DATA ANALYSIS

DS3114

PROJECT TITLE:

# FakeNews sentiment Analysis

COURSE REPRESENTS :
(DR.Omaima A. Fallatah)

SUBMITTED BY:


| Name | ID |
|---|---|
| Lamar Waleed Fattah | 444006719 |
| Rakha Matuq Nooh | 444001287 |

COLLEGE OF COMPUTING UMM AL-QURA UNIVERSITY

# Table of content

# 1. Introduction:

In this project, our goal is to build a model capable of predicting whether a news article is fake or real using a dataset from the Fake News competition. The dataset includes various features, such as the author, title, text, unique identifier (ID), and label that indicates the authenticity of the news articles. By analysing this dataset, we aim to explore techniques that can effectively distinguish between fake and real news content.

## 1.2 Data Description

- id: Unique identifier for the news
- title: Title of the news article
- author: Author of the news article
- text: The text of the article
- label: Target label where 0 indicates "Real" and 1 indicates "Fake"

## 1.3 Objectives:

1. Implement text preprocessing techniques to clean and prepare the data for analysis.
2. Apply **TF-IDF (Term Frequency-Inverse Document Frequency)** and **Bag of Words** vectorization methods to convert textual data into numerical features and compare.
3. Build and train a **Naive Bayes Multinomial model** to classify the news articles as either fake or real.
4. Evaluate the model's performance and fine-tune it to improve classification accuracy.
5. Analyse and interpret the model's ability to detect patterns indicative of fake news within the dataset.

## 2. Data Exploration

```
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px


from wordcloud import WordCloud
import nltk
import re
import string
from nltk.corpus import stopwords
nltk.download('punkt')
nltk.download('stopwords')
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer


stop_words = stopwords.words()
```

Installing the important data

We concatenated the three files (train, test, and submit) into one file named **fakenews**. After merging them, we removed the unnecessary columns, leaving us with only the following fields: **id**, **title**, **author**, **text**, and **label**.

This process helped streamline the data by focusing on the most relevant information, making it easier for further analysis or model training.

```
fakenews.head()
```

| | label | text | author | title | id |
|---|---|---|---|---|---|
| 0 | 1.0 | House Dem Aide: We Didn't Even See Comey's Let... | Darrell Lucus | House Dem Aide: We Didn't Even See Comey's Let... | 0 |
| 1 | 0.0 | Ever get the feeling your life circles the rou... | Daniel J. Flynn | FLYNN: Hillary Clinton, Big Woman on Campus - ... | 1 |
| 2 | 1.0 | Why the Truth Might Get You Fired October 29, ... | Consortiumnews.com | Why the Truth Might Get You Fired | 2 |
| 3 | 1.0 | Videos 15 Civilians Killed In Single US Airstr... | Jessica Purkiss | 15 Civilians Killed In Single US Airstrike Hav... | 3 |
| 4 | 1.0 | Print \nAn Iranian woman has been sentenced to... | Howard Portnoy | Iranian woman jailed for fictional unpublished... | 4 |

***The final dataset***

Number of columns (18285) number of rows (5), The dataset experiences issues with missing values (nulls), which can affect the quality of the analysis or model performance. However, on the bright side, it is free from duplicate entries.

```
[ ]: fakenews.isnull().sum()
```

[12]:

| | 0 |
|---|---|
| label | 5880 |
| text | 5668 |
| author | 7986 |
| title | 6091 |
| id | 1 |

**dtype:** int64

***To This***

```
fakenews.dropna(inplace=True)
fakenews.isnull().sum()
```

| | 0 |
|---|---|
| label | 0 |
| text | 0 |
| author | 0 |
| title | 0 |
| id | 0 |

**dtype:** int64

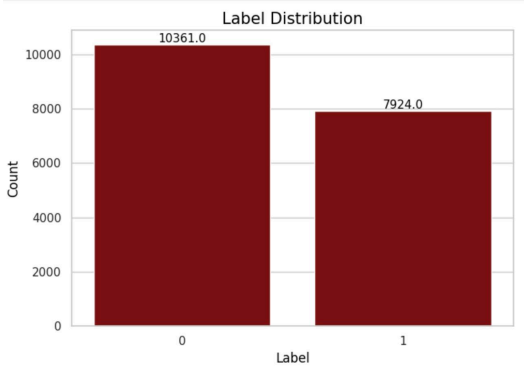
```
fakenews.duplicated().sum()
```

```
0
```

```
fakenews.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 18285 entries, 0 to 20799
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   label   18285 non-null  int64
 1   text    18285 non-null  object
 2   author  18285 non-null  object
 3   title   18285 non-null  object
 4   id      18285 non-null  object
dtypes: int64(1), object(4)
memory usage: 857.1+ KB
```

***Full info about the the dataset***




***Label Distribution / count***

```
label value count:
 label
0    10361
1     7924
Name: count, dtype: int64
```

# 3. Data Preprocessing

To improve the model's performance, we create a new feature **content** by combining the **author** and **title** columns.

```
0        Darrell Lucus House Dem Aide: We Didn’t Even S...
1        Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
2        Consortiumnews.com Why the Truth Might Get You...
3        Jessica Purkiss 15 Civilians Killed In Single ...
4        Howard Portnoy Iranian woman jailed for fictio...
                                ...
20795    Jerome Hudson Rapper T.I.: Trump a ’Poster Chi...
20796    Benjamin Hoffman N.F.L. Playoffs: Schedule, Ma...
20797    Michael J. de la Merced and Rachel Abrams Macy...
20798    Alex Ansary NATO, Russia To Hold Parallel Exer...
20799              David Swanson What Keeps the F-35 Alive
Name: content, Length: 18285, dtype: object
```

## 3.1 Data Cleaning

- **Lowercasing:** Converts text to lowercase for consistency.
- Link Removal: Deletes any URLs present in the text.
- **Number Removal:** Eliminates all numeric characters.
- **HTML Tag Removal:** Strips out HTML tags.
- **Punctuation Removal:** Removes all punctuation marks.
- **Whitespace Handling:** Condenses multiple spaces and removes newlines.
- **Emoji Removal:** Eliminates emojis using a predefined pattern.
- **Mentions and Hashtags:** Removes user mentions and hashtags.
- **Contraction Expansion:** Expands contractions to their full forms.
- **Tokenization and Lemmatization:** Breaks text into words and reduces words to their base forms, removing common stopwords.

Overall, the function effectively cleans and standardises text, making it suitable for further analysis.

We stored the cleaned text in a variable called fakenews_clean, and then we proceeded to remove stopwords from the cleaned_text.

| | cleaned_text | no_sw |
|---|---|---|
| 0 | house dem aide we didnt even see comeys letter... | house aide comeys letter jason chaffetz tweete... |
| 1 | ever get the feeling your life circles the rou... | feeling life circles roundabout heads straight... |
| 2 | why the truth might get you fired october th... | truth fired october tension intelligence analy... |
| 3 | videos civilians killed in single us airstrik... | videos civilians killed single airstrike ident... |
| 4 | print an iranian woman has been sentenced to s... | print iranian woman sentenced years prison ira... |
| ... | ... | ... |
| 20795 | rapper t i unloaded on black celebrities who m... | rapper unloaded black celebrities donald trump... |
| 20796 | when the green bay packers lost to the washing... | green bay packers lost washington redskins wee... |
| 20797 | the macys of today grew from the union of seve... | macys today grew union great names american re... |
| 20798 | nato russia to hold parallel exercises in balk... | nato russia hold parallel exercises balkans pr... |
| 20799 | david swanson is an author activist journali... | david swanson author activist journalist radio... |

18285 rows × 2 columns

To work with the most frequent words, we first identify and list the words that occur most often in the cleaned text. Once we have this list, we proceed to remove these frequently occurring words from our dataset, as they may not contribute significant value to our analysis. By eliminating these common words, we can focus on less frequent, more meaningful terms, which can enhance the overall quality of our text analysis.

***Most frequent***

| | word | count |
|---|---|---|
| 0 | mr | 65972 |
| 1 | — | 45456 |
| 2 | trump | 41770 |
| 3 | president | 21881 |
| 4 | clinton | 19930 |
| 5 | time | 19105 |
| 6 | years | 17656 |
| 7 | state | 17452 |
| 8 | states | 17394 |
| 9 | american | 14579 |

***After removing most frequent***

| | label | text | author | title | id | content | wo_stopfreq |
|---|---|---|---|---|---|---|---|
| 0 | 1 | House Dem Aide: We Didn't Even See Comey's Let... | Darrell Lucus | House Dem Aide: We Didn't Even See Comey's Let... | 0 | Darrell Lucus House Dem Aide: We Didn't Even S... | house aide comeys letter jason chaffetz tweete... |
| 1 | 0 | Ever get the feeling your life circles the rou... | Daniel J. Flynn | FLYNN: Hillary Clinton, Big Woman on Campus - ... | 1 | Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo... | feeling life circles roundabout heads straight... |
| 2 | 1 | Why the Truth Might Get You Fired October 29, ... | Consortiumnews.com | Why the Truth Might Get You Fired | 2 | Consortiumnews.com Why the Truth Might Get You... | truth fired october tension intelligence analy... |
| 3 | 1 | Videos 15 Civilians Killed In Single US Airstr... | Jessica Purkiss | 15 Civilians Killed In Single US Airstrike Hav... | 3 | Jessica Purkiss 15 Civilians Killed In Single ... | videos civilians killed single airstrike ident... |
| 4 | 1 | Print \nAn Iranian woman has been sentenced to... | Howard Portnoy | Iranian woman jailed for fictional unpublished... | 4 | Howard Portnoy Iranian woman jailed for fictio... | print iranian woman sentenced prison irans rev... |

Lemmatization converts words to their base forms by removing affixes from inflected variations. This process is vital for creating better features.

| | cleaned_text | no_sw | wo_stopfreq | wo_stopfreq_lem |
|---|---|---|---|---|
| 0 | house dem aide we didnt even see comeys letter... | house aide comeys letter jason chaffetz tweete... | house aide comeys letter jason chaffetz tweete... | house aide comeys letter jason chaffetz tweete... |
| 1 | ever get the feeling your life circles the rou... | feeling life circles roundabout heads straight... | feeling life circles roundabout heads straight... | feeling life circles roundabout heads straight... |
| 2 | why the truth might get you fired october th... | truth fired october tension intelligence analy... | truth fired october tension intelligence analy... | truth fired october tension intelligence analy... |
| 3 | videos civilians killed in single us airstrik... | videos civilians killed single airstrike ident... | videos civilians killed single airstrike ident... | videos civilians killed single airstrike ident... |
| 4 | print an iranian woman has been sentenced to s... | print iranian woman sentenced years prison ira... | print iranian woman sentenced prison irans rev... | print iranian woman sentenced prison irans rev... |
| ... | ... | ... | ... | ... |
| 20795 | rapper t i unloaded on black celebrities who m... | rapper unloaded black celebrities donald trump... | rapper unloaded black celebrities donald elect... | rapper unloaded black celebrities donald elect... |
| 20796 | when the green bay packers lost to the washing... | green bay packers lost washington redskins wee... | green bay packers lost washington redskins wee... | green bay packers lost washington redskins wee... |
| 20797 | the macys of today grew from the union of seve... | macys today grew union great names american re... | macys today grew union great names retailing i... | macys today grew union great names retailing i... |
| 20798 | nato russia to hold parallel exercises in balk... | nato russia hold parallel exercises balkans pr... | nato russia hold parallel exercises balkans pr... | nato russia hold parallel exercises balkans pr... |
| 20799 | david swanson is an author activist journali... | david swanson author activist journalist radio... | david swanson author activist journalist radio... | david swanson author activist journalist radio... |

18285 rows × 4 columns

Now, the data is clean! we can move to the next step: Naive Bayes Modelling

```
                                                text  \
0  House Dem Aide: We Didn’t Even See Comey’s Let...
1  Ever get the feeling your life circles the rou...
2  Why the Truth Might Get You Fired October 29, ...
3  Videos 15 Civilians Killed In Single US Airstr...
4  Print \nAn Iranian woman has been sentenced to...

                                              tokens
0  [House, Dem, Aide, :, We, Didn, ’, t, Even, Se...
1  [Ever, get, the, feeling, your, life, circles,...
2  [Why, the, Truth, Might, Get, You, Fired, Octo...
3  [Videos, 15, Civilians, Killed, In, Single, US...
4  [Print, An, Iranian, woman, has, been, sentenc...
```

***Tokenized text***

We uesed **port_stem=PorterStemmer()** , Stemming reduces words to their root forms, helping to simplify text data for analysis and model training. Then we applied it to the **content** column.

```
0        darrel lucu hous dem aid even see comey letter...
1        daniel j flynn flynn hillari clinton big woman...
2                   consortiumnew com truth might get fire
3        jessica purkiss civilian kill singl us airstri...
4        howard portnoy iranian woman jail fiction unpu...
                                  ...
20795    jerom hudson rapper trump poster child white s...
20796    benjamin hoffman n f l playoff schedul matchup...
20797    michael j de la merc rachel abram maci said re...
20798    alex ansari nato russia hold parallel exercis ...
20799                            david swanson keep f aliv
Name: content, Length: 18285, dtype: object
```

Now, we can move to the next step: **Implement The Model**

# 4. Implement Model:

## 4.1 Bag Of Words

The Bag of Words (BoW) model is a simple and widely used technique in natural language processing

After installing the important libraries we impliment **Bag Of Words**

```
# Bag of Words

vectorizer = CountVectorizer()
X = vectorizer.fit_transform(fakenews['text'])
y = fakenews['label']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Next, we applied the multinomial Naive Bayes model and subsequently evaluated its performance.

```
Accuracy: 90.57%
              precision    recall  f1-score   support

           0       0.89      0.95      0.92      2082
           1       0.93      0.85      0.89      1575

    accuracy                           0.91      3657
   macro avg       0.91      0.90      0.90      3657
weighted avg       0.91      0.91      0.90      3657
```

***It performed very well!***

## 4.2 TF-IDF

TF-IDF is widely used in text classification, clustering, and information retrieval tasks, as it provides a more nuanced understanding of text data compared to simpler models like Bag of Words.

```
# TF-IDF
vectorizer = TfidfVectorizer()
X_tfidf = vectorizer.fit_transform(fakenews['text'])
y = fakenews['label']

X_train, X_test, y_train, y_test = train_test_split(X_tfidf, y, test_size=0.2, random_state=42)
```

As the same as the previous step, we applied the multinomial Naive Bayes model and evaluated its performance.

```
Accuracy: 71.48%
              precision    recall  f1-score   support

           0       0.67      1.00      0.80      2082
           1       1.00      0.34      0.51      1575

    accuracy                           0.71      3657
   macro avg       0.83      0.67      0.65      3657
weighted avg       0.81      0.71      0.67      3657
```

***It performed very well , but Bag Of Words suited the dataset more!***

## 5. Conclusion

In this project, we successfully implemented various text preprocessing techniques and classification models to analyse a fake news dataset. We applied methods like Bag of Words (BoW) and TF-IDF for feature extraction and used the Naive Bayes Multinomial model to classify news articles as either real or fake. Our results indicated that the Bag of Words method was more suitable for this dataset compared to TF-IDF. Despite facing challenges related to the data, such as missing labels in the test dataset, Overall, the project provided valuable insights into detecting fake news and demonstrated the effectiveness of machine learning models in text classification tasks.

## 6. Challenges

We encountered challenges related to the train, submit and test files, as the core problem stemmed from the test file lacking the **label** indicating 1 for "reliable" and 0 for "unreliable." This absence led to errors in our analysis results and complicated our data analysis efforts. Despite attempting various solutions to address the issue, we found that the data was inconsistent. Ultimately, we resolved the problem by consolidating all the files using **concatenation** and creating a single file named "fake news," which effectively solved our data issues.