# TEXT ANALYSIS REPORT

By students :

| | |
|---|---|
| Hala faiz alharbi | 444001733 |
| Rawyah Alasmari | 443005517 |

## introduction

In this report, we will analyze text data for developing a machine learning program to determine when an article may be fake news. Our goal is to predict whether the news is reliable or not.

## Objective

Our goal is to predict and identify whether the news in question is reliable or not, based on the data provided to the model and the training it receives on that data.

## Data Overview

We used data taken from a website called Kaggle, which contains 5 columns and 20800 rows. The columns we have are:

- id: a unique identifier for a news article
- title: the title of the news article
- author: the author of the news article
- text: the text of the article; it may be incomplete
- label: a label indicating whether the article is unreliable

The results are as follows:

- 1: unreliable
- 0: reliable

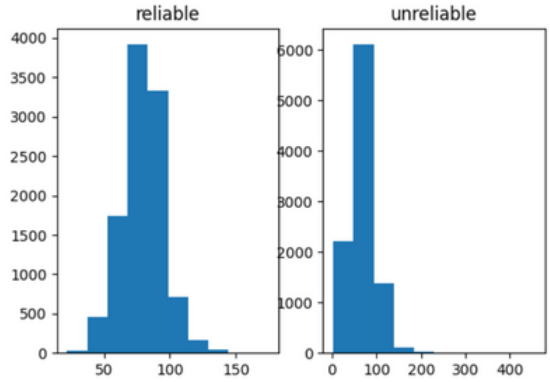
## Data Preprocessing

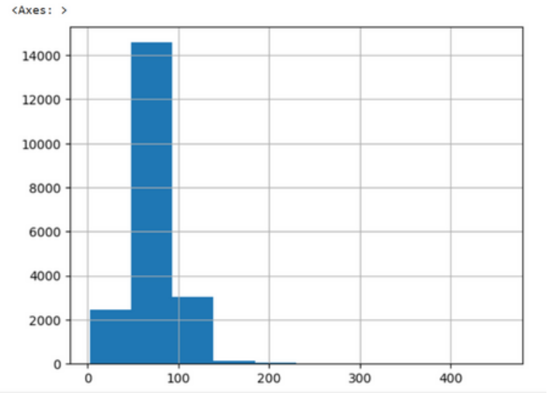
In this project, we imported important libraries such as pandas , matplotlib , and nltk for data analysis. We loaded the necessary resources from NLTK, including text segmentation tools and stop words. Then, we imported the data from a CSV file into a DataFrame using pandas , and we displayed the first 5 rows to check the content and structure of the data.

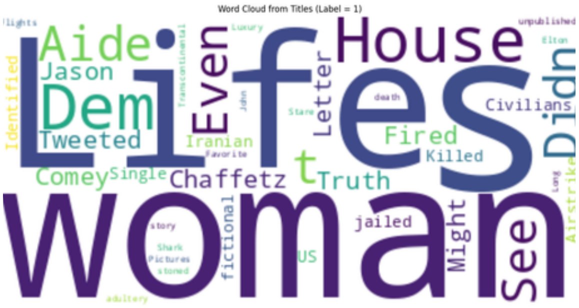
## Data Exploratory

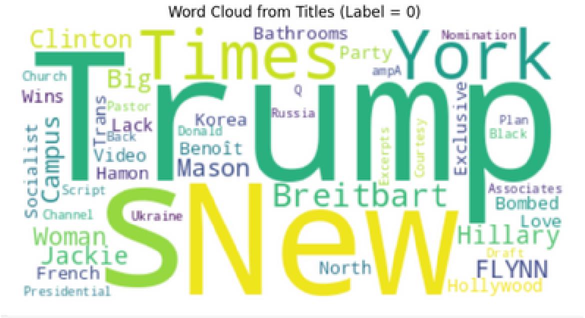
We identified all unique values in the label column of the DataFrame to understand the available classifications and the number of different categories. We verified the validity of the values in the column, then displayed some titles for both classifications 0 and 1 to analyze the content.

We counted the occurrences of each classification to understand sentiment distribution and plotted the distribution of title lengths for each classification to compare their characteristics. We also created word clouds for each classification to represent the most common words in the titles, which helped us analyze the main topics and understand general trends in the data









## Data preparation

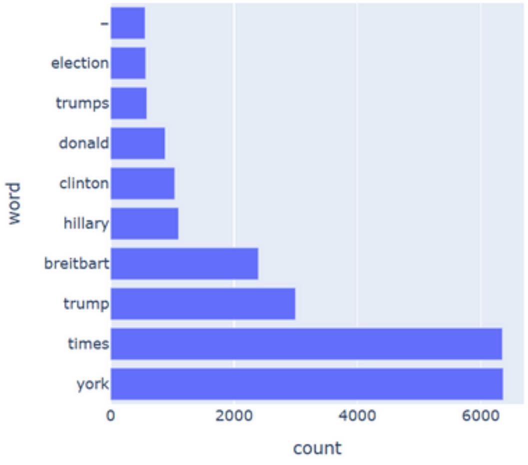
We counted the number of empty values in the title column and removed the rows with empty titles to ensure data cleanliness. We saved the cleaned dataset in a file called train_clean.csv and reloaded it to verify the operations, displaying the first 5 rows to confirm the desired format.

We cleaned the texts in the title column by removing links, numbers, emojis, and punctuation, and converting abbreviations to their full forms. We stored the cleaned titles along with their labels in a new DataFrame.

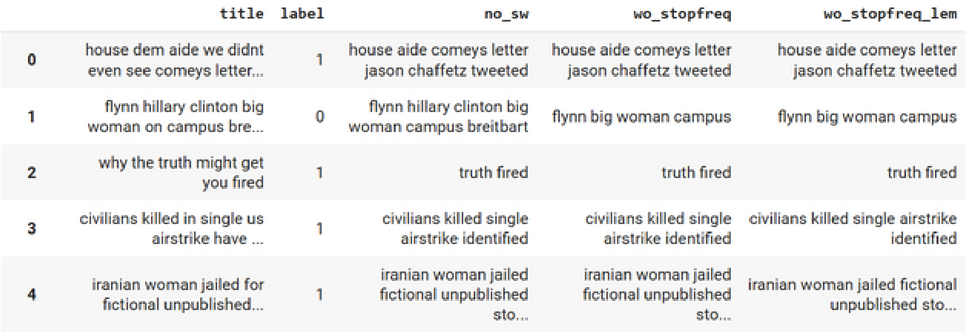
Next, we removed stop words and analyzed the most frequent words, then presented the results in a chart to better understand the content. We also applied lemmatization to improve data quality and provide better features for the models.

We conducted data preparation for a machine learning model by splitting it into two groups: training and testing. We started by dropping unnecessary columns, such as "no_sw" and "wo_stopfreq," while retaining the "label" and "title" columns. We converted numerical labels to categorical ones, replacing `0` with "reliable" and `1` with "unreliable." Next, we tokenized the text in the "title" column into words and used "CountVectorizer" to transform the text into a numerical matrix, ignoring common words. After that, we divided the data into 80% for training and 20% for testing, and reviewed the distribution of labels in the training set to ensure balance. In this way, we effectively prepared the data for model training and performance evaluation.

common words in text

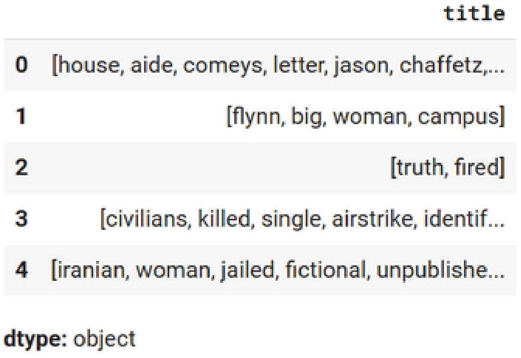


Remove the most frequent words



| | title | label | no_sw | wo_stopfreq | wo_stopfreq_lem |
|---|---|---|---|---|---|
| 0 | house dem aide we didnt even see comeys letter... | 1 | house aide comeys letter jason chaffetz tweeted | house aide comeys letter jason chaffetz tweeted | house aide comeys letter jason chaffetz tweeted |
| 1 | flynn hillary clinton big woman on campus bre... | 0 | flynn hillary clinton big woman campus breitbart | flynn big woman campus | flynn big woman campus |
| 2 | why the truth might get you fired | 1 | truth fired | truth fired | truth fired |
| 3 | civilians killed in single us airstrike have ... | 1 | civilians killed single airstrike identified | civilians killed single airstrike identified | civilians killed single airstrike identified |
| 4 | iranian woman jailed for fictional unpublished... | 1 | iranian woman jailed fictional unpublished sto... | iranian woman jailed fictional unpublished sto... | iranian woman jailed fictional unpublished sto... |

Tokenization



| | title |
|---|---|
| 0 | [house, aide, comeys, letter, jason, chaffetz,... |
| 1 | [flynn, big, woman, campus] |
| 2 | [truth, fired] |
| 3 | [civilians, killed, single, airstrike, identif... |
| 4 | [iranian, woman, jailed, fictional, unpublishe... |

**dtype:** object

## Navies Bayes Modelling

We compared the performance of three different Naive Bayes models (ComplementNB, BernoulliNB, MultinomialNB) on the same dataset. We evaluated each model by calculating its accuracy, displaying confusion matrices, and providing a classification report

Results:

- ComplementNB:
    - Model accuracy: 74.90%.
    - The confusion matrix:
    - True Positive: The actual value was 1640 unreliable news and the model predicted 1640 unreliable news
    - True Negative: The actual value was 1476 reliable news and the model predict 1476 unreliable news
    - False Positive: The actual values was 434 reliable news and the model predicted 434 unreliable news
    - False Negative: The actual values was 610 unreliable news and the model preditced 610 reliable news

- BernoulliNB:
    - Model accuracy: 78.22%.
    - The confusion matrix:
    - True Positive: The actual value was 1594 unreliable news and the model predicted 1594 unreliable news
    - True Negative: The actual value was 1660 reliable news and the model predict 1660 unreliable news
    - False Positive: The actual values was 480 reliable news and the model predicted 480 unreliable news
    - False Negative: The actual values was 426 unreliable news and the model preditced 426 reliable news
    - 
- MultinomialNB:
    - Model accuracy: 77.91%.
    - The confusion matrix:
    - True Positive: The actual value was 1639 unreliable news and the model predicted 1639 unreliable news
    - True Negative: The actual value was 1602 reliable news and the model predict 1602 unreliable news
    - False Positive: The actual values was 435 reliable news and the model predicted 435 unreliable news
    - False Negative: The actual values was 484 unreliable news and the model preditced 484 reliable news

## TF-IDF

We compared the performance of three different Naive Bayes models (ComplementNB, BernoulliNB, MultinomialNB) on the same dataset. We evaluated each model by calculating its accuracy, displaying confusion matrices, and providing a classification report.

Results:

- ComplementNB:
  - Model accuracy: 75.70%.

- BernoulliNB:
  - Model accuracy: 78.17%.

- MultinomialNB:
  - Model accuracy: 78.17%.

Conclusion:

From the results, we observe that the BernoulliNB model performed the best, achieving the highest accuracy at 78.22%, making it the most suitable choice for use.