Analysis of News Articles using Machine Learning Algorithms

Shashvat Kedia(1610110347)
Shivam Bansal(1610110351)

April 2019

ABSTRACT

Nowadays many machine learning algorithms are evolving it is very difficult task to select a particular algorithm for a specific problem. A multiple models have to be tried out on the given input data to select a particular algorithm. In this study a case study has been taken for selecting an algorithm for the classification of news articles. Multi- perspective analysis is performed on the data using various machine learning algorithms namely Random Forest Classifier, Decision tree, SVM. For the multi perspective analysis, features from the dataset are extracted and standard metrics are used. The metrics used are Accuracy, F- measure [10], Recall, and Precision. For the BBC news standard dataset [6], Random forest classifier [3] seems to be effective as its accuracy is **95.9**.

1 INTRODUCTION

Machine Learning is applied in solving many problems like Intrusion Detection Systems [4], Mobile Class Prediction [2], Music analytics [8] etc. In today's era, machine learning plays an important role in data analysis for which choosing an appropriate algorithm is important. To analyze data, multi- perspective analysis of data is to be performed. Further depending on the features of the data, machine learning is chosen. There are two broad domains of machine learning algorithms: supervised and unsupervised. Supervised learning can further be broken down into three categories Regres-

sion, Classification and Clustering the task that we are trying to solve is a classification one some of the well known classification algorithms are:-
1. Naive Bayes [5]

2. Logistic Regression [1]

3. Random Forests [3]

4. Decision Trees [9]

5. Support Vector Machine (SVM) [7]

Out of these well known classification algorithms some of which we are going to use for out task are:-
1. Random Forest [3]

2. Decision Trees [9]

3. Support Vector Machines (SVM) [7]

These algorithms are prominently used in Machine Learning Toolkit Provided with Python named scikit-learn
1. Random Forest Classifier: This classifier is built on an ensemble learning method for classification which gathers all types of decision trees at training time and classifies data based on the mode of the class.

2. Decision tree: It is a classifier which builds a tree-like structure based of probability of an event to happen based on its outcome, resource, cost and utility.

3.SVM with Linear SVC: It is Support Vector Classification with Linear Kernel.

2 PROPOSED WORK

This work proposes to use different classifiers on Standard Text data. The steps performed are as explained below:
1. Take input data for training classifiers,

2. Train the classifiers using various machine learning techniques,

3. Take the same input data for testing the trained classifiers,

4. Create a Confusion Matrix,

5. Calculate F-measure, Recall, Precision and Accuracy of each classifier,

3 IMPLEMENTATION

BBC news dataset consists of 2225 documents divided into five categories namely business, sports, entertainment, politics and tech. For multi-perspective analysis text features from the dataset were extracted which were 14788 words. The total dataset was divided into two categories of train and test data. The train data consisted of 1780 documents whereas the test data consisted of 445 documents. A confusion matrix was created for this data when it was executed with various machine learning algorithms.
The confusion matrix provided accuracy of the algorithm applied. Additionally, standard metrics like F-measure, Recall and Precision were calculated. The sequence of best suited algorithms for classification of BBC news data was derived using standard metrics. Using confusion matrix, False Positives and Classification Rate were calculated and the derived sequence was verified.

Importing Libraries and Data Set :
Using prebuilt libraries make our work simpler and improves our efficiency. We used major libraries such as :

Pandas : pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the

Python programming language.

numpy :NumPy is the fundamental package for scientific computing with Python

sklearn : Sklearn is a machine learning library for the Python programming language. It features various classification, regression and clustering algorithms.

nltk : Natural Language Toolkit is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language.

matplotlib : Matplotlib is a plotting library for the Python programming language.

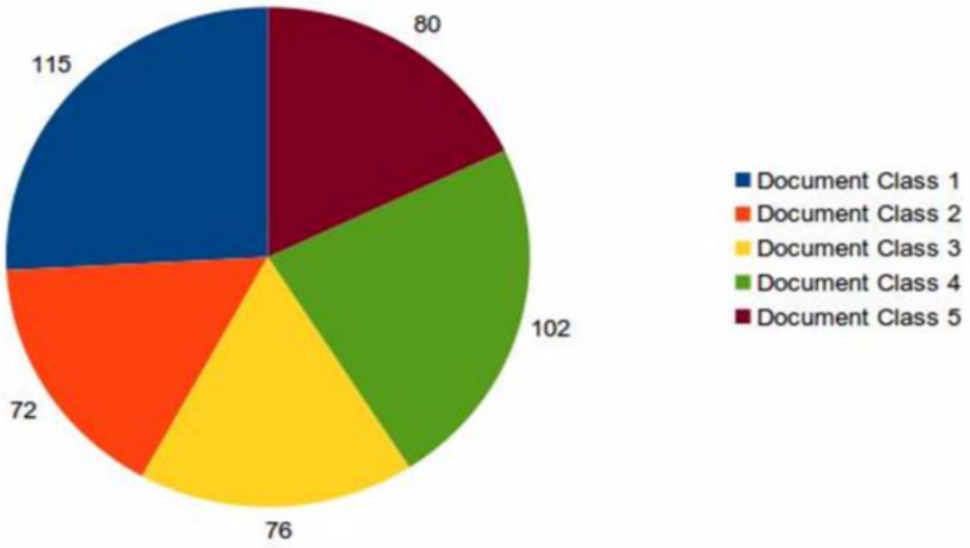
Preprocessing of Data / Cleaning of Data : Steps involved :
1. Removing Stopwords : Identify the stop words in any language using the nltk libraries and remove all such works in data set.

2. Tokenize: Tokenize the data into smaller tokens

3. Stemming: Apply stemming on the data sets

4. Generating Tf-Idf scores: It usually takes the feature values and stores them separately into a Matrix and gets relevant scores for it.

Splitting the Data into Test and Train :
We split the dataset into two: Test and train using preprocessing libraries .Our model was built on 80 percent train data and 20 percent test data.



4 RESULTS

Class 1, 2, 3, 4 and 5 represent 5 categories from the BBC news dataset. These are depicted in Figure 2. Class 1, class 2, class 3, class 4 and class 5 classified by various classifiers are shown in the confusion matrix generated for analysis. Standard metrics are calculated for Random Forrest Classifier, Decision tree, SVM. Further we have also calculated several metrics like Precision = (TP / TP + FP), Recall (TP / TP + FN), F1-Score = 2 * (Precision * Recall)/(Precision + Recall) with respect to each for the classes for each of the classification algorithms we have appliend on the dataset the results for the same is listed below.

Table 1 : Classifiers and Standard Metrics used in Machine Learning

	Random Forest Classifier	Decision Tree Classifier	SVM
Accuracy	0.959	0.847	0.919
F- measure	0.940	0.810	0.872
Precision	0.965	0.817	0.947
Recall	0.952	0.813	0.908

Table 2 : Precision and Recall for Random Forest Classifier

Class	Precision	Recall	F1-Score
1	0.98	0.94	0.96
2	0.97	0.95	0.96
3	0.95	0.99	0.97
4	0.97	0.95	0.97
5	0.96	0.96	0.96

Table 3 : Precision and Recall for Decision Tree Classifier

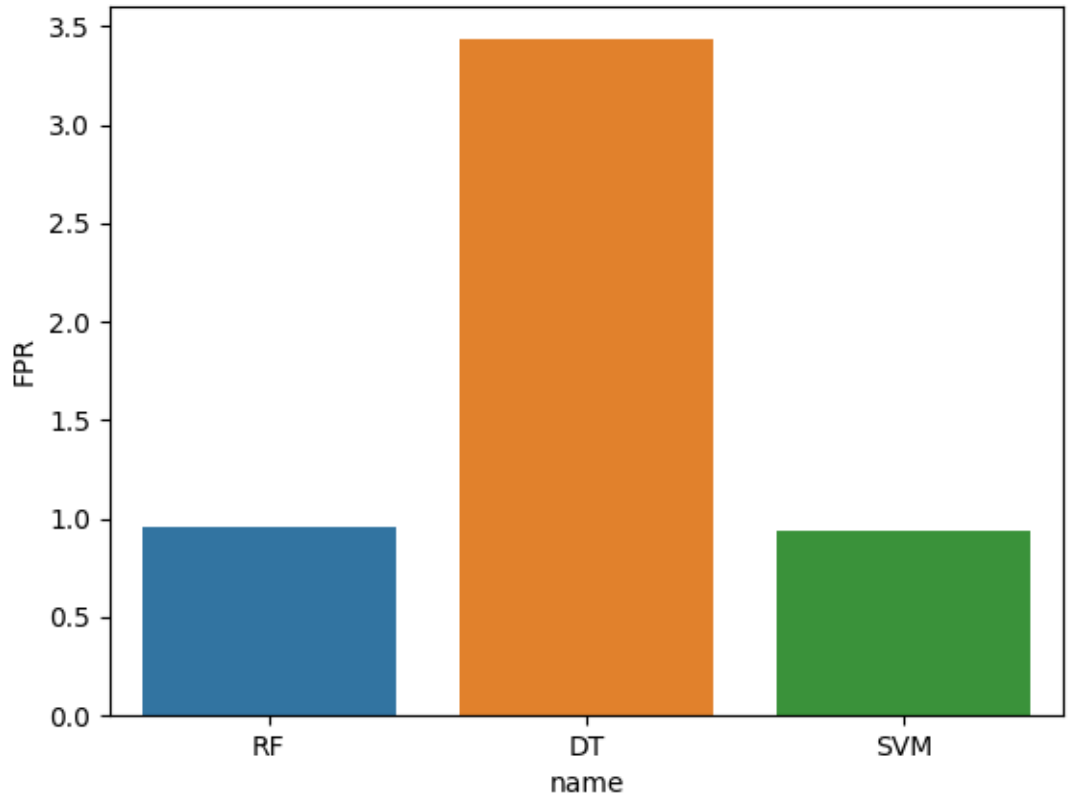
Class	Precision	Recall	F1-Score
1	0.89	0.82	0.85
2	0.83	0.85	0.82
3	0.85	0.85	0.85
4	0.89	0.95	0.92
5	0.86	0.82	0.83

Table 4 : Precision and Recall for SVM

Class	Precision	Recall	F1-Score
1	0.96	0.92	0.94
2	0.95	0.95	0.96
3	0.93	0.96	0.95
4	0.92	0.95	0.95
5	0.96	0.96	0.96

The comparison of various models can bes done based on the false positive rate as well as the classification rate for the following can be seen here:
1. Random Forrest Classifier
2. Decision Tree Classifier
3. SVM

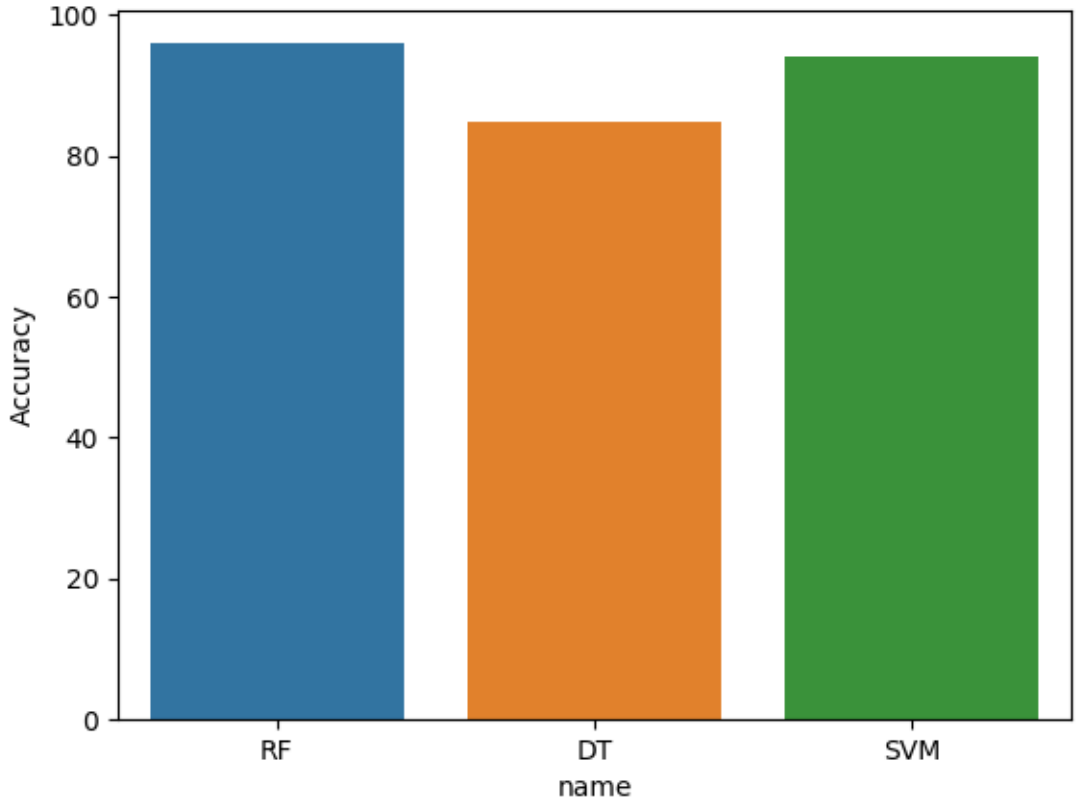
Figure 1: False Positive Rate for classifiers



False Positive Rate is given by FP / (FP + TN) and signifies the proportion of the no. of data points that have been classified as positive to no.of data points belonging to negative class(in this case any of the other class).

Classification Rate is measure of no. of data points that have been correctly classifier by the model i.e accuracy of the model which is calculate using (TP + TN) / (TP + FP + TN + FN).

Figure 2: Classification Rate for classifier



5 CONCLUSION

Table 5 Metrics for various classifiers

	Random Forest Classifier	Decision Tree Classifier	SVM
Accuracy	0.959	0.847	0.919
F- measure	0.940	0.810	0.872
Precision	0.965	0.817	0.947
Recall	0.952	0.813	0.908
FPR	0.908	3.433	0.757
Classification Rate	95.955	84.719	94.43

Based on the results that we have obtained Random forest classifier seems

to be the best fit as compared to other models as its provides the highest accuracy. The performance of the models can further be improved by replacing the way we represent words i.e accuracy can be improved by using word embeddings instead of Tf-idf as word embeddings provide more information regarding any particular word.

6 LIMITATIONS

One of the biggest limitations that our model has is that it has been trained on unbalanced dataset i.e the class probablities for each class are not equals as the dataset does not contain same no. of instances of each class this results in the model to be biased towards the class that appears majority number of times in the dataset thus affecting model performance.

References

[1] Paul Allison. *Logistic Regression Using Sas®: Theory and Application*. SAS Publishing, first edition, 1999.

[2] Muhammad Asim and Zafar Khan. Mobile price class prediction using machine learning techniques. *International Journal of Computer Applications*, 179:6–11, 03 2018.

[3] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[4] Roberto Di Pietro and Luigi V. Mancini. *Intrusion Detection Systems*. Springer Publishing Company, Incorporated, 1 edition, 2008.

[5] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Mach. Learn.*, 29(2-3):131–163, November 1997.

[6] Derek Greene and Pádraig Cunningham. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proc. 23rd International Conference on Machine learning (ICML'06)*, pages 377–384. ACM Press, 2006.

[7] Marti A. Hearst. Support vector machines. *IEEE Intelligent Systems*, 13(4):18–28, July 1998.

[8] David Meredith. *Computational Music Analysis*. Springer Publishing Company, Incorporated, 1st edition, 2015.

[9] J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, March 1986.

[10] Ethan Zhang and Yi Zhang. *F-Measure*, pages 1147–1147. Springer US, Boston, MA, 2009.