

Fake-News Detection Project

Submitted by:

Biranchi Narayan Nath

**ACKNOWLEDGMENT**

The dataset and to do instructions of the project was provided by my internship Company FlipRobo . During working with the project I have taken some help from my SME Nitin Mishra and from notes and previous projects to complete the project.

**INTRODUCTION**

* Business Problem Framing

The authenticity of Information has become a longstanding issue affecting businesses and society, both for printed and digital media. On social networks, the reach and effects of information spread occur at such a fast pace and so amplified that distorted, inaccurate, or false information acquires a tremendous potential to cause real-world impacts, within minutes, for millions of users. Recently, several public concerns about this problem and some approaches to mitigate the problem were expressed.

So our work is to build a model to detect the Fake news. This project is highly associated with real word. Because in the world we stay, work & move depending on news. And circulation of Fake news can effect a lot to everyone.

* Conceptual Background of the Domain Problem

This is a binary classification Problem.

* Review of Literature

The shape of the dataset is (20800,6) means it have 6 features and 20800 observations. In the dataset both the classes of Fake and True news are approximately 50-50%.

* Motivation for the Problem Undertaken

The main moto is to build a model which can predict Fake and True News correctly. Building a best model to predict the news classes is a great achievement for me and all the news users.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

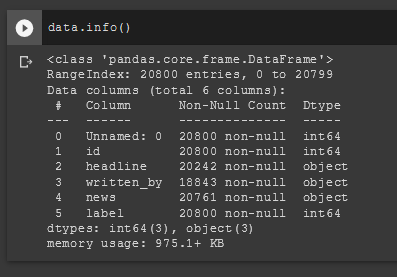
During working in the project, mathematics and statistics are used behind most of the technique and algorithms.

In .describe() method we find the statistical summary of the dataset so that one can know the data, in the technique we find the count, mean, standard deviation, minimum, maximum and percentiles like 25th,50th & 75th of all the features. Knowledge of statistical summary is highly required before working with any project.

Data Sources and their formats

The data is provided by my internship Company FlipRobo the origin of the data is news. The dataset have 6 features and 20800 rows/observations.In the 6 features the datatype of 3 features are integer and rest 3 features are object. This is a binary classification problem. The features are

Unnamed: 0,id,headline, written\_by,news and labels.



* Data Preprocessing Done

Mostly the data we get for model building is not cleaned. So data cleaning is done to get cleaned data for model building.

* Hardware and Software Requirements and Tools Used

**Hardware:** Laptop and Mouse.

**Software:**

i. Jupyter Notebook: All the libraries, packages and tools are called by it.

ii. Pandas: for reading csv file, using dataframe and dataframe related operations like checking shape, information and statistical summary. Saving the prediction in csv file.

iii.matplotlib & seaborn: both are used for plotting

iv.sklearn: It is used for splitting the dataset into train test sets, For vectorization, getting algorithms used for model building , metrics used for checking model capacity and saving the model for further use.

v.nltk is used to get the stopwords

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

For solving the problem the approaches are

i. Data Preprocessing:

In this step EDA is performed to know the dataset. So that to find clear vision about the data and perform as per required. In this step data is cleaned for model building.

ii. Dataset splitting: Dataset is splitted into train test split.

iii. Model Building: multiple models are created.

iv. Model Evaluating: Among all the models the best model is choose for Final model.

* Testing of Identified Approaches (Algorithms)

The algorithms are:

i.MultinomialNB

ii. LogisticRegression

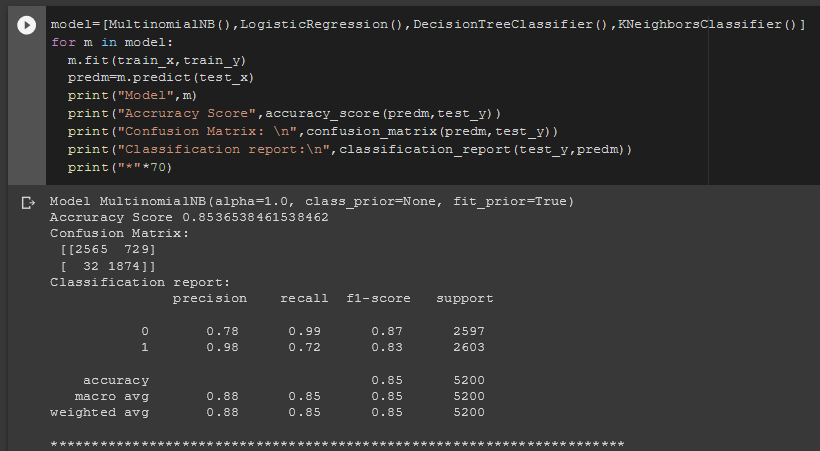
iii.DecisionTreeClassifier

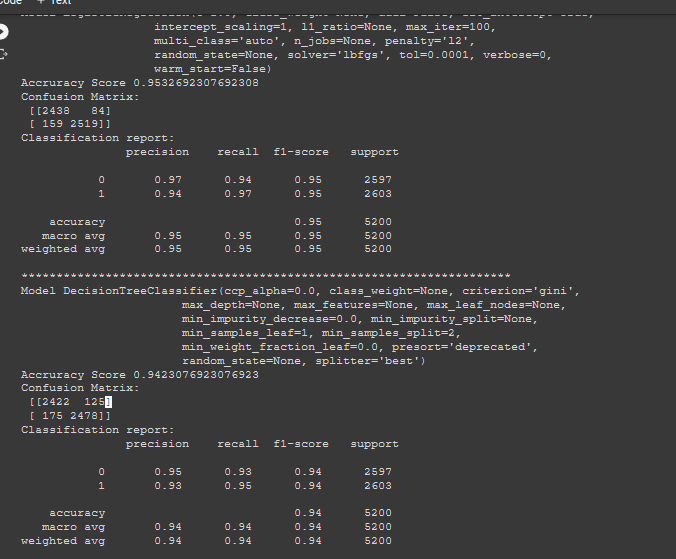
iv. KNeighborsClassifier

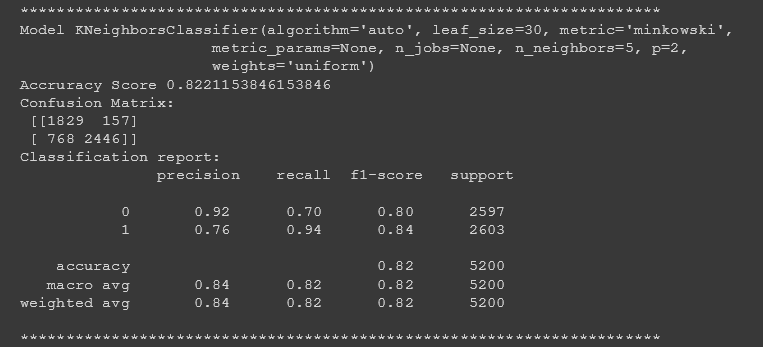
v. RandomForestClassifier

* Run and Evaluate selected models

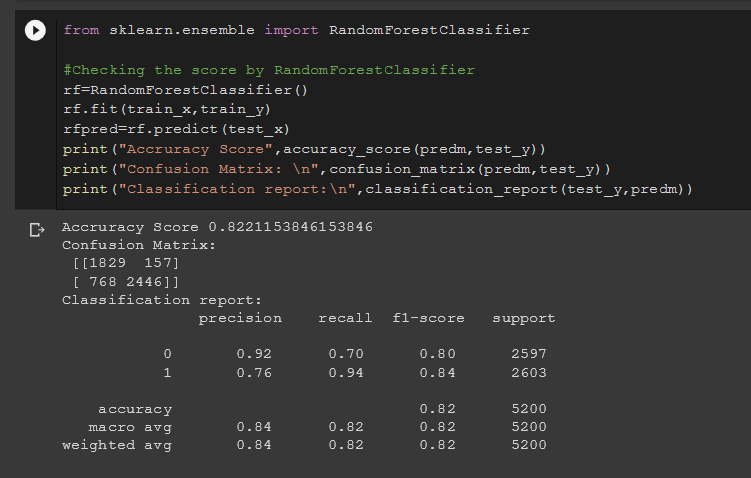
MultinomialNB,LogisticRegression, DecisionTreeClassifier & KNeighborsClassifier







RandomForest



* Key Metrics for success in solving problem under consideration

What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

i.accuracy\_score: As it is a classification problem accuracy score is required to check the accuracy score so it is used.

ii.confusion\_matrix: To see the number tp,fp,tn,fn.

iii)Classification\_report: To check all the metrics like precision,recall,f1-score it is used.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Among all the used 5 models LogisticRegression has highest f1-score, accuracy, recall and precision. So we finalize the model for final model building.