**CSE 587 Lab 3 Report**

**Suhit Datta**

**Sourav Ranu**

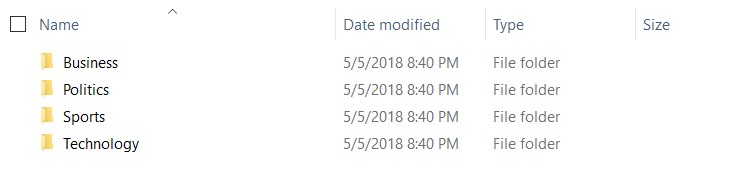
Data has been collected from NY Times using a Python code (***Data Collection.ipynb***) that has been created as a part of Lab2.

Four categories of data have been considered:

1. Politics
2. Sports
3. Business
4. Technology (own choice)

Multiple articles (greater than 50 articles) have been collected and are distributed into different folders with the same name.

This is contained in a folder called **DIC\_Project\_Data**

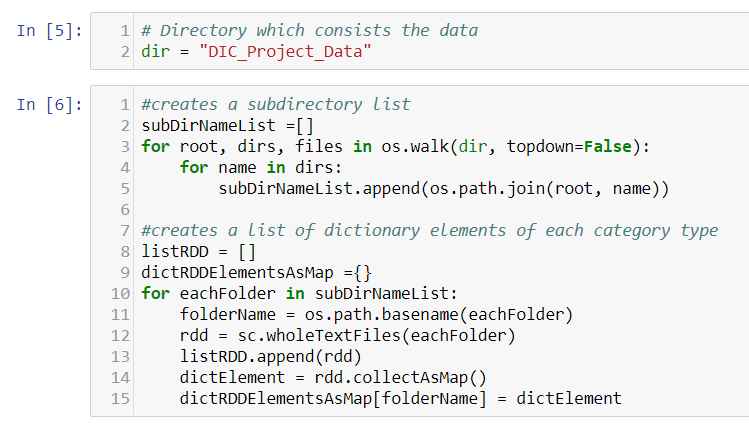


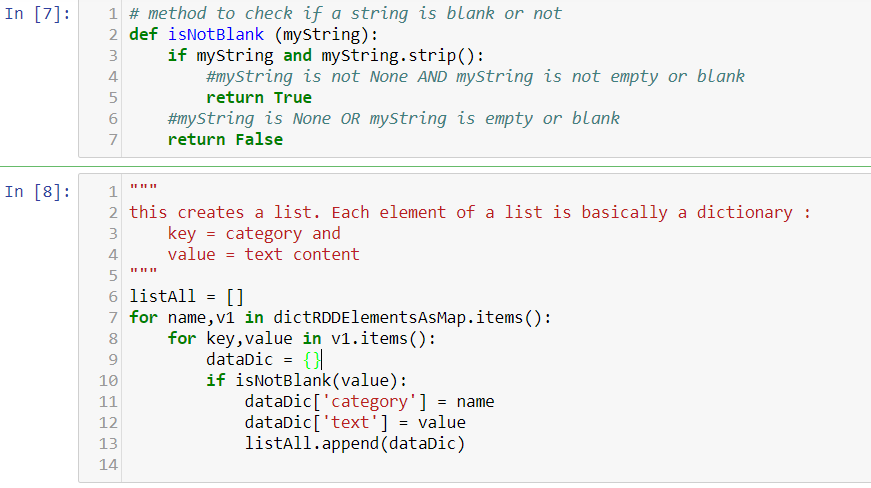
Code has been written in PySpark (Python) (***Lab3\_Code.ipynb***)

Data has been collected from the directory and stored in the form of a dictionary with key being the category name and the value being the content.

The elements of the dictionary are placed in a list in a specific format.

The code snippet exemplifies the process :





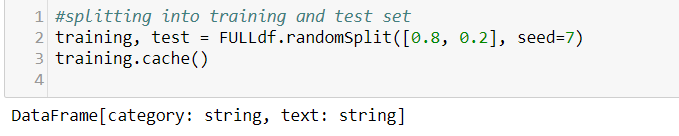
In this case, **listAll** consists of the contents.



Once the data has been extracted from these folders, it is collected to form a Spark Dataframe with two columns:

1. Text
2. Category

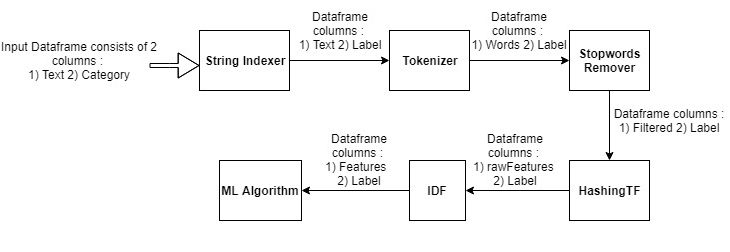
This dataframe is split into Train dataframe and Test dataframe in the ratio 80:20



The model is trained using the Training Dataframe.

The various processes in the pipeline is mentioned as below:

**Pipeline Flowchart:**

****

The model thus created is used to test on the Test dataframe that has been separated earlier from the initial dataframe.

The following are the **ML Algorithms** used:

1) Random Forest

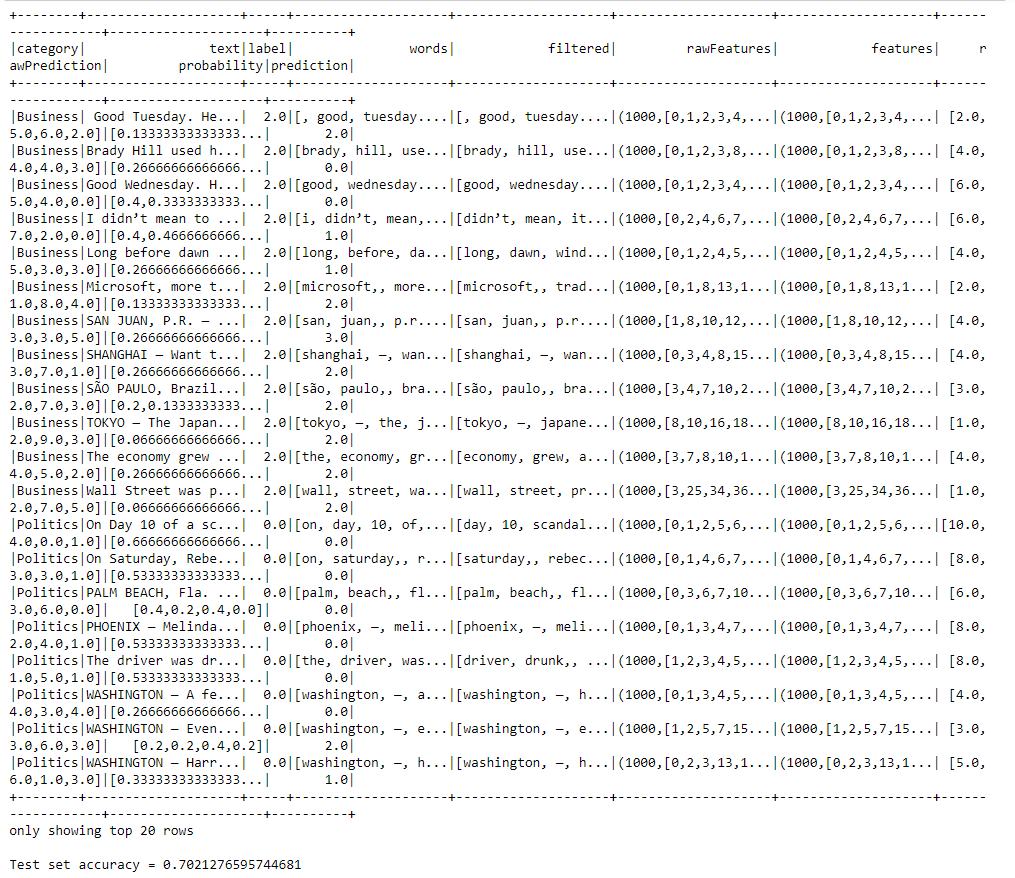
2) Naïve Bayes

3) Multiclass Logistic Regression

1. Code snippet for Random Forest



The output obtained is as follows:

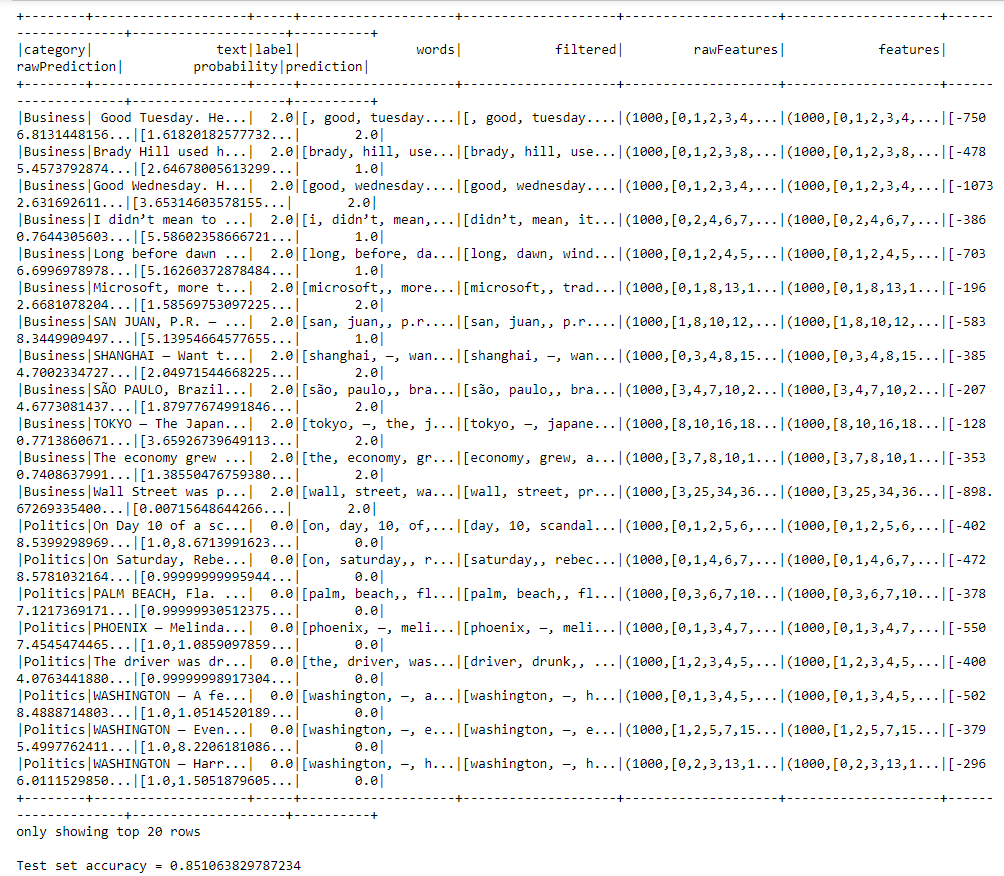


Accuracy of **70.2127 %** is obtained for **Random Forest** on the Test dataframe.

1. Code snippet for Naive Bayes :



The output obtained is as follows:

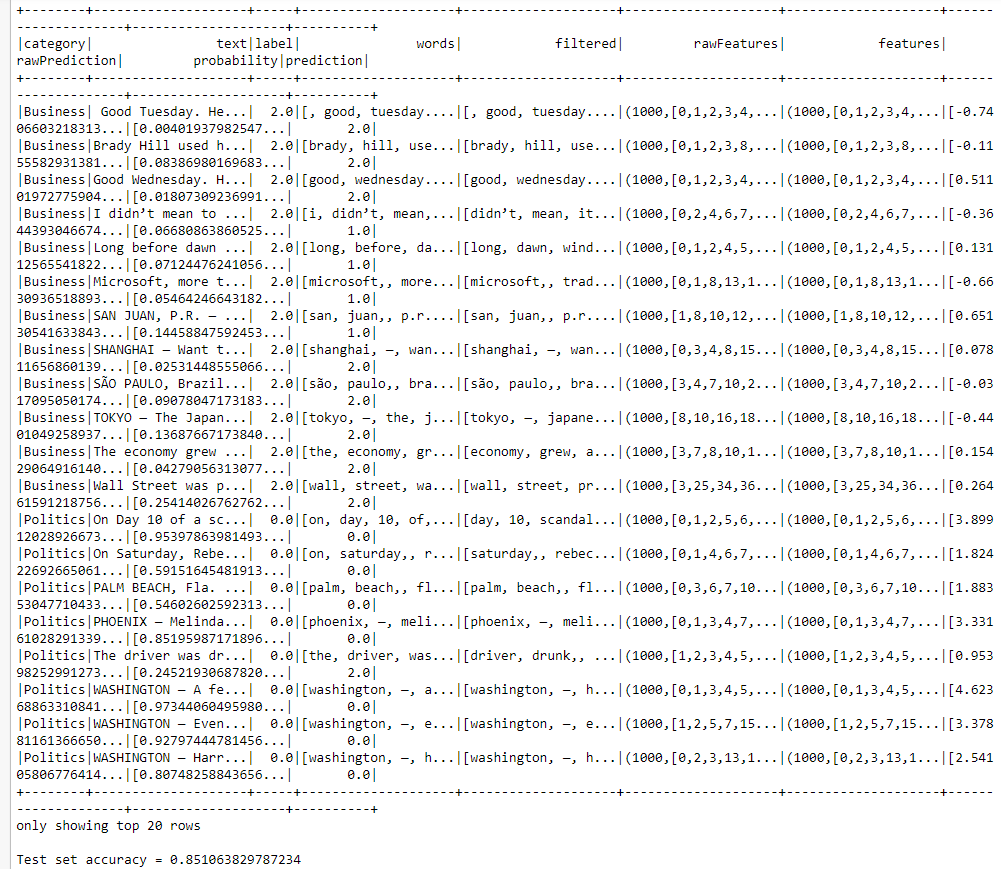


Accuracy of **85.1063 %** is obtained for **Naïve Bayes** on the Test dataframe.

1. Code snippet for Logistic Regression



The output obtained is as follows:

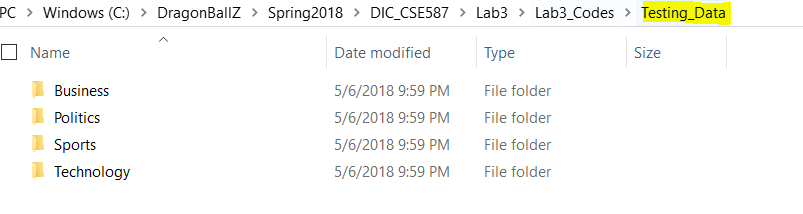


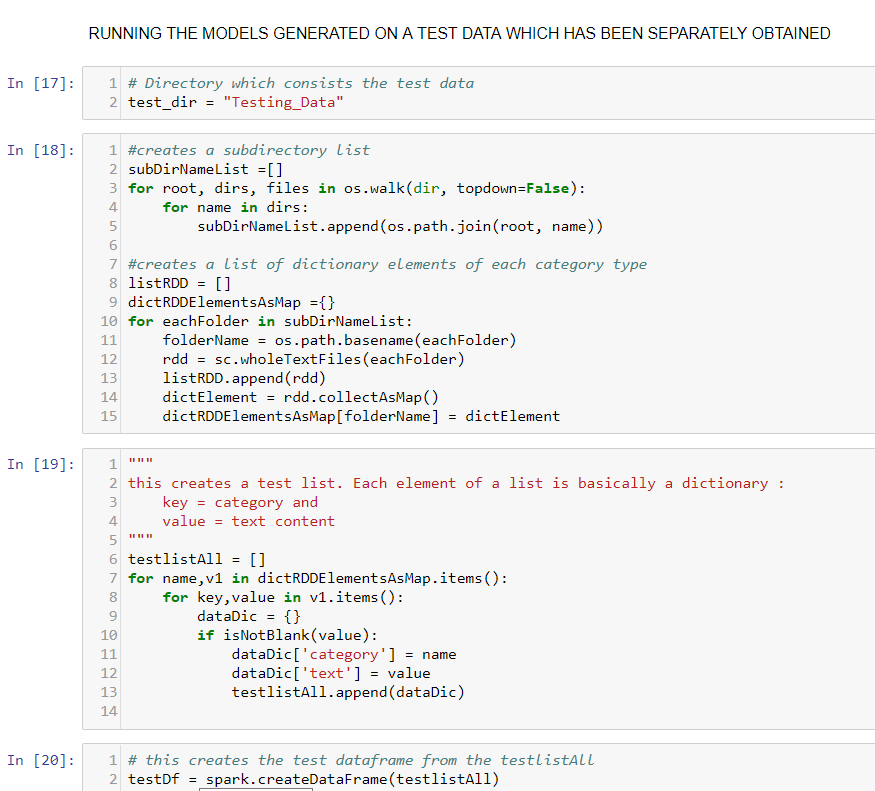
Accuracy of **85.10638 %** is obtained for Logistic Regression on the Test dataframe.

**Running the Model on Unknown Test Data**

A similar approach for data collection has been taken to collect Test data. Only this time lesser number of articles have been collected as compared to the previous step.

The data has been randomly collected. A folder Testing\_Data has been created which consists of the subfolders as shown below.

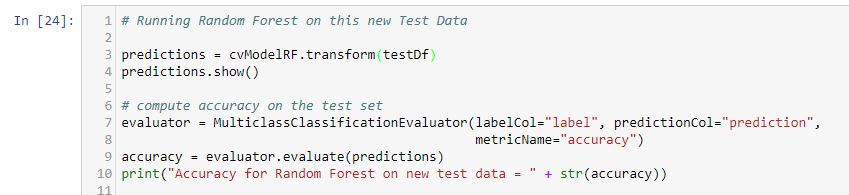




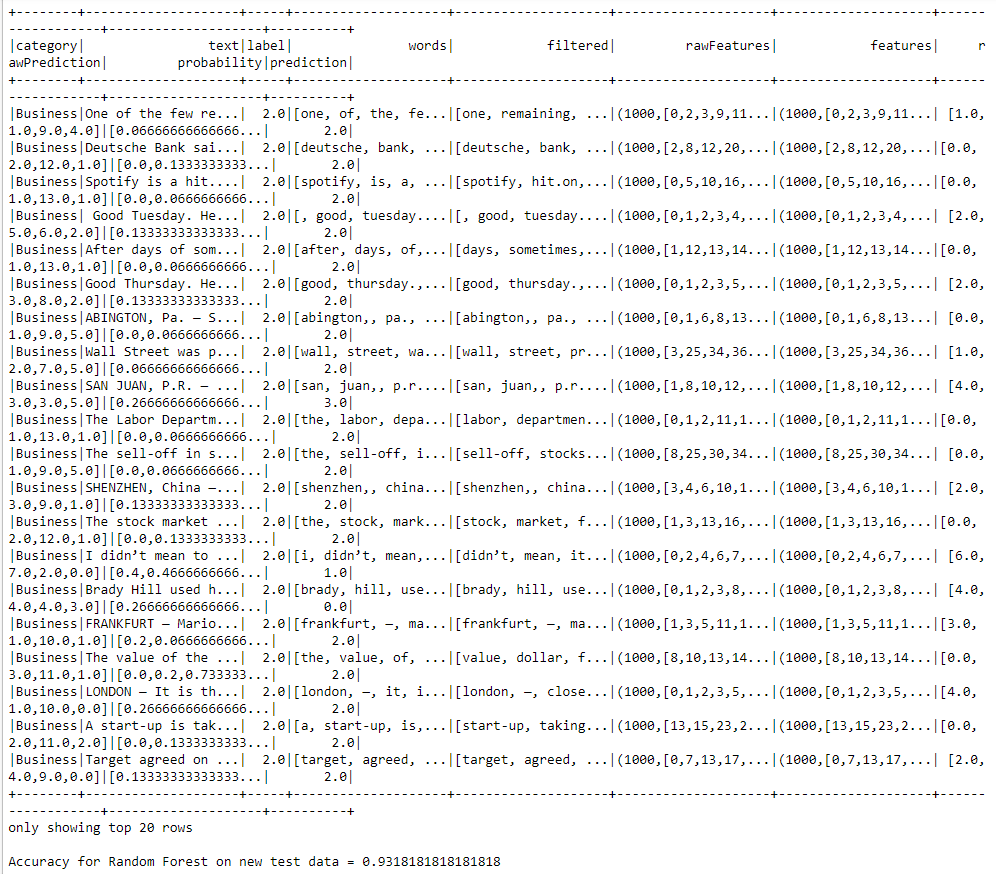
The same ML algorithms have been run in the following order:

1. **Random Forest**

Input snippet is as follows:



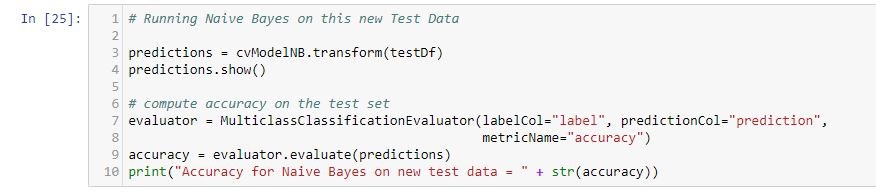
Output obtained is as follows:



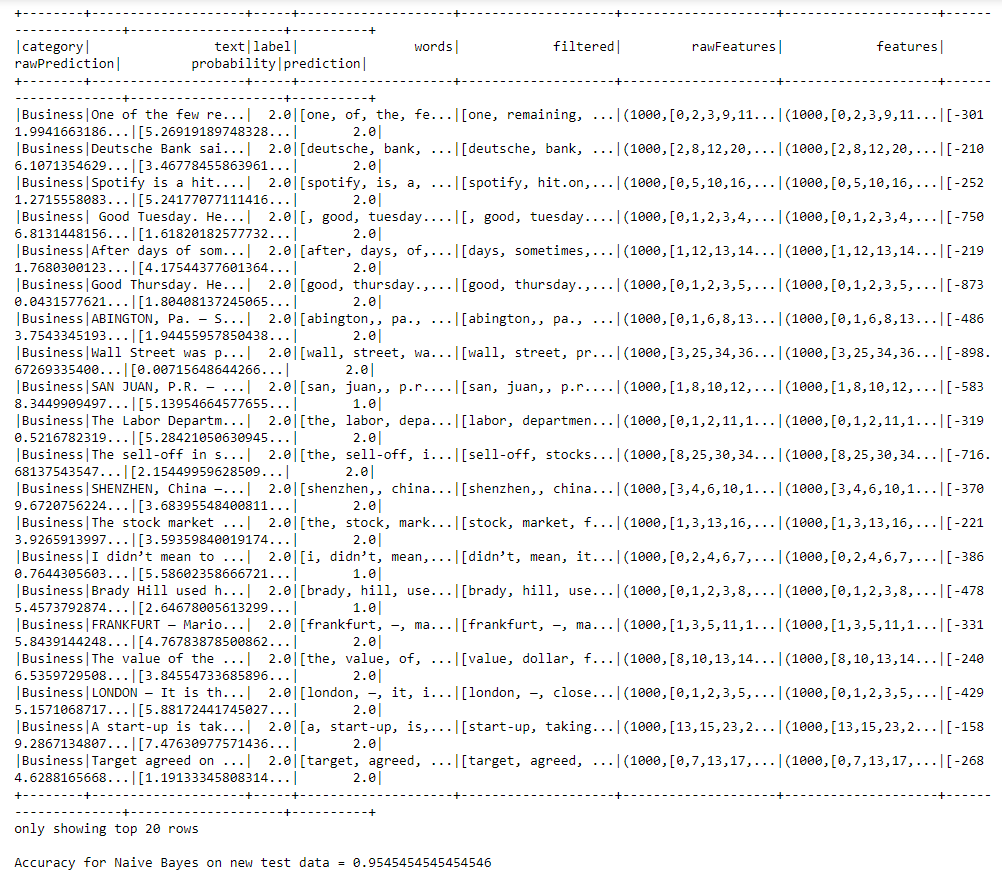
Accuracy % for this new test data = 93.1818 using Random Forest

1. **Naïve Bayes**

Input snippet is as follows:



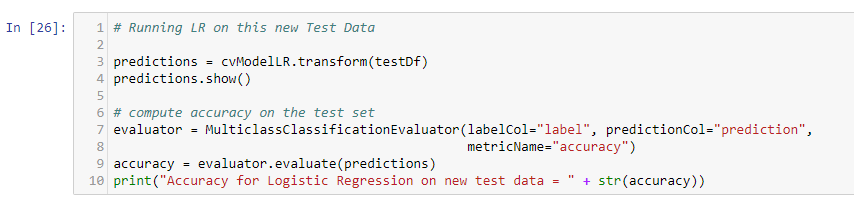
Output obtained is as follows:



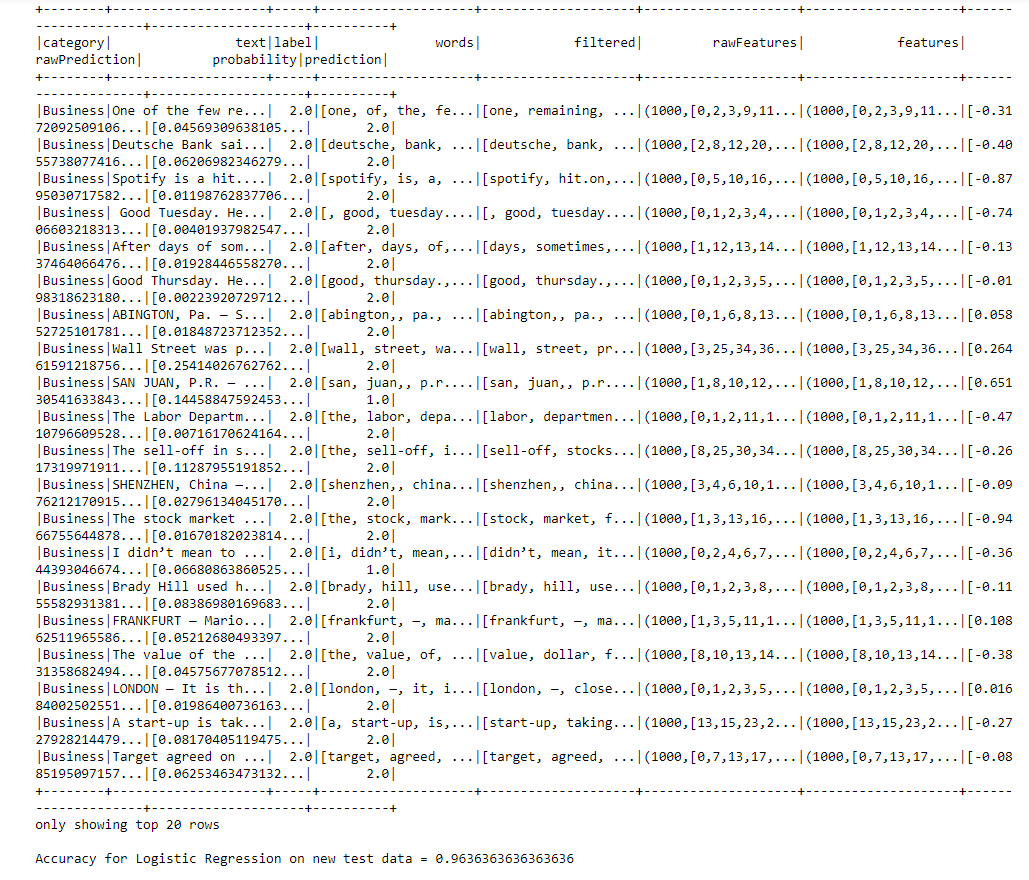
Accuracy % for this new test data = 95.4545 using Naïve Bayes

1. **Logistic Regression (Multiclass)**

Input snippet is as follows:



Output obtained is as follows:

Accuracy % for this new test data = 96.3636 using Logistic Regression