# Fake News Detection Model

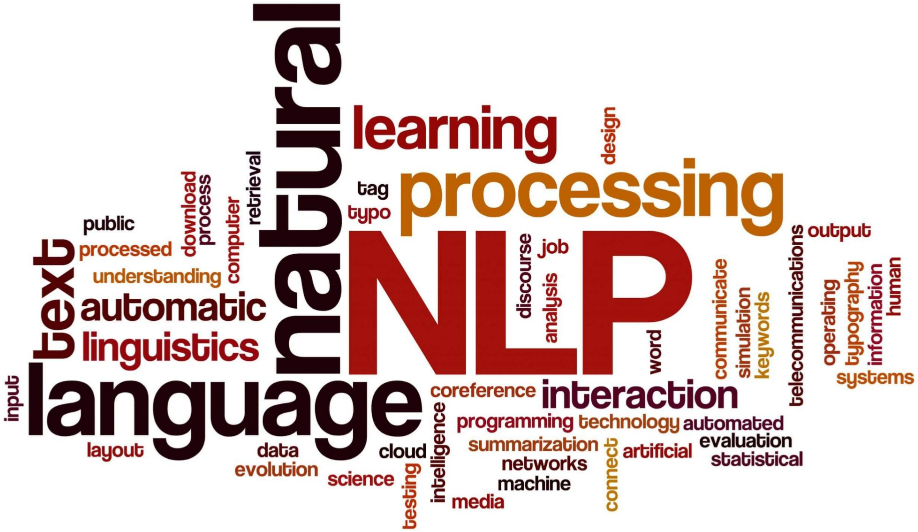
By Fabíolla Mayrink and Rael Guimaraes

https://github.com/rael-guimaraes/Fake_news_prediction

What is the Fake News Detection Model and how does it work?

What is NLP and how does it contribute to the Fake News Detection Model?

What are the advantages of using NLP for the fake news detection model?



- The Fake News Detection Model is a tool that uses NLP to analyze and classify news articles as either real or fake.

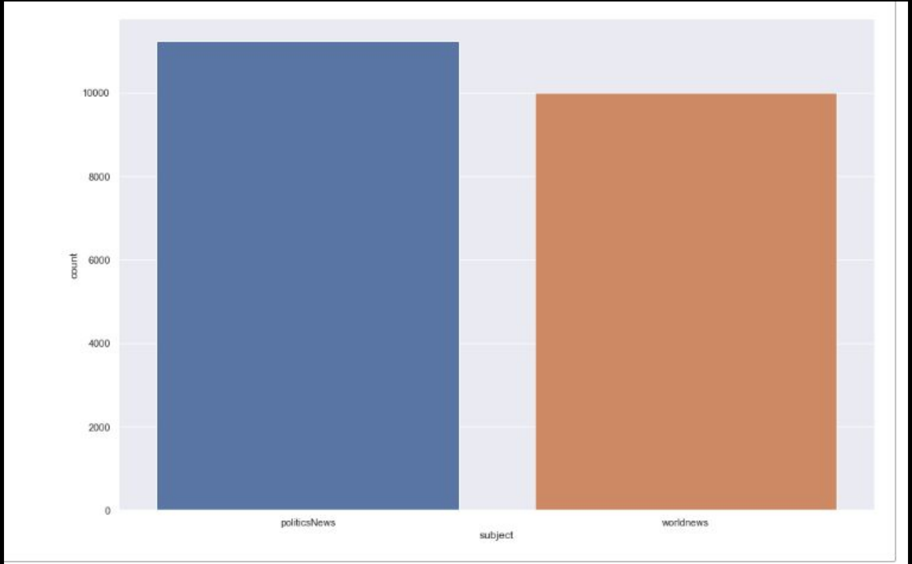
- NLP stands for Natural Language Processing, which is a part of Artificial Intelligence (AI) that gives computers the ability to understand text and spoken words in a way that is similar to humans. NLP is used in the Fake News Detection Model to analyze and classify news articles as real or fake.

- The advantages of using NLP for the fake news detection model include its ability to perform large-scale analysis, automate processes in real-time, and tailor the NLP algorithm to any company's needs and criteria, industry-specific language, sarcasm and misused words.

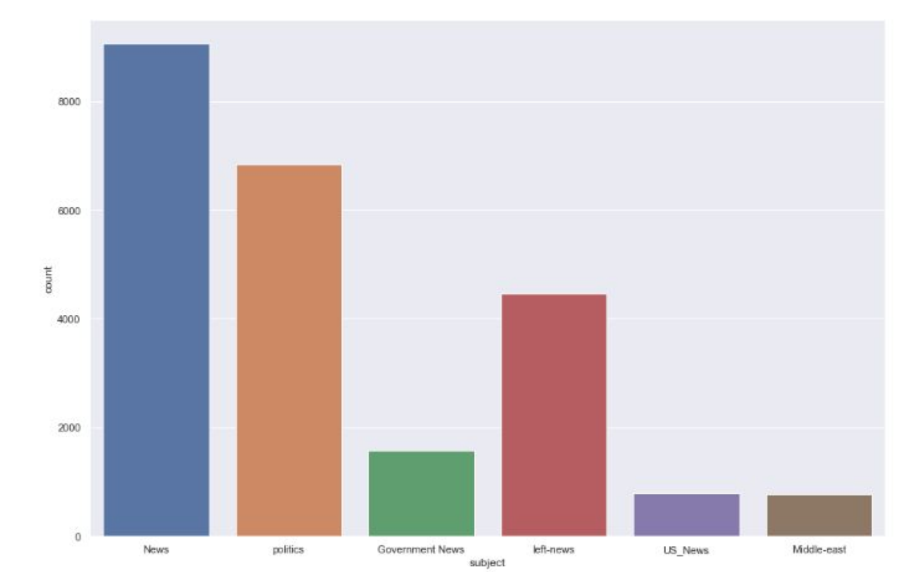
# Data Understanding

## **Fake News Detection Model Dataset**

- To develop our fake news detection model, we used 2 datasets: "True.csv" and "Fake.csv";
- News collected from 2015 to 2017;
- 2 category in true news and 6 in false news.



**Here we can see that there are only 2 categories of news in the real dataset. The top news is politicsNews.**

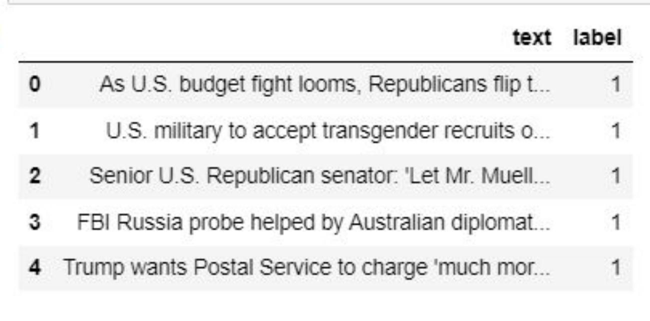


Here we can see that there are 6 categories of news in the fake dataset. The top news is News.

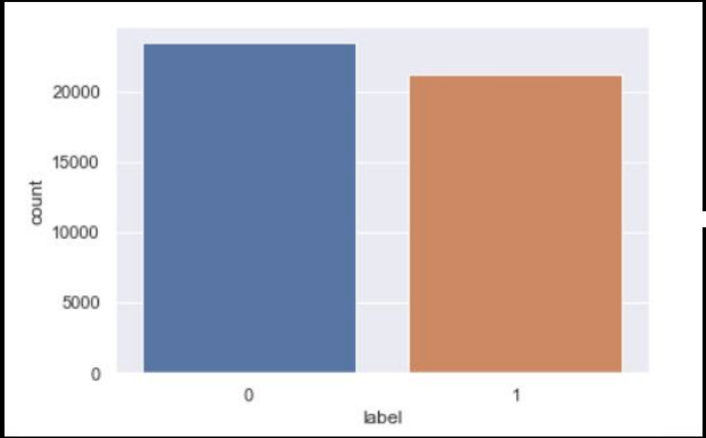
# Data Preparation

Combined the "True.csv" and "Fake.csv" datasets and created a new column called label to identify the source of each entry.

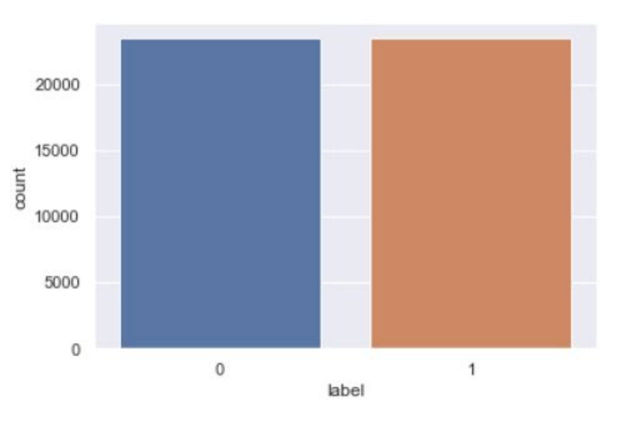
Removed noisy data, including the columns "day", "month", and "year".



| | text | label |
|---|---|---|
| **0** | As U.S. budget fight looms, Republicans flip t... | 1 |
| **1** | U.S. military to accept transgender recruits o... | 1 |
| **2** | Senior U.S. Republican senator: 'Let Mr. Muell... | 1 |
| **3** | FBI Russia probe helped by Australian diplomat... | 1 |
| **4** | Trump wants Postal Service to charge 'much mor... | 1 |

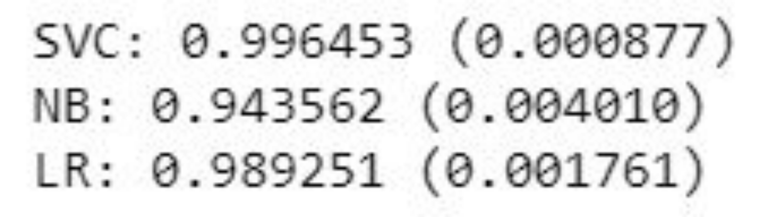


Balanced the data using oversampling technique.



# Modelling

- Description of test harness technique using stratified 10-fold cross-validation
- Importance of maintaining class distribution and fixed random seed
- Linear SVC's superiority in handling unstructured and semi-unstructured data
- Comparison with Logistic Regression model
- Linear SVC's better performance in linear and non-linear problems
- Comparison with Naive Bayes model

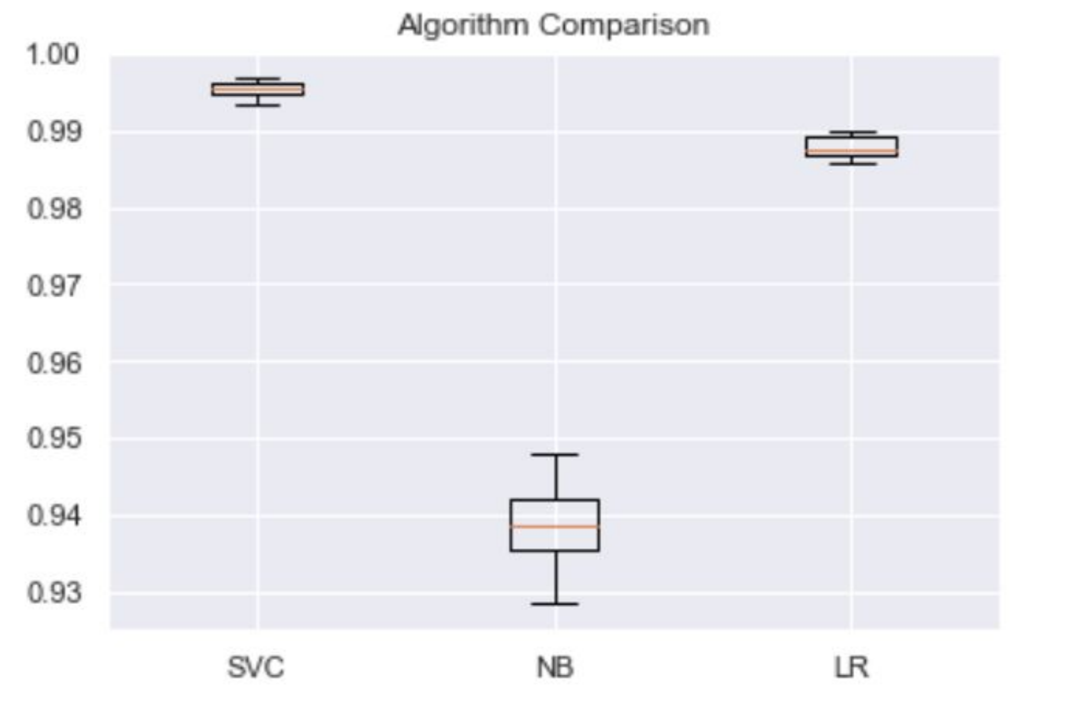


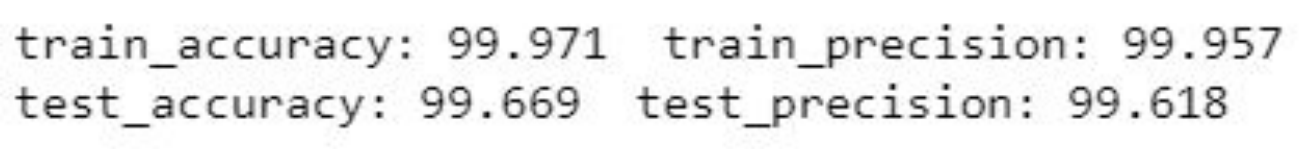
```
SVC: 0.996453 (0.000877)
NB: 0.943562 (0.004010)
LR: 0.989251 (0.001761)
```

# Evaluation of the Model

Best model for fake news detection is the Linear SVC model.

The model gave us accuracy around 99% and the precision results around 99%. It is considered to be an excellent result.

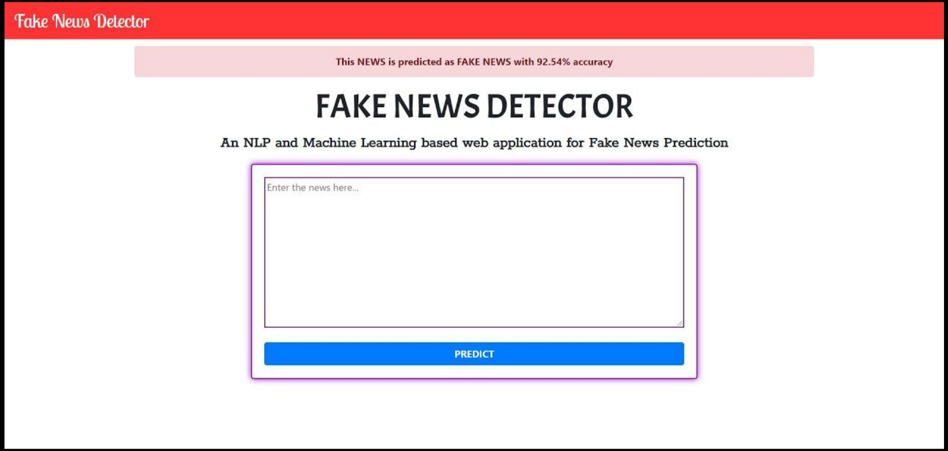




```
train_accuracy: 99.971  train_precision: 99.957
test_accuracy: 99.669  test_precision: 99.618
```

# Deployment of the Model

We plan to deploy our fake news detection model as a web application that users can subscribe to and access by entering a news article URL or copying and pasting the article. Subscription fees will vary based on the client profile and access will be limited based on the paid fee. We aim to constantly update the model with up-to-date data and train it to avoid model staleness, and deploy it on a smaller scale to overcome high expenses. At the start, we will be responsible for data collection and maintenance, and expand as necessary.



This NEWS is predicted as FAKE NEWS with 92.54% accuracy

FAKE NEWS DETECTOR

An NLP and Machine Learning based web application for Fake News Prediction

# References

Adachi, F.de P. (2021) *Deploying a fake news detector web application with Google Cloud Run and flask*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/deploying-a-fake-news-detector-web-application-with-google-cloud-run-and-flask-eb750cce986d (Accessed: April 25, 2023).

Bisaillon, C. (2020) *Fake and real news dataset*, *Kaggle*. Available at: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset (Accessed: April 25, 2023).

Borcan, M. (2020) *TF-IDF explained and python sklearn implementation*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/tf-idf-explained-and-python-sklearn-implementation-b020c5e83275 (Accessed: April 25, 2023).

DeepLearning.AI (2023) *Natural language processing (NLP) - A complete guide*, *(NLP) [A Complete Guide]*. Available at: https://www.deeplearning.ai/resources/natural-language-processing/ (Accessed: April 10, 2023).

Ganesan, K. (2023) *What are stop words?*, *Kavita Ganesan, PhD*. Available at: https://kavita-ganesan.com/what-are-stop-words/#.ZEeRBXbMI6k (Accessed: April 25, 2023).

Haskins, J. (2023) *Fake news: What laws are designed to protect*, *LegalZoom*. Legalzoom.com. Available at: https://www.legalzoom.com/articles/fake-news-what-laws-are-designed-to-protect (Accessed: April 10, 2023).

IBM (2023) *What is natural language processing?*, *IBM*. Available at: https://www.ibm.com/topics/natural-language-processing (Accessed: April 25, 2023).

INNOQ (2023) *ML-ops.org*, *ML Ops: Machine Learning Operations*. Available at: https://ml-ops.org/content/mlops-principles#:~:text=Model%20staleness%20test.,of%20prediction%20in%20intelligent%20software. (Accessed: April 25, 2023).

Jain, P. (2021) *Basics of countvectorizer*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/basics-of-countvectorizer-e26677900f9c#:~:text=Countvectorizer%20is%20a%20method%20to%20convert%20text%20to%20numerical%20data. (Accessed: April 25, 2023).

Javatpoint (2021) *What is noise in data mining - javatpoint*, *www.javatpoint.com*. Available at: https://www.javatpoint.com/what-is-noise-in-data-mining (Accessed: April 25, 2023).

Mazumder, S. (2022) *5 techniques to handle imbalanced data for a classification problem*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/ (Accessed: April 25, 2023).

Narkhede, S. (2018) *Understanding confusion matrix*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 (Accessed: April 25, 2023).

Oshikawa, R., Qian, J. and Wang, W.Y. (2020) *A survey on natural language processing for fake news detection*, *ACL Anthology*. Available at: https://aclanthology.org/2020.lrec-1.747/ (Accessed: April 25, 2023).

TechTarget (2017) *What is support Vector Machine (SVM)?: Definition from TechTarget*, *WhatIs.com*. TechTarget. Available at: https://www.techtarget.com/whatis/definition/support-vector-machine-SVM#:~:text=A%20support%20vector%20machine%20(SVM,which%20are%20labeled%20for%20classification. (Accessed: April 25, 2023).