Classification of Textual Data

Eric Anderson, Yawan Xu, Ali Tapan
McGill University
COMP 551 - Applied Machine Learning
Miniproject 2
Winter 2020

***Abstract*—This work applies to 5 classification models, logistic regression, decision trees, support vector machines, ada boost and random forest for the following datasets:**

1) **The *20 News Group Dataset*[1], this data set is a collection of newsgroup documents where the goal is to predict the category of the given news article.**
2) **The *IMDb Reviews Dataset* [2], this dataset contains movie reviews along with their associated binary sentiment polarity labels where the goal is to predict whether the review is positive or negative.**

We analyzed and compared the performance of all 5 classification models on both datasets in terms of accuracy and training time. We found that the support vector machine classifier performed the best (mean of 85.56% accuracy for both datasets), and the decision tree classifier performed the worst (mean of 62.44% accuracy for both datasets). Furthermore, the training time for support vector machines model was significantly faster (mean of 5.280s for both datasets) compared to all of the models while random forest classification model had the slowest training time (mean of 161.5s for both datasets).

I. Introduction

A. Task Summary

One of the most popular applications of machine learning is the analysis of categorical data, specifically text data. In this work, we implemented logistic regression, decision trees, support vector machines, ada boost and random forest and compared their performance on two textual datasets. The implementation of these models were done with SciKit learn package, a Python library. Default train subsets (subset='train', and remove=(['headers', 'footers', 'quotes']) in sklearn.datasets) were used to train the models and report the final performance on the test subset for the *20 News Group Dataset*[1] and the *IMDb Reviews Dataset*[2]. Finally, we ran grid search cross validation to optimize the hyperparameters. Grid search automatically tested all possible combinations of given hyperparameters, then used mean accuracy over 5 cross validation splits from the training set to help us choose the parameters which would give the highest accuracy on the test set.

B. Background

In this section we briefly discuss the background of the algorithms implemented in our classification.

- *Logistic Regression* is a variation of linear regression, where a model calculates the dependent variable based on the independent variable. The important difference compared to linear regression is that the output value modeled in logistic regression is a categorical value rather than a numerical value. Categorical values include binary (0 or 1), multinomial (three or more categories without ordering e.g. child, adult, elder) or ordinal values (three or more categories with ordering e.g. movie ratings from 1 to 10). Logistic regression utilizes the sigmoid function to fit the data.
- *Decision Trees* are a well known classification technique for predictive modeling to go from the observation of the item to the conclusion of the classification. Decision trees build regression or classification models as a form of tree of nodes by breaking down data into smaller subsets. The final results can be observed from the leaves of the tree. The main disadvantage of using a decision tree classification is they are generally unstable and have high variance.
- *Support Vector Machines* is a type of supervised learning model to analyze data for classification and regression. The goal of support vector machines is to find a hyperplane that

best divides the data to label it. Support vector machines are popularly used in text categorization, handwriting recognition and image classification because they can process unstructured data. They also scale well with higher dimensional data compared to other classification techniques.

- *Ada Boost* or Adaptive Boosting is a type of machine learning model that focuses on the misclassified data through assigning weights. It combines multiple "weak classifiers" such as a one-level decision trees (decision stumps) to form a single "strong classifier". The output of the "weak classifiers" determine how much weight should be assigned to each result in the dataset. Due to the combination and processing of multiple weak classifiers, the disadvantage of ada boost is its longer training times.
- *Random Forest* is an ensemble learning method that constructs a multitude of decision trees and uses their mean or median to predict the outcome. Some of the main advantages of using random forest are its lower likelihood of overfitting the data and its ability to process missing values in the dataset.

C. Important Findings

We found that decision trees struggle with higher dimensional data which we concluded from its lower accuracy in our models. This was expected due to the fact that decision trees are highly sensitive to small changes which results them to have high amounts of variance and overfitting the data.

Furthermore, with the IMDb dataset, we found that 1) removing stop words and 2) including bigrams and trigrams did not improve our accuracy. With the 20 News Group dataset, we found that 1) dropping words which occurred in more than 75% of training samples, 2) not inducing a lower limit as to how often a word must appear to be included, and 3) enabling sublinear term-frequency scaling in our transformer improved our accuracy.

II. Datasets

A. 20 News Group Dataset

The *20 News Group Dataset* contains 18846 samples of newsgroups posts with 20 topics (classes) on each sample in two subsets: one for training and one for testing and performance evaluation. The split between the train and test set is based upon the messages posted before and after a specific date. This data was used to predict predict the category of the given news article. To gain an understanding of the data within this news group, we plotted up to the 85th percentile of data length, according to their category. This plot shows that while there are some discrepancies in length between categories, those were usually caused by large outliars pulling the data, as all of the category length medians fell within a few hundred characters. (Figure 1)

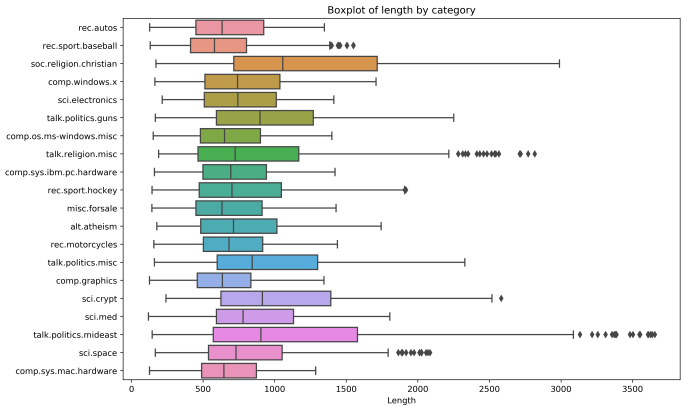

Fig. 1. Comparison of data lengths in 20 News Groups

B. IMDb Reviews Dataset

The *IMDb Reviews Dataset* contains movie reviews along with their associated binary sentiment polarity labels. It is intended to serve as a benchmark for sentiment classification by Stanford University. It contains 50000 reviews (samples) that are evenly split into train and test sets that are 25000 for each. IMDb user ratings are scaled from 1 to 10. To label the given reviews the custodian of the dataset labeled anything less then or equal 4 as negative and anything with greater or equal to 7 as positive. The review ratings that are in between 4 and 7 were left out.

III. Proposed Approach

A. Preprocessing

Our approach for feature extraction for the *20 News Group Dataset* consisted of including all of the 20 categories and removing 'headers', 'footers' and 'quotes', as well as a cleanup of the text within the training samples. This cleanup was done by removing any ASCII Linefeed, ASCII Carriage Return, ASCII Horizontal Tab and punctuation signs from the data.

For the *IMDb Reviews Dataset*, we combined the positive and negative examples from the 'pos' and 'neg' folders for both the training and testing dataset, and read into a dataframe with the appropriate labels: positive as '1' and negative as '0'.

B. Algorithm Implementation

The following algorithms were implemented using the standard SciKit library for Python:

- Logistic regression: sklearn.linear_model.LogisticRegression
- Decision tree: sklearn.tree.DecisionTreeClassifier
- Support vector machines: sklearn.svm.LinearSVC
- Ada boost: sklearn.ensemble.AdaBoostClassifier
- Random forest: sklearn.ensemble.RandomForestClassifier

C. Grid Search Cross Validation

To optimize the classifiers, we used Grid Search cross validation with 5 cross validation folds to study the effects of different hyperparameters and recorded the top three that gave the best performance, judged by model accuracy and standard deviation.

IV. Results

As we can see from Table 1, support vector machines achieved the highest accuracy of 83.1% on the 20 News Groups dataset, and logistic regression has a 0.1% advantage over support vector machines on the IMDb dataset. Furthermore, from Figure 2 and 3, we see that both logistic regression and support vector machines also have the lowest training time. The good performance of these linear

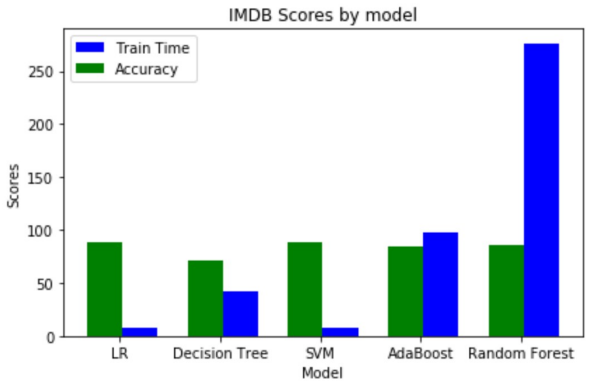

Fig. 2. Comparison of model performances on IMDb dataset

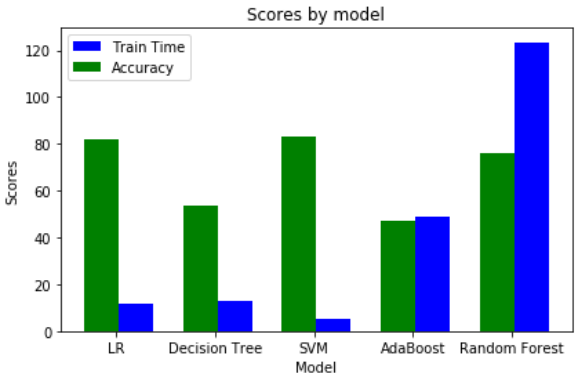

Fig. 3. Comparison of model performances on 20 News Groups dataset

models was expected since they often work best on high dimensional sparse data like ours. On the other hand, Ada boost and random forest take significantly longer to train, possibly due to their sensitivity to noise data.

We also saw that for the 20 News Groups dataset, increasing the depth of the base estimator decision trees increased training time and decreased accuracy nearly linearly. (Figure 4).

To improve the performance of our models, we tried a few things to process the data by reducing its noise. We removed the stop-words, and instead of just single-word tokens, we tried including word pairs (bigrams) and trigrams. However, both methods failed to improve the model performance.

V. Discussion and Conclusion

We concluded that linear models (logistic regression and support vector machine) are better

Word Rank	comp.sys.ibm.pc.hardware	misc.forsale	rec.autos	rec.sport.baseball	sci.crypt	sci.med	talk.politics.mideast
1	gateway	sale	car	baseball	clipper	doctor	israel
2	ide	forsale	cars	phillies	tapped	msg	israeli
3	pc	shipping	warning	cubs	encryption	disease	armenia
4	monitors	wanted	oil	sox	key	photography	armenian
5	bus	offer	automotive	stadium	pgp	cancer	hezbollah

TABLE I
THE 5 WORDS WITH THE HIGHEST SUPPORT VECTOR MACHINES WEIGHTS FOR AN ASSORTMENT OF CATEGORIES

Dataset	Best model	Performance
20 News Groups	**SVM**	**83.1%**
	LR	82.3%
	DT	53.4%
	AB	47.2%
	RF	75.8%
IMDb	**LR**	**87.5%**
	SVM	87.4%
	DT	70.4%
	AB	70.4%
	RF	84.3%

TABLE II
PERFORMANCE OF THE MODELS WITH THE BEST HYPERPARAMETERS FOR BOTH DATASETS (BEST MODELS IN BOLD)

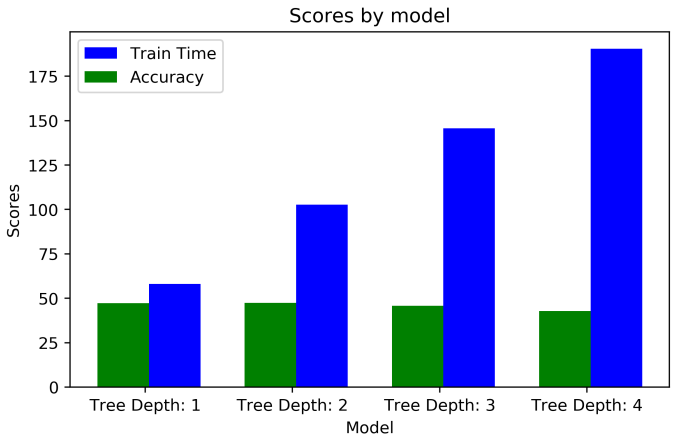

Fig. 4. Comparison of AdaBoost accuracies and training times

at classifying high-dimensional spaces such as text, than decision trees, ada boost, and random forest.

We realized that Grid Search might not have been the best method of validation, as it takes a long computing time. Notably for the random forest classifier on the IMDb dataset, we had to significantly scale down the range of parameters, because it was taking over five hours to compute. Due to the limited range of parameters, we might not have found the best performance of the random forest model.

VI. FUTURE WORK

To improve the performance of our models, we could experiment with other methods of text processing, such as stemming and lemmatizing. We could also improve the computing time of the hyperparameter tuning, as some classifiers, such as random forest, required hours of computation with Grid Search. One approach that claims to be more efficient is the one proposed by Bergstra and Bengio (2012)[5], who found that RandomizedSearch followed by Grid Search greatly reduces the computing time, while delivering equally good results. In the future, we could investigate and validate this approach.

Another thing we could look into is assessing the importance of specific words to the assignment of each class. For instance, Table I shows the weights of each word per class for the 20 News Group Dataset.

VII. STATEMENT OF CONTRIBUTIONS

- **Eric Anderson**: Built and implemented the models and evaluations for 20 News Group Dataset, generated graphs, and implemented grid search cross validation.
- **Ali Tapan**: Report formulation, dataset and code review
- **Yawan Xu**: Built and implemented the models and evaluations for IMDB Reviews Dataset. Wrote part of the report.

REFERENCES

[1] Lang, Ken (1995). 20 Newsgroups [http://qwone.com/ jason/20Newsgroups/].

[2] Maas, Andrew (2011). Large Movie Review Dataset [http://ai.stanford.edu/ amaas/data/sentiment/].
[3] Maas, Andrew L, et al. “Learning Word Vectors for Sentiment Analysis.” Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, edited by Raymond E Daly, Association for Computational Linguistics, 2011, pp. 142–150.
[4] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).
[5] James Bergstra, Yoshua Bengio. (2012). Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research 13 (2012) 281-305.

APPENDIX

```
-------------------LogisticRegression---------------------
Time to fit: 11.648s
LogisticRegression accuracy: 82.25%
                          precision    recall  f1-score   support

              alt.atheism       0.78      0.71      0.74       319
            comp.graphics       0.71      0.79      0.75       389
  comp.os.ms-windows.misc       0.76      0.72      0.74       394
 comp.sys.ibm.pc.hardware       0.74      0.74      0.74       392
    comp.sys.mac.hardware       0.81      0.84      0.82       385
           comp.windows.x       0.86      0.75      0.80       395
             misc.forsale       0.82      0.88      0.85       390
                rec.autos       0.89      0.87      0.88       396
          rec.motorcycles       0.91      0.94      0.92       398
       rec.sport.baseball       0.88      0.92      0.90       397
         rec.sport.hockey       0.92      0.95      0.94       399
                sci.crypt       0.93      0.89      0.91       396
          sci.electronics       0.73      0.75      0.74       393
                  sci.med       0.85      0.84      0.84       396
                sci.space       0.86      0.91      0.88       394
   soc.religion.christian       0.79      0.92      0.85       398
       talk.politics.guns       0.73      0.86      0.79       364
    talk.politics.mideast       0.96      0.86      0.91       376
       talk.politics.misc       0.77      0.61      0.68       310
       talk.religion.misc       0.72      0.50      0.59       251

                micro avg       0.82      0.82      0.82      7532
                macro avg       0.82      0.81      0.81      7532
             weighted avg       0.82      0.82      0.82      7532
```

Fig. 5. 20 News Group Dataset - Classification Report - Logistic Regression

```
-------------------LogisticRegression---------------------
Time to fit: 7.799s
LogisticRegression accuracy: 88.66%
              precision    recall  f1-score   support

           0       0.88      0.89      0.89     12500
           1       0.89      0.88      0.89     12499

    accuracy                           0.89     24999
   macro avg       0.89      0.89      0.89     24999
weighted avg       0.89      0.89      0.89     24999
```

Fig. 6. IMDb Reviews Dataset - Classification Report - Logistic Regression

```
-------------------DecisionTreeClassifier---------------------
Time to fit: 13.132s
DecisionTreeClassifier accuracy: 53.41%
                          precision    recall  f1-score   support

              alt.atheism       0.49      0.40      0.44       319
            comp.graphics       0.36      0.49      0.41       389
  comp.os.ms-windows.misc       0.50      0.49      0.50       394
 comp.sys.ibm.pc.hardware       0.35      0.46      0.40       392
    comp.sys.mac.hardware       0.51      0.48      0.49       385
           comp.windows.x       0.58      0.48      0.52       395
             misc.forsale       0.64      0.74      0.69       390
                rec.autos       0.56      0.54      0.55       396
          rec.motorcycles       0.84      0.70      0.76       398
       rec.sport.baseball       0.47      0.53      0.50       397
         rec.sport.hockey       0.78      0.66      0.71       399
                sci.crypt       0.81      0.64      0.72       396
          sci.electronics       0.24      0.31      0.27       393
                  sci.med       0.40      0.51      0.44       396
                sci.space       0.63      0.61      0.62       394
   soc.religion.christian       0.55      0.64      0.59       398
       talk.politics.guns       0.63      0.62      0.62       364
    talk.politics.mideast       0.80      0.60      0.68       376
       talk.politics.misc       0.43      0.37      0.40       310
       talk.religion.misc       0.50      0.26      0.34       251

                micro avg       0.53      0.53      0.53      7532
                macro avg       0.55      0.53      0.53      7532
             weighted avg       0.56      0.53      0.54      7532
```

Fig. 7. 20 News Group Dataset - Classification Report - Decision Tree

```
-------------------DecisionTreeClassifier---------------------
Time to fit: 41.853s
DecisionTreeClassifier accuracy: 71.29%
              precision    recall  f1-score   support

           0       0.71      0.73      0.72     12500
           1       0.72      0.70      0.71     12499

    accuracy                           0.71     24999
   macro avg       0.71      0.71      0.71     24999
weighted avg       0.71      0.71      0.71     24999
```

Fig. 8. IMDb Reviews Dataset - Classification Report - Decision Tree

```
-------------------LinearSVC---------------------
Time to fit: 5.266s
LinearSVC accuracy: 83.13%
                          precision    recall  f1-score   support

              alt.atheism       0.77      0.72      0.75       319
            comp.graphics       0.74      0.80      0.77       389
  comp.os.ms-windows.misc       0.75      0.72      0.74       394
 comp.sys.ibm.pc.hardware       0.74      0.74      0.74       392
    comp.sys.mac.hardware       0.82      0.85      0.84       385
           comp.windows.x       0.86      0.76      0.81       395
             misc.forsale       0.83      0.90      0.87       390
                rec.autos       0.90      0.88      0.89       396
          rec.motorcycles       0.91      0.94      0.92       398
       rec.sport.baseball       0.90      0.93      0.91       397
         rec.sport.hockey       0.94      0.96      0.95       399
                sci.crypt       0.91      0.91      0.91       396
          sci.electronics       0.75      0.73      0.74       393
                  sci.med       0.85      0.85      0.85       396
                sci.space       0.88      0.91      0.89       394
   soc.religion.christian       0.80      0.91      0.86       398
       talk.politics.guns       0.72      0.88      0.79       364
    talk.politics.mideast       0.97      0.88      0.92       376
       talk.politics.misc       0.78      0.61      0.68       310
       talk.religion.misc       0.71      0.57      0.63       251

                micro avg       0.83      0.83      0.83      7532
                macro avg       0.83      0.82      0.82      7532
             weighted avg       0.83      0.83      0.83      7532
```

Fig. 9. 20 News Group Dataset - Classification Report - Support Vector Machines

```
-------------------LinearSVC---------------------
Time to fit: 7.061s
LinearSVC accuracy: 87.98%
              precision    recall  f1-score   support

           0       0.87      0.89      0.88     12500
           1       0.89      0.87      0.88     12499

    accuracy                           0.88     24999
   macro avg       0.88      0.88      0.88     24999
weighted avg       0.88      0.88      0.88     24999
```

Fig. 10. IMDb Reviews Dataset - Classification Report - Support Vector Machines

```
------------------AdaBoostClassifier--------------------
Time to fit: 48.648s
AdaBoostClassifier accuracy: 47.23%
                          precision    recall  f1-score   support

            alt.atheism       0.68      0.34      0.46       319
          comp.graphics       0.48      0.38      0.42       389
comp.os.ms-windows.misc       0.60      0.43      0.50       394
comp.sys.ibm.pc.hardware      0.50      0.16      0.24       392
   comp.sys.mac.hardware      0.68      0.46      0.55       385
          comp.windows.x      0.76      0.44      0.56       395
            misc.forsale      0.86      0.69      0.76       390
               rec.autos      0.74      0.45      0.56       396
         rec.motorcycles      0.91      0.59      0.71       398
      rec.sport.baseball      0.53      0.46      0.50       397
        rec.sport.hockey      0.88      0.50      0.64       399
               sci.crypt      0.92      0.66      0.77       396
         sci.electronics      0.12      0.80      0.20       393
                 sci.med      0.71      0.17      0.27       396
               sci.space      0.68      0.51      0.58       394
  soc.religion.christian      0.54      0.57      0.56       398
      talk.politics.guns      0.62      0.62      0.62       364
   talk.politics.mideast      0.90      0.53      0.67       376
      talk.politics.misc      0.43      0.38      0.40       310
      talk.religion.misc      0.33      0.17      0.22       251

               micro avg      0.47      0.47      0.47      7532
               macro avg      0.64      0.47      0.51      7532
            weighted avg      0.65      0.47      0.52      7532
```

Fig. 11. 20 News Group Dataset - Classification Report - Ada Boost

```
GridSearchCV took 469.75 seconds for 5 candidates parameter settings.
Model with rank: 1
Mean validation score: 0.836 (std: 0.009)
Parameters: {'clf__C': 5.0, 'clf__penalty': 'l2', 'clf__solver': 'liblinear'}

Model with rank: 2
Mean validation score: 0.833 (std: 0.008)
Parameters: {'clf__C': 3.7750000000000004, 'clf__penalty': 'l2', 'clf__solver': 'liblinear'}

Model with rank: 3
Mean validation score: 0.828 (std: 0.008)
Parameters: {'clf__C': 2.5500000000000003, 'clf__penalty': 'l2', 'clf__solver': 'liblinear'}
```

Fig. 12. Sample output of Logistic Regression Grid Search over 5 candidate parameters

```
------------------AdaBoostClassifier--------------------
Time to fit: 97.153s
AdaBoostClassifier accuracy: 84.22%
              precision    recall  f1-score   support

           0       0.87      0.81      0.84     12500
           1       0.82      0.87      0.85     12499

    accuracy                           0.84     24999
   macro avg       0.84      0.84      0.84     24999
weighted avg       0.84      0.84      0.84     24999
```

Fig. 13. IMDb Reviews Dataset - Classification Report - Ada Boost

```
-----------------RandomForestClassifier--------------------
Time to fit: 123.403s
RandomForestClassifier accuracy: 75.82%
                          precision    recall  f1-score   support

            alt.atheism       0.70      0.54      0.61       319
          comp.graphics       0.67      0.70      0.68       389
comp.os.ms-windows.misc       0.67      0.76      0.71       394
comp.sys.ibm.pc.hardware      0.72      0.69      0.70       392
   comp.sys.mac.hardware      0.77      0.77      0.77       385
          comp.windows.x      0.81      0.71      0.75       395
            misc.forsale      0.77      0.92      0.84       390
               rec.autos      0.79      0.79      0.79       396
         rec.motorcycles      0.87      0.88      0.87       398
      rec.sport.baseball      0.75      0.90      0.82       397
        rec.sport.hockey      0.86      0.92      0.89       399
               sci.crypt      0.86      0.90      0.88       396
         sci.electronics      0.66      0.55      0.60       393
                 sci.med      0.84      0.71      0.77       396
               sci.space      0.78      0.87      0.82       394
  soc.religion.christian      0.62      0.91      0.74       398
      talk.politics.guns      0.66      0.85      0.74       364
   talk.politics.mideast      0.94      0.77      0.85       376
      talk.politics.misc      0.81      0.47      0.60       310
      talk.religion.misc      0.74      0.26      0.39       251

               micro avg      0.76      0.76      0.76      7532
               macro avg      0.76      0.74      0.74      7532
            weighted avg      0.76      0.76      0.75      7532
```

Fig. 14. 20 News Group Dataset - Classification Report - Random Forest

```
-------------------RandomForestClassifier---------------------
Time to fit: 275.989s
RandomForestClassifier accuracy: 85.28%
              precision    recall  f1-score   support

           0       0.85      0.86      0.85     12500
           1       0.86      0.84      0.85     12499

    accuracy                           0.85     24999
   macro avg       0.85      0.85      0.85     24999
weighted avg       0.85      0.85      0.85     24999
```

Fig. 15. IMDb Reviews Dataset - Classification Report - Random Forest