DATA MINING AND WAREHOUSING

REPORT
FAKE NEWS CLASSIFICATION

Introduction

This project tries to classify a news item (i.e a news statement) as fake or real using two machine learning algorithms: Random forest and Logistic Regression. A comparison of the two shows the following :

Data

The models have been trained with the data from the LIAR dataset which is a new benchmark dataset for fake news detection.

https://arxiv.org/abs/1705.00648

Implementation

Data preparation:

1. The train and test data is contained in the files Fake-new-classifier/Data/. Create a dictionary of the occurrence of each word in the training dataset.
2. Pick the top 3000 words as frequent words and create a feature matrix of these words and the corresponding labels
3. Train using **RandomForestClassifier**() and **LogisticRegression()**

Random forest

Random forest classifier creates a set of decision trees from a randomly selected subset of the training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

RandomForestClassifier(n_estimators,criterion). I have used n_estimators (No of trees)=300 , criterion (Branching criterion) = gini

Logistic regression

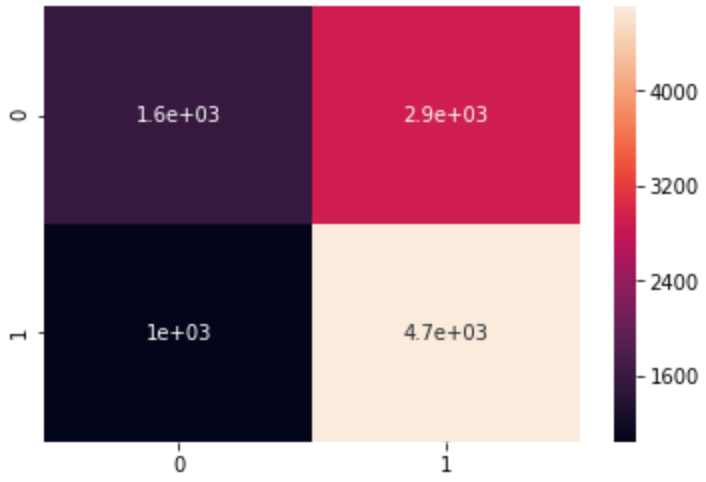
Logistic regression id a machine learning technique that can be used to classify data into discrete classes (in this case binary classes 0- fake and 1-true). Logistic regression uses the logistic function or the sigmoid function, 1 / (1 + e^-value) that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

Result

Logistic Regression:

Confusion matrix:

	0	1
0	1580	2908
1	1043	4709



Total statements classified: 10240

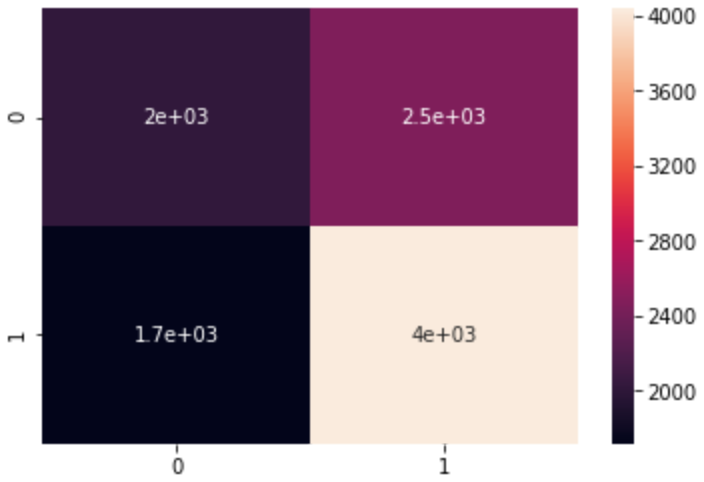
Score: 0.7044355553757985

score length 5

Random forest classifier

Confusion matrix:

	0	1
0	2022	2466
1	1717	4035



Total statements classified: 10240

Score: 0.6585569269368081

score length 5