DEMOKRITOS

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS

Department of Digital Systems
MSC in Artificial Intelligence

Fake News Prediction

Dimitrios Delikonstantis

14/02/2021

Reason for subject selection

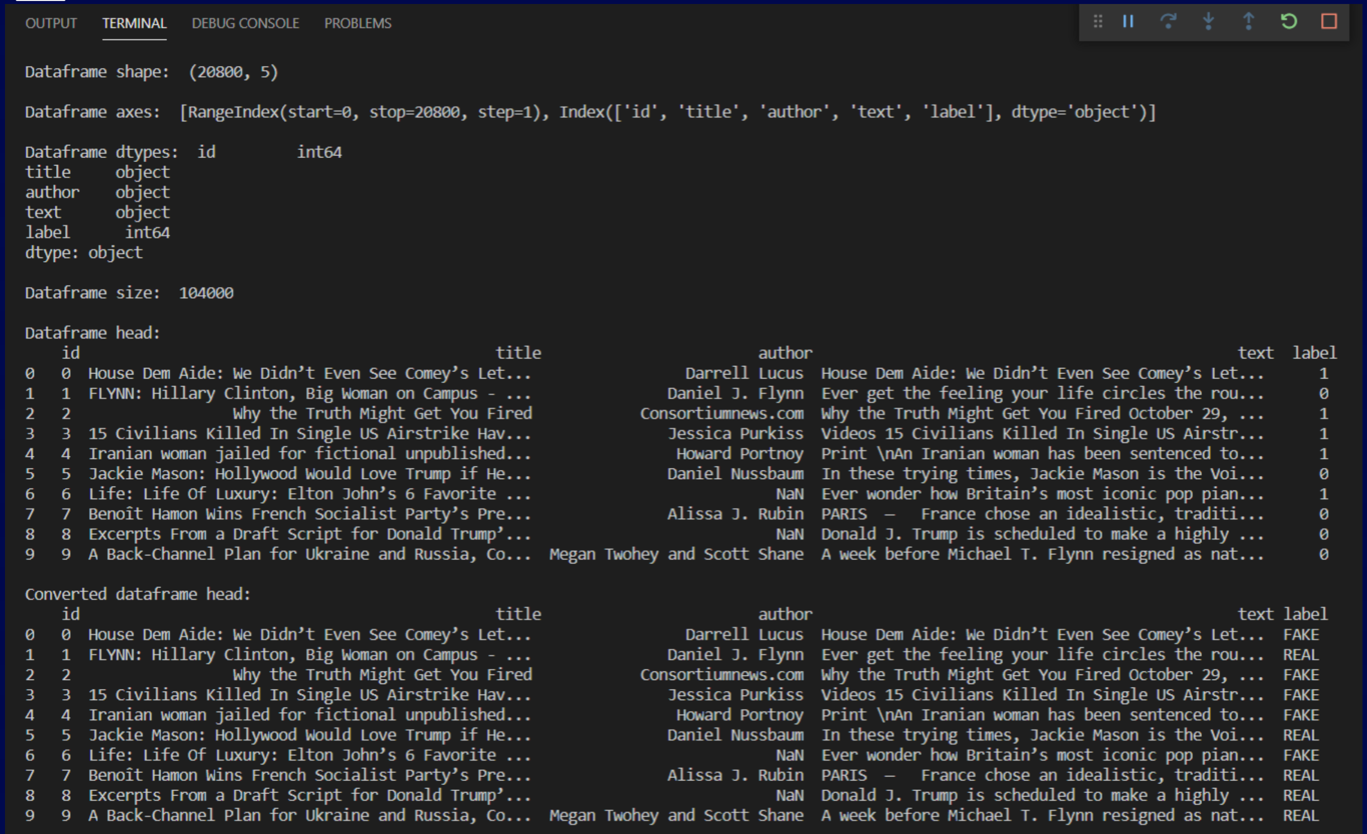
- Fake or misleading news can be dangerous
- It is also used for making money via advertising (clickbait)
- Prevalence has increased with the rise of social media
- Social media algorithms have been implicated in the spread of fake news
- Since coronavirus appeared fake news have been on the rise
- Anti-vaccination movement is growing vastly via fake news. WHO declared it one of the top ten global threats
- Therefore, vastly reducing fake news is crucial for the society’s own good

The data

- Dataset retrieved from kaggle
- Used data from 20800 article news labeled ‘Real’ or ‘Fake’

Feature extraction

- Features, samples and their properties extracted with pandas read_csv into a dataframe



```
Dataframe shape:  (20800, 5)

Dataframe axes:  [RangeIndex(start=0, stop=20800, step=1), Index(['id', 'title', 'author', 'text', 'label'], dtype='object')]

Dataframe dtypes:  id         int64
title     object
author    object
text      object
label      int64
dtype: object

Dataframe size:  104000

Dataframe head:
    id                                             title                         author                                               text  label
0   0  House Dem Aide: We Didn’t Even See Comey’s Let...                  Darrell Lucus  House Dem Aide: We Didn’t Even See Comey’s Let...      1
1   1  FLYNN: Hillary Clinton, Big Woman on Campus - ...                Daniel J. Flynn  Ever get the feeling your life circles the rou...      0
2   2                  Why the Truth Might Get You Fired             Consortiumnews.com  Why the Truth Might Get You Fired October 29, ...      1
3   3  15 Civilians Killed In Single US Airstrike Hav...                Jessica Purkiss  Videos 15 Civilians Killed In Single US Airstr...      1
4   4  Iranian woman jailed for fictional unpublished...                 Howard Portnoy  Print \nAn Iranian woman has been sentenced to...      1
5   5  Jackie Mason: Hollywood Would Love Trump if He...                Daniel Nussbaum  In these trying times, Jackie Mason is the Voi...      0
6   6  Life: Life Of Luxury: Elton John’s 6 Favorite ...                            NaN  Ever wonder how Britain’s most iconic pop pian...      1
7   7  Benoît Hamon Wins French Socialist Party’s Pre...                Alissa J. Rubin  PARIS  —  France chose an idealistic, traditi...      0
8   8  Excerpts From a Draft Script for Donald Trump’...                            NaN  Donald J. Trump is scheduled to make a highly ...      0
9   9  A Back-Channel Plan for Ukraine and Russia, Co...  Megan Twohey and Scott Shane  A week before Michael T. Flynn resigned as nat...      0

Converted dataframe head:
    id                                             title                         author                                               text label
0   0  House Dem Aide: We Didn’t Even See Comey’s Let...                  Darrell Lucus  House Dem Aide: We Didn’t Even See Comey’s Let...  FAKE
1   1  FLYNN: Hillary Clinton, Big Woman on Campus - ...                Daniel J. Flynn  Ever get the feeling your life circles the rou...  REAL
2   2                  Why the Truth Might Get You Fired             Consortiumnews.com  Why the Truth Might Get You Fired October 29, ...  FAKE
3   3  15 Civilians Killed In Single US Airstrike Hav...                Jessica Purkiss  Videos 15 Civilians Killed In Single US Airstr...  FAKE
4   4  Iranian woman jailed for fictional unpublished...                 Howard Portnoy  Print \nAn Iranian woman has been sentenced to...  FAKE
5   5  Jackie Mason: Hollywood Would Love Trump if He...                Daniel Nussbaum  In these trying times, Jackie Mason is the Voi...  REAL
6   6  Life: Life Of Luxury: Elton John’s 6 Favorite ...                            NaN  Ever wonder how Britain’s most iconic pop pian...  FAKE
7   7  Benoît Hamon Wins French Socialist Party’s Pre...                Alissa J. Rubin  PARIS  —  France chose an idealistic, traditi...  REAL
8   8  Excerpts From a Draft Script for Donald Trump’...                            NaN  Donald J. Trump is scheduled to make a highly ...  REAL
9   9  A Back-Channel Plan for Ukraine and Russia, Co...  Megan Twohey and Scott Shane  A week before Michael T. Flynn resigned as nat...  REAL
```

Cross validation

- Initially, sklearn train_test_split was used for splitting the dataset.
- Can we be sure that the splitting done by train_test_split is unbiased?
- Cross validate with sklearn RepeatedKFold where different splits each repetition are produced

Bag-of-words model

- Text is changed into vectors of numbers to be processed by machine learning algorithms (feature encoding)
- Describes the occurrence of words within a document
- Transform each document from a corpus of documents into a vector and use it as an input to a machine learning algorithm

TF-IDF

- Statistical measure that shows how important a word is in a document
- Term frequency shows the frequency of a word in a document.
- Inverse document frequency shows the frequency of a word within the corpus of documents.
- TF-IDF score = TF * IDF
- The higher the score, the more relevant that word is in that particular document
- Sklearn TfidfVectorizer used for different case scenarios.

Logistic regression metrics

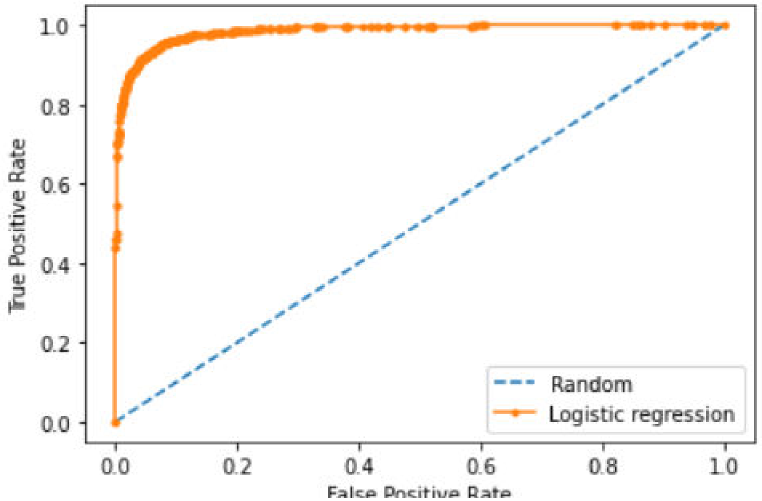
- Accuracy: 93.79%
- Confusion matrix:

```
                         Actual
          Predicted      [[1965  111]
                          [ 147 1937]]
```

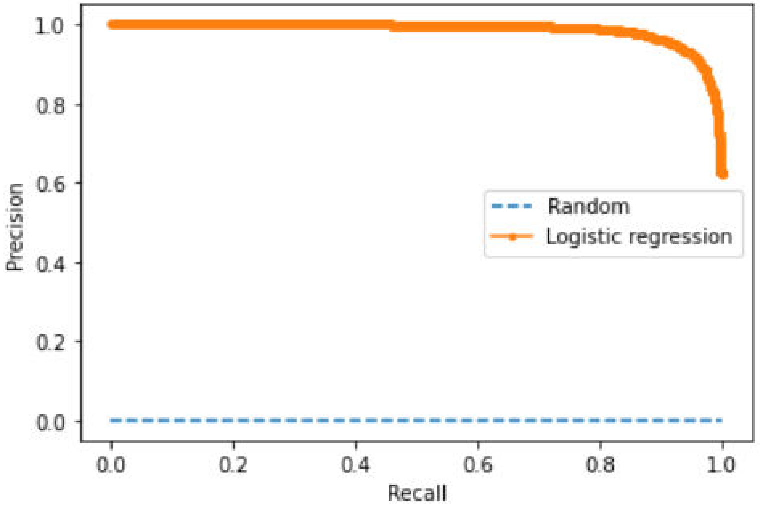
- Precision: 0.9458
- Recall: 0.9294
- F1-score: 0.9375

Logistic regression ROC/Precision recall curves

- ROC AUC: 0.9851



- ROC AUC: 0.9859



Naïve Bayes metrics

- Accuracy: 88.46%
- Confusion matrix:

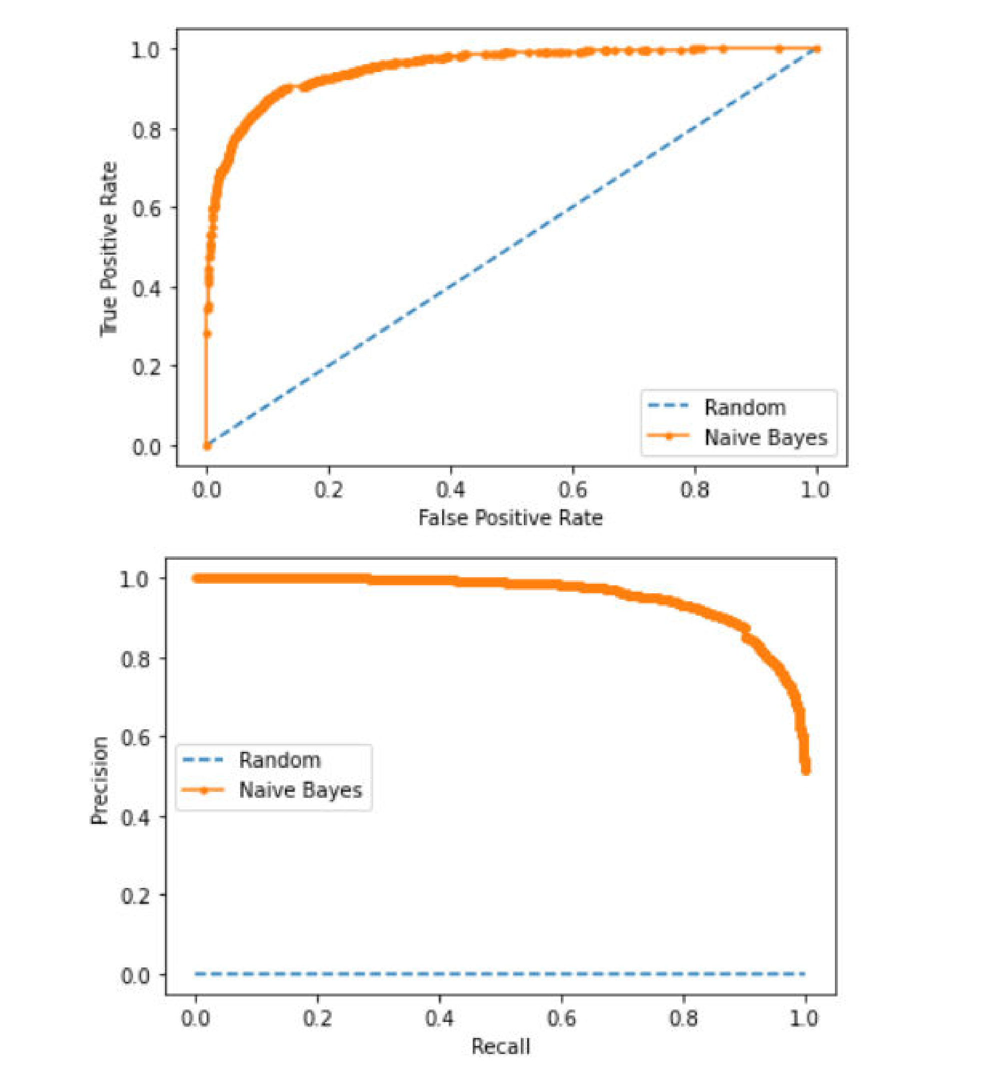
```
                    Actual
          Predicted [[1802  274]
                    [ 206 1878]]
```

- Precision: 0.8726
- Recall: 0.9011
- F1-score: 0.8866

Naïve Bayes ROC/Precision recall curves

- ROC AUC: 0.9535

- ROC AUC: 0.9574


True Positive Rate
False Positive Rate
Random
Naive Bayes
Precision
Recall
Random
Naive Bayes