Model Fit Process Results

Classification Reports:

The classification reports are generated after the model fitting process and during the visualization process:

Below is Multinomial Naïve Bayes Classification Report:

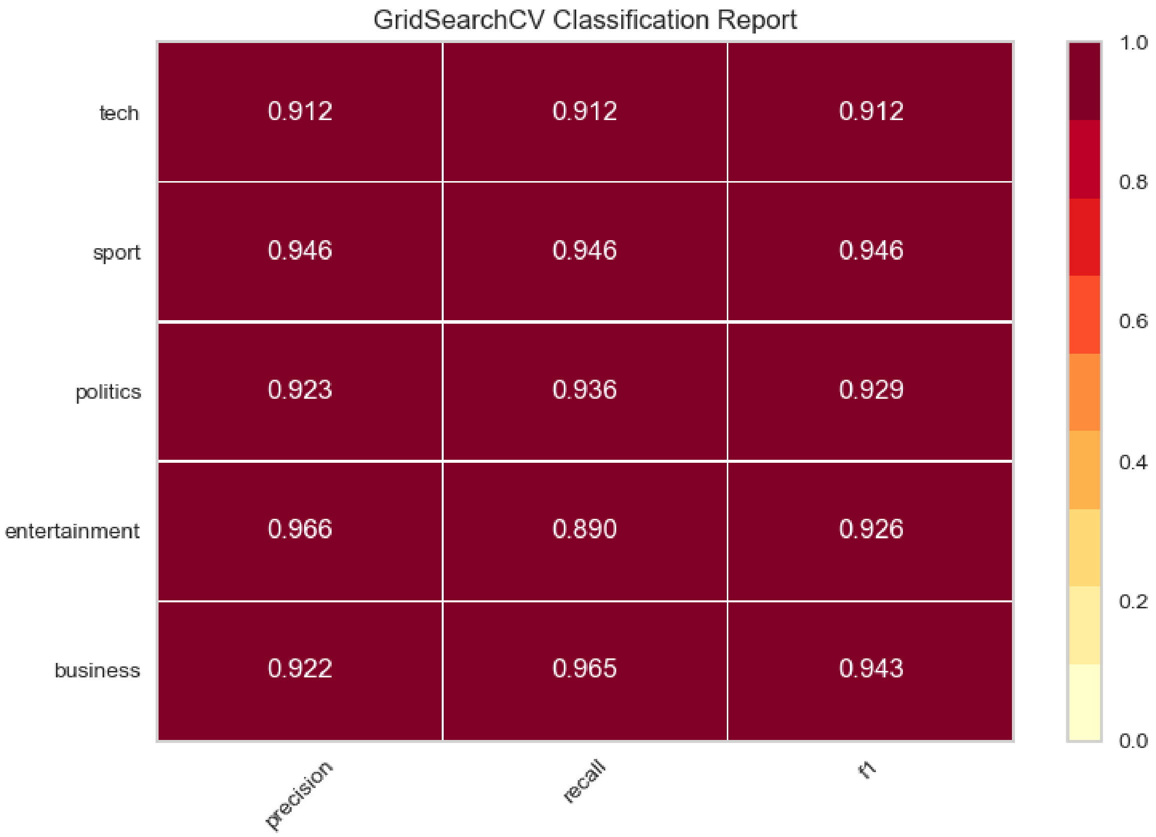


Figure 1:MultinomialNB Classification Report

Confusion Matrix:

Below is the Confusion Matrix Report:

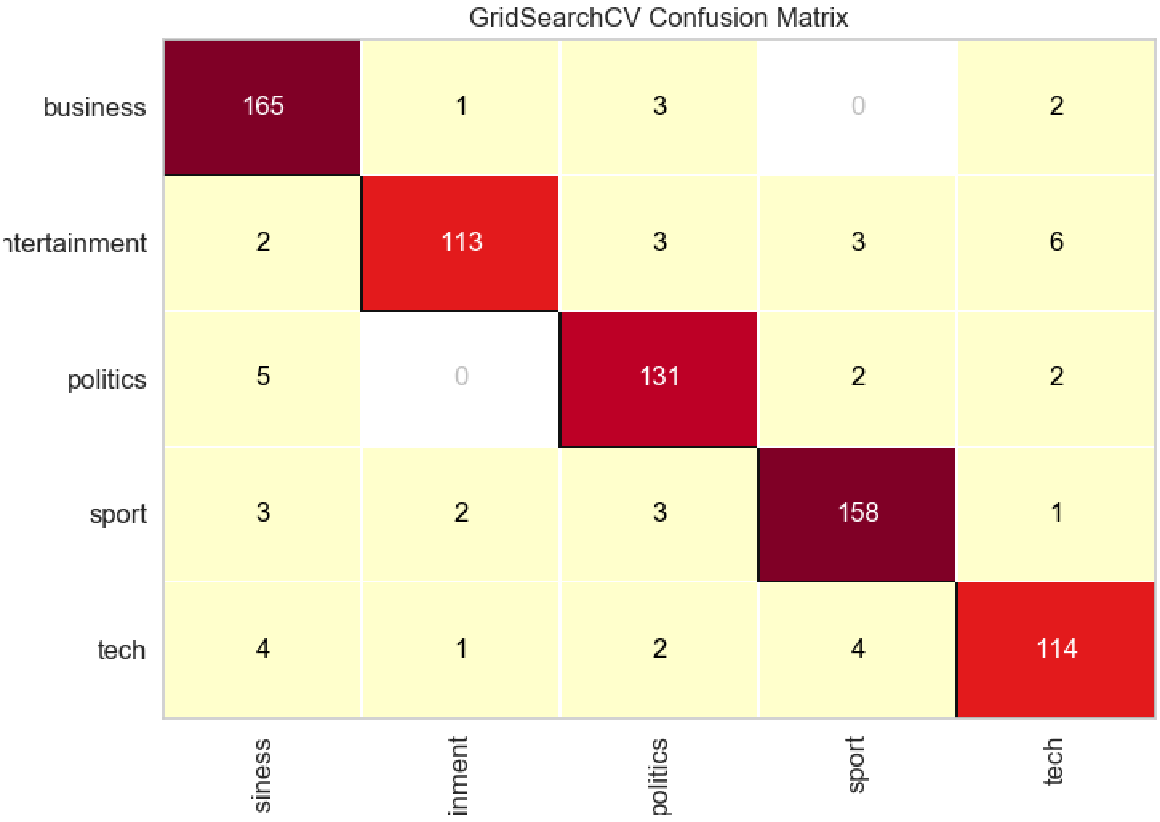


Figure 2:MultinomialNB Confusion Matrix

Frequency Distribution:

Below is the Horizontal Histogram for top 50 word tokens before and after stopword removal:

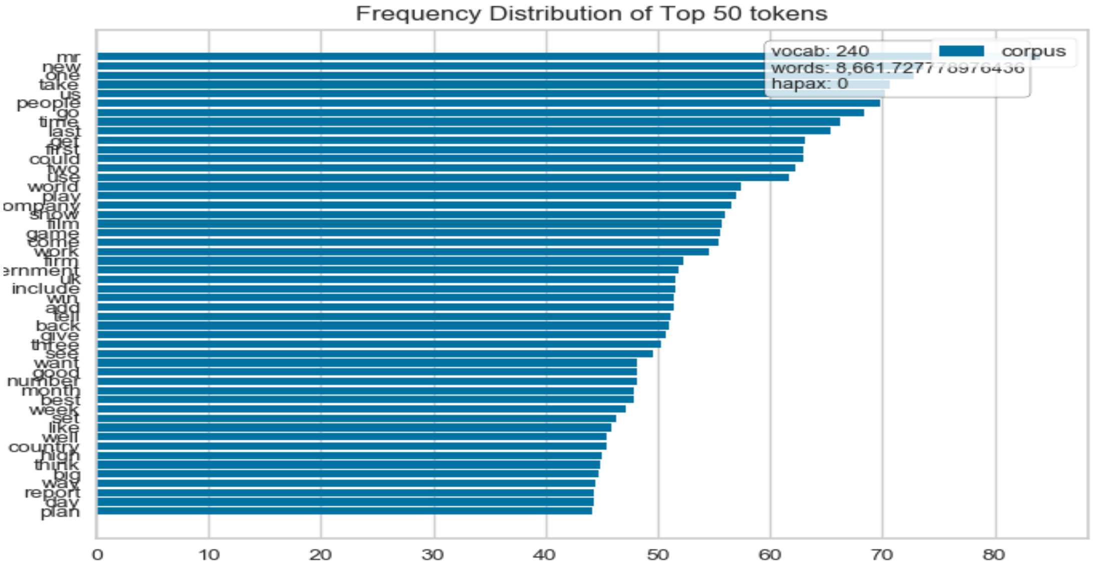


Figure 3: Before Stopword Removal

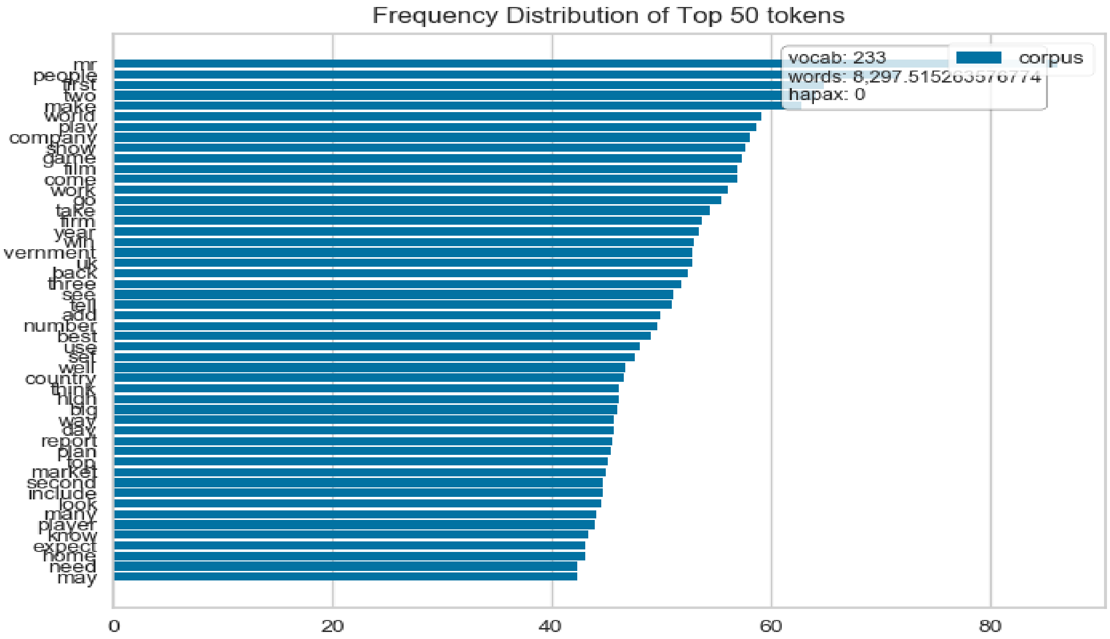
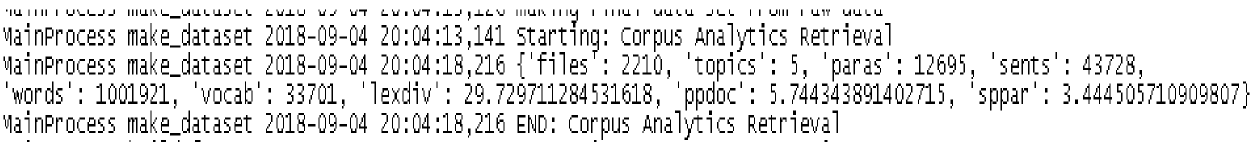


Figure 4: After Stopword Removal

Corpus Analytics

Corpus Analytics are generated after Corpus Extraction and analytics are output to the process log file which will be created in the site_packages folder. Below is the screenshot

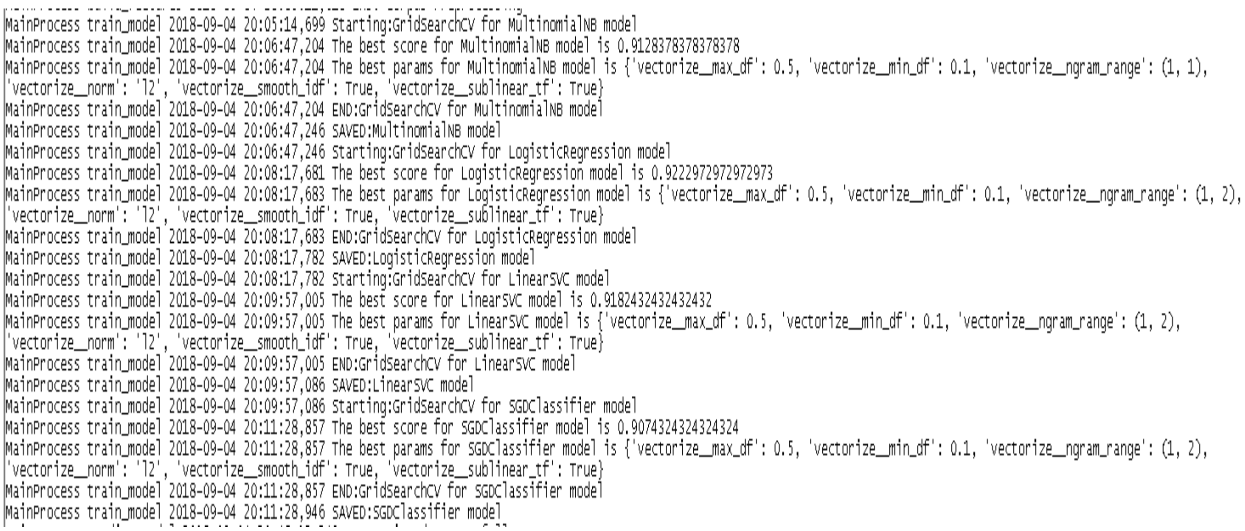


```
MainProcess make_dataset 2018-09-04 20:04:13,141 Starting: Corpus Analytics Retrieval
MainProcess make_dataset 2018-09-04 20:04:18,216 {'files': 2210, 'topics': 5, 'paras': 12695, 'sents': 43728,
'words': 1001921, 'vocab': 33701, 'lexdiv': 29.729711284531618, 'ppdoc': 5.744343891402715, 'sppar': 3.444505710909807}
MainProcess make_dataset 2018-09-04 20:04:18,216 END: Corpus Analytics Retrieval
```

- Files: 2210 are the number of raw text files in the Corpus.
- Topics: 5 are the number of News categories.
- Paras: 12695 are the number of paragraphs in the entire corpus.
- Sents: 43728 are the total number of sentences in the corpus.
- Words: 1001921 are the total number of words in the corpus.
- Vocab: 33701 are the total number of unique words in the corpus.
- Lexdiv: 29.729 is the lexical diversity of the corpus (number of times each word in vocabulary is used).
- Ppdoc: 5.74 number of paragraphs per document.
- Sppar: 3.44 number of sentences per paragraph

Model Best Parameters and Best Scores

These are also available in the process log file as shown in the screenshot below:



```
MainProcess train_model 2018-09-04 20:05:14,699 Starting:GridSearchCV for MultinomialNB model
MainProcess train_model 2018-09-04 20:06:47,204 The best score for MultinomialNB model is 0.9128378378378378
MainProcess train_model 2018-09-04 20:06:47,204 The best params for MultinomialNB model is {'vectorize__max_df': 0.5, 'vectorize__min_df': 0.1, 'vectorize__ngram_range': (1, 1),
'vectorize__norm': 'l2', 'vectorize__smooth_idf': True, 'vectorize__sublinear_tf': True}
MainProcess train_model 2018-09-04 20:06:47,204 END:GridSearchCV for MultinomialNB model
MainProcess train_model 2018-09-04 20:06:47,246 SAVED:MultinomialNB model
MainProcess train_model 2018-09-04 20:06:47,246 Starting:GridSearchCV for LogisticRegression model
MainProcess train_model 2018-09-04 20:08:17,681 The best score for LogisticRegression model is 0.9222972972972973
MainProcess train_model 2018-09-04 20:08:17,683 The best params for LogisticRegression model is {'vectorize__max_df': 0.5, 'vectorize__min_df': 0.1, 'vectorize__ngram_range': (1, 2),
'vectorize__norm': 'l2', 'vectorize__smooth_idf': True, 'vectorize__sublinear_tf': True}
MainProcess train_model 2018-09-04 20:08:17,683 END:GridSearchCV for LogisticRegression model
MainProcess train_model 2018-09-04 20:08:17,782 SAVED:LogisticRegression model
MainProcess train_model 2018-09-04 20:08:17,782 Starting:GridSearchCV for LinearSVC model
MainProcess train_model 2018-09-04 20:09:57,005 The best score for LinearSVC model is 0.9182432432432432
MainProcess train_model 2018-09-04 20:09:57,005 The best params for LinearSVC model is {'vectorize__max_df': 0.5, 'vectorize__min_df': 0.1, 'vectorize__ngram_range': (1, 2),
'vectorize__norm': 'l2', 'vectorize__smooth_idf': True, 'vectorize__sublinear_tf': True}
MainProcess train_model 2018-09-04 20:09:57,005 END:GridSearchCV for LinearSVC model
MainProcess train_model 2018-09-04 20:09:57,086 SAVED:LinearSVC model
MainProcess train_model 2018-09-04 20:09:57,086 Starting:GridSearchCV for SGDClassifier model
MainProcess train_model 2018-09-04 20:11:28,857 The best score for SGDClassifier model is 0.9074324324324324
MainProcess train_model 2018-09-04 20:11:28,857 The best params for SGDClassifier model is {'vectorize__max_df': 0.5, 'vectorize__min_df': 0.1, 'vectorize__ngram_range': (1, 2),
'vectorize__norm': 'l2', 'vectorize__smooth_idf': True, 'vectorize__sublinear_tf': True}
MainProcess train_model 2018-09-04 20:11:28,857 END:GridSearchCV for SGDClassifier model
MainProcess train_model 2018-09-04 20:11:28,946 SAVED:SGDClassifier model
```

For example the best score for Logistic Regression model is: 0.923

Best Parameters are:

Max_df:0.5,Min_df:0.1,Ngram_range(1,2),Norm:l2,Smooth_df:True,Sublinear_tf:True