

**Covid19 Fake News Detection**

***Embedded Project***

**Submitted**

**By**

**Sharmad Rajnish Lawande – 17BCE1073**

**Submitted**

**To**

**Dr. Shridevi S**

**Assistant Professor**

**VIT School of Computer Science and Engineering**

***Web Mining CSE3024***

***Winder Semester 2020***

**Abstract**

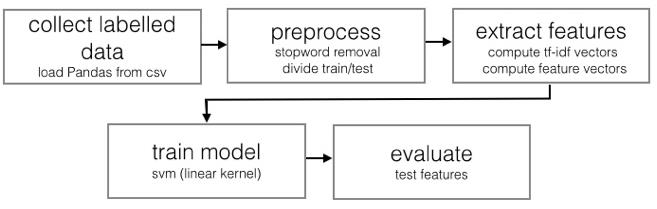
Fake news related to the coronavirus pandemic has now become a huge problem since false information can lead to worry and concerns regarding the disease. It is not possible to perfectly detect fake news unless the news has been labelled fake or real. Therefore, I have taken this issue as my problem and have developed a project that can detect fake news regarding Covid19 pandemic with the help of the dataset from Kaggle containing Covid19 public media information. I have also used different machine learning classifiers to check which classifier is best suited for the detection

***Keywords: fake, machine learning classifiers, Covid19, pandemic, detection, Kaggle***

**Objectives:**

* To detect fake news with the help of sklearn classifiers using confusion matrix and feature words from given Covid19 public media dataset
* To identify the words and token counts that lead to real or fake news
* To compare and analyze different classifiers used and to evaluate the best classifier for detecting the fake news

**Proposed Model:**

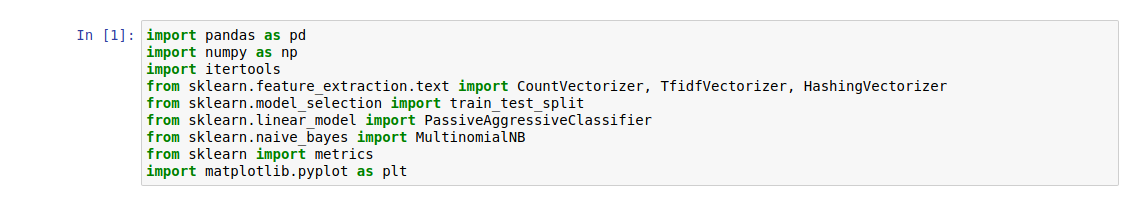


**Procedure:**

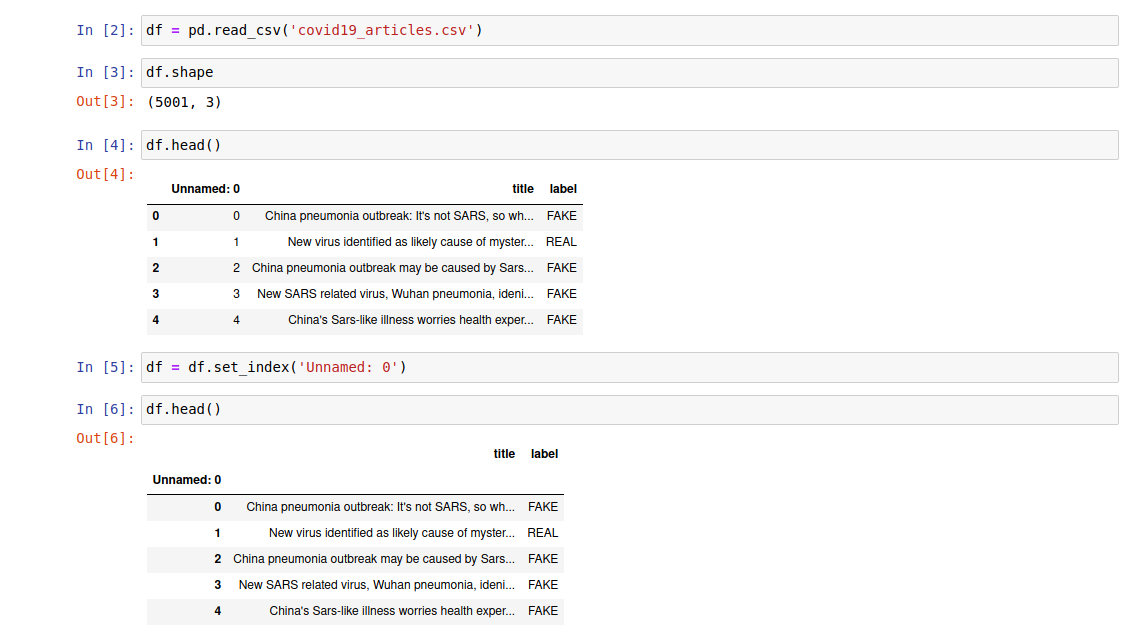
1. **Libraries used:**

* **Pandas and Numpy** - for importing and working on the dataset
* **Itertools** – functions to create iterators for efficient looping
* **Sklearn** - to make use of different machine learning classifiers
* **Matplotlib** – for plotting the graph of accuracies of classifiers and plotting the confusion matrix

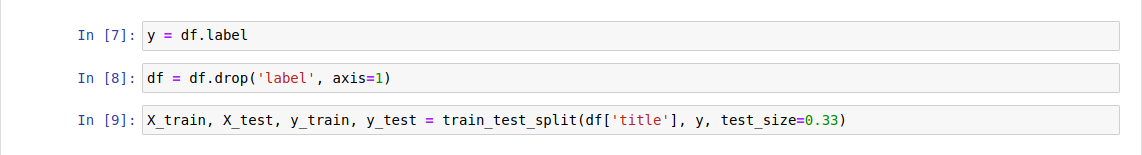
**Dataset Source:** <https://www.kaggle.com/jannalipenkova/covid19-public-media-dataset>



1. **Reading CSV Dataset:**

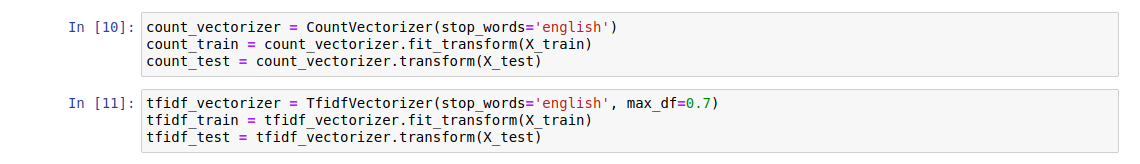
****

1. **Extracting the data for training:**

****

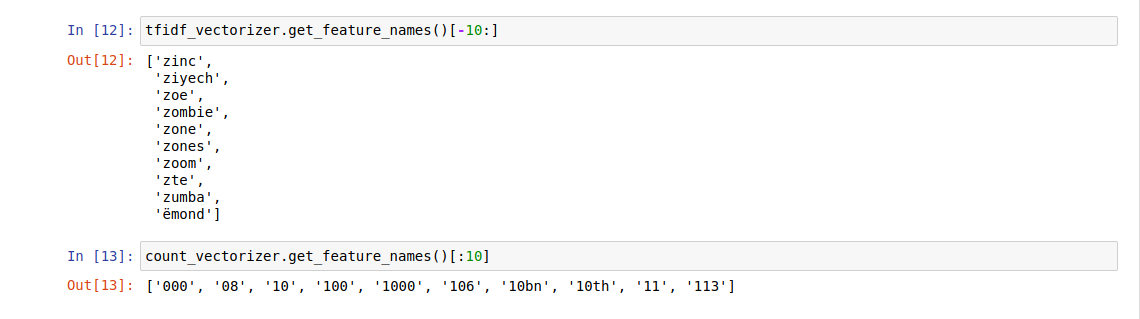
1. **Creating 2 Vectorizer Classifiers:**

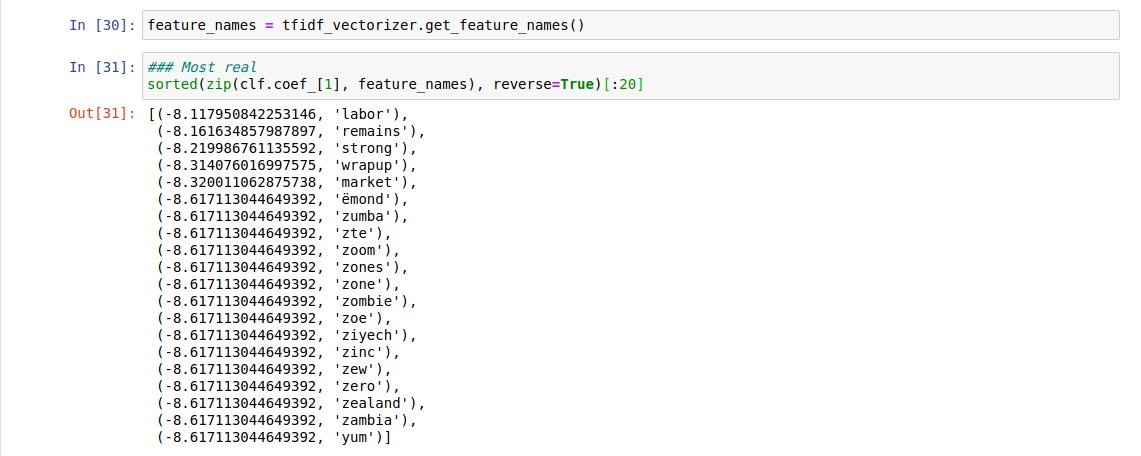
* **Count Vectorizer**
* **Tfidf Vectorizer**
* **Stop words** are just a list of words which I don’t want to use as features.
* I have set the parameter **stop\_words=’english’** to use a built-in list.
* The model’s performance is improved if I don’t take those words in my analysis

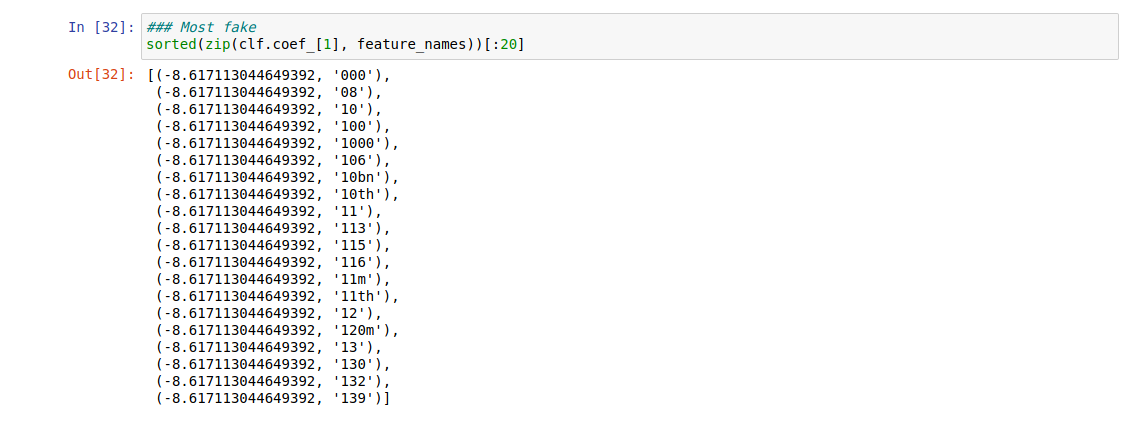
****

1. **Extracting feature words using Tfidf Vectorizer and Count Vectorizer:**

* The words identified by Tfidf Vectorizer are features that help to detect REAL news mostly
* The tokens identified by Count Vectorizer are features that help to detect FAKE news mostly

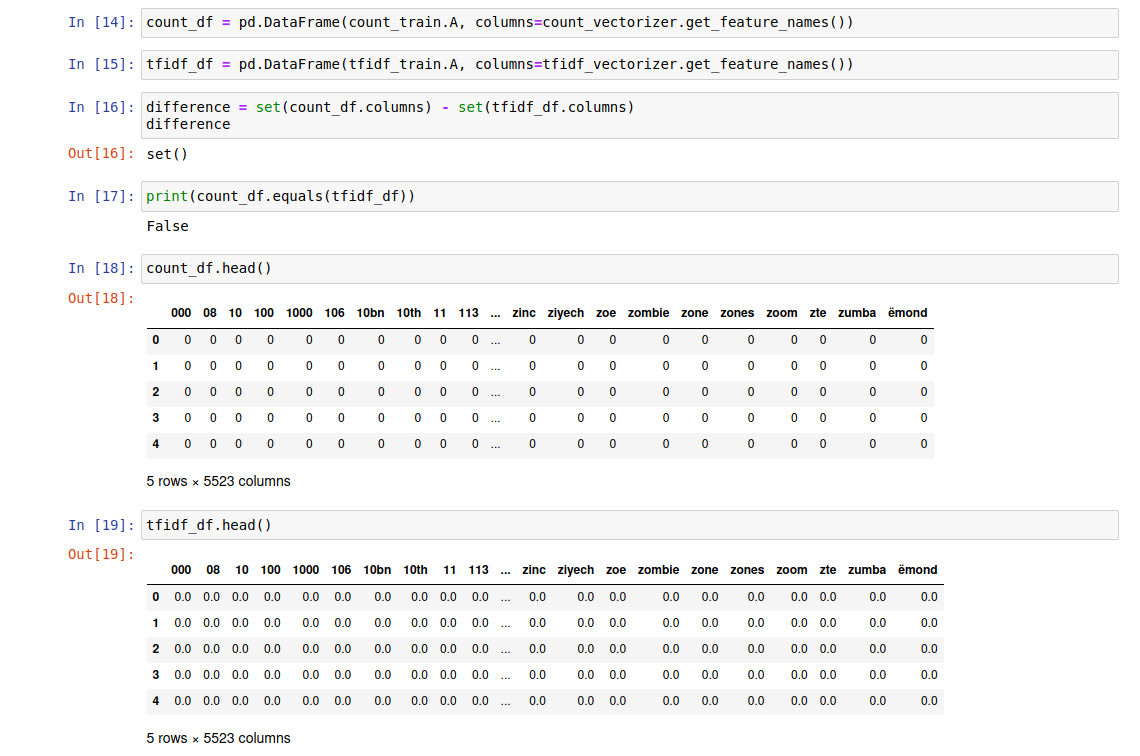
****

****

****

1. **Checking for Common Features from Both Classifiers**

* I have not found anything common using both classifiers since the set has returned NULL and it prints FALSE
* The values in the matrix are all found to be 0, so no common features

****

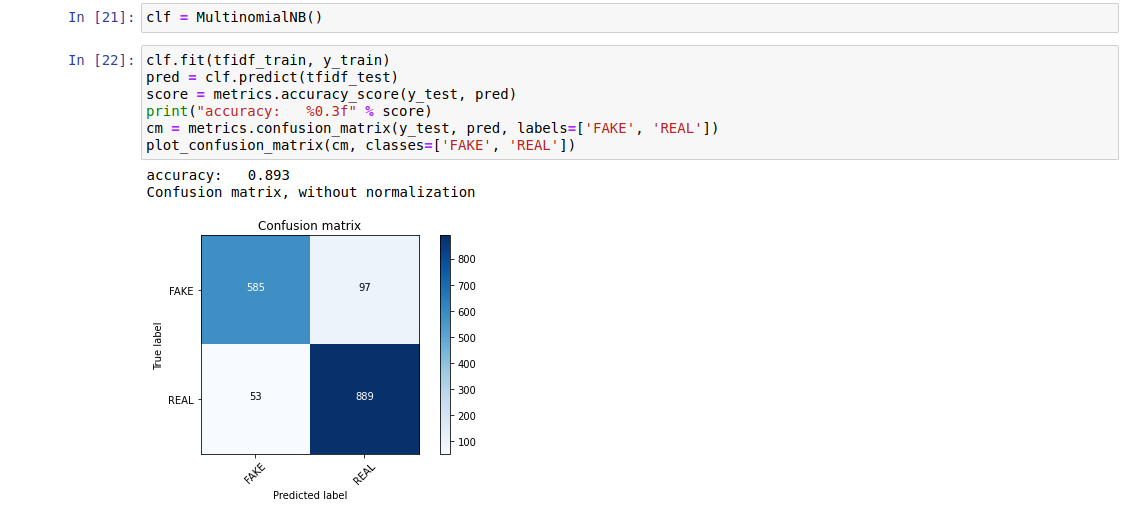
1. **Plotting the Confusion Matrix and Accuracy:**

* **Classifiers used**

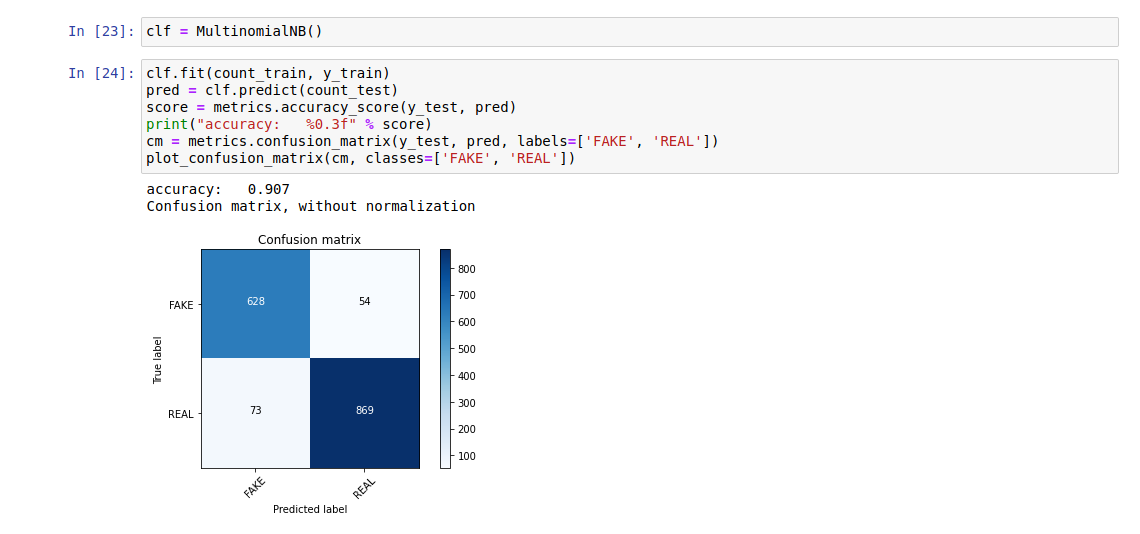
1. MultinomialNB with Tfidf Vectorizer
2. MultinomialNB with Count Vectorizer
3. Passive Aggressive Classifier with Tfidf Vectorizer
4. MultinomialNB with Hashing Vectorizer
5. Passive Aggressive Classifier with Hashing Vectorizer

****

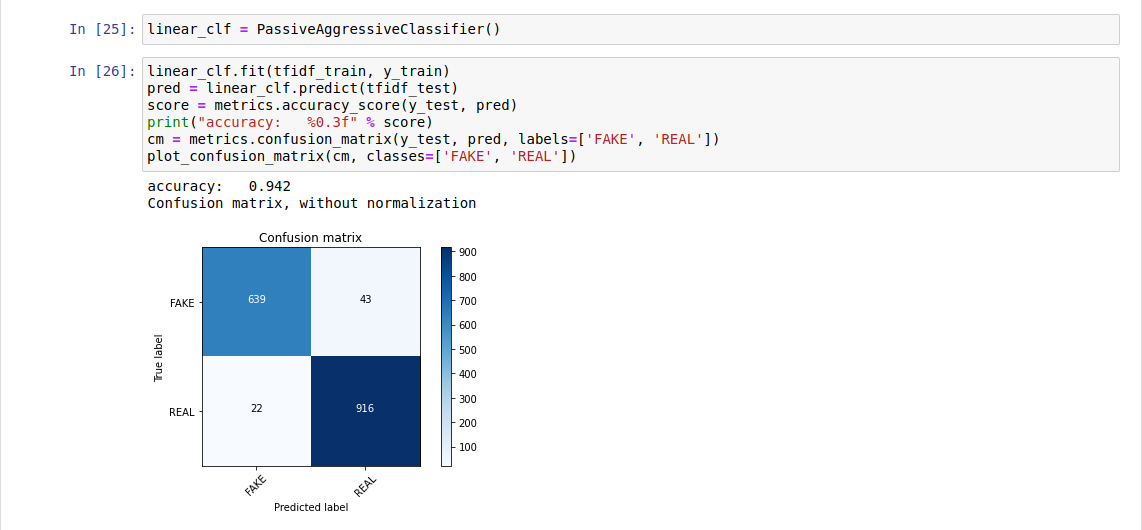
1. **MultinomialNB with Tfidf Vectorizer**

****

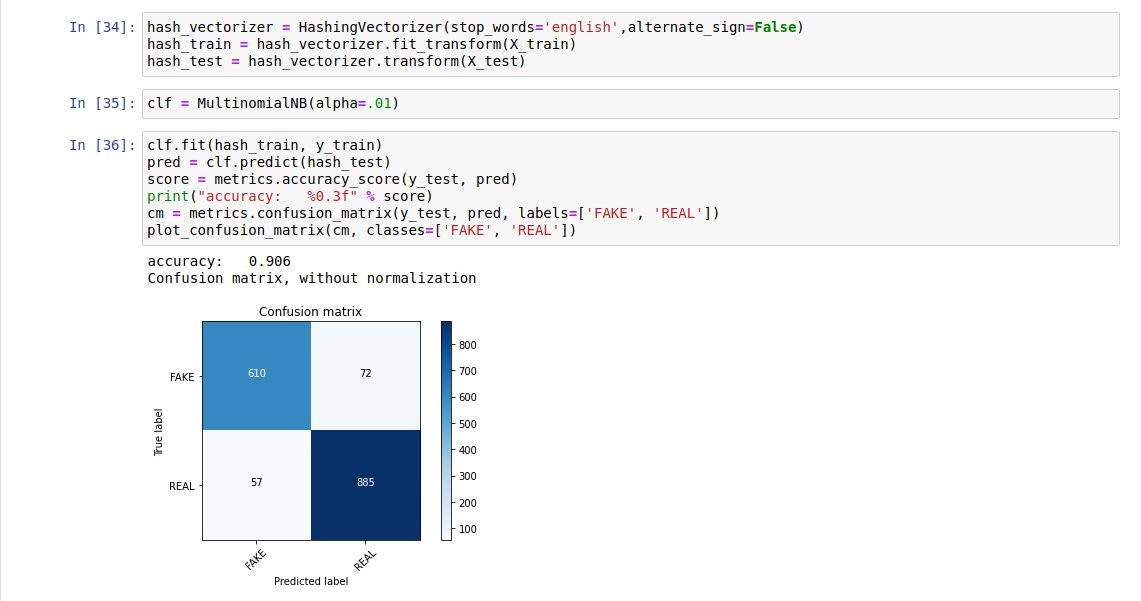
1. **MultinomialNB with Count Vectorizer**

****

1. **Passive Aggressive Classifier with Tfidf Vectorizer**

****

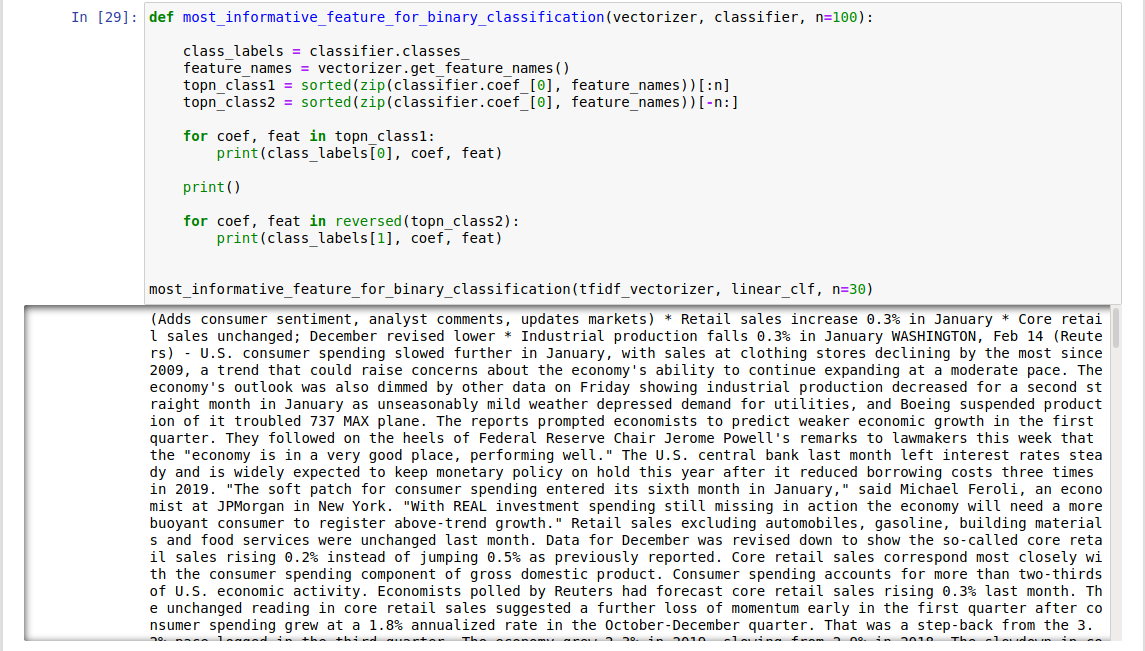
1. **MultinomialNB with Hashing Vectorizer**

****

1. **Passive Aggressive Classifier with Hashing Vectorizer**

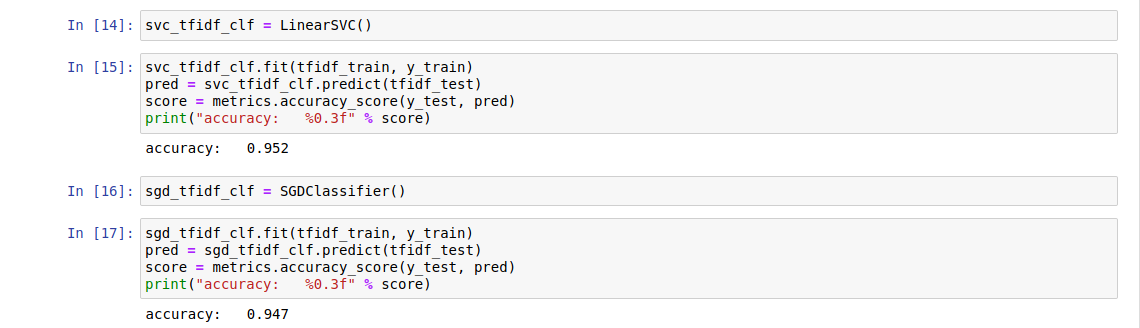
****

1. **Detecting Fake News using Passive Aggressive Classifier with Tfidf Vectorizer:**

****

1. **Comparison with Other Classifiers such as:**

* **Linear SVC (Support Vector Classification)**
* **SGD Classifier (Stochastic Gradient Descent)**
* Here I have set **alpha = 0.1** to increase performance of MultinomialNB Classifier

****

1. **Plotting Graph of Accuracies Obtained from Different Classifiers**

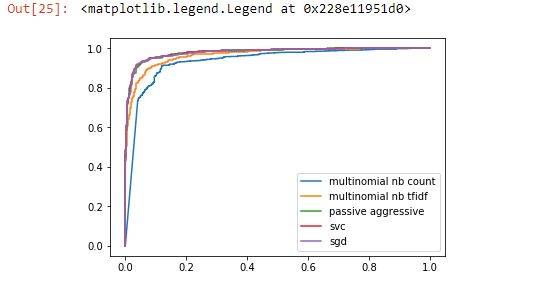
****

**Results and Analysis:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S. No** | **Classifier** | **Accuracy (%)** | **Best Classifier** |
| 1. | MultinomialNB with Tfidf Vectorizer | 89.3 | Passive Aggressive Classifier with Hashing Vectorizer |
| 2. | MultinomialNB with Tfidf Vectorizer | 90.7 |
| 3. | Passive Aggressive Classifier with Tfidf Vectorizer | 94.2 |
| 4. | MultinomialNB with Hashing Vectorizer | 90.6 |
| 5. | Passive Aggressive Classifier with Hashing Vectorizer | 94.5 |

* I have used **Passive Aggressive Classifier with Tfidf Vectorizer** for detecting the fake news

|  |  |  |  |
| --- | --- | --- | --- |
| **S. No** | **Classifier** | **Accuracy (%)** | **Best Classifier** |
| 1. | MultinomialNB (alpha = 0.1) with Tfidf Vectorizer | 89.0 | Linear SVC |
| 2. | MultinomialNB (alpha = 0.1) with Count Vectorizer | 91.3 |
| 3. | Passive Aggressive Classifier | 94.1 |
| 4. | Linear SVC | 95.2 |
| 5. | SGD Classifier | 94.7 |

****

**Conclusion:**

1. The best classifier for detection of Covid19 fake news was found to be **Passive Aggressive Classifier with Tfidf Vectorizer (accuracy = 94.2%)**
2. **Passive Aggressive Classifier with Tfidf Vectorizer** gave the value of
3. **TN = 639 and TP = 916** in the confusion matrixand could detect fake news more accurately
4. **Passive Aggressive Classifier with Hashing Vectorizer** gave highest value of **TN =** **642** **and TP = 919**
5. There was an improvement in performance (accuracy) of **MultinomialNB** after tuning the value of **alpha = 0.1**
6. After comparing with other classifiers such as Linear SVC and SGD, it was found that **Linear SVC (accuracy = 95.2%)** was the best among all of them