Fake News Detection Using Machine Learning Algorithms

SNAPSHOTS OF SYSTEM WORKING

A. Static System-

Figure : Static output (True)

Figure : Static Output (False) B.

Dynamic System-

Figure : Fake News Detector (Home Screen)

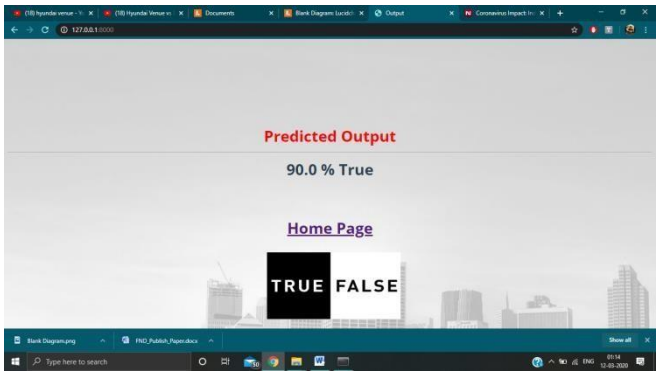

Figure : Fake News Detector (Output page)

RESULTS

Implementation was done using the above algorithms with Vector features- Count Vectors and Tf-Idf vectors at Word level and Ngramlevel. Accuracy was noted for all models. We used K-fold cross validation technique to improve the effectiveness of the models.

A. Dataset split using K-fold cross validation

This cross-validation technique was used for splitting the dataset randomly into k-folds. (k-1) folds were used for building the model while k^{th} fold was used to check the effectiveness of the model. This was repeated until each of the k-folds served as the test set. I used 3fold cross validation for this experiment where 67% of the data is used for training the model and remaining 33% for testing. B. Confusion Matrices for Static System

After applying various extracted features (Bag-of-words, Tf-Idf. N-grams) on three different classifiers (Naïve bayes, Logistic Regression and Random Forest), their confusion matrix showing actual set and predicted sets are mentioned below:

Table 2: Confusion Matrix for Naïve Bayes Classifier using Tf-Idf features-

Total= 10240	Naïve Bayes Classifier	
	Fake (Predicted)	*True (Predicted)*
Fake (Actual)	841	3647
True (Actual)	427	5325

Table 3: Confusion Matrix for Logistic Regresssion using Tf-Idf features-

Total= 10240	Logistic Regression	
	Fake (Predicted)	*True (Predicted)*
Fake (Actual)	1617	2871
True (Actual)	1097	4655

Table 4: Confusion Matrix for Random Forest Classifier using Tf-Idf features-

Total= 10240	Random Forest	
	Fake (Predicted)	*True (Predicted)*
Fake (Actual)	1979	2509
True (Actual)	1630	4122

Table 5: Comparison of Precision, Recall, F1-scores and Accuracy for all three classifiers-

Classifiers	Precision	Recall	F1-Score	Accuracy
Naïve Bayes	0.59	0.92	0.72	0.60
Random Forest	0.62	0.71	0.67	0.59
Logistic Regression	0.69	0.83	0.75	0.65

As evident above our best model came out to be Logistic Regression with an accuracy of 65%. Hence we then used grid search parameter optimization to increase the performance of logistic regression which then gave us the accuracy of 80%.

Hence we can say that if a user feed a particular news article or its headline in our model, there are 80% chances that it will be classified to its true nature.

C. Confusion Matrix for Dynamic System

We used real_or_fake.csv with passive aggressive classifier and obtained the following confusion matrix-

Table 6: Confusion Matrix for passive aggressive Classifier-

Total= 1267	Passive Aggressive Classifier	
	Fake (Predicted)	*True (Predicted)*
Fake (Actual)	588	50
True (Actual)	42	587

Table 7: Performance measures-

Classifier	Precision	Recall	F1-Score	Accuracy
PAC	0.93	0.9216	0.9257	0.9273