurrent workers.
[learning_curve] Training set sizes: [ 1720  3440  8601 17203]
[Parallel(n_jobs=10)]: Done   3 out of  20 | elapsed:    3.1s remaining:   1
7.5s
[Parallel(n_jobs=10)]: Done   9 out of  20 | elapsed:   10.7s remaining:   1
3.0s
[Parallel(n_jobs=10)]: Done  15 out of  20 | elapsed:   25.8s remaining:
8.6s
[Parallel(n_jobs=10)]: Done  20 out of  20 | elapsed:   41.1s finished
```

In [27]:
```
myExp.showBaseModelFeatureImportance(upperValue=0.025)
myExp.showFinalModelFeatureImportance(startValue=0.01,
                                      increment=0.001,
                                      upperValue=0.03)
```

```
  0%|          | 0/251 [00:00<?, ?it/s]
  0%|          | 0/22 [00:00<?, ?it/s]
```


Total Documents >= Importance Level
Number of Documents
Cumulative count of documents
300
250
200
150
100
50
0
0.000
0.005
0.010
0.015
0.020
0.025
Feature Importance





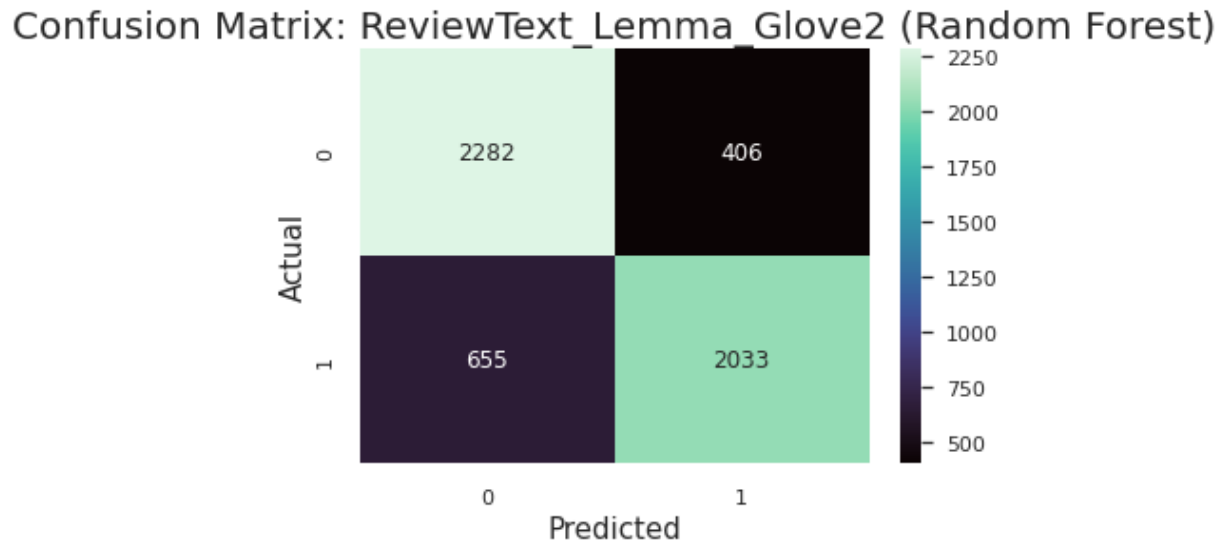
In [18]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Glove2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: True
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: True
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

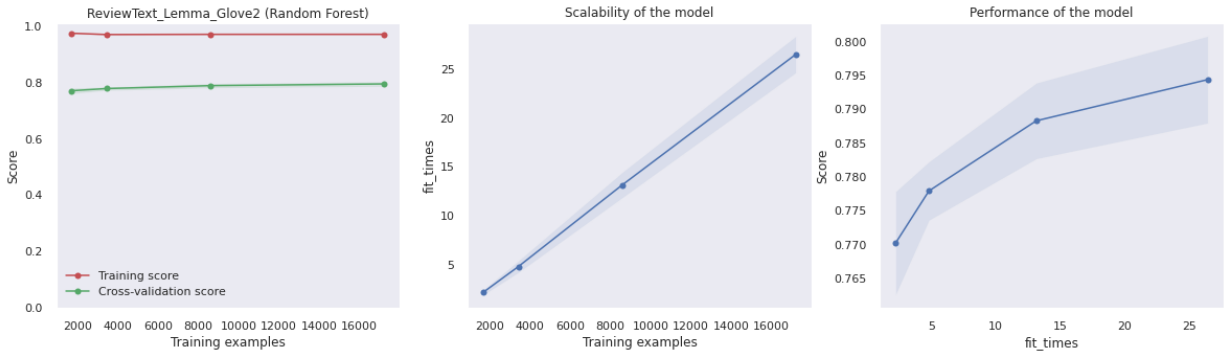
In [19]:

```
myExp.showBaseModelReport(axis_labels)
```

```
Base Model Stats:
Accuracy: 0.8
Precision: 0.81
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.61
              precision    recall  f1-score   support

           0       0.78      0.85      0.81      2688
           1       0.83      0.76      0.79      2688

    accuracy                           0.80      5376
   macro avg       0.81      0.80      0.80      5376
weighted avg       0.81      0.80      0.80      5376
```



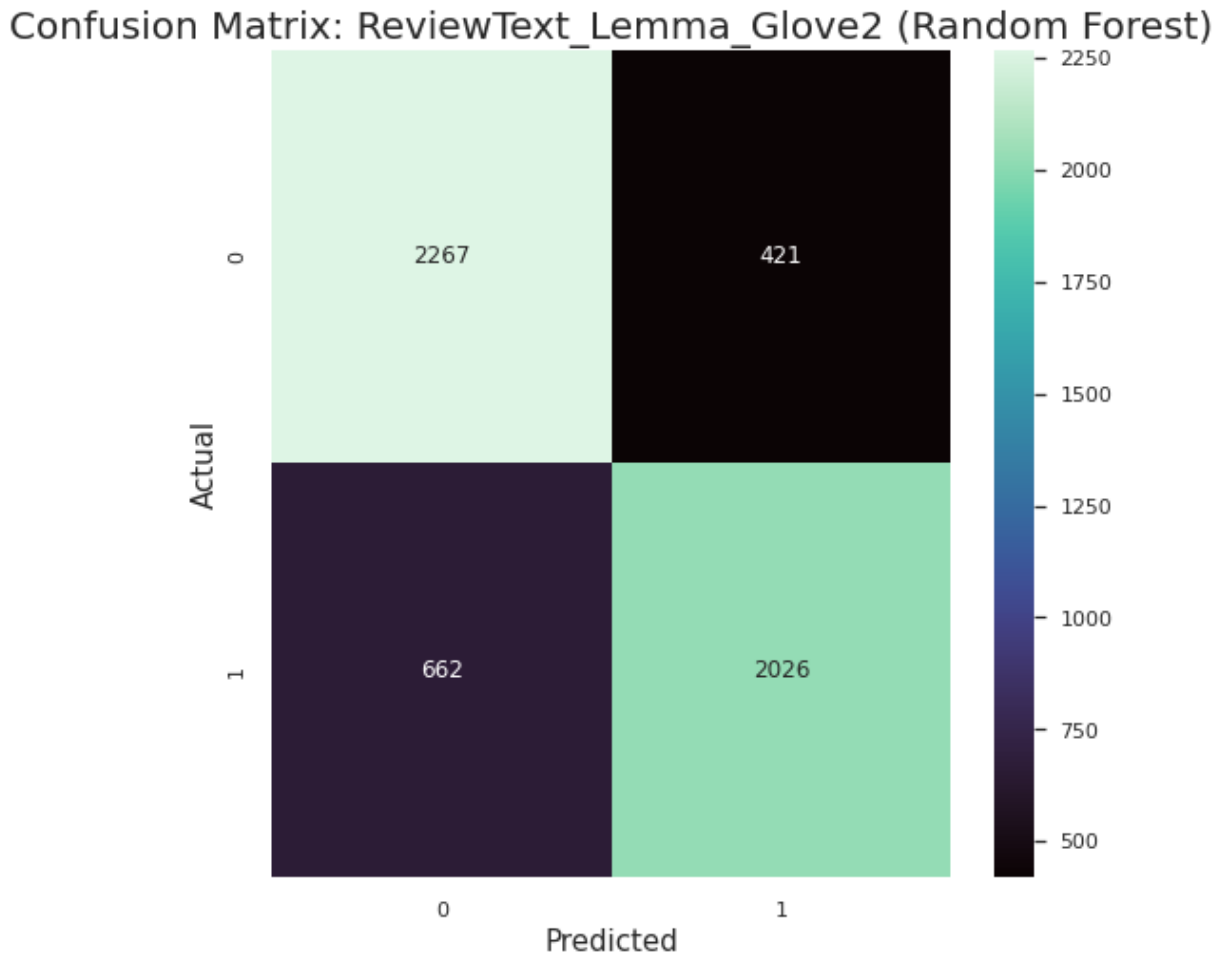
<Figure size 1440x576 with 0 Axes>



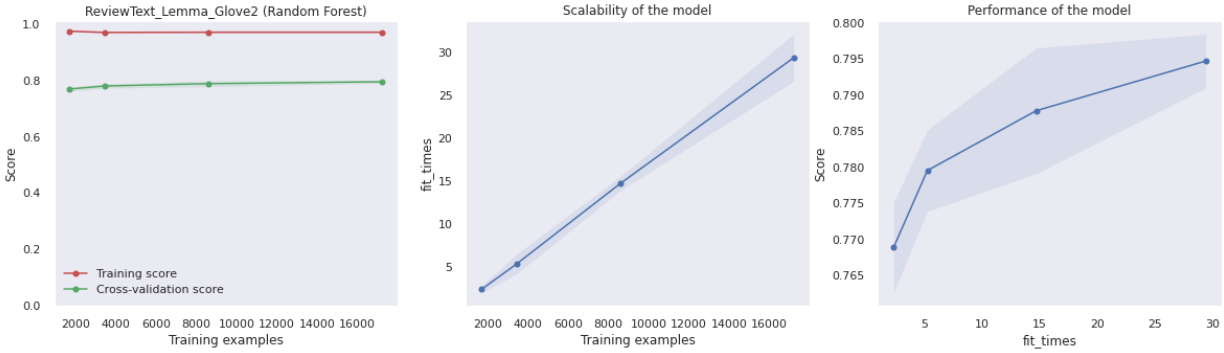
Base model ROCAUC not calculated. Starting now

In [26]:
```
myExp.showFinalModelReport(axis_labels)
```

```
Final Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.6
              precision    recall  f1-score   support

           0       0.77      0.84      0.81      2688
           1       0.83      0.75      0.79      2688

    accuracy                           0.80      5376
   macro avg       0.80      0.80      0.80      5376
weighted avg       0.80      0.80      0.80      5376
```



<Figure size 576x576 with 0 Axes>



Final model ROCAUC not calculated. Starting now

In [21]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Glove2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: True
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: True
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

Save Experiment

In [22]:
```
jarvis.saveExperiment(myExp, FILE_NAME)
```

```
[CV] END ....................., score=(train=0.969, test=0.773) total time=
4.6s
[CV] END ....................., score=(train=0.969, test=0.785) total time=
4.2s
[CV] END ....................., score=(train=0.969, test=0.785) total time=  1
2.8s
[CV] END ....................., score=(train=0.976, test=0.781) total time=
2.9s
[CV] END ....................., score=(train=0.974, test=0.762) total time=
1.9s
[CV] END ....................., score=(train=0.973, test=0.766) total time=
2.9s
[CV] END ....................., score=(train=0.970, test=0.777) total time=
7.5s
[CV] END ....................., score=(train=0.970, test=0.804) total time=  2
8.4s
[CV] END ....................., score=(train=0.971, test=0.801) total time=  1
```

```
6.6s
[CV] END ....................., score=(train=0.971, test=0.785) total time=  1
2.7s
[CV] END ....................., score=(train=0.974, test=0.767) total time=
2.3s
[CV] END ....................., score=(train=0.970, test=0.794) total time=  1
4.7s
[CV] END ....................., score=(train=0.976, test=0.775) total time=
2.7s
[CV] END ....................., score=(train=0.974, test=0.758) total time=
2.3s
[CV] END ....................., score=(train=0.973, test=0.765) total time=
2.1s
[CV] END ....................., score=(train=0.970, test=0.777) total time=
6.0s
[CV] END ....................., score=(train=0.969, test=0.773) total time=
4.6s
[CV] END ....................., score=(train=0.972, test=0.783) total time=  1
4.7s
[CV] END ....................., score=(train=0.971, test=0.794) total time=  1
1.9s
[CV] END ....................., score=(train=0.970, test=0.790) total time=  2
4.3s
[CV] END ....................., score=(train=0.969, test=0.784) total time=
4.4s
[CV] END ....................., score=(train=0.969, test=0.775) total time=
5.8s
[CV] END ....................., score=(train=0.969, test=0.786) total time=  1
4.7s
[CV] END ....................., score=(train=0.969, test=0.774) total time=
4.8s
[CV] END ....................., score=(train=0.972, test=0.781) total time=  1
5.7s
[CV] END ....................., score=(train=0.971, test=0.793) total time=  3
1.2s
[CV] END ....................., score=(train=0.971, test=0.786) total time=  2
9.2s
[CV] END ....................., score=(train=0.970, test=0.797) total time=  3
1.6s
[CV] END ....................., score=(train=0.974, test=0.772) total time=
2.6s
[CV] END ....................., score=(train=0.970, test=0.798) total time=  2
5.8s
[CV] END ....................., score=(train=0.972, test=0.788) total time=
4.8s
[CV] END ....................., score=(train=0.972, test=0.796) total time=  2
7.1s
[CV] END ....................., score=(train=0.972, test=0.780) total time=
5.1s
[CV] END ....................., score=(train=0.972, test=0.793) total time=  2
5.3s
[CV] END ....................., score=(train=0.974, test=0.769) total time=
2.5s
[CV] END ....................., score=(train=0.970, test=0.799) total time=  3
2.4s
[CV] END ....................., score=(train=0.974, test=0.780) total time=
2.0s
[CV] END ....................., score=(train=0.970, test=0.796) total time=  1
3.5s
```

```
[CV] END ....................., score=(train=0.971, test=0.775) total time=  1
4.2s
[CV] END ....................., score=(train=0.970, test=0.788) total time=  2
```

Scratchpad

In [ ]:

Configuration

In [1]:
```
# Parameters
ENABLE_COLAB = False

PROJECT_NAME = 'ML1010-Group-Project'
EXPERIMENT_NAME = 'ReviewText_Lemma_MPNet2 (Random Forest)'
FILE_NAME = '01_ML1010_GP_RF_MPNet2'
LOAD_FROM_EXP = False

#Root Machine Learning Directory. Projects appear underneath
GOOGLE_DRIVE_MOUNT = '/content/gdrive'
COLAB_ROOT_DIR = GOOGLE_DRIVE_MOUNT + '/MyDrive/Colab Notebooks'
COLAB_INIT_DIR = COLAB_ROOT_DIR + '/utility_files'

LOCAL_ROOT_DIR = '/home/magni//ML_Root/project_root'
LOCAL_INIT_DIR = LOCAL_ROOT_DIR + '/utility_files'
```

Bootstrap Environment

In [2]:
```
#add in support for utility file directory and importing
import sys
import os

if ENABLE_COLAB:
  #Need access to drive
  from google.colab import drive
  drive.mount(GOOGLE_DRIVE_MOUNT, force_remount=True)

  #add in utility directory to syspath to import
  INIT_DIR = COLAB_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = COLAB_ROOT_DIR

else:
  #add in utility directory to syspath to import
  INIT_DIR = LOCAL_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = LOCAL_ROOT_DIR

#Import Utility Support
from jarvis import Jarvis
jarvis = Jarvis(ROOT_DIR, PROJECT_NAME)

import mv_python_utils as mvutils
```

```
Wha...where am I?
I am awake now.
```

```
I have set your current working directory to /home/magni/ML_Root/project_root
/ML1010-Group-Project
The current time is 10:30
Hello sir. Extra caffeine may help.
```

Setup Runtime Environment

In [3]:

```
if ENABLE_COLAB:
  #!pip install scipy -q
  #!pip install scikit-learn -q
  #!pip install pycaret -q
  #!pip install matplotlib -q
  #!pip install joblib -q
  #!pip install pandasql -q
  !pip install umap_learn -q
  !pip install sentence_transformers -q
  !pip install spacytextblob -q
  !pip install flair -q
  display('Google Colab enabled')
else:
  display('Google Colab not enabled')

#Common imports
import json
import pandas as pd
import numpy as np
import matplotlib
import re
import nltk
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn import metrics
from sklearn.datasets import load_digits
from sklearn.model_selection import train_test_split as tts
#from yellowbrick.classifier import ConfusionMatrix
#from sklearn.linear_model import LogisticRegression
from yellowbrick.target import ClassBalance
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier

nltk.download('stopwords')
%matplotlib inline
```

```
'Google Colab not enabled'
```

```
[nltk_data] Downloading package stopwords to /home/magni/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [4]:
```
import importlib
import cw_df_metric_utils as cwutils
import DataPackage as dp
import DataPackageSupport as dps
import DataExperiment
import DataExperimentSupport
```

```
2022-01-15 10:31:01.507502: W tensorflow/stream_executor/platform/default/dso
_loader.cc:64] Could not load dynamic library 'libcudart.so.11.0'; dlerror: l
ibcudart.so.11.0: cannot open shared object file: No such file or directory
2022-01-15 10:31:01.507529: I tensorflow/stream_executor/cuda/cudart_stub.cc:
29] Ignore above cudart dlerror if you do not have a GPU set up on your machi
ne.
```

In [26]:
```
importlib.reload(dp)
importlib.reload(dps)
importlib.reload(DataExperiment)
importlib.reload(DataExperimentSupport)
```

Out[26]:
```
<module 'DataExperimentSupport' from '/home/magni/ML_Root/project_root/utilit
y_files/DataExperimentSupport.py'>
```

Load Data

In [5]:
```
#axis_labels=[1,2,3,4,5]
axis_labels=[0,1]
classifier = RandomForestClassifier()
ANALSYSIS_COL = 'reviewText_lemma_mpnet'
UNIQUE_COL = 'uuid'
TARGET_COL = 'overall_posneg'
```

In [6]:
```
if LOAD_FROM_EXP:
    #start from saved state
    myExp = jarvis.loadExperiment(FILE_NAME)
    myExp.display()

else:
    #start from source file and regenerate
    testDf = pd.read_pickle(jarvis.DATA_DIR_WORK + "/01_NL_ReviewText_All(new

    testDfBert = cwutils.getBertEncodeFrame(df=testDf,
                                            bertColumn=ANALSYSIS_COL,
                                            uniqueColumn=UNIQUE_COL,
                                            otherColumns=[TARGET_COL]
                                            )



    myExp = DataExperiment.DataExperiment(projectName=PROJECT_NAME,
                                          experimentName=EXPERIMENT_NAME,
                                          origData=testDfBert,
                                          uniqueColumn=UNIQUE_COL,
                                          targetColumn=TARGET_COL,
                                          classifier=classifier)
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_MPNet2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: False
---> isBaseModelPredicted: False
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: False
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: False
    ---> isTrainTestSplit: False
    Data:
    ---> isOrigDataLoaded: True
    ---> isTrainDataLoaded: False
    ---> isTestDataLoaded: False
```

In [7]:
```
myExp.processDataPackage()
```


```
Undersampling data to match min class: 0 of size: 13440
```


```
Completed train/test split (test_size = 0.2):
---> Original data size: 26880
---> Training data size: 21504
---> Testing data size: 5376
---> Stratified on column: overall_posneg
```

In [8]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_MPNet2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: False
---> isBaseModelPredicted: False
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: False
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
```

```
---> targetColumn: overall_posneg
Process:
---> isBalanced: True
---> isTrainTestSplit: True
Data:
---> isOrigDataLoaded: False
---> isTrainDataLoaded: True
---> isTestDataLoaded: True
```

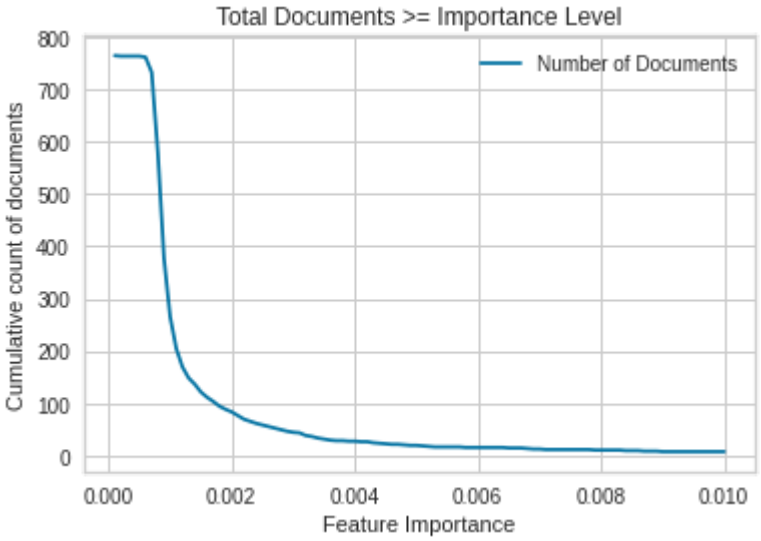
In [9]:
```
myExp.createBaseModel()
```

In [10]:
```
myExp.predictBaseModel()
```

```
Base Model Stats:
Accuracy: 0.81
Precision: 0.81
Recalll: 0.81
F1 Score: 0.81
Cohen kappa:: 0.62
```

In [11]:
```
impFeatures = myExp.analyzeBaseModelFeatureImportance(returnAbove=0.002)
```

```
  0%|          | 0/101 [00:00<?, ?it/s]
Feature Importance Summary:
---> Original feature count: 768
---> Returned feature count: 83
---> Removed feature count: 685
---> Return items above (including): 0.002
```



In [12]:
```
myExp.createFinalModel(featureImportanceThreshold=0.002)
```

```
  0%|          | 0/101 [00:00<?, ?it/s]
  0%|          | 0/101 [00:00<?, ?it/s]
```

In [13]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
```

```
---> experimentName: ReviewText_Lemma_MPNet2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: True
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

In [14]:

```
myExp.predictFinalModel()
myExp.display()
```

```
Final Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.61
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_MPNet2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

In [15]:
```
myExp.createBaseModelLearningCurve(n_jobs=10)
```

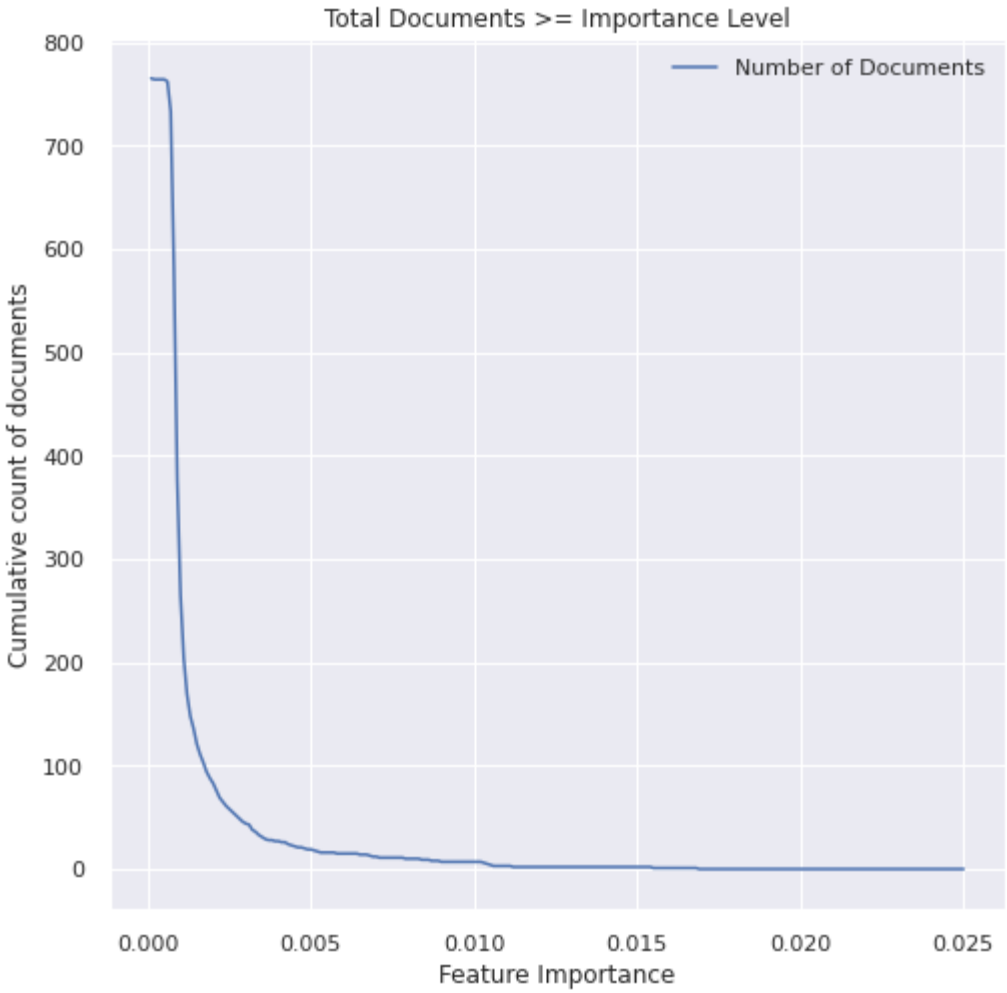
```
[learning_curve] Training set sizes: [ 1720  3440  8601 17203]
[Parallel(n_jobs=10)]: Using backend LokyBackend with 10 concurrent workers.
[Parallel(n_jobs=10)]: Done   3 out of  20 | elapsed:    6.2s remaining:   3
5.0s
[Parallel(n_jobs=10)]: Done   9 out of  20 | elapsed:   21.3s remaining:   2
6.0s
[Parallel(n_jobs=10)]: Done  15 out of  20 | elapsed:   42.0s remaining:   1
4.0s
[Parallel(n_jobs=10)]: Done  20 out of  20 | elapsed:  1.1min finished
```

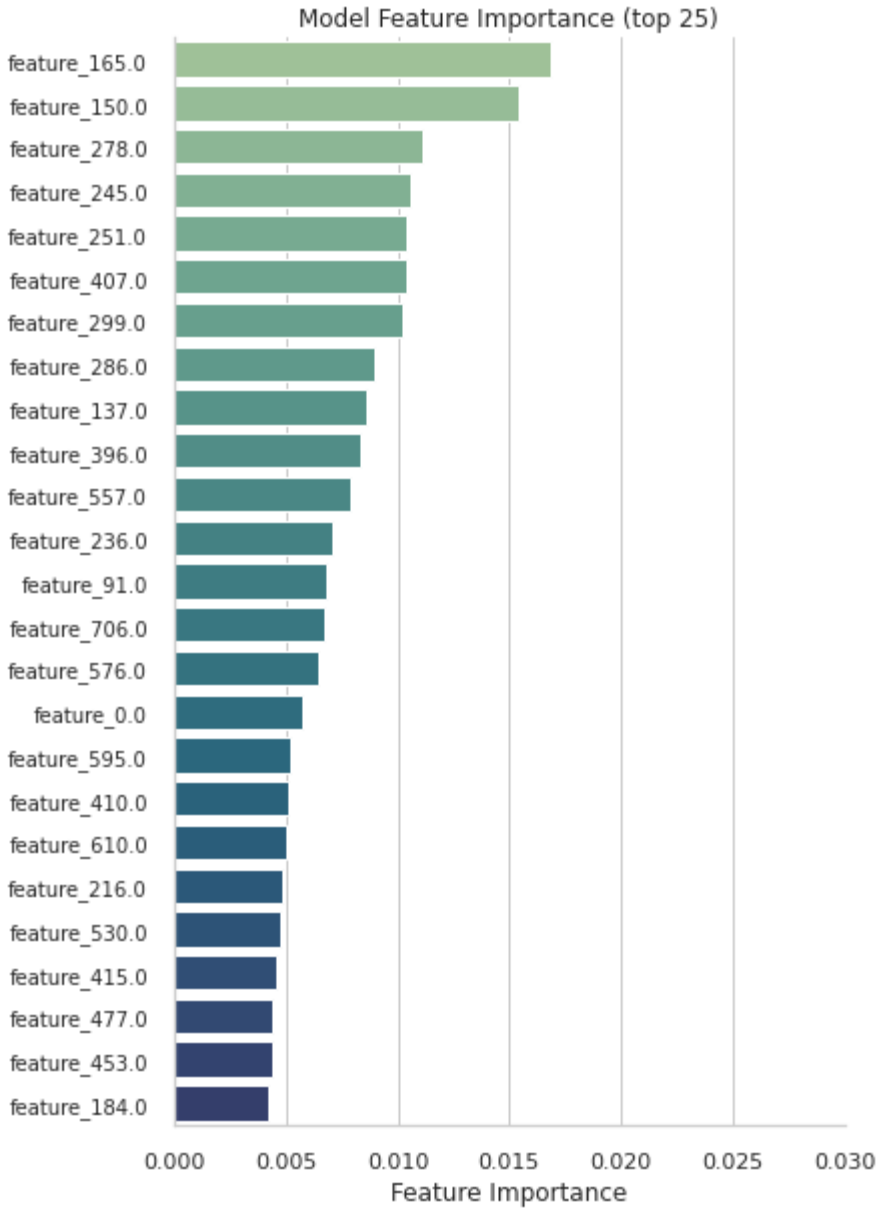
In [16]:
```
myExp.createFinalModelLearningCurve(n_jobs=10)
```

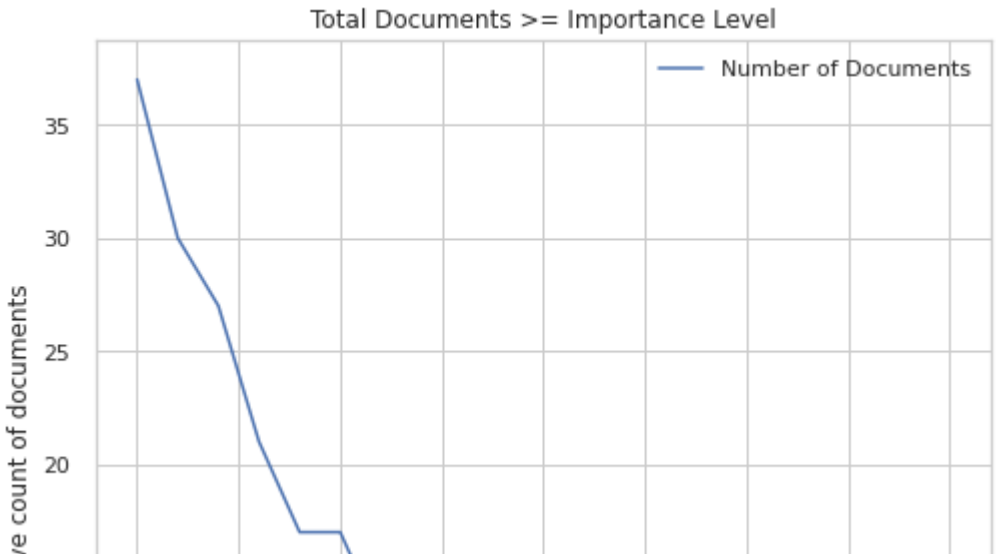
```
[Parallel(n_jobs=10)]: Using backend LokyBackend with 10 concurrent workers.
[learning_curve] Training set sizes: [ 1720  3440  8601 17203]
[Parallel(n_jobs=10)]: Done   3 out of  20 | elapsed:    1.1s remaining:
6.0s
[Parallel(n_jobs=10)]: Done   9 out of  20 | elapsed:    4.1s remaining:
5.0s
[Parallel(n_jobs=10)]: Done  15 out of  20 | elapsed:    9.5s remaining:
3.2s
[Parallel(n_jobs=10)]: Done  20 out of  20 | elapsed:   14.4s finished
```

In [29]:
```
myExp.showBaseModelFeatureImportance(upperValue=0.025)
myExp.showFinalModelFeatureImportance(startValue=0.01,
                                      increment=0.001,
                                      upperValue=0.03)
```

```
  0%|          | 0/251 [00:00<?, ?it/s]
  0%|          | 0/22 [00:00<?, ?it/s]
```


Total Documents >= Importance Level
Number of Documents
Cumulative count of documents
800
700
600
500
400
300
200
100
0
0.000
0.005
0.010
0.015
0.020
0.025
Feature Importance





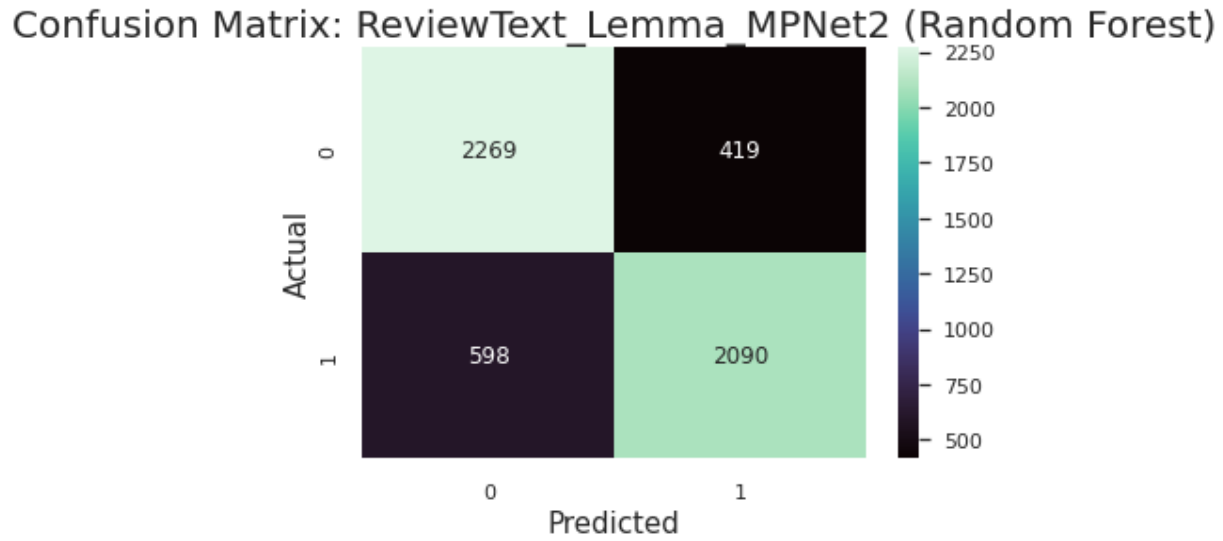
In [18]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_MPNet2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: True
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: True
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

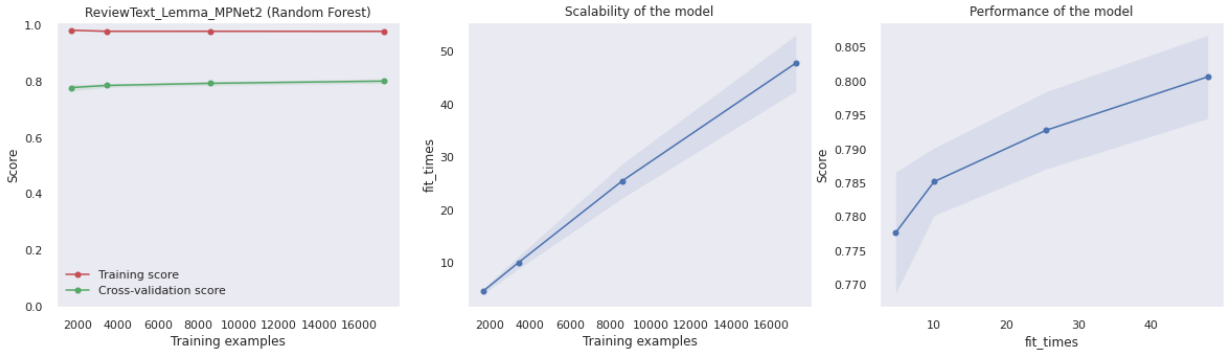
In [19]:

```
myExp.showBaseModelReport(axis_labels)
```

```
Base Model Stats:
Accuracy: 0.81
Precision: 0.81
Recalll: 0.81
F1 Score: 0.81
Cohen kappa:: 0.62
              precision    recall  f1-score   support

           0       0.79      0.84      0.82      2688
           1       0.83      0.78      0.80      2688

    accuracy                           0.81      5376
   macro avg       0.81      0.81      0.81      5376
weighted avg       0.81      0.81      0.81      5376
```



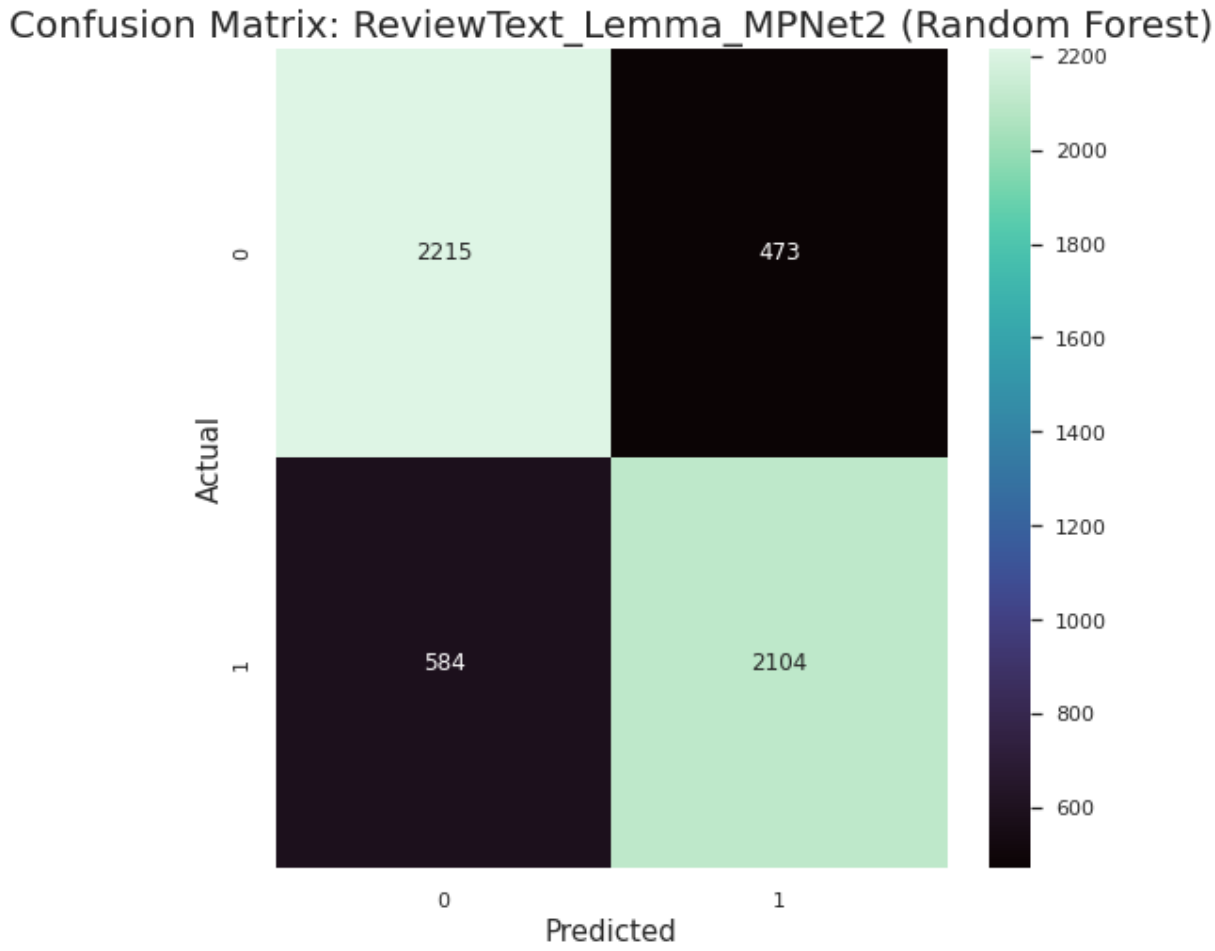
<Figure size 1440x576 with 0 Axes>



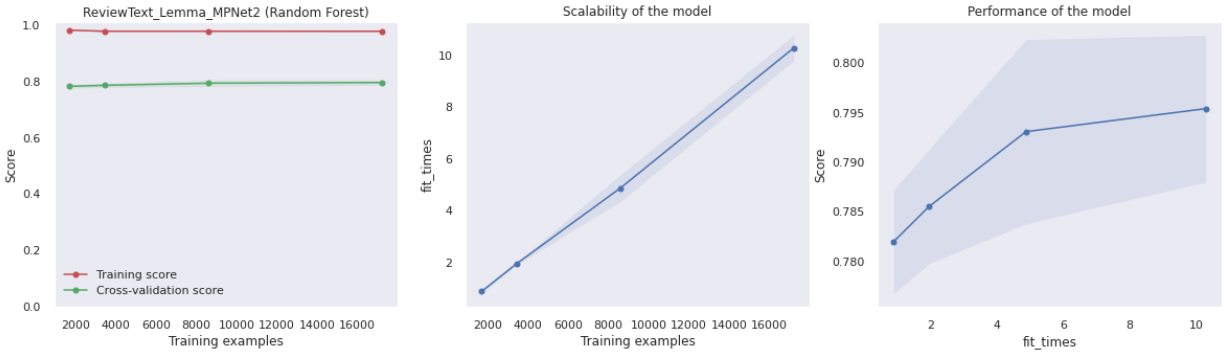
Base model ROCAUC not calculated. Starting now

In [28]:
```
myExp.showFinalModelReport(axis_labels)
```

```
Final Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.61
              precision    recall  f1-score   support

           0       0.79      0.82      0.81      2688
           1       0.82      0.78      0.80      2688

    accuracy                           0.80      5376
   macro avg       0.80      0.80      0.80      5376
weighted avg       0.80      0.80      0.80      5376
```



<Figure size 576x576 with 0 Axes>



Final model ROCAUC not calculated. Starting now

In [21]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_MPNet2 (Random Forest)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: True
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: True
---> isClassifierLoaded: True
RandomForestClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

Save Experiment

In [22]:
```
jarvis.saveExperiment(myExp, FILE_NAME)
```

```
[CV] END ....................., score=(train=0.978, test=0.807) total time=  5
4.0s
[CV] END ....................., score=(train=0.978, test=0.777) total time=
4.4s
[CV] END ....................., score=(train=0.978, test=0.793) total time=  1
1.7s
[CV] END ....................., score=(train=0.980, test=0.769) total time=
4.3s
[CV] END ....................., score=(train=0.978, test=0.784) total time=
8.5s
[CV] END ....................., score=(train=0.977, test=0.801) total time=  3
8.3s
[CV] END ....................., score=(train=0.978, test=0.790) total time=
2.1s
[CV] END ....................., score=(train=0.980, test=0.784) total time=
1.0s
[CV] END ....................., score=(train=0.978, test=0.783) total time=
2.0s
```

```
[CV] END ....................., score=(train=0.977, test=0.779) total time=  1
0.9s
[CV] END ....................., score=(train=0.978, test=0.782) total time=
9.1s
[CV] END ....................., score=(train=0.977, test=0.790) total time=  2
0.3s
[CV] END ....................., score=(train=0.981, test=0.774) total time=
0.7s
[CV] END ....................., score=(train=0.978, test=0.803) total time=
5.3s
[CV] END ....................., score=(train=0.978, test=0.783) total time=  2
8.9s
[CV] END ....................., score=(train=0.977, test=0.794) total time=
2.0s
[CV] END ....................., score=(train=0.979, test=0.796) total time=
5.5s
[CV] END ....................., score=(train=0.977, test=0.793) total time=  5
0.5s
[CV] END ....................., score=(train=0.977, test=0.777) total time=
1.9s
[CV] END ....................., score=(train=0.978, test=0.784) total time=
2.0s
[CV] END ....................., score=(train=0.977, test=0.789) total time=
5.2s
[CV] END ....................., score=(train=0.983, test=0.785) total time=
4.9s
[CV] END ....................., score=(train=0.977, test=0.808) total time=  4
9.8s
[CV] END ....................., score=(train=0.977, test=0.785) total time=  1
0.0s
[CV] END ....................., score=(train=0.981, test=0.769) total time=
5.1s
[CV] END ....................., score=(train=0.981, test=0.774) total time=
5.8s
[CV] END ....................., score=(train=0.979, test=0.793) total time=  2
3.5s
[CV] END ....................., score=(train=0.978, test=0.804) total time=  1
0.3s
[CV] END ....................., score=(train=0.977, test=0.798) total time=  2
8.1s
[CV] END ....................., score=(train=0.981, test=0.789) total time=
0.9s
[CV] END ....................., score=(train=0.977, test=0.801) total time=  1
1.2s
[CV] END ....................., score=(train=0.977, test=0.788) total time=  1
1.1s
[CV] END ....................., score=(train=0.978, test=0.794) total time=  4
6.8s
[CV] END ....................., score=(train=0.983, test=0.785) total time=
1.0s
[CV] END ....................., score=(train=0.981, test=0.778) total time=
1.0s
[CV] END ....................., score=(train=0.978, test=0.789) total time=  1
0.6s
[CV] END ....................., score=(train=0.981, test=0.791) total time=
4.4s
[CV] END ....................., score=(train=0.978, test=0.799) total time=  2
7.5s
[CV] END ....................., score=(train=0.977, test=0.801) total time=
```

```
4.2s
[CV] END ....................., score=(train=0.977, test=0.798) total time=
0.7s
```

Scratchpad

In [ ]:

Configuration

In [1]:
```
# Parameters
ENABLE_COLAB = False

PROJECT_NAME = 'ML1010-Group-Project'
EXPERIMENT_NAME = 'ReviewText_Lemma_Bert2 (XGB)'
FILE_NAME = '01_ML1010_GP_XGB_Bert2'
LOAD_FROM_EXP = False

#Root Machine Learning Directory. Projects appear underneath
GOOGLE_DRIVE_MOUNT = '/content/gdrive'
COLAB_ROOT_DIR = GOOGLE_DRIVE_MOUNT + '/MyDrive/Colab Notebooks'
COLAB_INIT_DIR = COLAB_ROOT_DIR + '/utility_files'

LOCAL_ROOT_DIR = '/home/magni//ML_Root/project_root'
LOCAL_INIT_DIR = LOCAL_ROOT_DIR + '/utility_files'
```

Bootstrap Environment

In [2]:
```
#add in support for utility file directory and importing
import sys
import os

if ENABLE_COLAB:
  #Need access to drive
  from google.colab import drive
  drive.mount(GOOGLE_DRIVE_MOUNT, force_remount=True)

  #add in utility directory to syspath to import
  INIT_DIR = COLAB_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = COLAB_ROOT_DIR

else:
  #add in utility directory to syspath to import
  INIT_DIR = LOCAL_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = LOCAL_ROOT_DIR

#Import Utility Support
from jarvis import Jarvis
jarvis = Jarvis(ROOT_DIR, PROJECT_NAME)

import mv_python_utils as mvutils
```

```
Wha...where am I?
I am awake now.
```

```
I have set your current working directory to /home/magni/ML_Root/project_root
/ML1010-Group-Project
The current time is 10:36
Hello sir. Extra caffeine may help.
```

Setup Runtime Environment

In [3]:

```
if ENABLE_COLAB:
  #!pip install scipy -q
  #!pip install scikit-learn -q
  #!pip install pycaret -q
  #!pip install matplotlib -q
  #!pip install joblib -q
  #!pip install pandasql -q
  !pip install umap_learn -q
  !pip install sentence_transformers -q
  !pip install spacytextblob -q
  !pip install flair -q
  display('Google Colab enabled')
else:
  display('Google Colab not enabled')

#Common imports
import json
import pandas as pd
import numpy as np
import matplotlib
import re
import nltk
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn import metrics
from sklearn.datasets import load_digits
from sklearn.model_selection import train_test_split as tts
#from yellowbrick.classifier import ConfusionMatrix
#from sklearn.linear_model import LogisticRegression
from yellowbrick.target import ClassBalance
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier

nltk.download('stopwords')
%matplotlib inline
```

```
'Google Colab not enabled'
[nltk_data] Downloading package stopwords to /home/magni/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [4]:
```
import importlib
import cw_df_metric_utils as cwutils
import DataPackage as dp
import DataPackageSupport as dps
import DataExperiment
import DataExperimentSupport
```

```
2022-01-15 10:36:13.008934: W tensorflow/stream_executor/platform/default/dso
_loader.cc:64] Could not load dynamic library 'libcudart.so.11.0'; dlerror: l
ibcudart.so.11.0: cannot open shared object file: No such file or directory
2022-01-15 10:36:13.008963: I tensorflow/stream_executor/cuda/cudart_stub.cc:
29] Ignore above cudart dlerror if you do not have a GPU set up on your machi
ne.
```

In [23]:
```
importlib.reload(dp)
importlib.reload(dps)
importlib.reload(DataExperiment)
importlib.reload(DataExperimentSupport)
```

Out[23]:
```
<module 'DataExperimentSupport' from '/home/magni/ML_Root/project_root/utilit
y_files/DataExperimentSupport.py'>
```

Load Data

In [5]:
```
#axis_labels=[1,2,3,4,5]
axis_labels=[0,1]
#classifier = RandomForestClassifier()
classifier = XGBClassifier(eval_metric='mlogloss')
ANALSYSIS_COL = 'reviewText_lemma_bert'
UNIQUE_COL = 'uuid'
TARGET_COL = 'overall_posneg'
```

In [6]:

```
if LOAD_FROM_EXP:
    #start from saved state
    myExp = jarvis.loadExperiment(FILE_NAME)
    myExp.display()

else:
    #start from source file and regenerate
    testDf = pd.read_pickle(jarvis.DATA_DIR_WORK + "/01_NL_ReviewText_All(new

    testDfBert = cwutils.getBertEncodeFrame(df=testDf,
                                            bertColumn=ANALSYSIS_COL,
                                            uniqueColumn=UNIQUE_COL,
                                            otherColumns=[TARGET_COL]
                                            )



    myExp = DataExperiment.DataExperiment(projectName=PROJECT_NAME,
                                          experimentName=EXPERIMENT_NAME,
                                          origData=testDfBert,
                                          uniqueColumn=UNIQUE_COL,
                                          targetColumn=TARGET_COL,
                                          classifier=classifier)
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Bert2 (XGB)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: False
---> isBaseModelPredicted: False
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: False
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
XGBClassifier(base_score=None, booster=None, colsample_bylevel=None,
              colsample_bynode=None, colsample_bytree=None,
              enable_categorical=False, eval_metric='mlogloss', gamma=None,
              gpu_id=None, importance_type=None, interaction_constraints=Non
e,
              learning_rate=None, max_delta_step=None, max_depth=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=None, reg_alpha=None,
              reg_lambda=None, scale_pos_weight=None, subsample=None,
              tree_method=None, validate_parameters=None, verbosity=None)

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: False
    ---> isTrainTestSplit: False
    Data:
    ---> isOrigDataLoaded: True
    ---> isTrainDataLoaded: False
```

In [7]:
```
myExp.processDataPackage()
```


```
Undersampling data to match min class: 0 of size: 13440
```


```
Completed train/test split (test_size = 0.2):
---> Original data size: 26880
---> Training data size: 21504
---> Testing data size: 5376
---> Stratified on column: overall_posneg
```

In [8]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Bert2 (XGB)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: False
---> isBaseModelPredicted: False
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: False
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
XGBClassifier(base_score=None, booster=None, colsample_bylevel=None,
```

```
                colsample_bynode=None, colsample_bytree=None,
                enable_categorical=False, eval_metric='mlogloss', gamma=None,
                gpu_id=None, importance_type=None, interaction_constraints=Non
e,
                learning_rate=None, max_delta_step=None, max_depth=None,
                min_child_weight=None, missing=nan, monotone_constraints=None,
                n_estimators=100, n_jobs=None, num_parallel_tree=None,
                predictor=None, random_state=None, reg_alpha=None,
                reg_lambda=None, scale_pos_weight=None, subsample=None,
                tree_method=None, validate_parameters=None, verbosity=None)

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
```

In [9]:
```
myExp.createBaseModel()
```

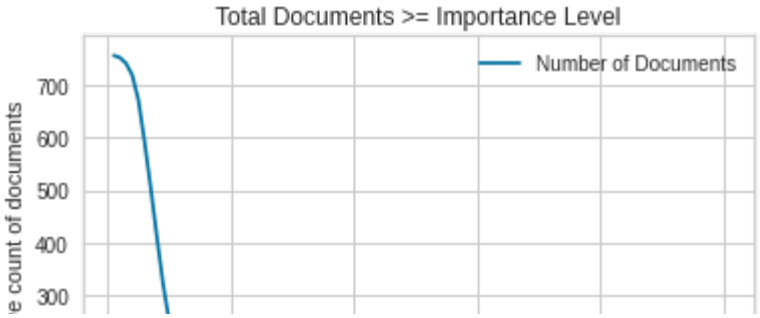
```
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
```

In [10]:
```
myExp.predictBaseModel()
```

```
Base Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.6
```

In [11]:
```
impFeatures = myExp.analyzeBaseModelFeatureImportance(returnAbove=0.002)
```

```
  0%|          | 0/101 [00:00<?, ?it/s]
Feature Importance Summary:
---> Original feature count: 768
---> Returned feature count: 40
---> Removed feature count: 728
---> Return items above (including): 0.002
```



In [12]:
```
myExp.createFinalModel(featureImportanceThreshold=0.002)
```

```
  0%|          | 0/101 [00:00<?, ?it/s]
  0%|          | 0/101 [00:00<?, ?it/s]
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
```

In [13]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Bert2 (XGB)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: True
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
XGBClassifier(base_score=None, booster=None, colsample_bylevel=None,
              colsample_bynode=None, colsample_bytree=None,
              enable_categorical=False, eval_metric='mlogloss', gamma=None,
              gpu_id=None, importance_type=None, interaction_constraints=Non
e,
              learning_rate=None, max_delta_step=None, max_depth=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=None, reg_alpha=None,
              reg_lambda=None, scale_pos_weight=None, subsample=None,
              tree_method=None, validate_parameters=None, verbosity=None)

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
```

```
    ---> isTrainDataLoaded: True
```

In [14]:

```
myExp.predictFinalModel()
myExp.display()
```

```
Final Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.6
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Bert2 (XGB)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
XGBClassifier(base_score=None, booster=None, colsample_bylevel=None,
              colsample_bynode=None, colsample_bytree=None,
              enable_categorical=False, eval_metric='mlogloss', gamma=None,
              gpu_id=None, importance_type=None, interaction_constraints=Non
e,
              learning_rate=None, max_delta_step=None, max_depth=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=None, reg_alpha=None,
              reg_lambda=None, scale_pos_weight=None, subsample=None,
              tree_method=None, validate_parameters=None, verbosity=None)

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

In [15]:
```
myExp.createBaseModelLearningCurve(n_jobs=10)
```

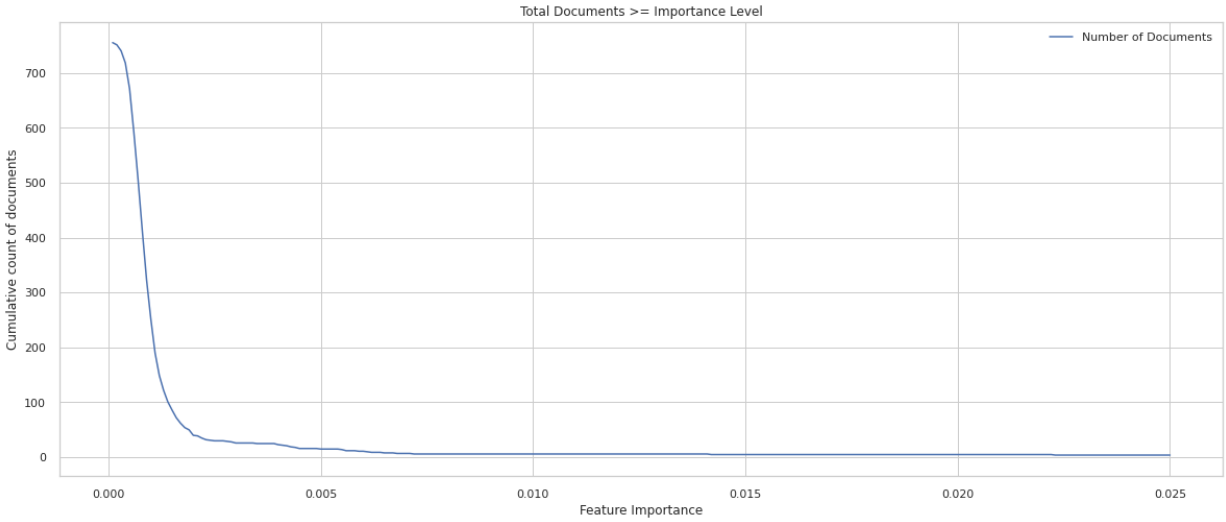
```
[learning_curve] Training set sizes: [ 1720  3440  8601 17203]
[Parallel(n_jobs=10)]: Using backend LokyBackend with 10 concurrent workers.
[Parallel(n_jobs=10)]: Done   3 out of  20 | elapsed:  2.2min remaining: 12.2
min
[Parallel(n_jobs=10)]: Done   9 out of  20 | elapsed:  4.4min remaining:  5.4
min
[Parallel(n_jobs=10)]: Done  15 out of  20 | elapsed:  6.5min remaining:  2.2
min
[Parallel(n_jobs=10)]: Done  20 out of  20 | elapsed:  7.1min finished
```

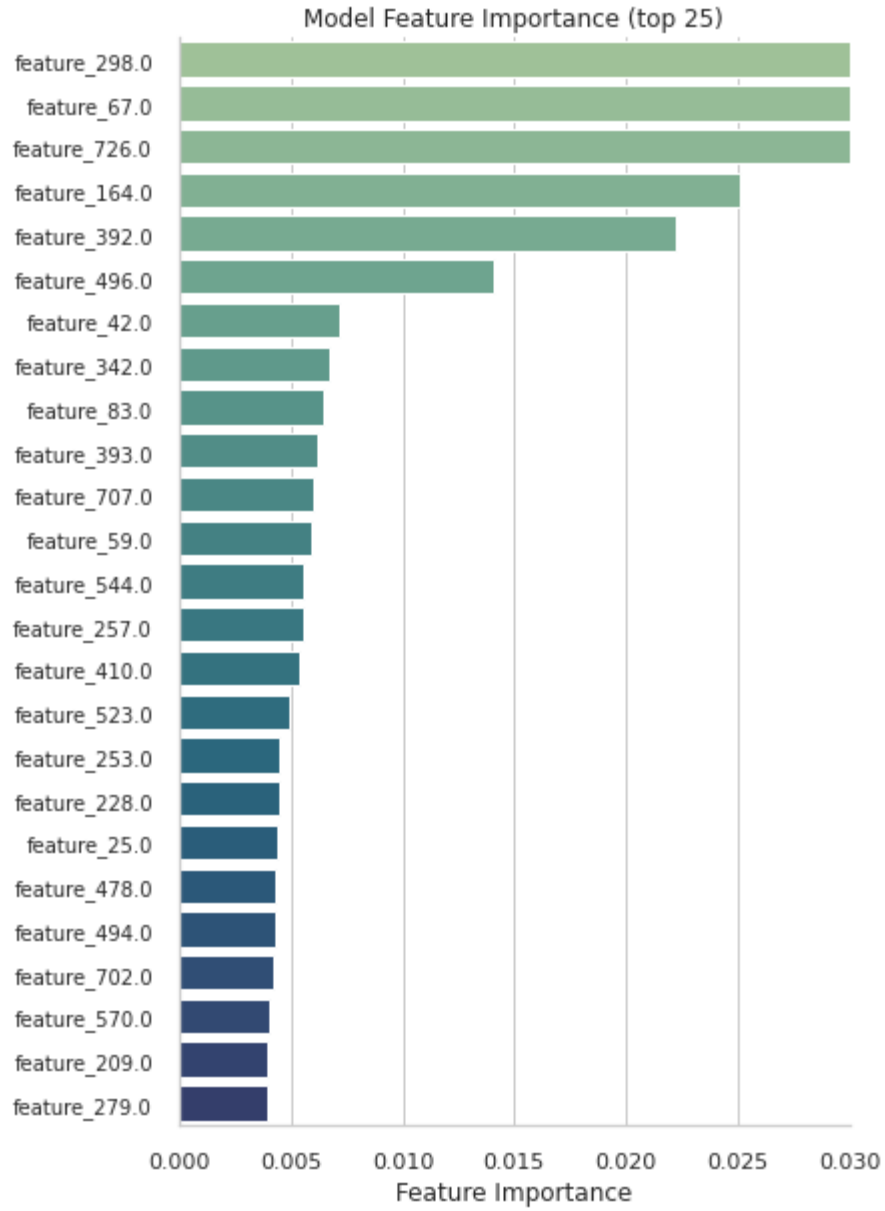
In [16]:
```
myExp.createFinalModelLearningCurve(n_jobs=10)
```

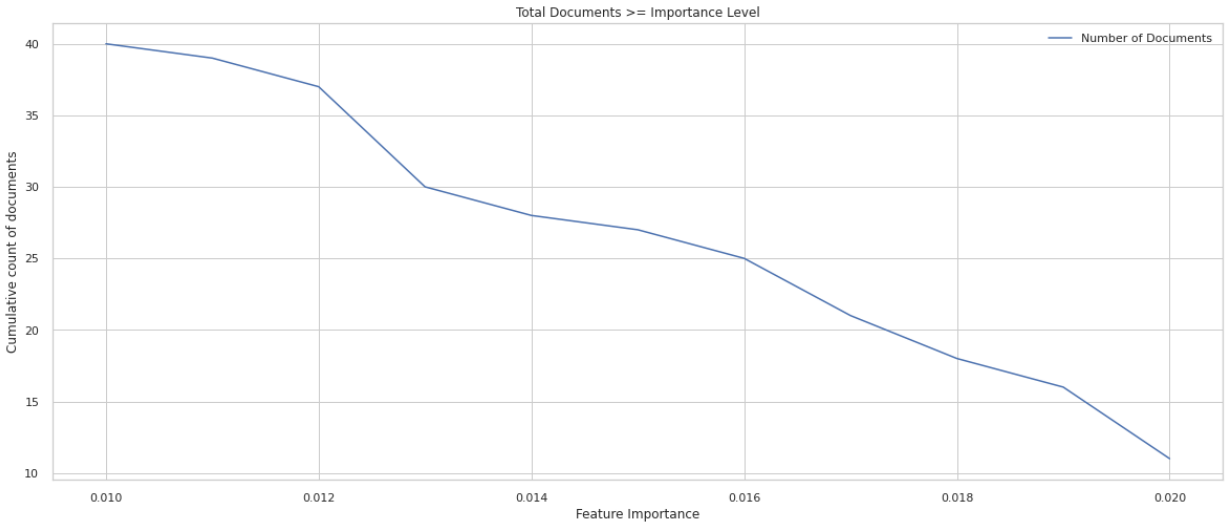
```
[Parallel(n_jobs=10)]: Using backend LokyBackend with 10 concurrent workers.
[learning_curve] Training set sizes: [ 1720  3440  8601 17203]
[Parallel(n_jobs=10)]: Done   3 out of  20 | elapsed:  1.8min remaining: 10.0
min
[Parallel(n_jobs=10)]: Done   9 out of  20 | elapsed:  2.4min remaining:  3.0
min
[Parallel(n_jobs=10)]: Done  15 out of  20 | elapsed:  4.1min remaining:  1.4
min
[Parallel(n_jobs=10)]: Done  20 out of  20 | elapsed:  4.2min finished
```

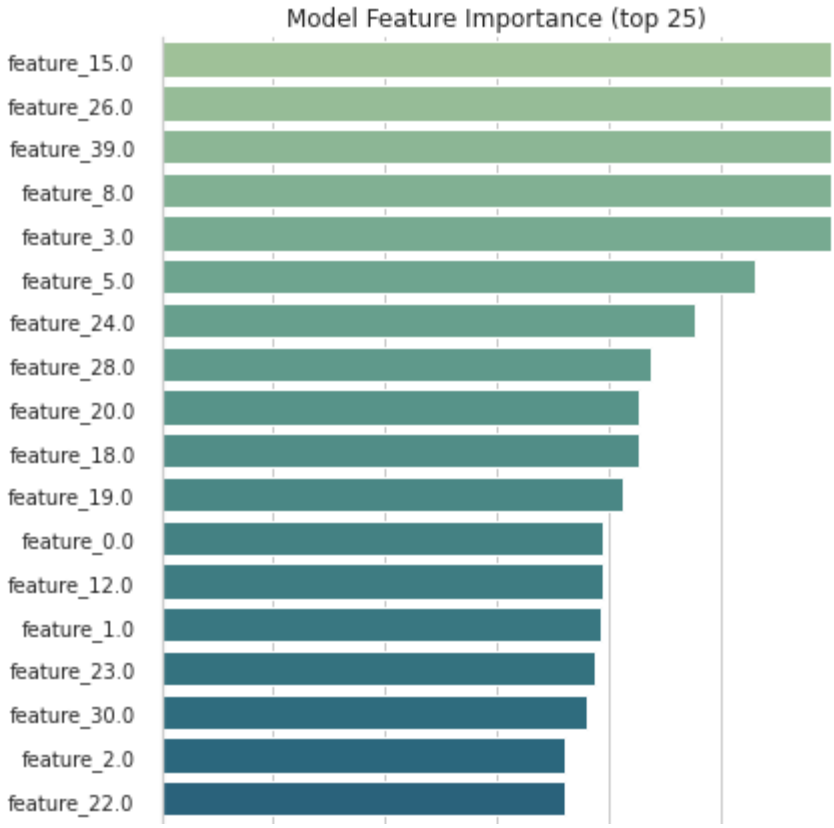
In [25]:
```
myExp.showBaseModelFeatureImportance(upperValue=0.025)
myExp.showFinalModelFeatureImportance(startValue=0.01,
                                      increment=0.001,
                                      upperValue=0.02)
```

```
  0%|          | 0/251 [00:00<?, ?it/s]
  0%|          | 0/12 [00:00<?, ?it/s]
```




Model Feature Importance (top 25)
feature_298.0
feature_67.0
feature_726.0
feature_164.0
feature_392.0
feature_496.0
feature_42.0
feature_342.0
feature_83.0
feature_393.0
feature_707.0
feature_59.0
feature_544.0
feature_257.0
feature_410.0
feature_523.0
feature_253.0
feature_228.0
feature_25.0
feature_478.0
feature_494.0
feature_702.0
feature_570.0
feature_209.0
feature_279.0
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Feature Importance


Total Documents >= Importance Level
Number of Documents
Cumulative count of documents
40
35
30
25
20
15
10
0.010
0.012
0.014
0.016
0.018
0.020
Feature Importance



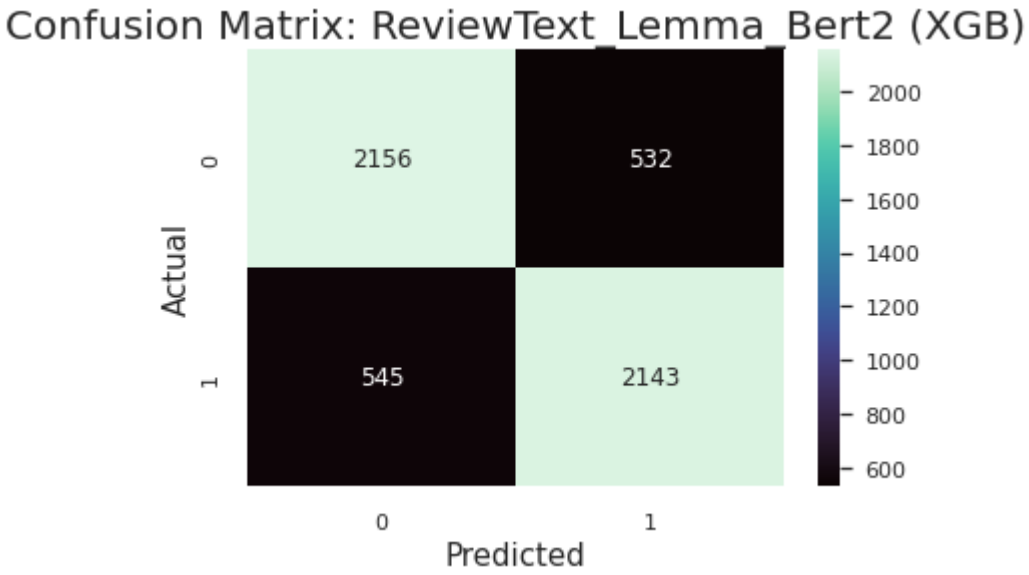
In [18]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Bert2 (XGB)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: True
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: True
---> isClassifierLoaded: True
XGBClassifier(base_score=None, booster=None, colsample_bylevel=None,
              colsample_bynode=None, colsample_bytree=None,
              enable_categorical=False, eval_metric='mlogloss', gamma=None,
              gpu_id=None, importance_type=None, interaction_constraints=Non
e,
              learning_rate=None, max_delta_step=None, max_depth=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=None, reg_alpha=None,
              reg_lambda=None, scale_pos_weight=None, subsample=None,
              tree_method=None, validate_parameters=None, verbosity=None)

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
```

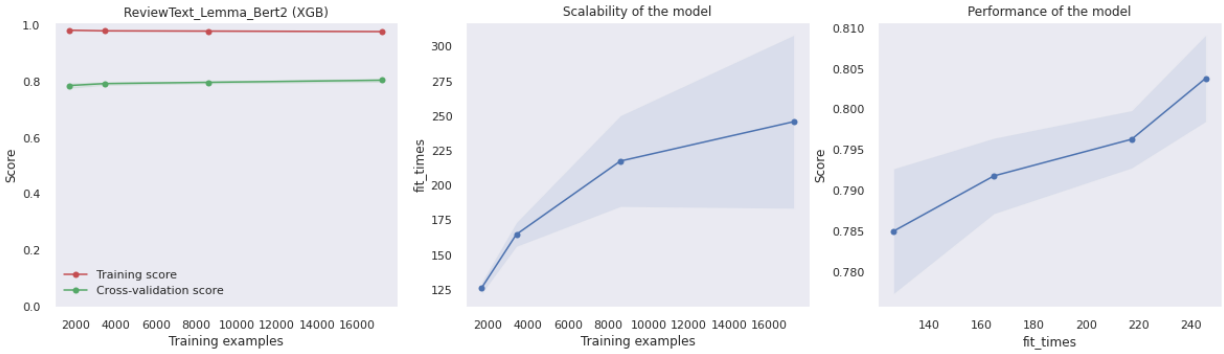
```
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
```

In [19]:
```
myExp.showBaseModelReport(axis_labels)
```

```
Base Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.6
              precision    recall  f1-score   support

           0       0.80      0.80      0.80      2688
           1       0.80      0.80      0.80      2688

    accuracy                           0.80      5376
   macro avg       0.80      0.80      0.80      5376
weighted avg       0.80      0.80      0.80      5376
```

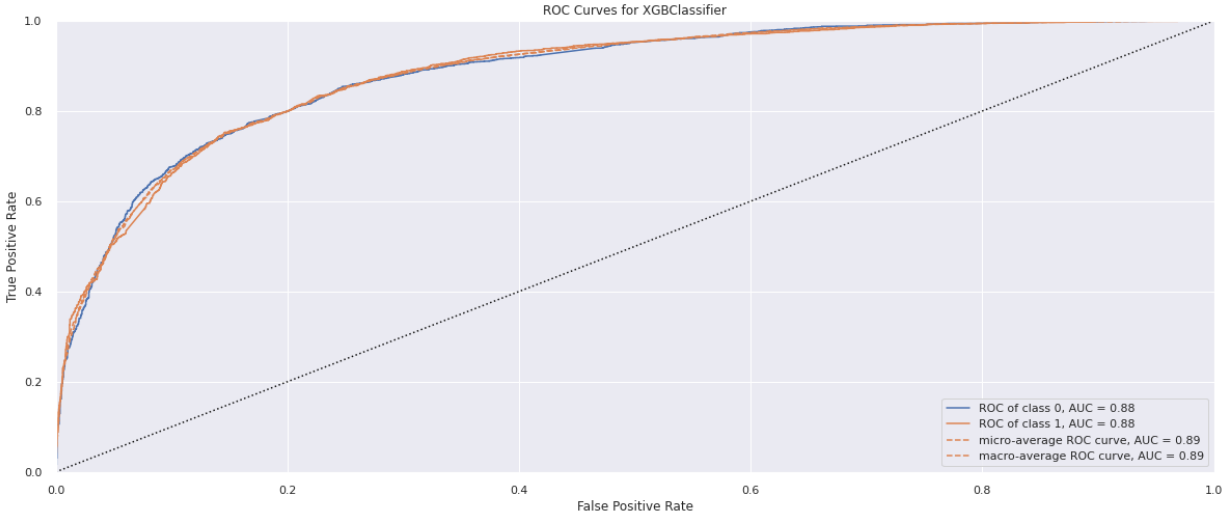



```
<Figure size 1440x576 with 0 Axes>
```



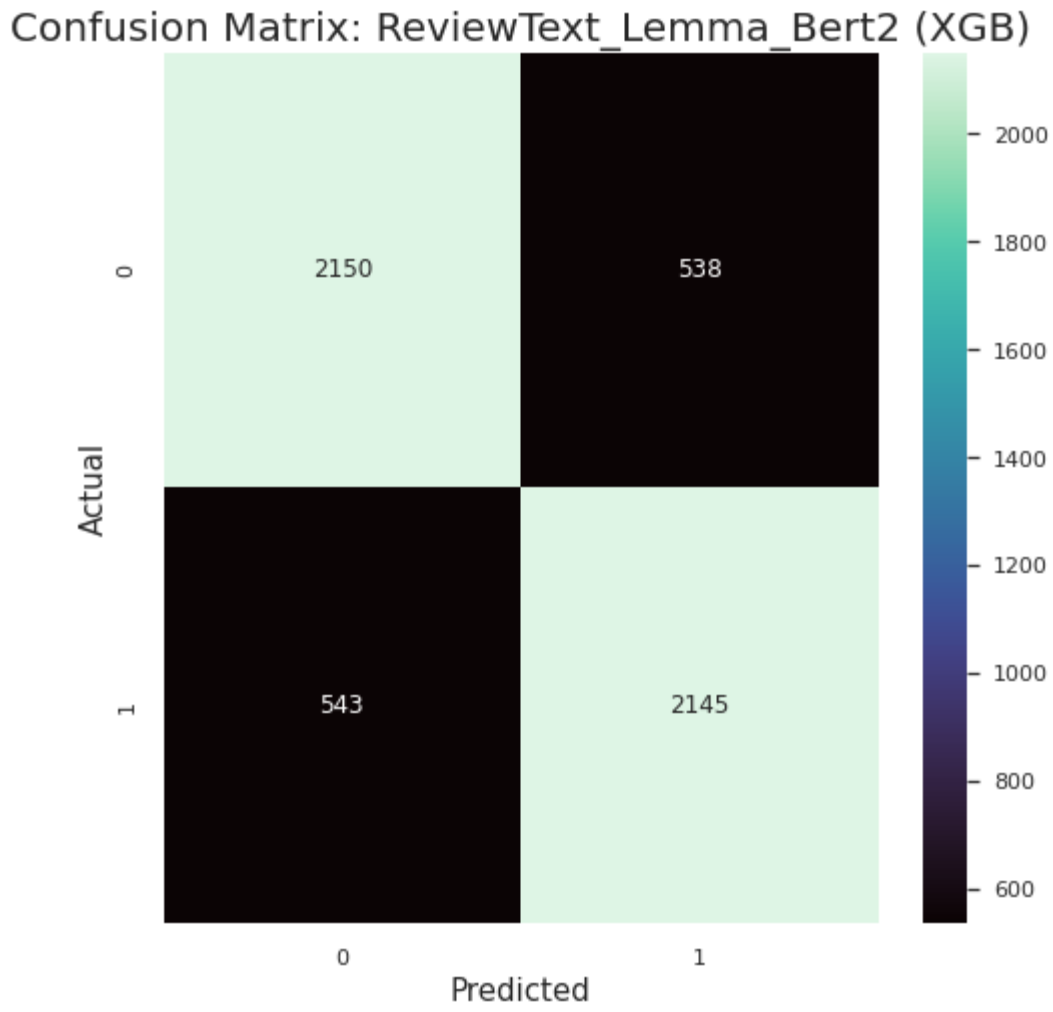

```
Base model ROCAUC not calculated. Starting now
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/sklearn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprecated and will be removed in a future release. To remove this warning, do the following: 1) Pass option use_label_encoder=False when constructing XGBClassi
```

```
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
```

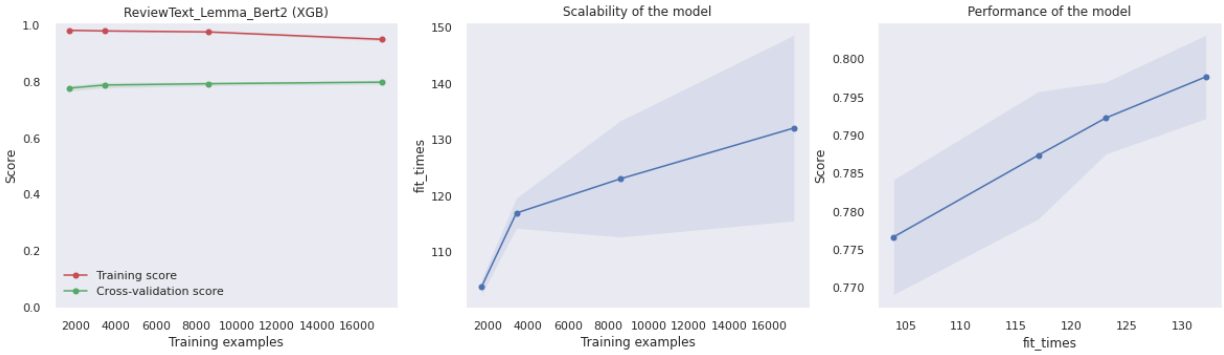


In [27]:
```
myExp.showFinalModelReport(axis_labels)
```

```
Final Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.6
              precision    recall  f1-score   support

           0       0.80      0.80      0.80      2688
           1       0.80      0.80      0.80      2688

    accuracy                           0.80      5376
   macro avg       0.80      0.80      0.80      5376
weighted avg       0.80      0.80      0.80      5376
```




```
<Figure size 576x576 with 0 Axes>
```




```
Final model ROCAUC not calculated. Starting now
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/sklearn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprecated and will be removed in a future release. To remove this warning, do the following: 1) Pass option use_label_encoder=False when constructing XGBClassifier object; and 2) Encode your labels (y) as integers starting with 0, i.e. 0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
```

In [21]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Bert2 (XGB)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: True
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: True
---> isClassifierLoaded: True
XGBClassifier(base_score=None, booster=None, colsample_bylevel=None,
              colsample_bynode=None, colsample_bytree=None,
              enable_categorical=False, eval_metric='mlogloss', gamma=None,
              gpu_id=None, importance_type=None, interaction_constraints=Non
e,
              learning_rate=None, max_delta_step=None, max_depth=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=None, reg_alpha=None,
              reg_lambda=None, scale_pos_weight=None, subsample=None,
              tree_method=None, validate_parameters=None, verbosity=None)

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

Save Experiment

In [22]:
```
jarvis.saveExperiment(myExp, FILE_NAME)
```

```
[CV] END ....................., score=(train=0.980, test=0.790) total time= 2.1
min
[CV] END ....................., score=(train=0.981, test=0.777) total time= 2.2
min
[CV] END ....................., score=(train=0.978, test=0.796) total time= 2.6
min
[CV] END ....................., score=(train=0.981, test=0.777) total time= 1.8
min
[CV] END ....................., score=(train=0.981, test=0.780) total time= 1.7
min
```

```
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.981, test=0.781) total time= 2.0
min
[CV] END ..................., score=(train=0.978, test=0.797) total time= 3.9
min
[CV] END ..................., score=(train=0.976, test=0.785) total time= 2.1
min
[CV] END ..................., score=(train=0.981, test=0.788) total time= 1.7
min
```

```
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
```

```
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.979, test=0.790) total time= 2.8
min
[CV] END ..................., score=(train=0.980, test=0.793) total time= 2.9
min
[CV] END ..................., score=(train=0.980, test=0.779) total time= 1.9
min
[CV] END ..................., score=(train=0.980, test=0.779) total time= 2.0
min
```

```
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.980, test=0.784) total time= 2.8
min
[CV] END ..................., score=(train=0.977, test=0.803) total time= 3.7
min
[CV] END ..................., score=(train=0.980, test=0.774) total time= 1.7
min
[CV] END ..................., score=(train=0.977, test=0.797) total time= 2.2
min
```

```
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
```

```
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
[CV] END ..................., score=(train=0.977, test=0.799) total time= 4.9
min
[CV] END ..................., score=(train=0.984, test=0.764) total time= 1.8
min
[CV] END ..................., score=(train=0.948, test=0.802) total time= 2.3
min
```

```
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.980, test=0.792) total time= 4.0
min
[CV] END ..................., score=(train=0.978, test=0.797) total time= 2.5
min
[CV] END ..................., score=(train=0.979, test=0.790) total time= 2.0
```

```
min
[CV] END .................., score=(train=0.975, test=0.795) total time= 2.1

/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END .................., score=(train=0.984, test=0.780) total time= 2.1
min
[CV] END .................., score=(train=0.977, test=0.805) total time= 4.5
min
[CV] END .................., score=(train=0.976, test=0.788) total time= 2.2
min
[CV] END .................., score=(train=0.978, test=0.801) total time= 1.9
min

/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
```

```
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.977, test=0.797) total time= 4.9
min
[CV] END ..................., score=(train=0.977, test=0.807) total time= 2.1
min
[CV] END ..................., score=(train=0.979, test=0.788) total time= 2.0
min
[CV] END ..................., score=(train=0.955, test=0.802) total time= 2.1
min
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.979, test=0.795) total time= 2.9
min
[CV] END ..................., score=(train=0.975, test=0.811) total time= 4.0
min
[CV] END ..................., score=(train=0.950, test=0.793) total time= 2.4
min
[CV] END ..................., score=(train=0.976, test=0.797) total time= 1.7
min
[CV] END ..................., score=(train=0.978, test=0.794) total time= 4.0
min
[CV] END ..................., score=(train=0.981, test=0.798) total time= 2.0
min
[CV] END ..................., score=(train=0.951, test=0.789) total time= 2.5
min
```

```
[CV] END ..................., score=(train=0.943, test=0.802) total time= 1.7
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
```

Scratchpad

In [ ]:

Configuration

In [1]:
```
# Parameters
ENABLE_COLAB = False

PROJECT_NAME = 'ML1010-Group-Project'
EXPERIMENT_NAME = 'ReviewText_Lemma_Glove2 (XGB)'
FILE_NAME = '01_ML1010_GP_XGB_Glove2'
LOAD_FROM_EXP = False

#Root Machine Learning Directory. Projects appear underneath
GOOGLE_DRIVE_MOUNT = '/content/gdrive'
COLAB_ROOT_DIR = GOOGLE_DRIVE_MOUNT + '/MyDrive/Colab Notebooks'
COLAB_INIT_DIR = COLAB_ROOT_DIR + '/utility_files'

LOCAL_ROOT_DIR = '/home/magni//ML_Root/project_root'
LOCAL_INIT_DIR = LOCAL_ROOT_DIR + '/utility_files'
```

Bootstrap Environment

In [2]:
```
#add in support for utility file directory and importing
import sys
import os

if ENABLE_COLAB:
  #Need access to drive
  from google.colab import drive
  drive.mount(GOOGLE_DRIVE_MOUNT, force_remount=True)

  #add in utility directory to syspath to import
  INIT_DIR = COLAB_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = COLAB_ROOT_DIR

else:
  #add in utility directory to syspath to import
  INIT_DIR = LOCAL_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = LOCAL_ROOT_DIR

#Import Utility Support
from jarvis import Jarvis
jarvis = Jarvis(ROOT_DIR, PROJECT_NAME)

import mv_python_utils as mvutils
```

```
Wha...where am I?
I am awake now.
```

```
I have set your current working directory to /home/magni/ML_Root/project_root
/ML1010-Group-Project
The current time is 10:49
Hello sir. Extra caffeine may help.
```

Setup Runtime Environment

In [3]:

```
if ENABLE_COLAB:
  #!pip install scipy -q
  #!pip install scikit-learn -q
  #!pip install pycaret -q
  #!pip install matplotlib -q
  #!pip install joblib -q
  #!pip install pandasql -q
  !pip install umap_learn -q
  !pip install sentence_transformers -q
  !pip install spacytextblob -q
  !pip install flair -q
  display('Google Colab enabled')
else:
  display('Google Colab not enabled')

#Common imports
import json
import pandas as pd
import numpy as np
import matplotlib
import re
import nltk
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn import metrics
from sklearn.datasets import load_digits
from sklearn.model_selection import train_test_split as tts
#from yellowbrick.classifier import ConfusionMatrix
#from sklearn.linear_model import LogisticRegression
from yellowbrick.target import ClassBalance
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier

nltk.download('stopwords')
%matplotlib inline
```

```
'Google Colab not enabled'
```

```
[nltk_data] Downloading package stopwords to /home/magni/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [4]:
```
import importlib
import cw_df_metric_utils as cwutils
import DataPackage as dp
import DataPackageSupport as dps
import DataExperiment
import DataExperimentSupport
```

```
2022-01-15 10:49:04.747345: W tensorflow/stream_executor/platform/default/dso
_loader.cc:64] Could not load dynamic library 'libcudart.so.11.0'; dlerror: l
ibcudart.so.11.0: cannot open shared object file: No such file or directory
2022-01-15 10:49:04.747372: I tensorflow/stream_executor/cuda/cudart_stub.cc:
29] Ignore above cudart dlerror if you do not have a GPU set up on your machi
ne.
```

In [23]:
```
importlib.reload(dp)
importlib.reload(dps)
importlib.reload(DataExperiment)
importlib.reload(DataExperimentSupport)
```

Out[23]:
```
<module 'DataExperimentSupport' from '/home/magni/ML_Root/project_root/utilit
y_files/DataExperimentSupport.py'>
```

Load Data

In [5]:
```
#axis_labels=[1,2,3,4,5]
axis_labels=[0,1]
#classifier = RandomForestClassifier()
classifier = XGBClassifier(eval_metric='mlogloss')
ANALSYSIS_COL = 'reviewText_lemma_glove'
UNIQUE_COL = 'uuid'
TARGET_COL = 'overall_posneg'
```

In [6]:

```
if LOAD_FROM_EXP:
    #start from saved state
    myExp = jarvis.loadExperiment(FILE_NAME)
    myExp.display()

else:
    #start from source file and regenerate
    testDf = pd.read_pickle(jarvis.DATA_DIR_WORK + "/01_NL_ReviewText_All(new

    testDfBert = cwutils.getBertEncodeFrame(df=testDf,
                                            bertColumn=ANALSYSIS_COL,
                                            uniqueColumn=UNIQUE_COL,
                                            otherColumns=[TARGET_COL]
                                            )



    myExp = DataExperiment.DataExperiment(projectName=PROJECT_NAME,
                                          experimentName=EXPERIMENT_NAME,
                                          origData=testDfBert,
                                          uniqueColumn=UNIQUE_COL,
                                          targetColumn=TARGET_COL,
                                          classifier=classifier)
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Glove2 (XGB)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: False
---> isBaseModelPredicted: False
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: False
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
XGBClassifier(base_score=None, booster=None, colsample_bylevel=None,
              colsample_bynode=None, colsample_bytree=None,
              enable_categorical=False, eval_metric='mlogloss', gamma=None,
              gpu_id=None, importance_type=None, interaction_constraints=Non
e,
              learning_rate=None, max_delta_step=None, max_depth=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=None, reg_alpha=None,
              reg_lambda=None, scale_pos_weight=None, subsample=None,
              tree_method=None, validate_parameters=None, verbosity=None)

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: False
    ---> isTrainTestSplit: False
    Data:
    ---> isOrigDataLoaded: True
    ---> isTrainDataLoaded: False
```

In [7]:
```
myExp.processDataPackage()
```


```
Undersampling data to match min class: 0 of size: 13440
```


```
Completed train/test split (test_size = 0.2):
---> Original data size: 26880
---> Training data size: 21504
---> Testing data size: 5376
---> Stratified on column: overall_posneg
```

In [8]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Glove2 (XGB)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: False
---> isBaseModelPredicted: False
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: False
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
XGBClassifier(base_score=None, booster=None, colsample_bylevel=None,
```

```
                colsample_bynode=None, colsample_bytree=None,
                enable_categorical=False, eval_metric='mlogloss', gamma=None,
                gpu_id=None, importance_type=None, interaction_constraints=Non
e,
                learning_rate=None, max_delta_step=None, max_depth=None,
                min_child_weight=None, missing=nan, monotone_constraints=None,
                n_estimators=100, n_jobs=None, num_parallel_tree=None,
                predictor=None, random_state=None, reg_alpha=None,
                reg_lambda=None, scale_pos_weight=None, subsample=None,
                tree_method=None, validate_parameters=None, verbosity=None)

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
```

In [9]:
```
myExp.createBaseModel()
```

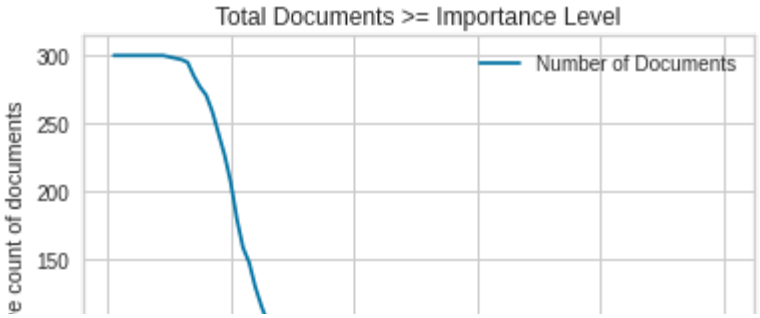
```
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
```

In [10]:
```
myExp.predictBaseModel()
```

```
Base Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.61
```

In [11]:
```
impFeatures = myExp.analyzeBaseModelFeatureImportance(returnAbove=0.002)
```

```
  0%|          | 0/101 [00:00<?, ?it/s]
Feature Importance Summary:
---> Original feature count: 300
---> Returned feature count: 207
---> Removed feature count: 93
---> Return items above (including): 0.002
```



In [12]:
```
myExp.createFinalModel(featureImportanceThreshold=0.002)
```

```
  0%|          | 0/101 [00:00<?, ?it/s]
  0%|          | 0/101 [00:00<?, ?it/s]
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
```

In [13]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Glove2 (XGB)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: True
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
XGBClassifier(base_score=None, booster=None, colsample_bylevel=None,
              colsample_bynode=None, colsample_bytree=None,
              enable_categorical=False, eval_metric='mlogloss', gamma=None,
              gpu_id=None, importance_type=None, interaction_constraints=Non
e,
              learning_rate=None, max_delta_step=None, max_depth=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=None, reg_alpha=None,
              reg_lambda=None, scale_pos_weight=None, subsample=None,
              tree_method=None, validate_parameters=None, verbosity=None)

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
```

```
    ---> isTrainDataLoaded: True
```

In [14]:

```
myExp.predictFinalModel()
myExp.display()
```

```
Final Model Stats:
Accuracy: 0.81
Precision: 0.81
Recalll: 0.81
F1 Score: 0.81
Cohen kappa:: 0.62
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Glove2 (XGB)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
XGBClassifier(base_score=None, booster=None, colsample_bylevel=None,
              colsample_bynode=None, colsample_bytree=None,
              enable_categorical=False, eval_metric='mlogloss', gamma=None,
              gpu_id=None, importance_type=None, interaction_constraints=Non
e,
              learning_rate=None, max_delta_step=None, max_depth=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=None, reg_alpha=None,
              reg_lambda=None, scale_pos_weight=None, subsample=None,
              tree_method=None, validate_parameters=None, verbosity=None)

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

In [15]:
```
myExp.createBaseModelLearningCurve(n_jobs=10)
```

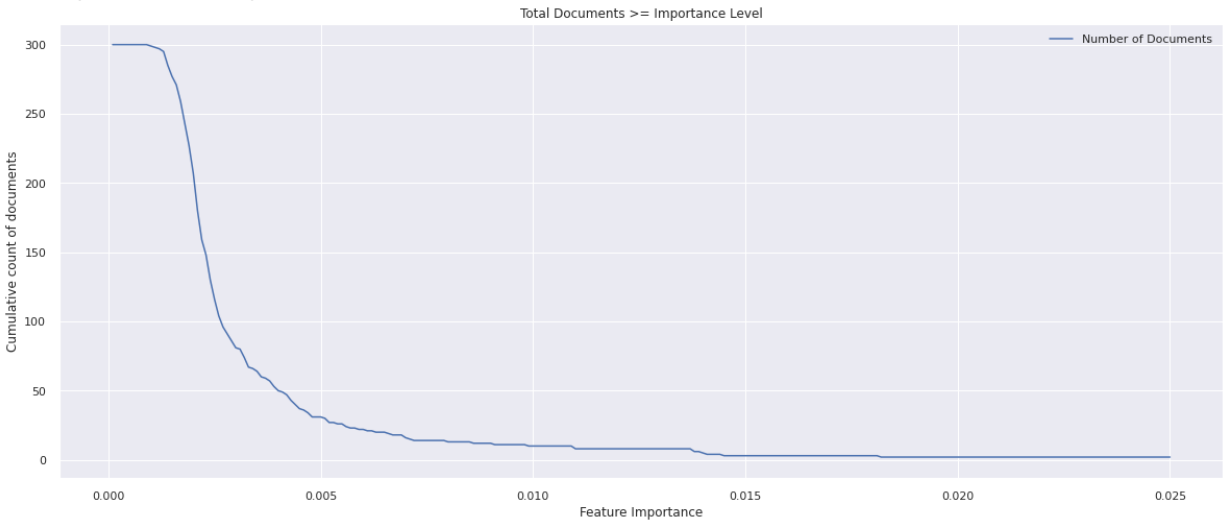
```
[learning_curve] Training set sizes: [ 1720  3440  8601 17203]
[Parallel(n_jobs=10)]: Using backend LokyBackend with 10 concurrent workers.
[Parallel(n_jobs=10)]: Done   3 out of  20 | elapsed:  2.0min remaining: 11.3
min
[Parallel(n_jobs=10)]: Done   9 out of  20 | elapsed:  4.0min remaining:  4.9
min
[Parallel(n_jobs=10)]: Done  15 out of  20 | elapsed:  5.9min remaining:  2.0
min
[Parallel(n_jobs=10)]: Done  20 out of  20 | elapsed:  6.2min finished
```

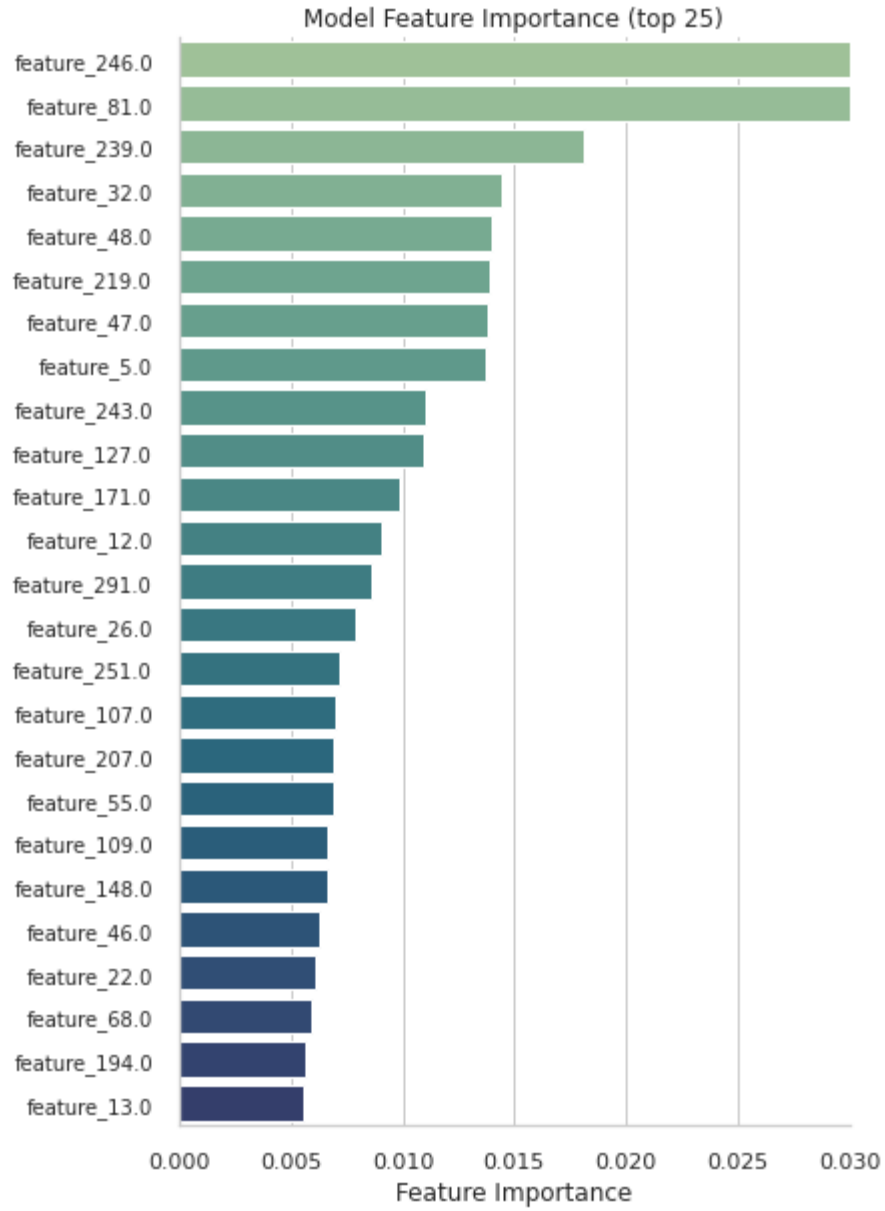
In [16]:
```
myExp.createFinalModelLearningCurve(n_jobs=10)
```

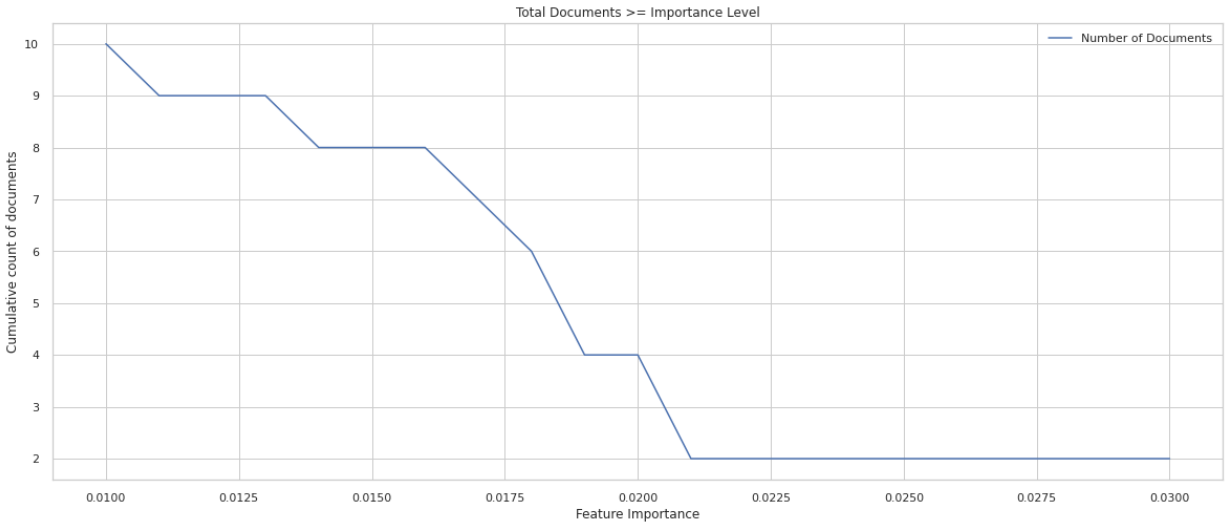
```
[Parallel(n_jobs=10)]: Using backend LokyBackend with 10 concurrent workers.
[learning_curve] Training set sizes: [ 1720  3440  8601 17203]
[Parallel(n_jobs=10)]: Done   3 out of  20 | elapsed:  2.0min remaining: 11.1
min
[Parallel(n_jobs=10)]: Done   9 out of  20 | elapsed:  3.9min remaining:  4.7
min
[Parallel(n_jobs=10)]: Done  15 out of  20 | elapsed:  5.5min remaining:  1.8
min
[Parallel(n_jobs=10)]: Done  20 out of  20 | elapsed:  5.9min finished
```

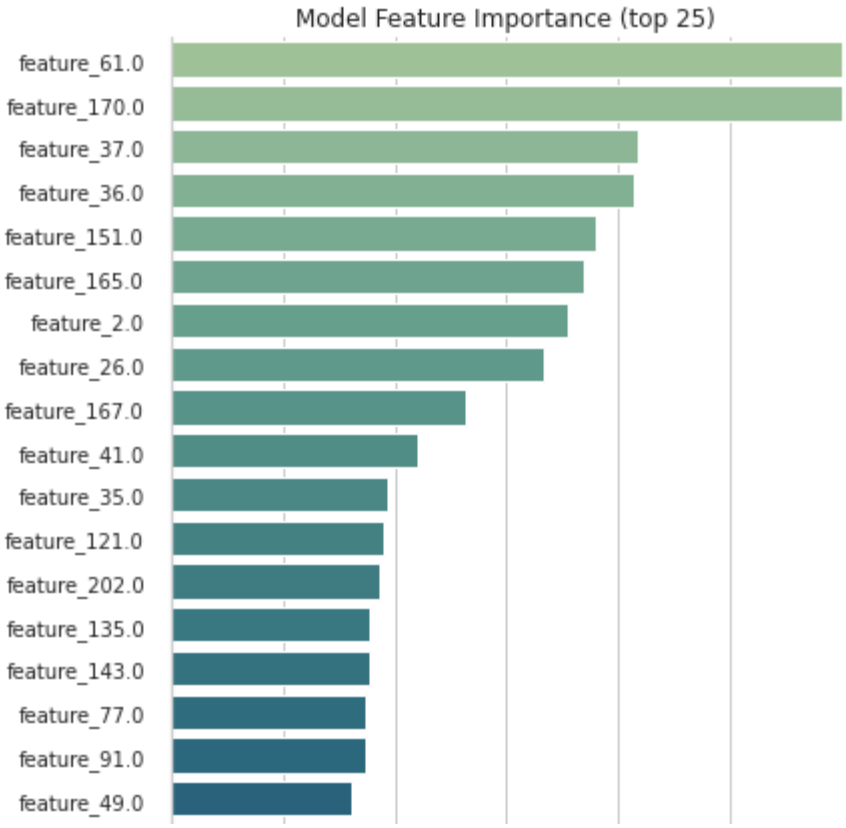
In [24]:
```
myExp.showBaseModelFeatureImportance(upperValue=0.025)
myExp.showFinalModelFeatureImportance(startValue=0.01,
                                      increment=0.001,
                                      upperValue=0.03)
```

```
  0%|          | 0/251 [00:00<?, ?it/s]
  0%|          | 0/22 [00:00<?, ?it/s]
```




Model Feature Importance (top 25)
feature_246.0
feature_81.0
feature_239.0
feature_32.0
feature_48.0
feature_219.0
feature_47.0
feature_5.0
feature_243.0
feature_127.0
feature_171.0
feature_12.0
feature_291.0
feature_26.0
feature_251.0
feature_107.0
feature_207.0
feature_55.0
feature_109.0
feature_148.0
feature_46.0
feature_22.0
feature_68.0
feature_194.0
feature_13.0
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Feature Importance


Total Documents >= Importance Level
Number of Documents
Cumulative count of documents
10
9
8
7
6
5
4
3
2
0.0100
0.0125
0.0150
0.0175
0.0200
0.0225
0.0250
0.0275
0.0300
Feature Importance



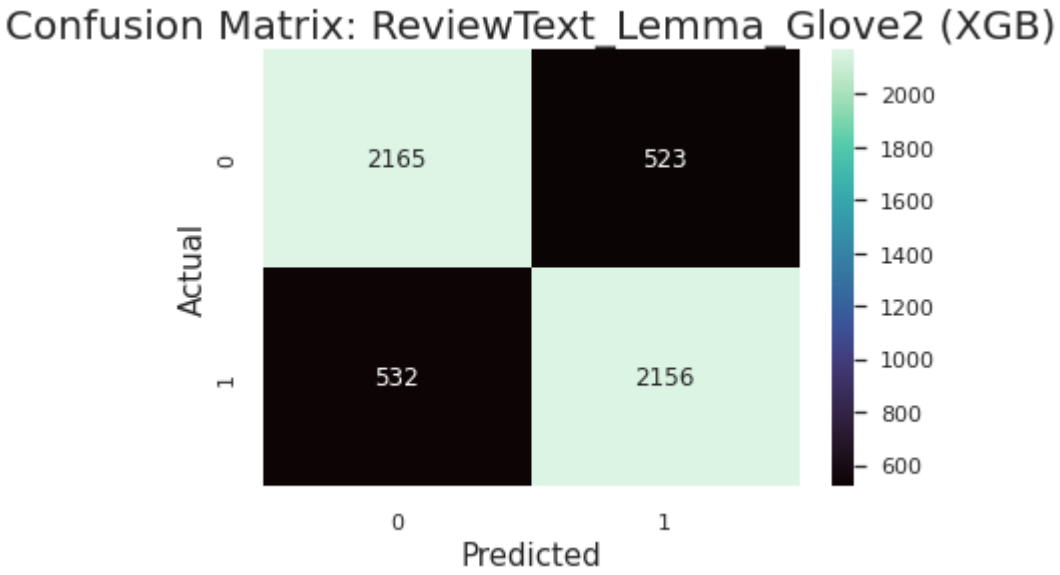
In [18]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Glove2 (XGB)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: True
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: True
---> isClassifierLoaded: True
XGBClassifier(base_score=None, booster=None, colsample_bylevel=None,
              colsample_bynode=None, colsample_bytree=None,
              enable_categorical=False, eval_metric='mlogloss', gamma=None,
              gpu_id=None, importance_type=None, interaction_constraints=Non
e,
              learning_rate=None, max_delta_step=None, max_depth=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=None, reg_alpha=None,
              reg_lambda=None, scale_pos_weight=None, subsample=None,
              tree_method=None, validate_parameters=None, verbosity=None)

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
```

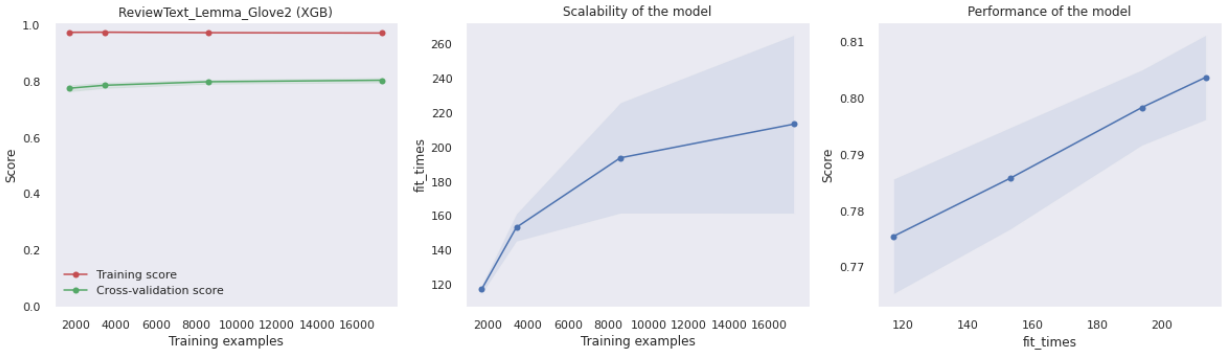
```
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
```

In [19]:
```
myExp.showBaseModelReport(axis_labels)
```

```
Base Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.61
              precision    recall  f1-score   support

           0       0.80      0.81      0.80      2688
           1       0.80      0.80      0.80      2688

    accuracy                           0.80      5376
   macro avg       0.80      0.80      0.80      5376
weighted avg       0.80      0.80      0.80      5376
```

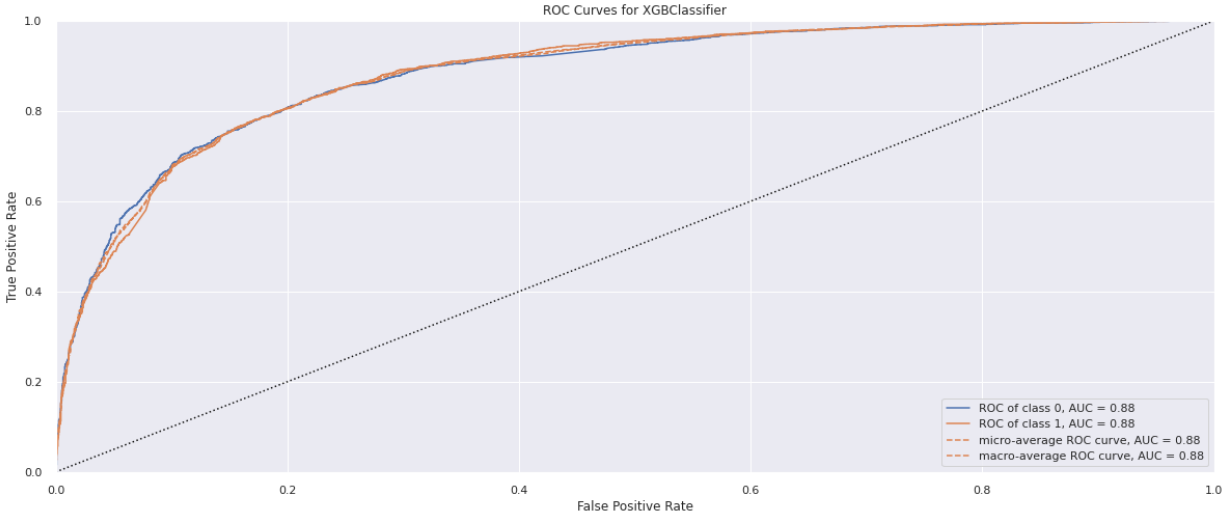



```
<Figure size 1440x576 with 0 Axes>
```



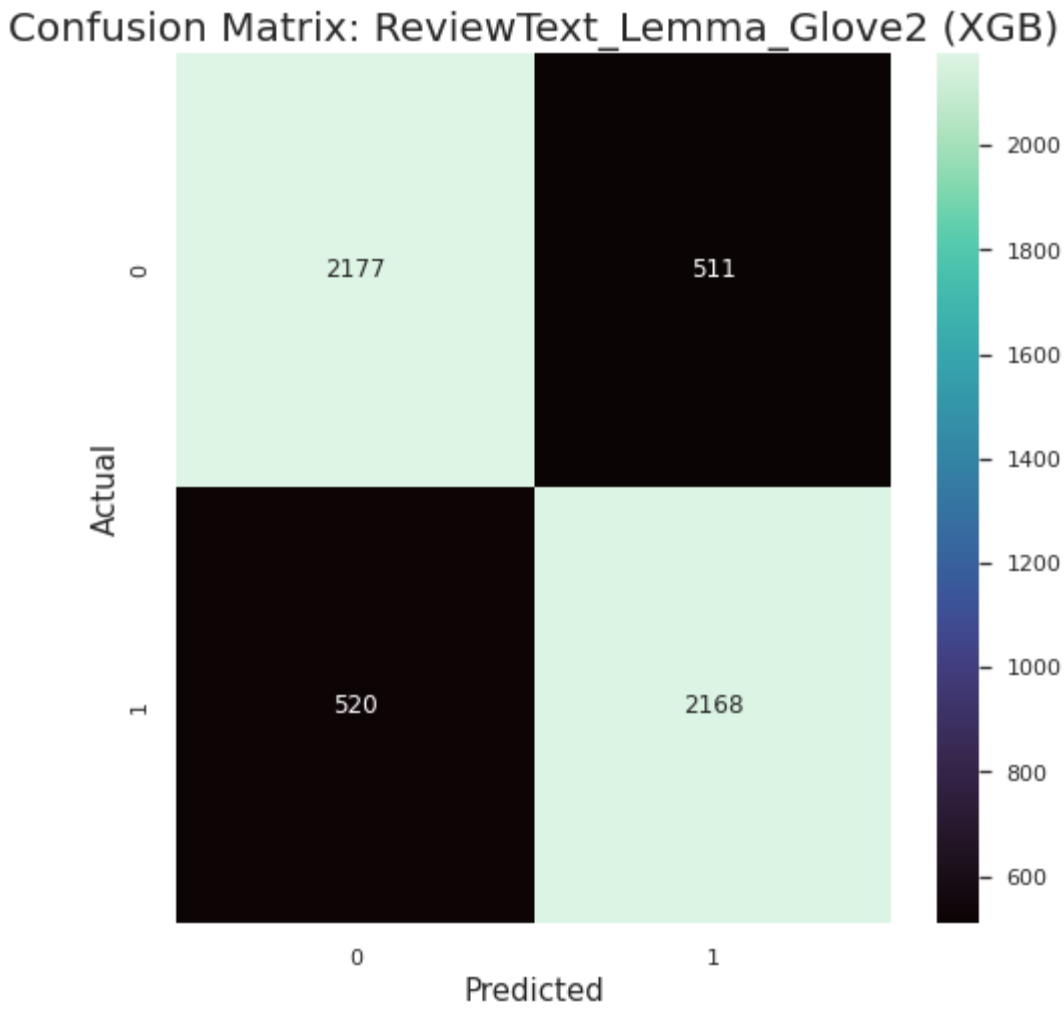

```
Base model ROCAUC not calculated. Starting now
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/sklearn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprecated and will be removed in a future release. To remove this warning, do the following: 1) Pass option use_label_encoder=False when constructing XGBClassi
```

```
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
```

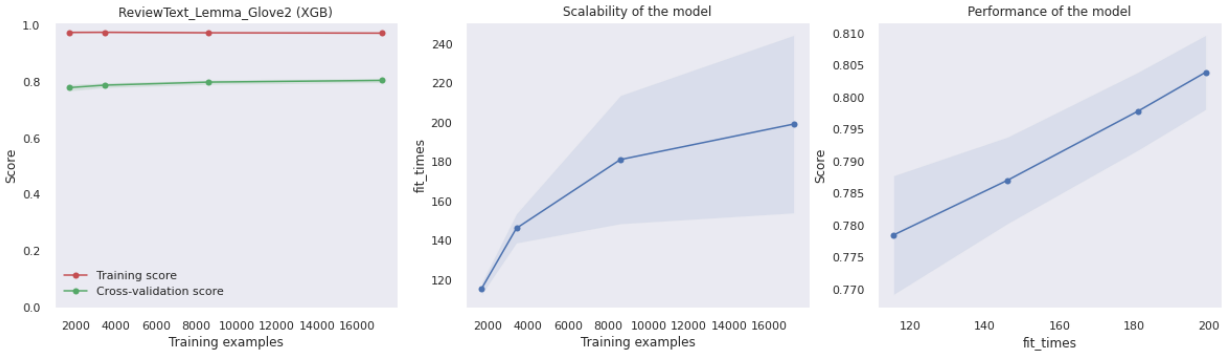


In [26]:
```
myExp.showFinalModelReport(axis_labels)
```

```
Final Model Stats:
Accuracy: 0.81
Precision: 0.81
Recalll: 0.81
F1 Score: 0.81
Cohen kappa:: 0.62
              precision    recall  f1-score   support

           0       0.81      0.81      0.81      2688
           1       0.81      0.81      0.81      2688

    accuracy                           0.81      5376
   macro avg       0.81      0.81      0.81      5376
weighted avg       0.81      0.81      0.81      5376
```




```
<Figure size 576x576 with 0 Axes>
```




```
Final model ROCAUC not calculated. Starting now
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/sklearn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprecated and will be removed in a future release. To remove this warning, do the following: 1) Pass option use_label_encoder=False when constructing XGBClassifier object; and 2) Encode your labels (y) as integers starting with 0, i.e. 0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
```

In [21]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Glove2 (XGB)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: True
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: True
---> isClassifierLoaded: True
XGBClassifier(base_score=None, booster=None, colsample_bylevel=None,
              colsample_bynode=None, colsample_bytree=None,
              enable_categorical=False, eval_metric='mlogloss', gamma=None,
              gpu_id=None, importance_type=None, interaction_constraints=Non
e,
              learning_rate=None, max_delta_step=None, max_depth=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=None, reg_alpha=None,
              reg_lambda=None, scale_pos_weight=None, subsample=None,
              tree_method=None, validate_parameters=None, verbosity=None)

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

Save Experiment

In [22]:
```
jarvis.saveExperiment(myExp, FILE_NAME)
```

```
[CV] END ....................., score=(train=0.973, test=0.791) total time= 3.6
min
[CV] END ....................., score=(train=0.977, test=0.778) total time= 1.9
min
[CV] END ....................., score=(train=0.972, test=0.799) total time= 4.1
min
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
```

```
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.975, test=0.798) total time= 2.6
min
[CV] END ..................., score=(train=0.975, test=0.778) total time= 2.6
min
[CV] END ..................., score=(train=0.975, test=0.779) total time= 2.5
min
[CV] END ..................., score=(train=0.975, test=0.784) total time= 2.5
min
```

```
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.973, test=0.761) total time= 2.0
min
[CV] END ..................., score=(train=0.970, test=0.808) total time= 3.9
min
[CV] END ..................., score=(train=0.975, test=0.790) total time= 1.9
min
```

```
[CV] END ..................., score=(train=0.972, test=0.799) total time= 3.3
min
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.972, test=0.802) total time= 3.6
min
[CV] END ..................., score=(train=0.976, test=0.784) total time= 2.3
min
[CV] END ..................., score=(train=0.972, test=0.807) total time= 3.4
min
[CV] END ..................., score=(train=0.977, test=0.780) total time= 1.8
min
```

```
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
```

```
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
[CV] END ..................., score=(train=0.971, test=0.811) total time= 4.3
min
[CV] END ..................., score=(train=0.973, test=0.767) total time= 2.0
min
[CV] END ..................., score=(train=0.970, test=0.805) total time= 3.6
min
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.972, test=0.794) total time= 2.6
min
[CV] END ..................., score=(train=0.973, test=0.789) total time= 3.3
min
[CV] END ..................., score=(train=0.975, test=0.797) total time= 2.5
min
[CV] END ..................., score=(train=0.973, test=0.793) total time= 3.1
min
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
```

```
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.975, test=0.789) total time= 1.9
min
[CV] END ..................., score=(train=0.972, test=0.805) total time= 3.5
min
[CV] END ..................., score=(train=0.973, test=0.789) total time= 3.4
min
[CV] END ..................., score=(train=0.976, test=0.783) total time= 2.2
min
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.972, test=0.794) total time= 4.3
min
[CV] END ..................., score=(train=0.972, test=0.810) total time= 2.0
min
[CV] END ..................., score=(train=0.972, test=0.793) total time= 2.5
min
[CV] END ..................., score=(train=0.972, test=0.796) total time= 3.2
min
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
```

```
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.975, test=0.774) total time= 2.7
min
[CV] END ..................., score=(train=0.972, test=0.795) total time= 3.4
min
[CV] END ..................., score=(train=0.971, test=0.813) total time= 3.9
min
[CV] END ..................., score=(train=0.973, test=0.801) total time= 2.0
min
```

```
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
```

```
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.973, test=0.782) total time= 2.0
min
[CV] END ..................., score=(train=0.973, test=0.767) total time= 2.0
min
[CV] END ..................., score=(train=0.973, test=0.804) total time= 2.2
min
[CV] END ..................., score=(train=0.973, test=0.787) total time= 2.0
min
[CV] END ..................., score=(train=0.973, test=0.769) total time= 2.0
min
[CV] END ..................., score=(train=0.972, test=0.807) total time= 1.9
min
```

```
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
```

Scratchpad

In [ ]:

Configuration

In [1]:
```
# Parameters
ENABLE_COLAB = False

PROJECT_NAME = 'ML1010-Group-Project'
EXPERIMENT_NAME = 'ReviewText_Lemma_MPNet2 (XGB)'
FILE_NAME = '01_ML1010_GP_XGB_MPNet2'
LOAD_FROM_EXP = False

#Root Machine Learning Directory. Projects appear underneath
GOOGLE_DRIVE_MOUNT = '/content/gdrive'
COLAB_ROOT_DIR = GOOGLE_DRIVE_MOUNT + '/MyDrive/Colab Notebooks'
COLAB_INIT_DIR = COLAB_ROOT_DIR + '/utility_files'

LOCAL_ROOT_DIR = '/home/magni//ML_Root/project_root'
LOCAL_INIT_DIR = LOCAL_ROOT_DIR + '/utility_files'
```

Bootstrap Environment

In [2]:
```
#add in support for utility file directory and importing
import sys
import os

if ENABLE_COLAB:
  #Need access to drive
  from google.colab import drive
  drive.mount(GOOGLE_DRIVE_MOUNT, force_remount=True)

  #add in utility directory to syspath to import
  INIT_DIR = COLAB_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = COLAB_ROOT_DIR

else:
  #add in utility directory to syspath to import
  INIT_DIR = LOCAL_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = LOCAL_ROOT_DIR

#Import Utility Support
from jarvis import Jarvis
jarvis = Jarvis(ROOT_DIR, PROJECT_NAME)

import mv_python_utils as mvutils
```

```
Wha...where am I?
I am awake now.
```

```
I have set your current working directory to /home/magni/ML_Root/project_root
/ML1010-Group-Project
The current time is 11:02
Hello sir. Extra caffeine may help.
```

Setup Runtime Environment

In [3]:

```
if ENABLE_COLAB:
  #!pip install scipy -q
  #!pip install scikit-learn -q
  #!pip install pycaret -q
  #!pip install matplotlib -q
  #!pip install joblib -q
  #!pip install pandasql -q
  !pip install umap_learn -q
  !pip install sentence_transformers -q
  !pip install spacytextblob -q
  !pip install flair -q
  display('Google Colab enabled')
else:
  display('Google Colab not enabled')

#Common imports
import json
import pandas as pd
import numpy as np
import matplotlib
import re
import nltk
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn import metrics
from sklearn.datasets import load_digits
from sklearn.model_selection import train_test_split as tts
#from yellowbrick.classifier import ConfusionMatrix
#from sklearn.linear_model import LogisticRegression
from yellowbrick.target import ClassBalance
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier

nltk.download('stopwords')
%matplotlib inline
```

```
'Google Colab not enabled'
```

```
[nltk_data] Downloading package stopwords to /home/magni/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [23]:
```
import importlib
import cw_df_metric_utils as cwutils
import DataPackage as dp
import DataPackageSupport as dps
import DataExperiment
import DataExperimentSupport
```

In [24]:
```
importlib.reload(dp)
importlib.reload(dps)
importlib.reload(DataExperiment)
importlib.reload(DataExperimentSupport)
```

Out[24]:
```
<module 'DataExperimentSupport' from '/home/magni/ML_Root/project_root/utilit
y_files/DataExperimentSupport.py'>
```

Load Data

In [5]:
```
#axis_labels=[1,2,3,4,5]
axis_labels=[0,1]
#classifier = RandomForestClassifier()
classifier = XGBClassifier(eval_metric='mlogloss')
ANALSYSIS_COL = 'reviewText_lemma_mpnet'
UNIQUE_COL = 'uuid'
TARGET_COL = 'overall_posneg'
```

In [6]:
```
if LOAD_FROM_EXP:
    #start from saved state
    myExp = jarvis.loadExperiment(FILE_NAME)
    myExp.display()

else:
    #start from source file and regenerate
    testDf = pd.read_pickle(jarvis.DATA_DIR_WORK + "/01_NL_ReviewText_All(new

    testDfBert = cwutils.getBertEncodeFrame(df=testDf,
                                            bertColumn=ANALSYSIS_COL,
                                            uniqueColumn=UNIQUE_COL,
                                            otherColumns=[TARGET_COL]
                                            )



    myExp = DataExperiment.DataExperiment(projectName=PROJECT_NAME,
                                          experimentName=EXPERIMENT_NAME,
                                          origData=testDfBert,
                                          uniqueColumn=UNIQUE_COL,
                                          targetColumn=TARGET_COL,
                                          classifier=classifier)
```

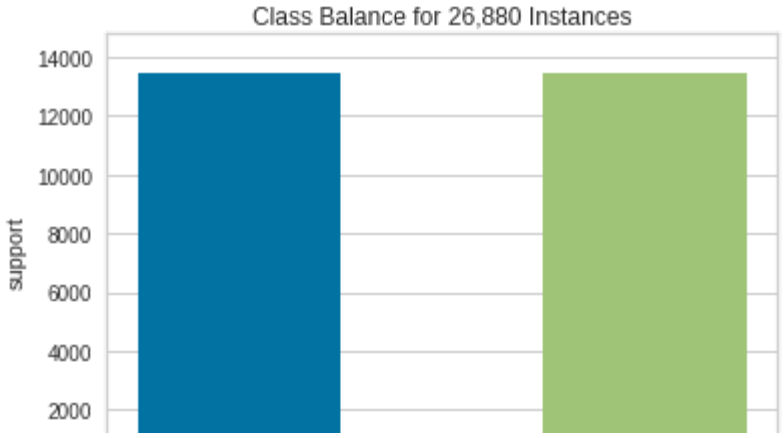
```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_MPNet2 (XGB)
---> isDataPackageLoaded: True
```

```
---> isBaseModelLoaded: False
---> isBaseModelPredicted: False
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: False
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
XGBClassifier(base_score=None, booster=None, colsample_bylevel=None,
              colsample_bynode=None, colsample_bytree=None,
              enable_categorical=False, eval_metric='mlogloss', gamma=None,
              gpu_id=None, importance_type=None, interaction_constraints=Non
e,
              learning_rate=None, max_delta_step=None, max_depth=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=None, reg_alpha=None,
              reg_lambda=None, scale_pos_weight=None, subsample=None,
              tree_method=None, validate_parameters=None, verbosity=None)

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: False
    ---> isTrainTestSplit: False
    Data:
    ---> isOrigDataLoaded: True
    ---> isTrainDataLoaded: False
    ---> isTestDataLoaded: False
```

In [7]:

```
myExp.processDataPackage()
```


```
Undersampling data to match min class: 0 of size: 13440
```



In [8]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_MPNet2 (XGB)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: False
---> isBaseModelPredicted: False
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: False
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
XGBClassifier(base_score=None, booster=None, colsample_bylevel=None,
              colsample_bynode=None, colsample_bytree=None,
              enable_categorical=False, eval_metric='mlogloss', gamma=None,
              gpu_id=None, importance_type=None, interaction_constraints=Non
e,
              learning_rate=None, max_delta_step=None, max_depth=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=None, reg_alpha=None,
              reg_lambda=None, scale_pos_weight=None, subsample=None,
              tree_method=None, validate_parameters=None, verbosity=None)

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

In [9]:
```
myExp.createBaseModel()
```

```
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
```

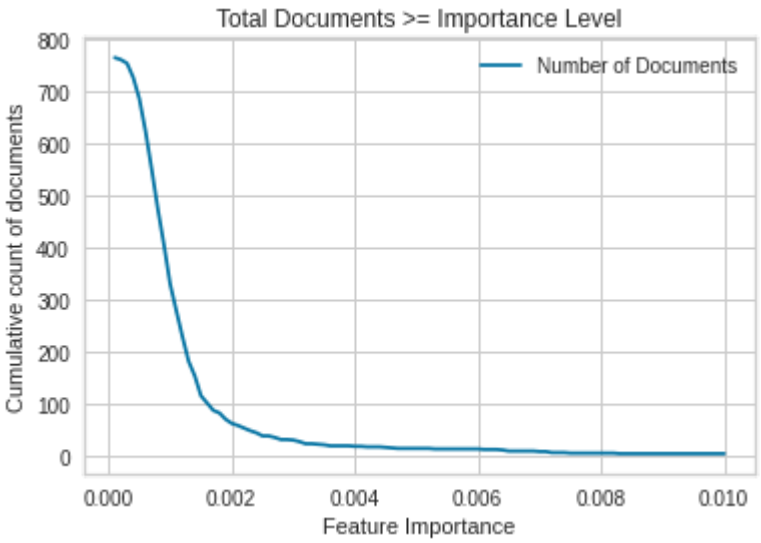
```
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
```

In [10]:
```
myExp.predictBaseModel()
```

```
Base Model Stats:
Accuracy: 0.81
Precision: 0.81
Recalll: 0.81
F1 Score: 0.81
Cohen kappa:: 0.63
```

In [11]:
```
impFeatures = myExp.analyzeBaseModelFeatureImportance(returnAbove=0.002)
```

```
  0%|          | 0/101 [00:00<?, ?it/s]
Feature Importance Summary:
---> Original feature count: 768
---> Returned feature count: 63
---> Removed feature count: 705
---> Return items above (including): 0.002
```



In [12]:
```
myExp.createFinalModel(featureImportanceThreshold=0.002)
```

```
  0%|          | 0/101 [00:00<?, ?it/s]
  0%|          | 0/101 [00:00<?, ?it/s]
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
```

In [13]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
```

```
---> experimentName: ReviewText_Lemma_MPNet2 (XGB)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: True
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
XGBClassifier(base_score=None, booster=None, colsample_bylevel=None,
              colsample_bynode=None, colsample_bytree=None,
              enable_categorical=False, eval_metric='mlogloss', gamma=None,
              gpu_id=None, importance_type=None, interaction_constraints=Non
e,
              learning_rate=None, max_delta_step=None, max_depth=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=None, reg_alpha=None,
              reg_lambda=None, scale_pos_weight=None, subsample=None,
              tree_method=None, validate_parameters=None, verbosity=None)

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
```

In [14]:

```
myExp.predictFinalModel()
myExp.display()
```

```
Final Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.61
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_MPNet2 (XGB)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
XGBClassifier(base_score=None, booster=None, colsample_bylevel=None,
              colsample_bynode=None, colsample_bytree=None,
              enable_categorical=False, eval_metric='mlogloss', gamma=None,
              gpu_id=None, importance_type=None, interaction_constraints=Non
e,
              learning_rate=None, max_delta_step=None, max_depth=None,
```

```
                min_child_weight=None, missing=nan, monotone_constraints=None,
                n_estimators=100, n_jobs=None, num_parallel_tree=None,
                predictor=None, random_state=None, reg_alpha=None,
                reg_lambda=None, scale_pos_weight=None, subsample=None,
                tree_method=None, validate_parameters=None, verbosity=None)

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

In [15]:
```
myExp.createBaseModelLearningCurve(n_jobs=10)
```

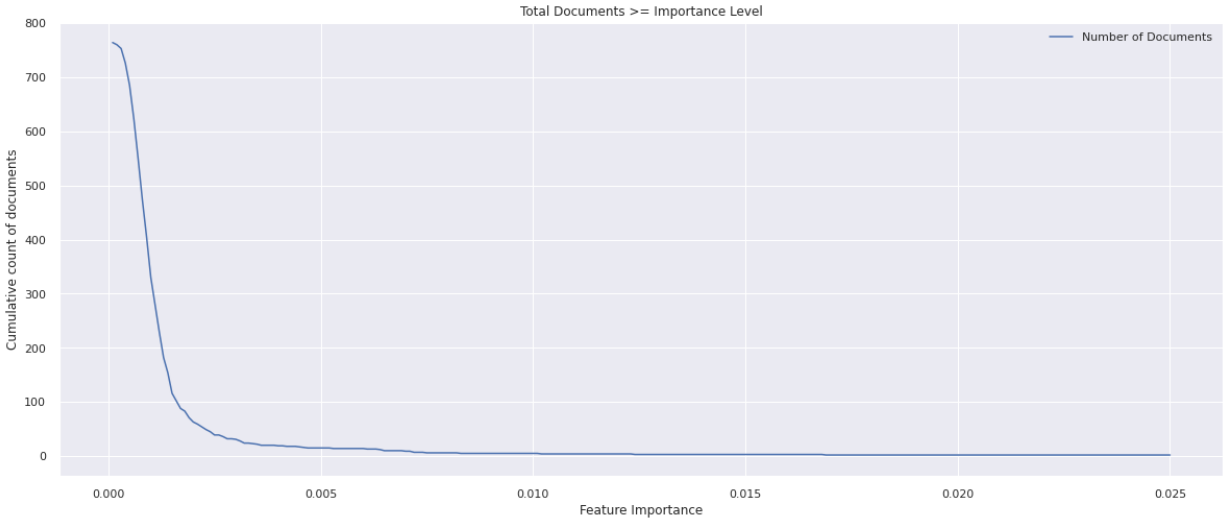
```
[learning_curve] Training set sizes: [ 1720  3440  8601 17203]
[Parallel(n_jobs=10)]: Using backend LokyBackend with 10 concurrent workers.
[Parallel(n_jobs=10)]: Done   3 out of  20 | elapsed:  2.2min remaining: 12.4
min
[Parallel(n_jobs=10)]: Done   9 out of  20 | elapsed:  4.6min remaining:  5.7
min
[Parallel(n_jobs=10)]: Done  15 out of  20 | elapsed:  6.9min remaining:  2.3
min
[Parallel(n_jobs=10)]: Done  20 out of  20 | elapsed:  7.6min finished
```

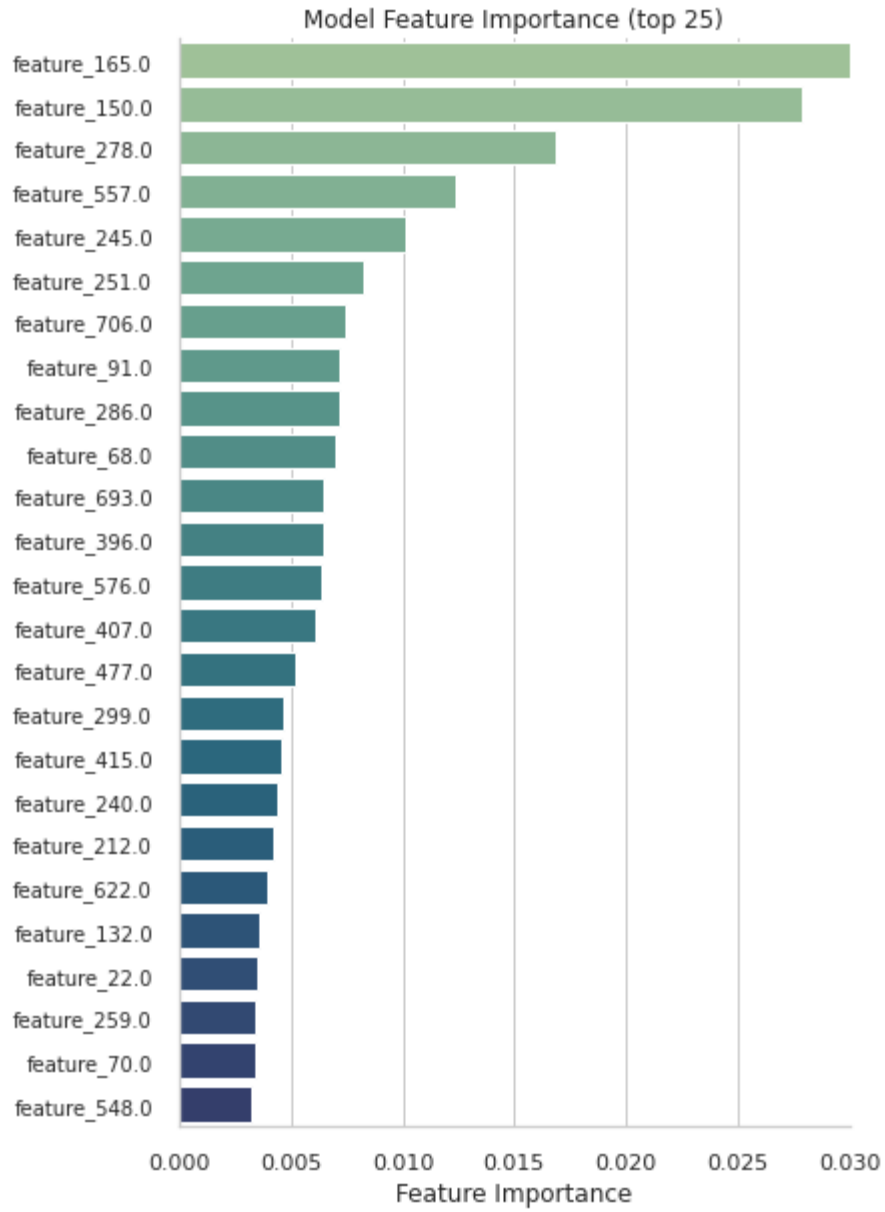
In [16]:
```
myExp.createFinalModelLearningCurve(n_jobs=10)
```

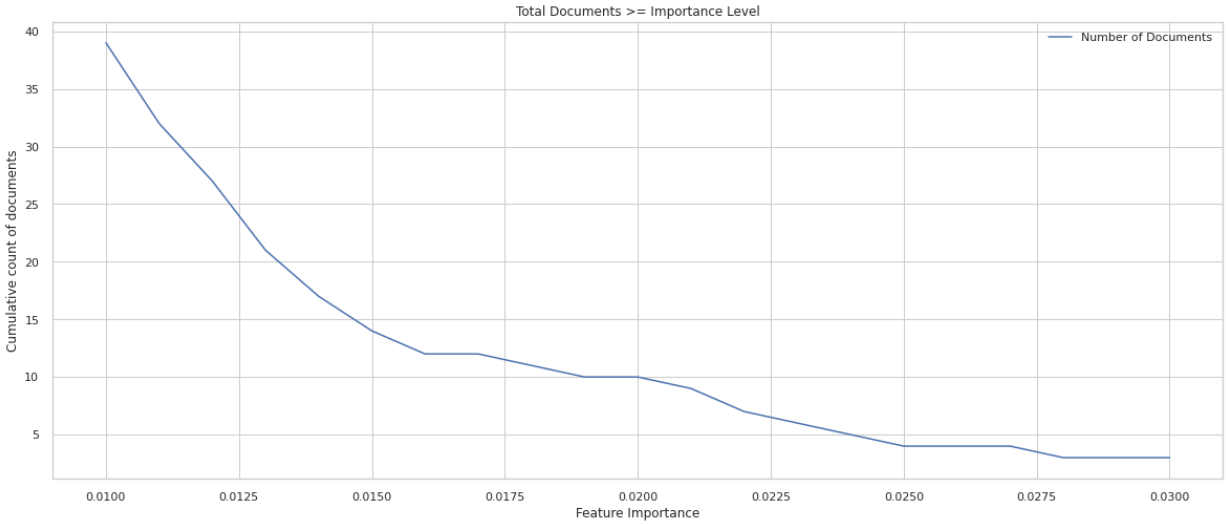
```
[Parallel(n_jobs=10)]: Using backend LokyBackend with 10 concurrent workers.
[learning_curve] Training set sizes: [ 1720  3440  8601 17203]
[Parallel(n_jobs=10)]: Done   3 out of  20 | elapsed:  1.8min remaining: 10.1
min
[Parallel(n_jobs=10)]: Done   9 out of  20 | elapsed:  2.7min remaining:  3.3
min
[Parallel(n_jobs=10)]: Done  15 out of  20 | elapsed:  4.2min remaining:  1.4
min
[Parallel(n_jobs=10)]: Done  20 out of  20 | elapsed:  4.4min finished
```

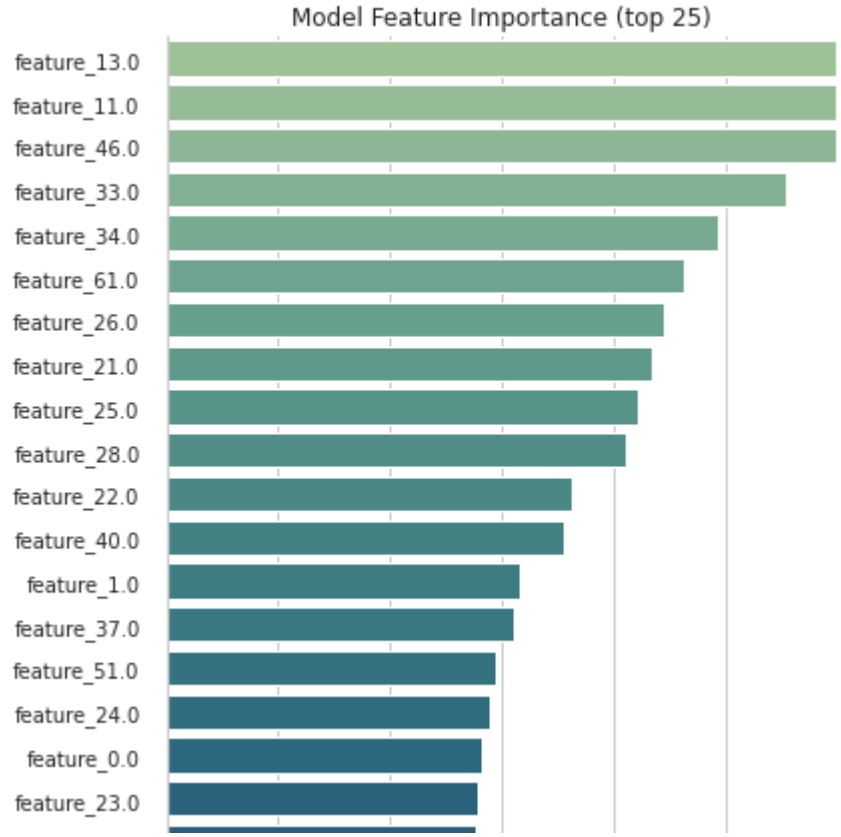
In [25]:
```
myExp.showBaseModelFeatureImportance(upperValue=0.025)
myExp.showFinalModelFeatureImportance(startValue=0.01,
                                      increment=0.001,
                                      upperValue=0.03)
```

```
  0%|          | 0/251 [00:00<?, ?it/s]
  0%|          | 0/22 [00:00<?, ?it/s]
```


Total Documents >= Importance Level
Number of Documents
Cumulative count of documents
800
700
600
500
400
300
200
100
0
0.000
0.005
0.010
0.015
0.020
0.025
Feature Importance


Model Feature Importance (top 25)
feature_165.0
feature_150.0
feature_278.0
feature_557.0
feature_245.0
feature_251.0
feature_706.0
feature_91.0
feature_286.0
feature_68.0
feature_693.0
feature_396.0
feature_576.0
feature_407.0
feature_477.0
feature_299.0
feature_415.0
feature_240.0
feature_212.0
feature_622.0
feature_132.0
feature_22.0
feature_259.0
feature_70.0
feature_548.0
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Feature Importance





In [18]:
```
myExp.display()
```

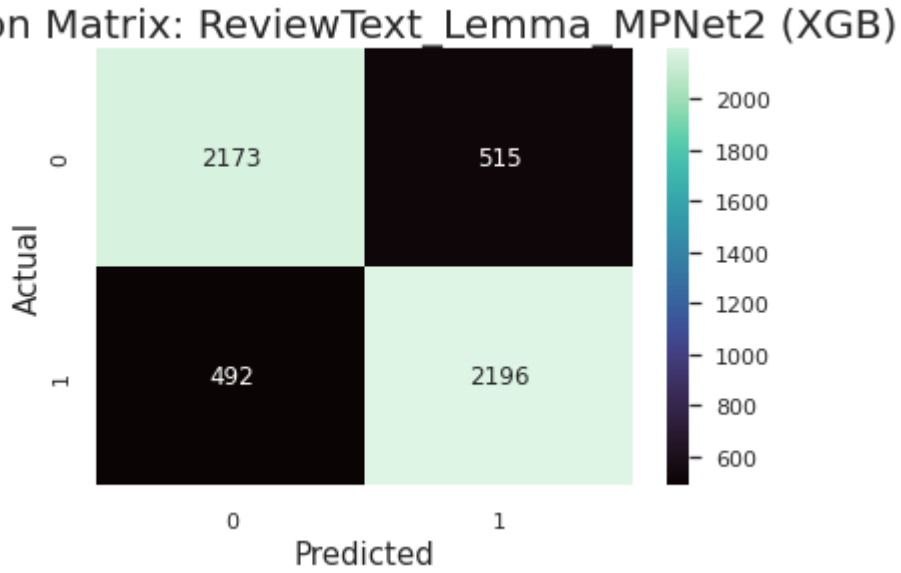
```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_MPNet2 (XGB)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: True
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: True
---> isClassifierLoaded: True
XGBClassifier(base_score=None, booster=None, colsample_bylevel=None,
```

```
              colsample_bynode=None, colsample_bytree=None,
              enable_categorical=False, eval_metric='mlogloss', gamma=None,
              gpu_id=None, importance_type=None, interaction_constraints=Non
e,
              learning_rate=None, max_delta_step=None, max_depth=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=None, reg_alpha=None,
              reg_lambda=None, scale_pos_weight=None, subsample=None,
              tree_method=None, validate_parameters=None, verbosity=None)

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
```

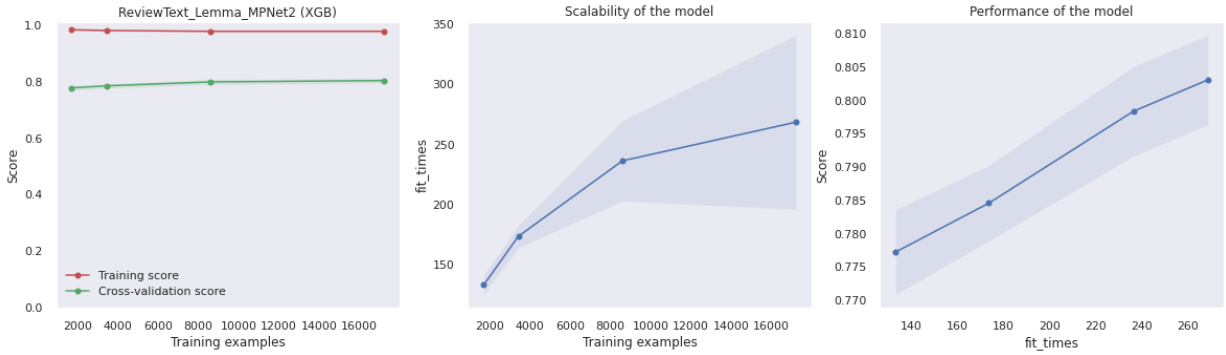
In [19]:

```
myExp.showBaseModelReport(axis_labels)
```

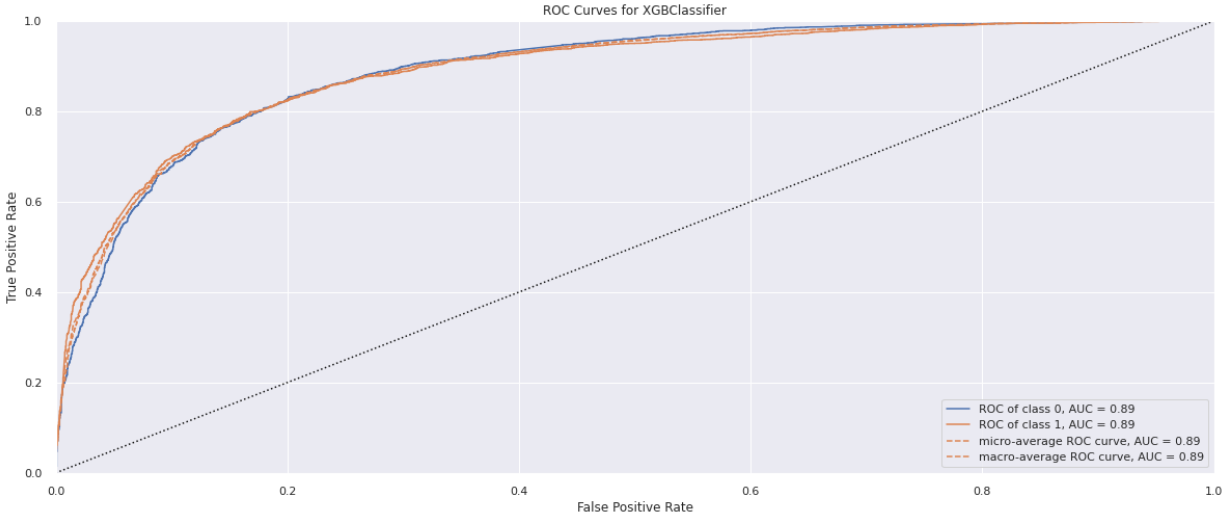
```
Base Model Stats:
Accuracy: 0.81
Precision: 0.81
Recalll: 0.81
F1 Score: 0.81
Cohen kappa:: 0.63
              precision    recall  f1-score   support

           0       0.82      0.81      0.81      2688
           1       0.81      0.82      0.81      2688

    accuracy                           0.81      5376
   macro avg       0.81      0.81      0.81      5376
weighted avg       0.81      0.81      0.81      5376
```




```
<Figure size 1440x576 with 0 Axes>
```

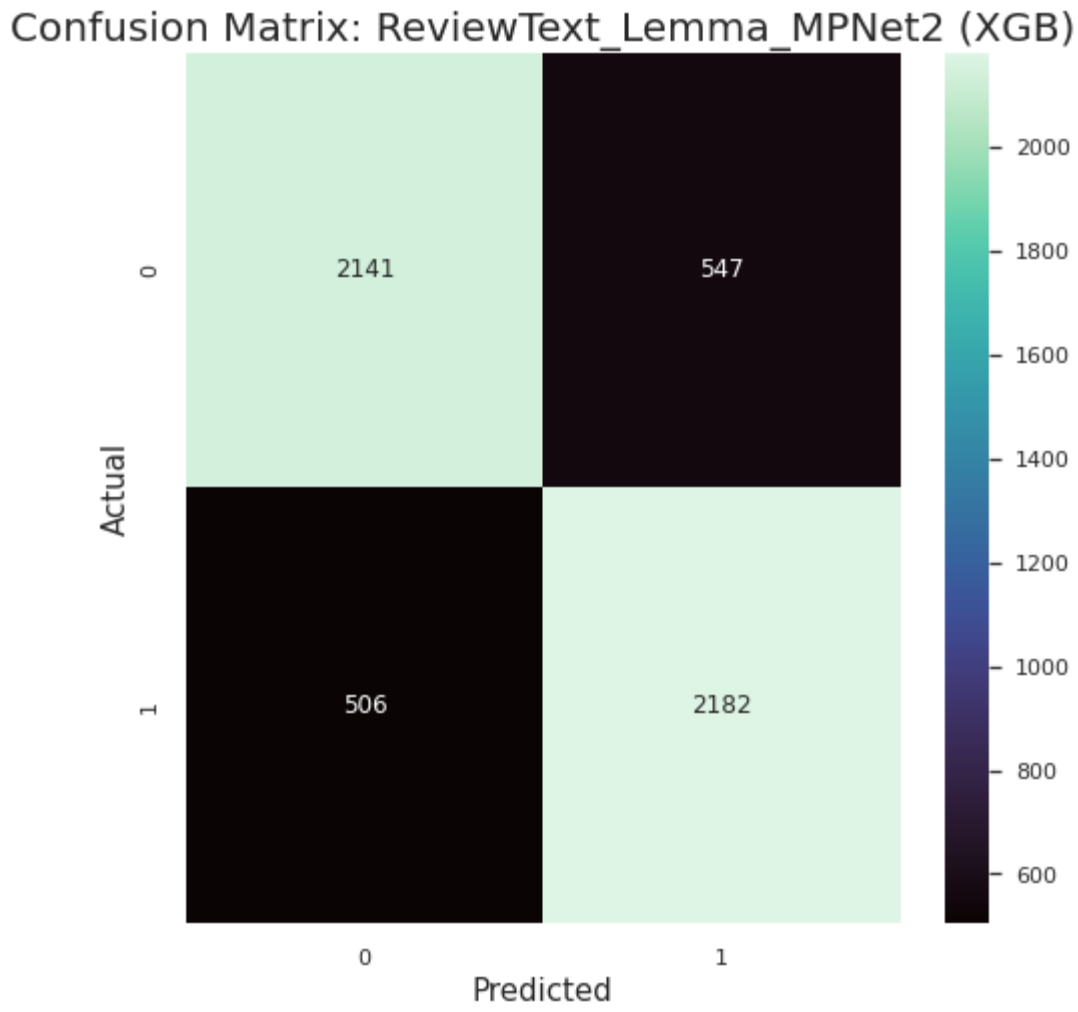



```
Base model ROCAUC not calculated. Starting now
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/sklearn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprecated and will be removed in a future release. To remove this warning, do the following: 1) Pass option use_label_encoder=False when constructing XGBClassifier object; and 2) Encode your labels (y) as integers starting with 0, i.e. 0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
```

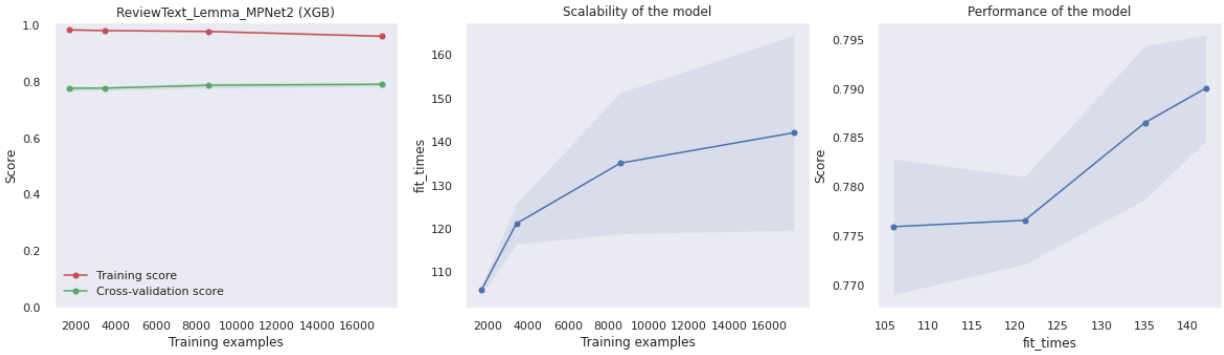


In [27]:
```
myExp.showFinalModelReport(axis_labels)
```

```
Final Model Stats:
Accuracy: 0.8
Precision: 0.8
Recalll: 0.8
F1 Score: 0.8
Cohen kappa:: 0.61
              precision    recall  f1-score   support

           0       0.81      0.80      0.80      2688
           1       0.80      0.81      0.81      2688

    accuracy                           0.80      5376
   macro avg       0.80      0.80      0.80      5376
weighted avg       0.80      0.80      0.80      5376
```

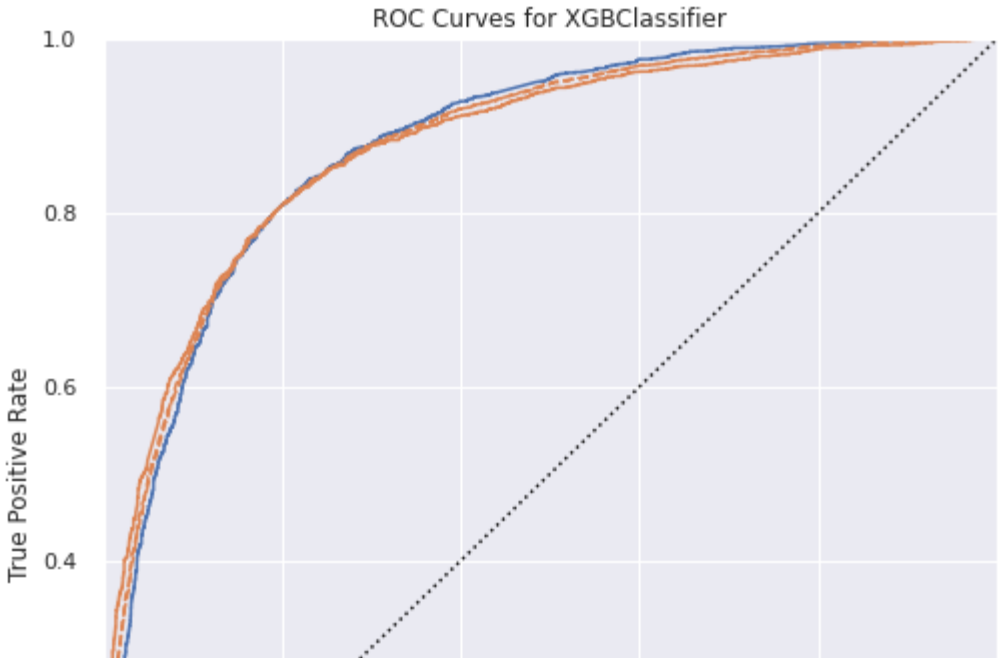


<Figure size 576x576 with 0 Axes>



Final model ROCAUC not calculated. Starting now

/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/sklearn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprecated and will be removed in a future release. To remove this warning, do the following: 1) Pass option use_label_encoder=False when constructing XGBClassifier object; and 2) Encode your labels (y) as integers starting with 0, i.e. 0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)



In [21]:
```
myExp.display()
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_MPNet2 (XGB)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: True
---> isBaseModelPredicted: True
---> isBaseModelLearningCurveCreated: True
---> isFinalModelLoaded: True
---> isFinalModelPredicted: True
---> isFinalModelLearningCurveCreated: True
---> isClassifierLoaded: True
XGBClassifier(base_score=None, booster=None, colsample_bylevel=None,
              colsample_bynode=None, colsample_bytree=None,
              enable_categorical=False, eval_metric='mlogloss', gamma=None,
              gpu_id=None, importance_type=None, interaction_constraints=Non
e,
              learning_rate=None, max_delta_step=None, max_depth=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=None, reg_alpha=None,
              reg_lambda=None, scale_pos_weight=None, subsample=None,
              tree_method=None, validate_parameters=None, verbosity=None)

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: True
    ---> isTrainTestSplit: True
    Data:
    ---> isOrigDataLoaded: False
    ---> isTrainDataLoaded: True
    ---> isTestDataLoaded: True
```

Save Experiment

In [22]:

```
jarvis.saveExperiment(myExp, FILE_NAME)
```

```
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.983, test=0.787) total time= 2.2
min
[CV] END ..................., score=(train=0.983, test=0.768) total time= 2.5
min
[CV] END ..................., score=(train=0.977, test=0.805) total time= 2.8
min
[CV] END ..................., score=(train=0.984, test=0.766) total time= 1.8
min
[CV] END ..................., score=(train=0.983, test=0.770) total time= 1.8
min
```

```
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
```

```
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.978, test=0.806) total time= 5.4
min
[CV] END ..................., score=(train=0.977, test=0.812) total time= 2.2
min
[CV] END ..................., score=(train=0.979, test=0.774) total time= 2.0
min
[CV] END ..................., score=(train=0.981, test=0.772) total time= 2.0
min

/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.978, test=0.806) total time= 5.5
min
[CV] END ..................., score=(train=0.983, test=0.781) total time= 1.7
min
[CV] END ..................., score=(train=0.977, test=0.783) total time= 2.4
min
[CV] END ..................., score=(train=0.978, test=0.799) total time= 4.2
min
[CV] END ..................., score=(train=0.983, test=0.777) total time= 2.3
```

```
min
[CV] END ..................., score=(train=0.977, test=0.792) total time= 2.4
min
[CV] END ..................., score=(train=0.983, test=0.785) total time= 1.8
min
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.977, test=0.807) total time= 4.2
min
[CV] END ..................., score=(train=0.980, test=0.792) total time= 2.7
min
[CV] END ..................., score=(train=0.984, test=0.778) total time= 1.8
min
[CV] END ..................., score=(train=0.957, test=0.782) total time= 2.5
min
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
```

```
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.981, test=0.791) total time= 2.8
min
[CV] END ..................., score=(train=0.981, test=0.782) total time= 3.2
min
[CV] END ..................., score=(train=0.981, test=0.781) total time= 2.1
min
[CV] END ..................., score=(train=0.978, test=0.774) total time= 2.2
min

/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.984, test=0.774) total time= 2.1
min
[CV] END ..................., score=(train=0.977, test=0.794) total time= 4.9
min
[CV] END ..................., score=(train=0.978, test=0.788) total time= 2.5
min
[CV] END ..................., score=(train=0.980, test=0.783) total time= 1.9
min

/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
```

```
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
[CV] END ..................., score=(train=0.979, test=0.780) total time= 2.9
min
[CV] END ..................., score=(train=0.978, test=0.792) total time= 4.1
min
[CV] END ..................., score=(train=0.982, test=0.773) total time= 2.1
min
[CV] END ..................., score=(train=0.961, test=0.786) total time= 2.3
```

```
min
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
[CV] END ..................., score=(train=0.984, test=0.781) total time= 2.1
min
[CV] END ..................., score=(train=0.977, test=0.790) total time= 4.4
min
[CV] END ..................., score=(train=0.961, test=0.792) total time= 2.7
min
[CV] END ..................., score=(train=0.977, test=0.797) total time= 1.7
min
[CV] END ..................., score=(train=0.982, test=0.778) total time= 2.9
min
[CV] END ..................., score=(train=0.978, test=0.797) total time= 4.4
min
[CV] END ..................., score=(train=0.962, test=0.792) total time= 2.7
min
[CV] END ..................., score=(train=0.961, test=0.798) total time= 1.7
min
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
```

```
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
/home/magni/python_env/ML1010_env2/lib64/python3.7/site-packages/xgboost/skle
arn.py:1224: UserWarning: The use of label encoder in XGBClassifier is deprec
ated and will be removed in a future release. To remove this warning, do the
following: 1) Pass option use_label_encoder=False when constructing XGBClassi
fier object; and 2) Encode your labels (y) as integers starting with 0, i.e.
0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
```

Scratchpad

In [ ]:

Configuration

In [1]:
```
# Parameters
ENABLE_COLAB = False

PROJECT_NAME = 'ML1010-Group-Project'
EXPERIMENT_NAME = 'ReviewText_Lemma_Bert2 (LSTM)'
FILE_NAME = '01_ML1010_GP_LSTM_Bert2'
LOAD_FROM_EXP = False

#Root Machine Learning Directory. Projects appear underneath
GOOGLE_DRIVE_MOUNT = '/content/gdrive'
COLAB_ROOT_DIR = GOOGLE_DRIVE_MOUNT + '/MyDrive/Colab Notebooks'
COLAB_INIT_DIR = COLAB_ROOT_DIR + '/utility_files'

LOCAL_ROOT_DIR = '/home/magni//ML_Root/project_root'
LOCAL_INIT_DIR = LOCAL_ROOT_DIR + '/utility_files'
```

Bootstrap Environment

In [2]:
```
#add in support for utility file directory and importing
import sys
import os

if ENABLE_COLAB:
  #Need access to drive
  from google.colab import drive
  drive.mount(GOOGLE_DRIVE_MOUNT, force_remount=True)

  #add in utility directory to syspath to import
  INIT_DIR = COLAB_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = COLAB_ROOT_DIR

else:
  #add in utility directory to syspath to import
  INIT_DIR = LOCAL_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = LOCAL_ROOT_DIR

#Import Utility Support
from jarvis import Jarvis
jarvis = Jarvis(ROOT_DIR, PROJECT_NAME)

import mv_python_utils as mvutils
```

```
Wha...where am I?
I am awake now.
```

```
I have set your current working directory to /home/magni/ML_Root/project_root
/ML1010-Group-Project
The current time is 11:24
Hello sir. Extra caffeine may help.
```

Setup Runtime Environment

In [3]:

```
if ENABLE_COLAB:
  #!pip install scipy -q
  #!pip install scikit-learn -q
  #!pip install pycaret -q
  #!pip install matplotlib -q
  #!pip install joblib -q
  #!pip install pandasql -q
  !pip install umap_learn -q
  !pip install sentence_transformers -q
  !pip install spacytextblob -q
  !pip install flair -q
  display('Google Colab enabled')
else:
  display('Google Colab not enabled')

#Common imports
import json
import pandas as pd
import numpy as np
import matplotlib
import re
import nltk
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn import metrics
from sklearn.datasets import load_digits
from sklearn.model_selection import train_test_split as tts
#from yellowbrick.classifier import ConfusionMatrix
#from sklearn.linear_model import LogisticRegression
from yellowbrick.target import ClassBalance
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.dummy import DummyClassifier

nltk.download('stopwords')
%matplotlib inline
```

```
'Google Colab not enabled'
```

```
[nltk_data] Downloading package stopwords to /home/magni/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [4]:
```
import cw_df_metric_utils as cwutils
import importlib
import DataPackage as dp
import DataPackageSupport as dps
import DataExperiment
import DataExperimentSupport as des
```

```
2022-01-15 11:24:26.143562: W tensorflow/stream_executor/platform/default/dso
_loader.cc:64] Could not load dynamic library 'libcudart.so.11.0'; dlerror: l
ibcudart.so.11.0: cannot open shared object file: No such file or directory
2022-01-15 11:24:26.143588: I tensorflow/stream_executor/cuda/cudart_stub.cc:
29] Ignore above cudart dlerror if you do not have a GPU set up on your machi
ne.
```

Load Data

In [5]:
```
#axis_labels=[1,2,3,4,5]
axis_labels=[0,1]

# using a dummyclassifier as DataExperiment requires a classifier to load
# and doesn't fully support Tensorflow models yet
classifier = DummyClassifier()
ANALSYSIS_COL = 'reviewText_lemma_bert'
UNIQUE_COL = 'uuid'
TARGET_COL = 'overall_posneg'
```

In [6]:
```
if LOAD_FROM_EXP:
    #start from saved state
    myExp = jarvis.loadExperiment(FILE_NAME)
    myExp.display()

else:
    #start from source file and regenerate
    testDf = pd.read_pickle(jarvis.DATA_DIR_WORK + "/01_NL_ReviewText_All(new

    testDfBert = cwutils.getBertEncodeFrame(df=testDf,
                                            bertColumn=ANALSYSIS_COL,
                                            uniqueColumn=UNIQUE_COL,
                                            otherColumns=[TARGET_COL]
                                            )


    myExp = DataExperiment.DataExperiment(projectName=PROJECT_NAME,
                                          experimentName=EXPERIMENT_NAME,
                                          origData=testDfBert,
                                          uniqueColumn=UNIQUE_COL,
                                          targetColumn=TARGET_COL,
                                          classifier=classifier)
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Bert2 (LSTM)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: False
```

```
---> isBaseModelPredicted: False
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: False
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
DummyClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: False
    ---> isTrainTestSplit: False
    Data:
    ---> isOrigDataLoaded: True
    ---> isTrainDataLoaded: False
```

In [7]:

```
#get the train data and downsample to 2900
tDf = myExp.dataPackage.getOrigData()
dps.displayClassBalance(tDf, myExp.dataPackage.targetColumn, verbose=True)
```

	overall_posneg	ttlCol
0	0	13440
1	1	49973

In [8]:

```
myExp.processDataPackage()
```


```
Undersampling data to match min class: 0 of size: 13440
```


```
Completed train/test split (test_size = 0.2):
---> Original data size: 26880
---> Training data size: 21504
---> Testing data size: 5376
---> Stratified on column: overall_posneg
```

In [9]:

```
tDf2 = myExp.dataPackage.getTrainData()
dps.displayClassBalance(tDf2, myExp.dataPackage.targetColumn, verbose=True)
```

	overall_posneg	ttlCol
0	0	10752
1	1	10752

In [10]:
```
SAMPLE_DOWN_SIZE=10700
# Do the sampling
tDf2 = tDf2.groupby(myExp.dataPackage.targetColumn, group_keys=False).apply(l
tDf2.reset_index(drop=True, inplace=True)
```

In [11]:
```
dps.displayClassBalance(tDf2, myExp.dataPackage.targetColumn, verbose=True)
```

	overall_posneg	ttlCol
0	0	10700
1	1	10700

In [12]:
```
from xgboost import XGBClassifier
from keras.layers.core import SpatialDropout1D
from keras.layers import Dropout, Dense, Flatten, LSTM, Input, Conv1D, MaxPoo
from keras.models import Sequential
from keras.backend import clear_session
from keras.layers.embeddings import Embedding
import keras

print(keras.__version__)
from keras import backend as K
K._get_available_gpus()
```

```
2.7.0
2022-01-15 11:25:38.346075: W tensorflow/stream_executor/platform/default/dso
_loader.cc:64] Could not load dynamic library 'libcuda.so.1'; dlerror: libcud
a.so.1: cannot open shared object file: No such file or directory
2022-01-15 11:25:38.346103: W tensorflow/stream_executor/cuda/cuda_driver.cc:
269] failed call to cuInit: UNKNOWN ERROR (303)
2022-01-15 11:25:38.346120: I tensorflow/stream_executor/cuda/cuda_diagnostic
s.cc:156] kernel driver does not appear to be running on this host (localhos
t.localdomain): /proc/driver/nvidia/version does not exist
2022-01-15 11:25:38.346419: I tensorflow/core/platform/cpu_feature_guard.cc:1
51] This TensorFlow binary is optimized with oneAPI Deep Neural Network Libra
ry (oneDNN) to use the following CPU instructions in performance-critical ope
rations:  AVX2 FMA
To enable them in other operations, rebuild TensorFlow with the appropriate c
ompiler flags.
```

Out[12]:
```
[]
```

In [20]:
```
from tensorflow.keras.metrics import AUC, Precision, Recall
```

In [33]:
```
tDf3 = tDf2.copy()
Y_train = np.array(tDf3[myExp.dataPackage.targetColumn])
tDf3.drop(myExp.dataPackage.uniqueColumn, axis=1, inplace=True)
tDf3.drop(myExp.dataPackage.targetColumn, axis=1, inplace=True)
X_train = np.array(tDf3)

#Are the numbers what we think they are?
print(len(tDf3.columns))
print(Y_train.shape)
print(X_train.shape)


EPOCHS=5
VAL_SPLIT=0.1

BATCH_SIZE=100
NUMBER_FEATURES=len(tDf3.columns)

DROPOUT_RATE=0.2
INTERNAL_LAYERS=100
LSTM_OUTPUT_UNITS=100
```

```
768
```

```
(21400,)
```

In [34]:

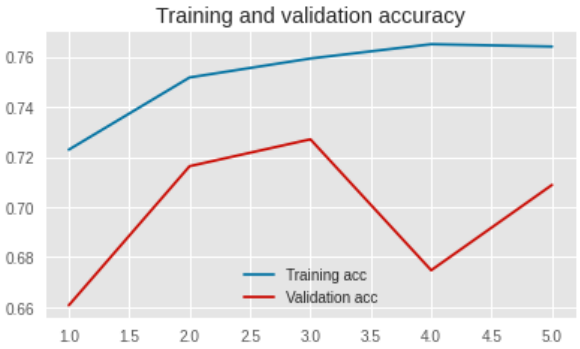
```
# Neural network
keras.backend.clear_session()

model2 = None
model2 = Sequential()

#model2.add(Input(shape=(NUMBER_FEATURES, 1)))
#model2.add(Dense(INTERNAL_LAYERS, activation='relu'))
#model2.add(Dropout(DROPOUT_RATE))

model2.add(LSTM(units=LSTM_OUTPUT_UNITS,
                input_shape=(NUMBER_FEATURES, 1),
                return_sequences=False
               )
          )
#model2.add(Dropout(DROPOUT_RATE))
#model2.add(Conv1D(filters=LSTM_OUTPUT_UNITS, kernel_size=3, padding='same',
#model2.add(MaxPooling1D(pool_size=2))
#model2.add(LSTM(units=LSTM_OUTPUT_UNITS))
#model2.add(Dropout(DROPOUT_RATE))


#model2.add(LSTM(units=LSTM_OUTPUT_UNITS))
#model2.add(Dropout(DROPOUT_RATE))

model2.add(Dense(INTERNAL_LAYERS, activation='relu'))
model2.add(Dropout(DROPOUT_RATE))

#model2.add(Dense(LSTM_OUTPUT_UNITS, activation='relu'))
#model2.add(Dropout(DROPOUT_RATE))

#model2.add(Dense(INTERNAL_LAYERS, activation='relu'))
#model2.add(Dropout(DROPOUT_RATE))

#model2.add(Dense(LSTM_OUTPUT_UNITS, activation='relu'))
#model2.add(Dropout(DROPOUT_RATE))


#softmax is for multiclass
#model2.add(Dense(1, activation='softmax'))
#----
#sigmoid is not for multiclass
model2.add(Dense(1, activation='sigmoid'))

model2.compile(loss='binary_crossentropy',
               optimizer='adam',
               metrics=['accuracy',
                        'mse',
                        AUC(),
                        Precision(),
                        Recall()
                       ]
              )

history = model2.fit(x=X_train,
                     y=Y_train,
                     epochs=EPOCHS,
                     batch_size=BATCH_SIZE,
```

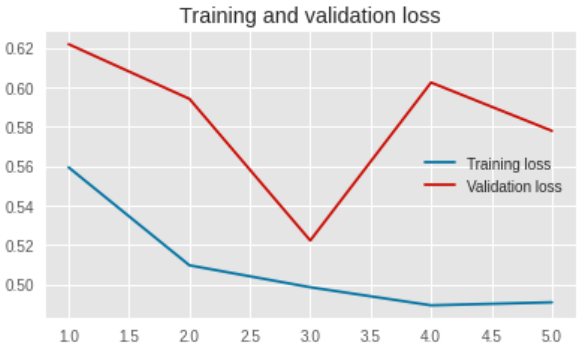
```
Epoch 1/5
193/193 [==============================] - 101s 514ms/step - loss: 0.5593 - a
ccuracy: 0.7229 - mse: 0.1890 - auc: 0.7815 - precision: 0.7012 - recall: 0.6
562 - val_loss: 0.6218 - val_accuracy: 0.6607 - val_mse: 0.2163 - val_auc: 0.
0000e+00 - val_precision: 1.0000 - val_recall: 0.6607
Epoch 2/5
193/193 [==============================] - 98s 509ms/step - loss: 0.5096 - ac
curacy: 0.7518 - mse: 0.1683 - auc: 0.8256 - precision: 0.7343 - recall: 0.69
19 - val_loss: 0.5940 - val_accuracy: 0.7164 - val_mse: 0.1993 - val_auc: 0.0
000e+00 - val_precision: 1.0000 - val_recall: 0.7164
Epoch 3/5
193/193 [==============================] - 96s 498ms/step - loss: 0.4985 - ac
curacy: 0.7594 - mse: 0.1642 - auc: 0.8342 - precision: 0.7456 - recall: 0.69
63 - val_loss: 0.5222 - val_accuracy: 0.7271 - val_mse: 0.1752 - val_auc: 0.0
000e+00 - val_precision: 1.0000 - val_recall: 0.7271
Epoch 4/5
193/193 [==============================] - 97s 501ms/step - loss: 0.4893 - ac
curacy: 0.7651 - mse: 0.1608 - auc: 0.8407 - precision: 0.7547 - recall: 0.69
85 - val_loss: 0.6023 - val_accuracy: 0.6748 - val_mse: 0.2085 - val_auc: 0.0
000e+00 - val_precision: 1.0000 - val_recall: 0.6748
Epoch 5/5
193/193 [==============================] - 97s 503ms/step - loss: 0.4909 - ac
curacy: 0.7642 - mse: 0.1611 - auc: 0.8397 - precision: 0.7561 - recall: 0.69
30 - val_loss: 0.5778 - val_accuracy: 0.7089 - val_mse: 0.1956 - val_auc: 0.0
000e+00 - val_precision: 1.0000 - val_recall: 0.7089
```

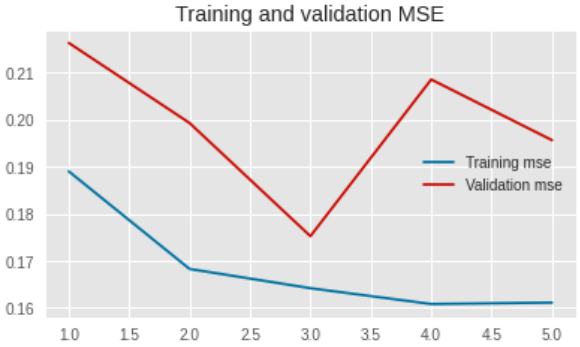
In [35]:

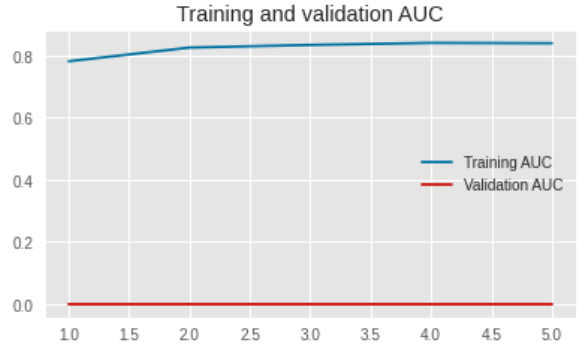
```
import matplotlib.pyplot as plt
from matplotlib.pyplot import xticks


plt.style.use('ggplot')

def plot_history(history):
    print(history.history.keys())
    acc = history.history['accuracy']
    val_acc = history.history['val_accuracy']
    loss = history.history['loss']
    val_loss = history.history['val_loss']

    auc = history.history['auc']
    val_auc = history.history['val_auc']
    mse = history.history['mse']
    val_mse = history.history['val_mse']

    precision = history.history['precision']
    val_precision = history.history['val_precision']
    recall = history.history['recall']
    val_recall = history.history['val_recall']

    x = range(1, len(acc) + 1)

    plt.figure(figsize=(14, 12))

    plt.subplot(3, 2, 1)
    plt.plot(x, acc, 'b', label='Training acc')
    plt.plot(x, val_acc, 'r', label='Validation acc')
    plt.title('Training and validation accuracy')
    plt.legend()

    plt.subplot(3, 2, 2)
    plt.plot(x, loss, 'b', label='Training loss')
    plt.plot(x, val_loss, 'r', label='Validation loss')
    plt.title('Training and validation loss')
    plt.legend()

    plt.subplot(3, 2, 3)
    plt.plot(x, mse, 'b', label='Training mse')
    plt.plot(x, val_mse, 'r', label='Validation mse')
    plt.title('Training and validation MSE')
    plt.legend()

    plt.subplot(3, 2, 4)
    plt.plot(x, auc, 'b', label='Training AUC')
    plt.plot(x, val_auc, 'r', label='Validation AUC')
    plt.title('Training and validation AUC')
    plt.legend()

    plt.subplot(3, 2, 5)
    plt.plot(x, precision, 'b', label='Training precision')
    plt.plot(x, val_precision, 'r', label='Validation precision')
    plt.title('Training and validation precision')
    plt.legend()

    plt.subplot(3, 2, 6)
```

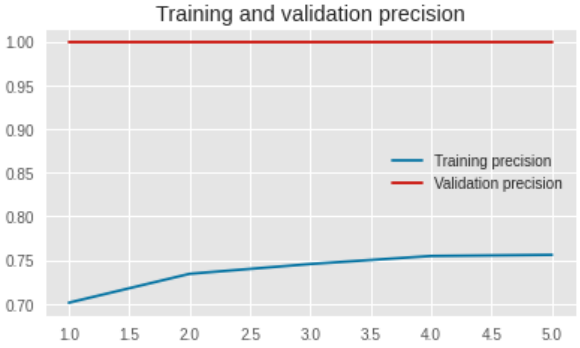
```
dict_keys(['loss', 'accuracy', 'mse', 'auc', 'precision', 'recall', 'val_loss
', 'val_accuracy', 'val_mse', 'val_auc', 'val_precision', 'val_recall'])
```

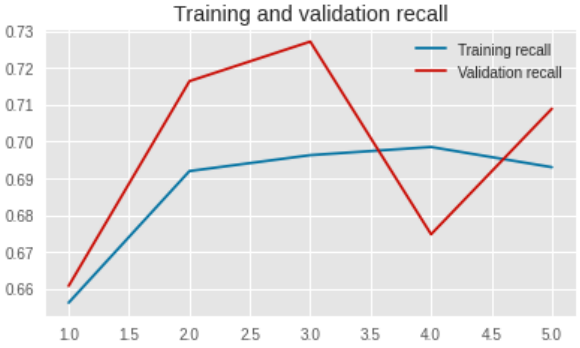












Save Experiment

In [16]:
```
jarvis.saveExperiment(myExp, FILE_NAME)
```

Scratchpad

In [ ]:
```
```

Configuration

In [1]:
```
# Parameters
ENABLE_COLAB = False

PROJECT_NAME = 'ML1010-Group-Project'
EXPERIMENT_NAME = 'ReviewText_Lemma_Bert2 (LSTM)'
FILE_NAME = '01_ML1010_GP_LSTM_Bert2_scratch'
LOAD_FROM_EXP = False

#Root Machine Learning Directory. Projects appear underneath
GOOGLE_DRIVE_MOUNT = '/content/gdrive'
COLAB_ROOT_DIR = GOOGLE_DRIVE_MOUNT + '/MyDrive/Colab Notebooks'
COLAB_INIT_DIR = COLAB_ROOT_DIR + '/utility_files'

LOCAL_ROOT_DIR = '/home/magni//ML_Root/project_root'
LOCAL_INIT_DIR = LOCAL_ROOT_DIR + '/utility_files'
```

Bootstrap Environment

In [2]:
```
#add in support for utility file directory and importing
import sys
import os

if ENABLE_COLAB:
  #Need access to drive
  from google.colab import drive
  drive.mount(GOOGLE_DRIVE_MOUNT, force_remount=True)

  #add in utility directory to syspath to import
  INIT_DIR = COLAB_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = COLAB_ROOT_DIR

else:
  #add in utility directory to syspath to import
  INIT_DIR = LOCAL_INIT_DIR
  sys.path.append(os.path.abspath(INIT_DIR))

  #Config environment variables
  ROOT_DIR = LOCAL_ROOT_DIR

#Import Utility Support
from jarvis import Jarvis
jarvis = Jarvis(ROOT_DIR, PROJECT_NAME)

import mv_python_utils as mvutils
```

```
Wha...where am I?
I am awake now.
```

```
I have set your current working directory to /home/magni/ML_Root/project_root
/ML1010-Group-Project
The current time is 12:33
Hello sir. Extra caffeine may help.
```

Setup Runtime Environment

In [3]:

```
if ENABLE_COLAB:
  #!pip install scipy -q
  #!pip install scikit-learn -q
  #!pip install pycaret -q
  #!pip install matplotlib -q
  #!pip install joblib -q
  #!pip install pandasql -q
  !pip install umap_learn -q
  !pip install sentence_transformers -q
  !pip install spacytextblob -q
  !pip install flair -q
  display('Google Colab enabled')
else:
  display('Google Colab not enabled')

#Common imports
import json
import pandas as pd
import numpy as np
import matplotlib
import re
import nltk
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn import metrics
from sklearn.datasets import load_digits
from sklearn.model_selection import train_test_split as tts
#from yellowbrick.classifier import ConfusionMatrix
#from sklearn.linear_model import LogisticRegression
from yellowbrick.target import ClassBalance
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.dummy import DummyClassifier

nltk.download('stopwords')
%matplotlib inline
```

```
'Google Colab not enabled'
```

```
[nltk_data] Downloading package stopwords to /home/magni/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [4]:
```
import cw_df_metric_utils as cwutils
import importlib
import DataPackage as dp
import DataPackageSupport as dps
import DataExperiment
import DataExperimentSupport as des
```

```
2022-01-15 12:33:32.641239: W tensorflow/stream_executor/platform/default/dso
_loader.cc:64] Could not load dynamic library 'libcudart.so.11.0'; dlerror: l
ibcudart.so.11.0: cannot open shared object file: No such file or directory
2022-01-15 12:33:32.641270: I tensorflow/stream_executor/cuda/cudart_stub.cc:
29] Ignore above cudart dlerror if you do not have a GPU set up on your machi
ne.
```

Load Data

In [5]:
```
#axis_labels=[1,2,3,4,5]
axis_labels=[0,1]

# using a dummyclassifier as DataExperiment requires a classifier to load
# and doesn't fully support Tensorflow models yet
classifier = DummyClassifier()
ANALYSIS_COL = 'reviewText_lemma_bert'
UNIQUE_COL = 'uuid'
TARGET_COL = 'overall_posneg'
```

In [6]:
```
if LOAD_FROM_EXP:
    #start from saved state
    myExp = jarvis.loadExperiment(FILE_NAME)
    myExp.display()

else:
    #start from source file and regenerate
    testDf = pd.read_pickle(jarvis.DATA_DIR_WORK + "/01_NL_ReviewText_All(new

    testDfBert = cwutils.getBertEncodeFrame(df=testDf,
                                            bertColumn=ANALYSIS_COL,
                                            uniqueColumn=UNIQUE_COL,
                                            otherColumns=[TARGET_COL]
                                            )


    myExp = DataExperiment.DataExperiment(projectName=PROJECT_NAME,
                                          experimentName=EXPERIMENT_NAME,
                                          origData=testDfBert,
                                          uniqueColumn=UNIQUE_COL,
                                          targetColumn=TARGET_COL,
                                          classifier=classifier)
```

```
DataExperiment summary:
---> projectName: ML1010-Group-Project
---> experimentName: ReviewText_Lemma_Bert2 (LSTM)
---> isDataPackageLoaded: True
---> isBaseModelLoaded: False
```

```
---> isBaseModelPredicted: False
---> isBaseModelLearningCurveCreated: False
---> isFinalModelLoaded: False
---> isFinalModelPredicted: False
---> isFinalModelLearningCurveCreated: False
---> isClassifierLoaded: True
DummyClassifier()

    DataPackage summary:
    Attributes:
    ---> uniqueColumn: uuid
    ---> targetColumn: overall_posneg
    Process:
    ---> isBalanced: False
    ---> isTrainTestSplit: False
    Data:
    ---> isOrigDataLoaded: True
    ---> isTrainDataLoaded: False
```

In [7]:

```
#get the train data and downsample to 2900
tDf = myExp.dataPackage.getOrigData()
dps.displayClassBalance(tDf, myExp.dataPackage.targetColumn, verbose=True)
```

	overall_posneg	ttlCol
0	0	13440
1	1	49973

In [8]:

```
myExp.processDataPackage()
```


```
Undersampling data to match min class: 0 of size: 13440
```


```
Completed train/test split (test_size = 0.2):
---> Original data size: 26880
---> Training data size: 21504
---> Testing data size: 5376
---> Stratified on column: overall_posneg
```

In [9]:

```
tDf2 = myExp.dataPackage.getTrainData()
dps.displayClassBalance(tDf2, myExp.dataPackage.targetColumn, verbose=True)
```

	overall_posneg	ttlCol
0	0	10752
1	1	10752

In [10]:
```
SAMPLE_DOWN_SIZE=10700
# Do the sampling
tDf2 = tDf2.groupby(myExp.dataPackage.targetColumn, group_keys=False).apply(l
tDf2.reset_index(drop=True, inplace=True)
```

In [11]:
```
dps.displayClassBalance(tDf2, myExp.dataPackage.targetColumn, verbose=True)
```

	overall_posneg	ttlCol
0	0	10700
1	1	10700

In [12]:
```
from xgboost import XGBClassifier
from keras.layers.core import SpatialDropout1D
from keras.layers import Dropout, Dense, Flatten, LSTM, Input, Conv1D, MaxPoo
from keras.models import Sequential
from keras.backend import clear_session
from keras.layers.embeddings import Embedding
import keras

print(keras.__version__)
from keras import backend as K
K._get_available_gpus()
```

```
2.7.0
2022-01-15 12:33:44.787035: W tensorflow/stream_executor/platform/default/dso
_loader.cc:64] Could not load dynamic library 'libcuda.so.1'; dlerror: libcud
a.so.1: cannot open shared object file: No such file or directory
2022-01-15 12:33:44.787067: W tensorflow/stream_executor/cuda/cuda_driver.cc:
269] failed call to cuInit: UNKNOWN ERROR (303)
2022-01-15 12:33:44.787082: I tensorflow/stream_executor/cuda/cuda_diagnostic
s.cc:156] kernel driver does not appear to be running on this host (localhos
t.localdomain): /proc/driver/nvidia/version does not exist
2022-01-15 12:33:44.787345: I tensorflow/core/platform/cpu_feature_guard.cc:1
51] This TensorFlow binary is optimized with oneAPI Deep Neural Network Libra
ry (oneDNN) to use the following CPU instructions in performance-critical ope
rations:  AVX2 FMA
To enable them in other operations, rebuild TensorFlow with the appropriate c
ompiler flags.
```

Out[12]:
```
[]
```

In [19]:
```
###Notes section:
#model = Sequential()
#model.add(LSTM(NumberOfLSTM, return_sequences=True,
#               input_shape=(YourSequenceLenght, YourWord2VecLenght)))
tExp = jarvis.loadExperiment('01_ML1010_GP_XGB_Bert2')
tFeat = tExp.getFinalFeatures()
print (len(tFeat))
print (tFeat[1])
```

```
40
c42
```

In [100…
```
from tensorflow.keras.metrics import AUC, Precision, Recall
```

In [127…

```
tDf3 = tDf2.copy()
Y_train = np.array(tDf3[myExp.dataPackage.targetColumn])
#tDf3.drop(myExp.dataPackage.uniqueColumn, axis=1, inplace=True)
#tDf3.drop(myExp.dataPackage.targetColumn, axis=1, inplace=True)
X_train = np.array(tDf3[tFeat])

print(Y_train.shape)
print(X_train.shape)
EPOCHS=5
VAL_SPLIT=0.1

BATCH_SIZE=100
NUMBER_FEATURES=len(tFeat)

DROPOUT_RATE=0.2
INTERNAL_LAYERS=5
LSTM_OUTPUT_UNITS=5
```

```
(21400,)
(21400, 40)
```

In [128…

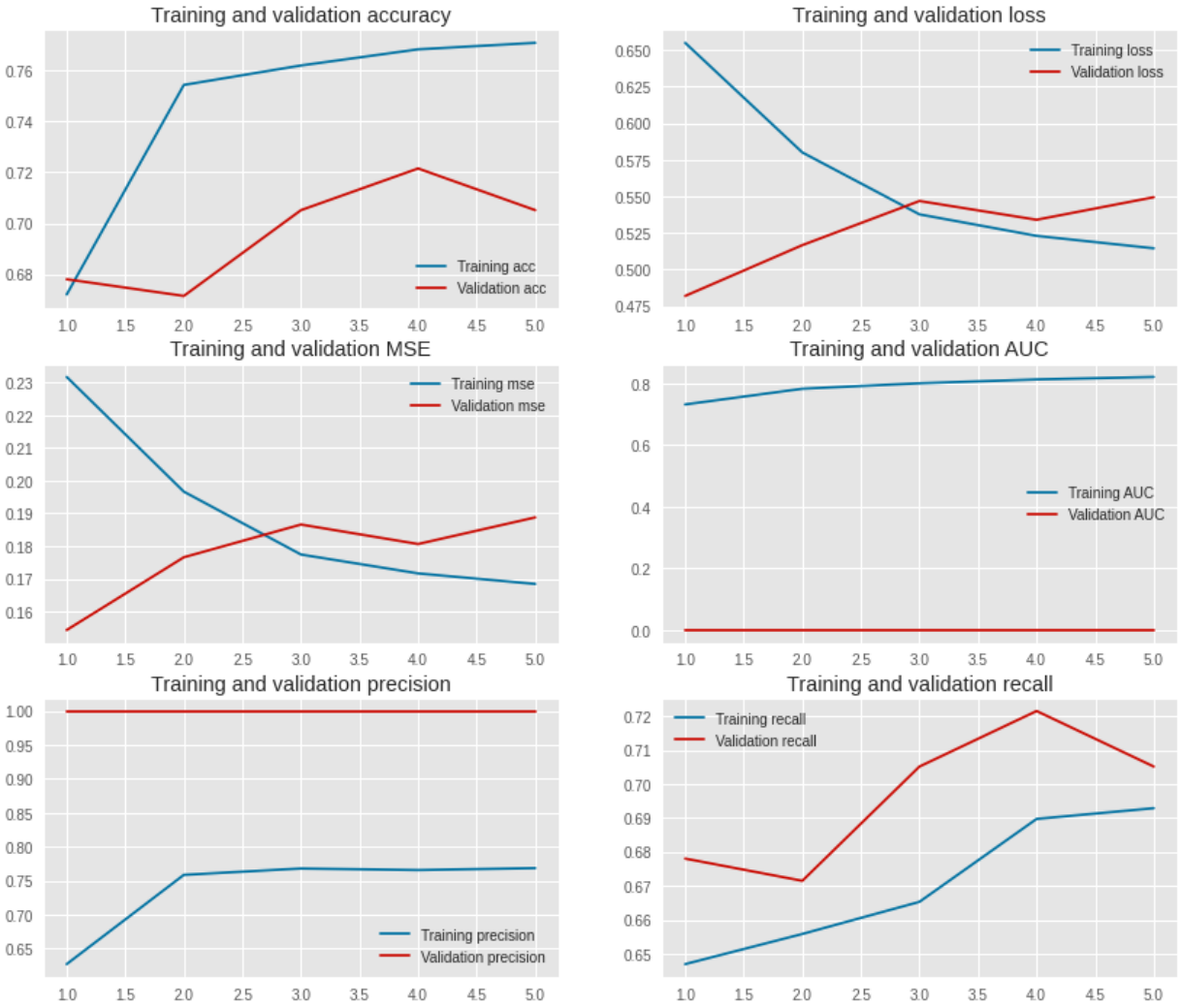
```
# Neural network
keras.backend.clear_session()

model2 = None
model2 = Sequential()

#model2.add(Input(shape=(NUMBER_FEATURES, 1)))
#model2.add(Dense(INTERNAL_LAYERS, activation='relu'))
#model2.add(Dropout(DROPOUT_RATE))

model2.add(LSTM(units=LSTM_OUTPUT_UNITS,
                input_shape=(NUMBER_FEATURES, 1),
                return_sequences=False
               )
          )
#model2.add(Dropout(DROPOUT_RATE))
#model2.add(Conv1D(filters=LSTM_OUTPUT_UNITS, kernel_size=3, padding='same',
#model2.add(MaxPooling1D(pool_size=2))
#model2.add(LSTM(units=LSTM_OUTPUT_UNITS))
#model2.add(Dropout(DROPOUT_RATE))


#model2.add(LSTM(units=LSTM_OUTPUT_UNITS))
#model2.add(Dropout(DROPOUT_RATE))

model2.add(Dense(INTERNAL_LAYERS, activation='relu'))
model2.add(Dropout(DROPOUT_RATE))

#model2.add(Dense(LSTM_OUTPUT_UNITS, activation='relu'))
#model2.add(Dropout(DROPOUT_RATE))

#model2.add(Dense(INTERNAL_LAYERS, activation='relu'))
#model2.add(Dropout(DROPOUT_RATE))

#model2.add(Dense(LSTM_OUTPUT_UNITS, activation='relu'))
#model2.add(Dropout(DROPOUT_RATE))


#softmax is for multiclass
#model2.add(Dense(1, activation='softmax'))
#----
#sigmoid is not for multiclass
model2.add(Dense(1, activation='sigmoid'))

model2.compile(loss='binary_crossentropy',
               optimizer='adam',
               metrics=['accuracy',
                        'mse',
                        AUC(),
                        Precision(),
                        Recall()
                       ]
              )

history = model2.fit(x=X_train,
                     y=Y_train,
                     epochs=EPOCHS,
                     batch_size=BATCH_SIZE,
```

```
Epoch 1/5
193/193 [==============================] - 4s 13ms/step - loss: 0.6550 - accu
racy: 0.6721 - mse: 0.2317 - auc: 0.7317 - precision: 0.6270 - recall: 0.6470
- val_loss: 0.4820 - val_accuracy: 0.6780 - val_mse: 0.1544 - val_auc: 0.0000
e+00 - val_precision: 1.0000 - val_recall: 0.6780
Epoch 2/5
193/193 [==============================] - 2s 10ms/step - loss: 0.5801 - accu
racy: 0.7543 - mse: 0.1967 - auc: 0.7826 - precision: 0.7585 - recall: 0.6558
- val_loss: 0.5167 - val_accuracy: 0.6715 - val_mse: 0.1766 - val_auc: 0.0000
e+00 - val_precision: 1.0000 - val_recall: 0.6715
Epoch 3/5
193/193 [==============================] - 2s 10ms/step - loss: 0.5378 - accu
racy: 0.7619 - mse: 0.1774 - auc: 0.8003 - precision: 0.7679 - recall: 0.6653
- val_loss: 0.5469 - val_accuracy: 0.7051 - val_mse: 0.1866 - val_auc: 0.0000
e+00 - val_precision: 1.0000 - val_recall: 0.7051
Epoch 4/5
193/193 [==============================] - 2s 10ms/step - loss: 0.5231 - accu
racy: 0.7682 - mse: 0.1717 - auc: 0.8127 - precision: 0.7656 - recall: 0.6897
- val_loss: 0.5341 - val_accuracy: 0.7215 - val_mse: 0.1806 - val_auc: 0.0000
e+00 - val_precision: 1.0000 - val_recall: 0.7215
Epoch 5/5
193/193 [==============================] - 2s 10ms/step - loss: 0.5146 - accu
racy: 0.7708 - mse: 0.1684 - auc: 0.8209 - precision: 0.7686 - recall: 0.6929
- val_loss: 0.5495 - val_accuracy: 0.7051 - val_mse: 0.1888 - val_auc: 0.0000
e+00 - val_precision: 1.0000 - val_recall: 0.7051
```

In [144…

```
import matplotlib.pyplot as plt
from matplotlib.pyplot import xticks


plt.style.use('ggplot')

def plot_history(history):
    print(history.history.keys())
    acc = history.history['accuracy']
    val_acc = history.history['val_accuracy']
    loss = history.history['loss']
    val_loss = history.history['val_loss']

    auc = history.history['auc']
    val_auc = history.history['val_auc']
    mse = history.history['mse']
    val_mse = history.history['val_mse']

    precision = history.history['precision']
    val_precision = history.history['val_precision']
    recall = history.history['recall']
    val_recall = history.history['val_recall']

    x = range(1, len(acc) + 1)

    plt.figure(figsize=(14, 12))

    plt.subplot(3, 2, 1)
    plt.plot(x, acc, 'b', label='Training acc')
    plt.plot(x, val_acc, 'r', label='Validation acc')
    plt.title('Training and validation accuracy')
    plt.legend()

    plt.subplot(3, 2, 2)
    plt.plot(x, loss, 'b', label='Training loss')
    plt.plot(x, val_loss, 'r', label='Validation loss')
    plt.title('Training and validation loss')
    plt.legend()

    plt.subplot(3, 2, 3)
    plt.plot(x, mse, 'b', label='Training mse')
    plt.plot(x, val_mse, 'r', label='Validation mse')
    plt.title('Training and validation MSE')
    plt.legend()

    plt.subplot(3, 2, 4)
    plt.plot(x, auc, 'b', label='Training AUC')
    plt.plot(x, val_auc, 'r', label='Validation AUC')
    plt.title('Training and validation AUC')
    plt.legend()

    plt.subplot(3, 2, 5)
    plt.plot(x, precision, 'b', label='Training precision')
    plt.plot(x, val_precision, 'r', label='Validation precision')
    plt.title('Training and validation precision')
    plt.legend()

    plt.subplot(3, 2, 6)
```

```
dict_keys(['loss', 'accuracy', 'mse', 'auc', 'precision', 'recall', 'val_loss
', 'val_accuracy', 'val_mse', 'val_auc', 'val_precision', 'val_recall'])
```



In [ ]:

```
myExp.
#classifier = XGBClassifier(eval_metric='mlogloss')
#classifier = SVC(gamma=0.001, verbose=True)
#classifier = RandomForestClassifier()

#print(model.summary())
```

Save Experiment

In [ ]:

```
jarvis.saveExperiment(myExp, FILE_NAME)
```

Scratchpad

In [ ]: