

Fake News Prediction

Submitted by:

Mandar M Joshi

**ACKNOWLEDGMENT**

In this project, I have built a classifier model that can identify news as real or fake. Nowadays, fake news has become a common trend. Even trusted media houses are known to spread fake news and are losing their credibility.

**INTRODUCTION**

* Business Problem Framing

Nowadays, fake news has become a common trend. Even trusted media houses are known to spread fake news and are losing their credibility. So, how can we trust any news to be real or fake

* Conceptual Background of the Domain Problem

In this world of media and digital marketing systems, its very much difficult to find out real or fake news for any of individuals or organization to stop hazardous effects of fake news on business.

* Review of Literature

Go through a lot of online reaserch which leds to take over the fake news predictions and reactions on the people and organization on the circumstances and understand the importance to analyse the news to save a lot of hazardous effects onto the society.

* Motivation for the Problem Undertaken

The main motivation is to save the entire society from fake news and prevent a lots of losses.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

There are total 5columns in the dataset which contains “id”,”headline”,”written\_by”,”news” and “label”.the feature which is known as “label” contains the exact figure of 0 as a no fake news and “1” as a fake news so to deal with the text my approach is to collect all these text into one single feature in order to perform necessary actions.

* Data Sources and their formats

Features such as “label” and “id” contains the numerical values.”label” gives the information about either the news is a fake or real and “id” represents the unique number to every news.so “id” is the feature which is non important for the modelling and remaining features we can convert it into single feature in order to supply the model.

* Data Preprocessing Done

In data preprocessing extensively used stopwords , tokenization, lemmatization, count vectorizer ,TF-IDF vectorizer.

* Data Inputs- Logic- Output Relationships

Provided with the csv file to read into jupyter notebook ,basically data input doesn’t directly affect on data output because all the necessary data cleaning or data preprocessing can be done manually.

* State the set of assumptions (if any) related to the problem under consideration

When I check the shape of dataset which contains 20800 rows my assumption was model will take a time to train.

* Hardware and Software Requirements and Tools Used

Laptop with 8gb ram and dual core processor will be a nice selection for this project. Extensively used jupter notebook for data preprocessing and training.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

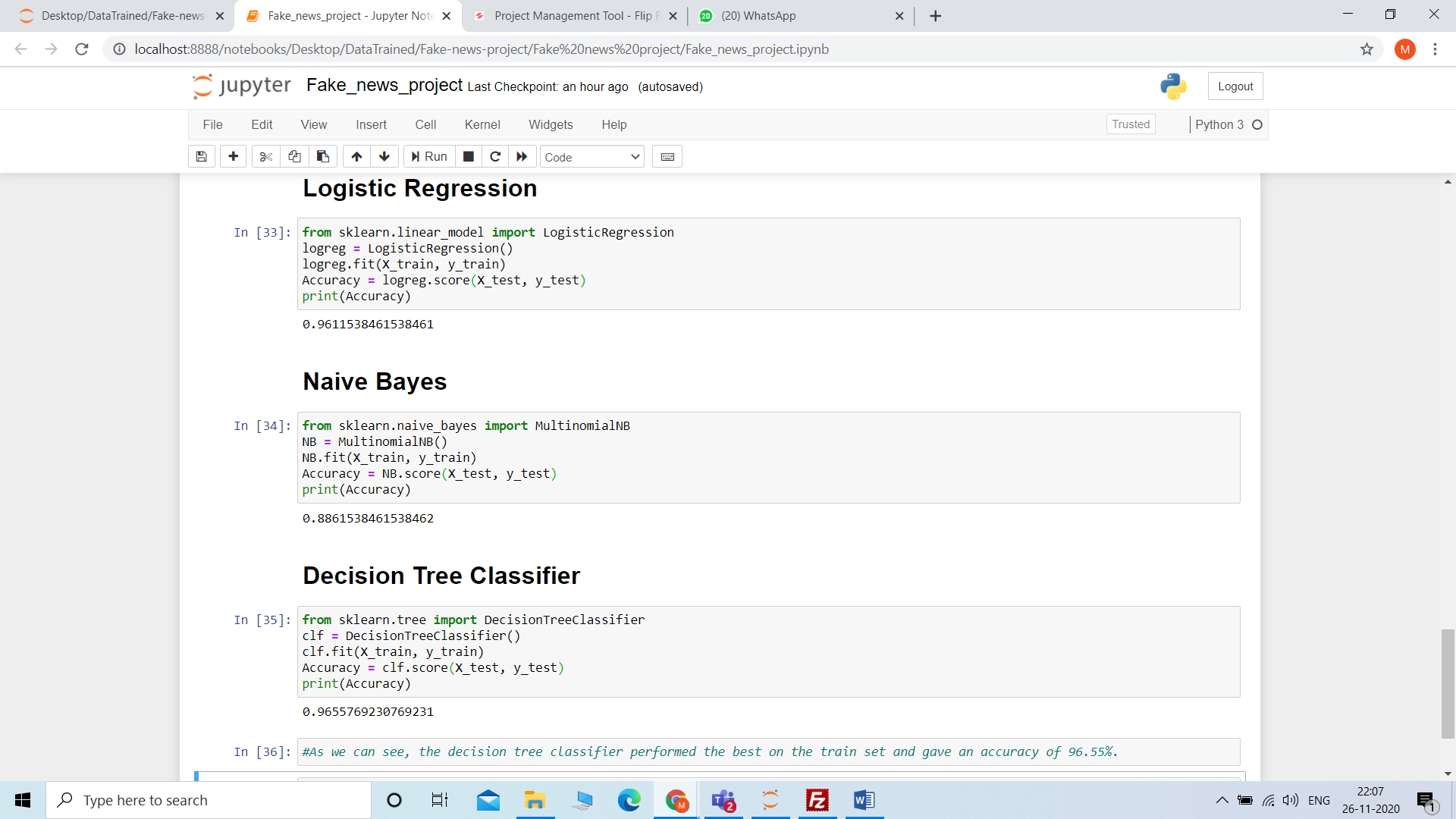
My only single approach for this problem is to collect every workds into single column and then moving ahead with the data preprocessing and model creation.

* Testing of Identified Approaches (Algorithms)

I have supplied a data to 3 models names are logistic regression,naïve bayes and decision tree classifier.

* Run and Evaluate selected models

I have supplied a data to 3 models names are logistic regression,naïve bayes and decision tree classifier

.

* Key Metrics for success in solving problem under consideration

accuracy criteria is the main parameter that need to check with the model predictions.

* Visualizations

No need of visualization.

* Interpretation of the Results

After doing data cleaning,data preprocessing , applying statistical methods such as used stopwords , tokenization, lemmatization, count vectorizer ,TF-IDF vectorizer.achieved accuracy logisctic regression – 96.11%,naïve bayes – 88.61,decision tree classifier-96.55%.

**CONCLUSION**

* Key Findings and Conclusions of the Study

By achieving the accuracy such as 96.55% we can use this model to predict fake or real news.

* Learning Outcomes of the Study in respect of Data Science

Learn a lot of things such as combining a text columns into single column apppying multiple techniques such as stopwords,lemmatization,count vectorizer ,TF-IDF model, and after converting all the data and supplied to the model with achieving sustainable accuracy.

* Limitations of this work and Scope for Future Work

The data is a small means only have 20800 rows so if the data is bigger then have to perform more datapreprocessing and used more algorithms.