**Introduction:**

The proliferation of fake news has been a major issue in recent years. The spread of misinformation can have severe consequences, such as influencing public opinion, causing panic, and even inciting violence. Therefore, it is essential to develop methods to detect fake news accurately. In this project, we compare the performance of two popular classification algorithms, Naive Bayes and Random Forest, in detecting fake news.

**About Data Collection:**

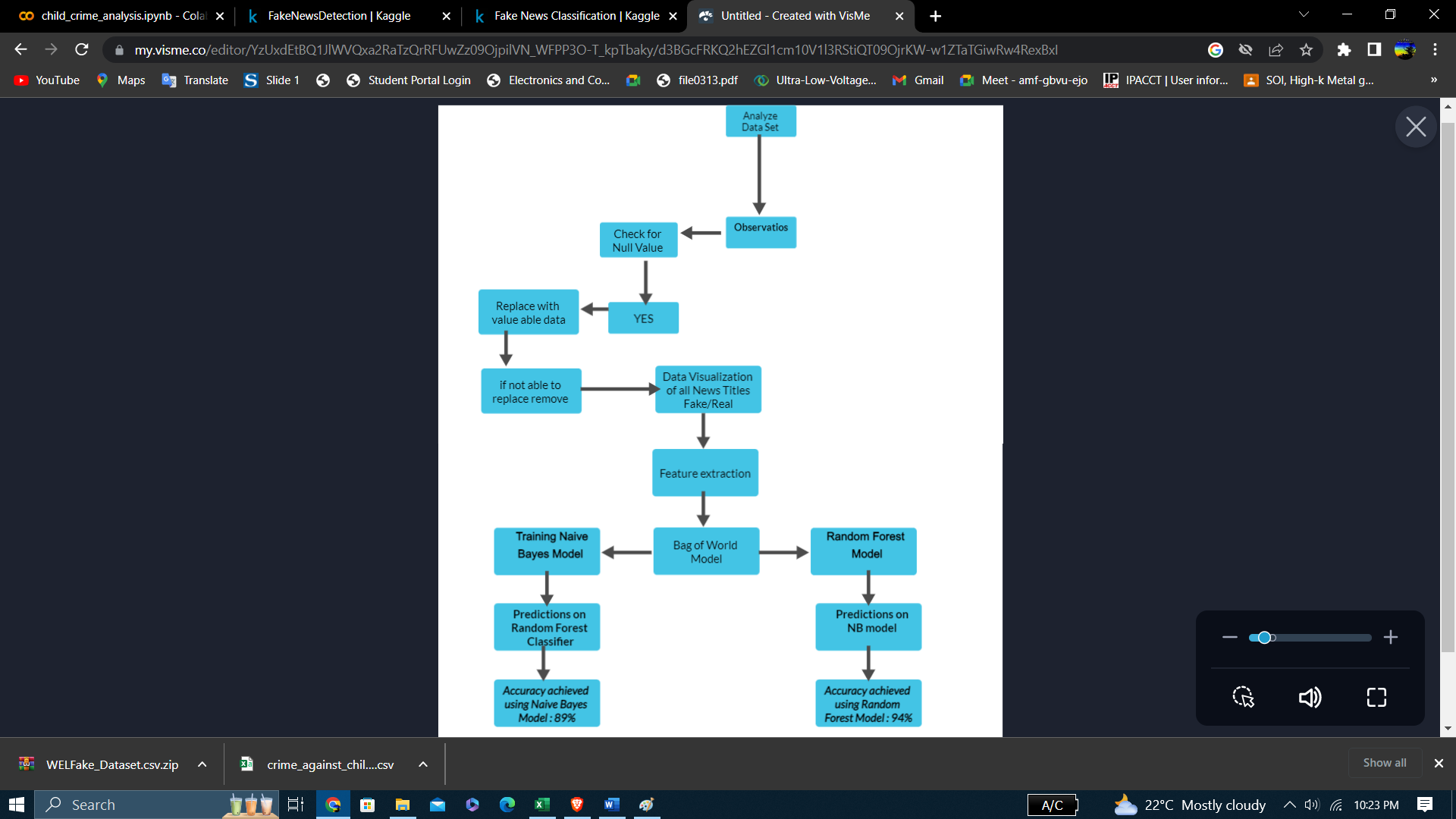
(WELFake) is a dataset of 72,134 news articles with 35,028 real and 37,106 fake news. For this, authors merged four popular news datasets (i.e. Kaggle, McIntire, Reuters, BuzzFeed Political) to prevent over-fitting of classifiers and to provide more text data for better ML training.

Dataset contains four columns: Serial number (starting from 0); Title (about the text news heading); Text (about the news content); and Label (0 = fake and 1 = real).

There are 78098 data entries in csv file out of which only 72134 entries are accessed as per the data frame.

**Published in:**

IEEE Transactions on Computational Social Systems: pp. 1-13 (doi: 10.1109/TCSS.2021.3068519).

**Work flow:**

**Observations:**

There are a total of 4 columns and 72134 rows in the data

Label is the target variable

Percentage of Real and fake News articles:

real: 51%

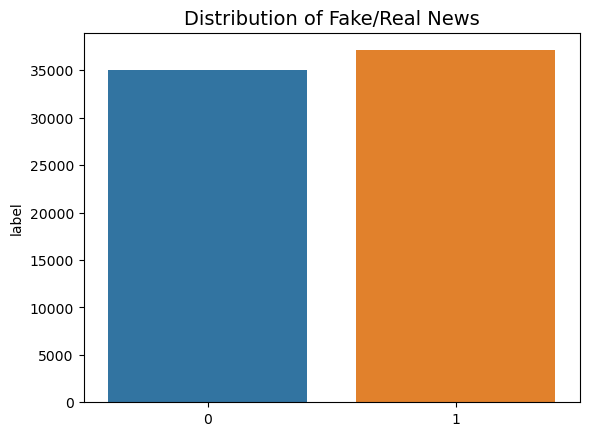
fake :49%

Missing values are present in the dataset

**Insights:**

Fake news seems to be quite a bit longer than real news.

**Distribution:**



**Feature Extraction:**

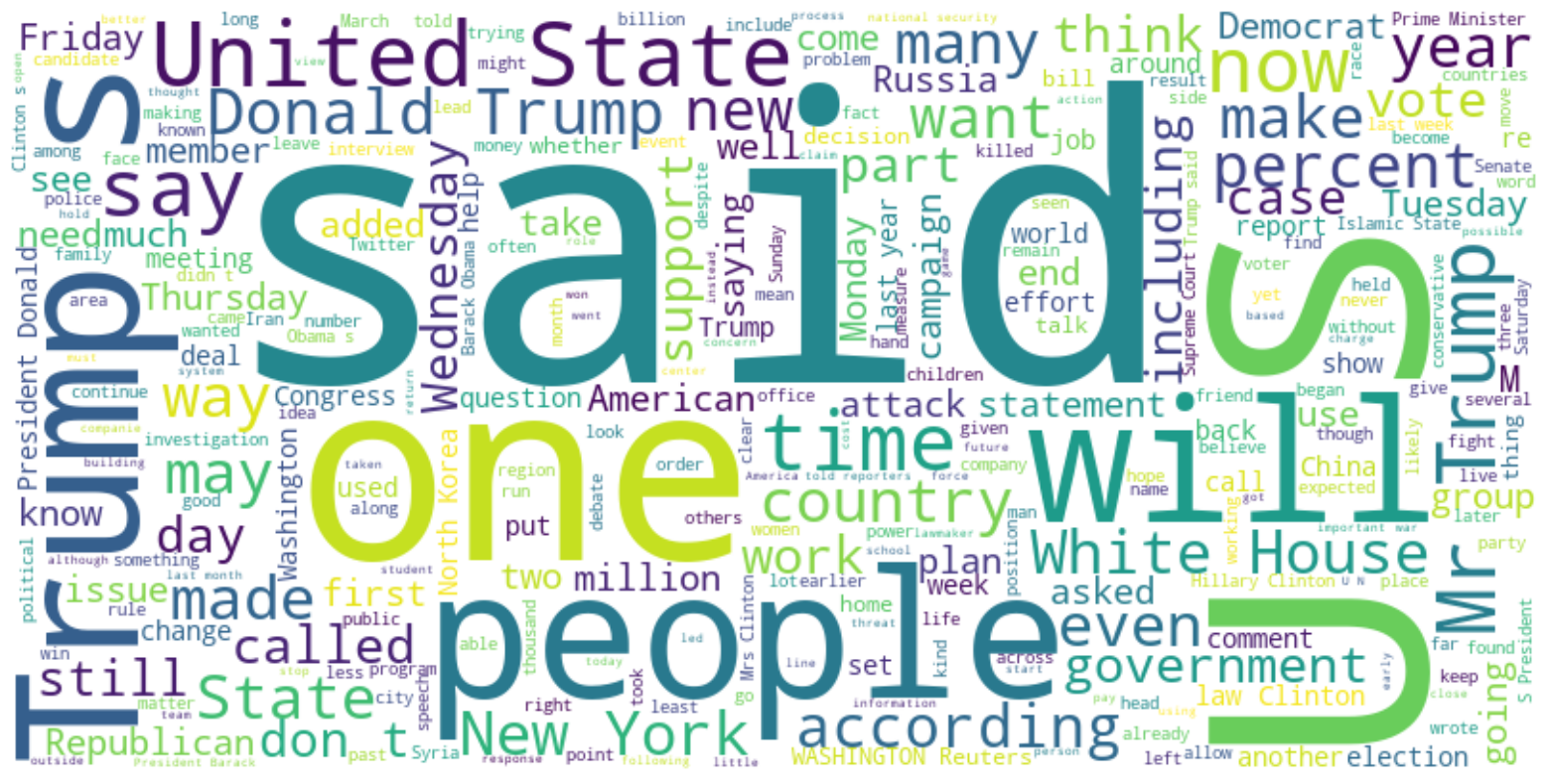
We used the Bag-of-Words approach to extract features from the text. In this approach, we represent each news article as a vector of word frequencies. We also used TF-IDF weighting to give more weight to important words and reduce the weight of common words.

**Data Visualization of all News Titles:**

**Quick Insights:**

many of available news articles are about elections and presidents of USA

**Data Visualization of texts in fake news:**



**Bag of Words Model:**

The Bag-of-Words model is a commonly used technique for representing text data in natural language processing (NLP). In this model, a document is represented as a multiset, or "bag," of its words, ignoring grammar and word order but retaining information about the frequency of each word in the document.

The process of creating a Bag-of-Words model involves the following steps:

Tokenization: The text is first broken down into individual words or tokens.

Vocabulary creation: A vocabulary of all unique words/tokens in the corpus is created.

Counting word occurrences: For each document in the corpus, a vector is created with the length of the vocabulary, where each element represents the frequency of a particular word in the document.

**Model Training and Evaluation:**

We split the dataset into training and testing sets with a 70:30 ratio. We trained the Naive Bayes and Random Forest models on the training set and evaluated their performance on the testing set using accuracy, precision, recall, and F1-score metrics.

**Results:**

The Naive Bayes model achieved:

precision recall f1-score support

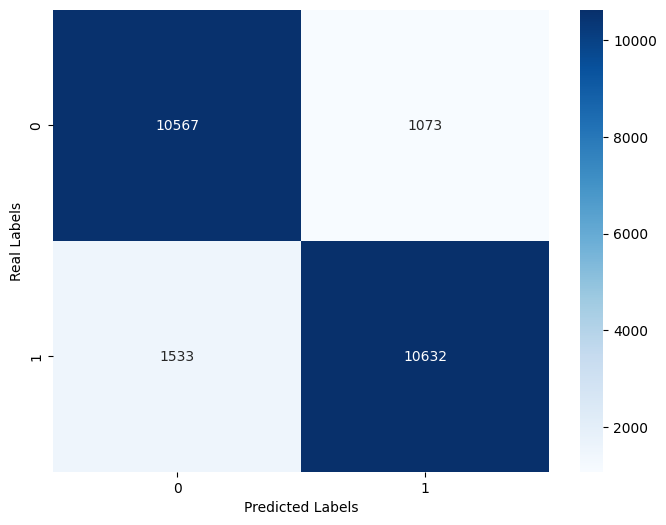
0 0.87 0.91 0.89 11640

1 0.91 0.87 0.89 12165

accuracy 0.89 23805

macro avg 0.89 0.89 0.89 23805

weighted avg 0.89 0.89 0.89 23805



The Random Forest model achieved:

precision recall f1-score support

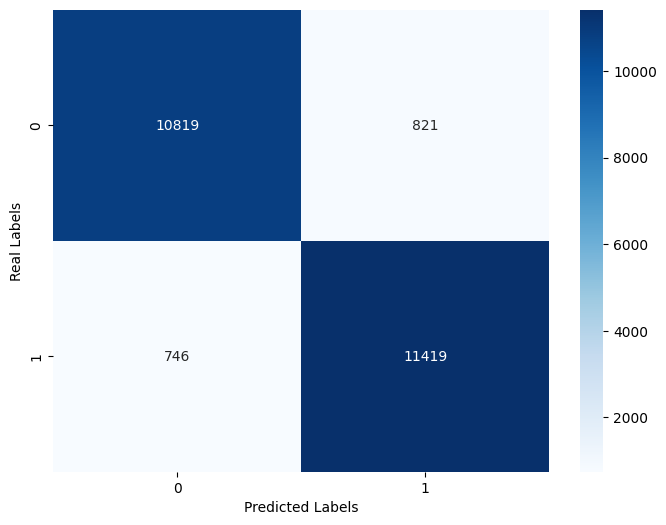
0 0.94 0.93 0.93 11640

1 0.93 0.94 0.94 12165

accuracy 0.93 23805

macro avg 0.93 0.93 0.93 23805

weighted avg 0.93 0.93 0.93 23805



**Conclusion:**

Both Naive Bayes and Random Forest models performed well in detecting fake news. However, the Random Forest model outperformed the Naive Bayes model in terms of accuracy and other evaluation metrics. Therefore, we recommend using the Random Forest model for fake news detection. Further research can be done to improve the accuracy of the models by incorporating other features, such as user information and metadata.

Coding link-

<https://colab.research.google.com/drive/1Rb5c1Q3v-Cgaow0iYTKsmmuONach4Mmi?usp=sharing>

dataset link –

<https://zenodo.org/record/4561253/files/WELFake_Dataset.csv?download=1>