# Fake and Real News Classification

## Final Term Project

DATS_6312 NLP for Data Science

Instructor: Amir Jafari

by

## Group

Meghana Gantla

Kohisha Aruganti

# Table of Contents

- Scope of the project
- Features implemented
- Data sources
- Expected outcomes
- Logical architecture
- Data flow
- Cloud services
- Inputs and outputs
- Conclusion

# Scope of the project

- Build classical and non-classical machine learning models to classify fake and real news.
- The models will be trained using nearly 45000 articles containing both real (1) and fake (0) data.
- Finally, we understand which of the machine learning models is best performing

# Data Source

kaggle

## Fake and Real News Dataset

https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

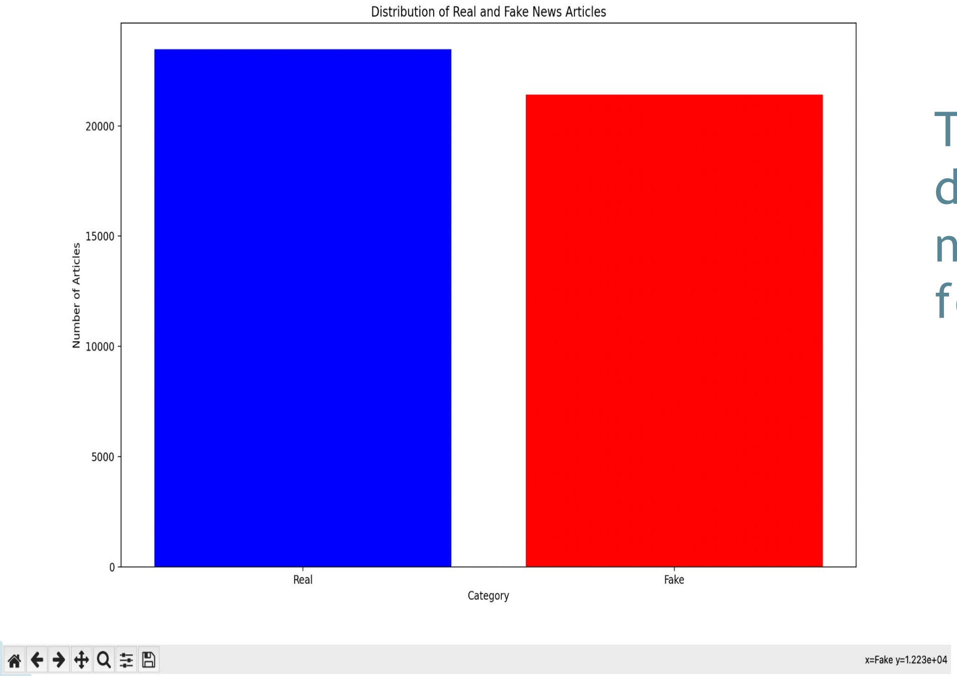
# About Data

- The dataset is classified into fake and real news. The “True” articles are sourced from Reuters.com and “Fake” articles are collected from unverified websites and Wikipedia.
- We have ‘Date’, ‘title’, ‘Subject’, and ‘Text’ in both of the CSV files initially. We combined them by adding a ‘Target’ variable of 0’s and 1’s.
- There are a total of 44,919 observations, out of which 21,417 are from True.csv and 23,502 are from Fake.csv.

# Data Preprocessing

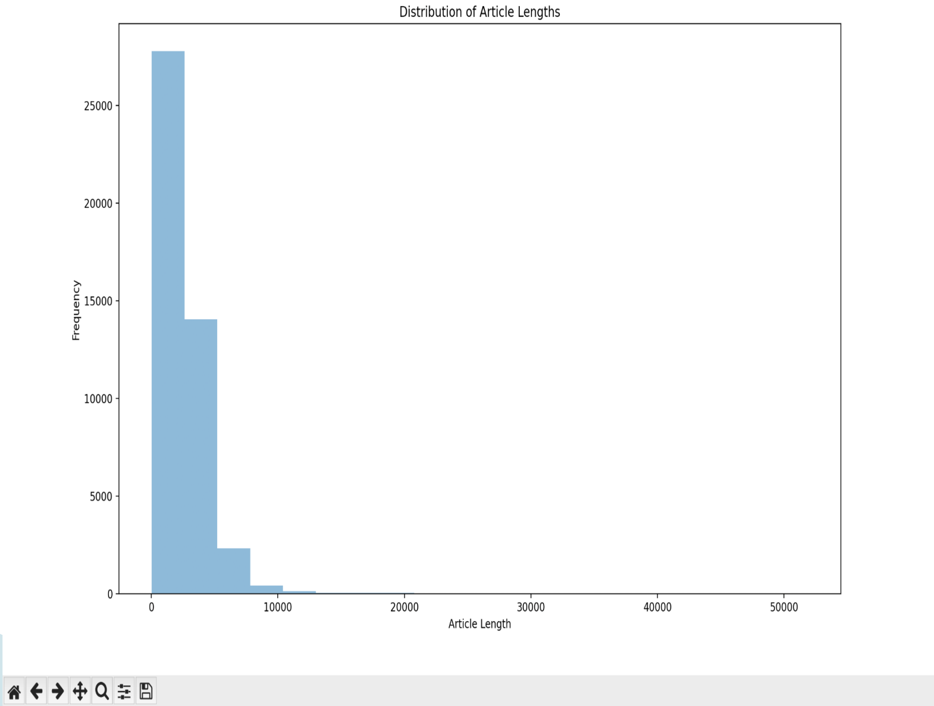
- We combined the ‘title’ and ‘text’ to make our text column.
- We then removed URLs
- Applied lower casing
- Removed contractions
- Removed punctuations
- Removed stop words
- Performed Lemmatization

# EDA - Fake vs Real news counts



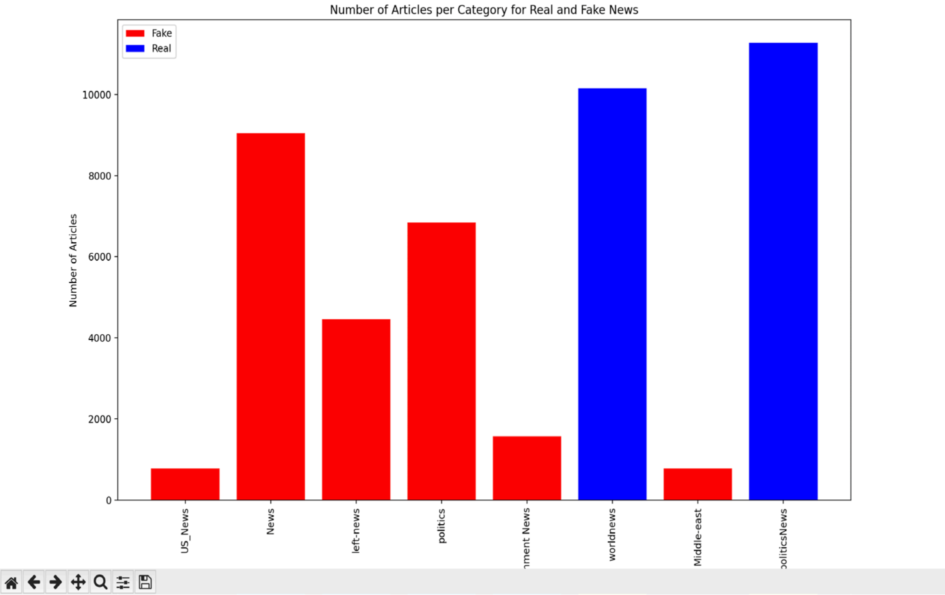
The data is almost equally distributed. Therefore, we do not need to upsample or downsample for the analysis

# Article length



Here we can observe the distribution of article lengths. There’s a sufficient amount of data in each row.

# Subject



We can see that the subject column for real articles and fake articles can cause problems in modeling because of this distribution. Therefore, we will not be including it in the analysis.

# Modeling

**In classical machine learning models, we employed,**

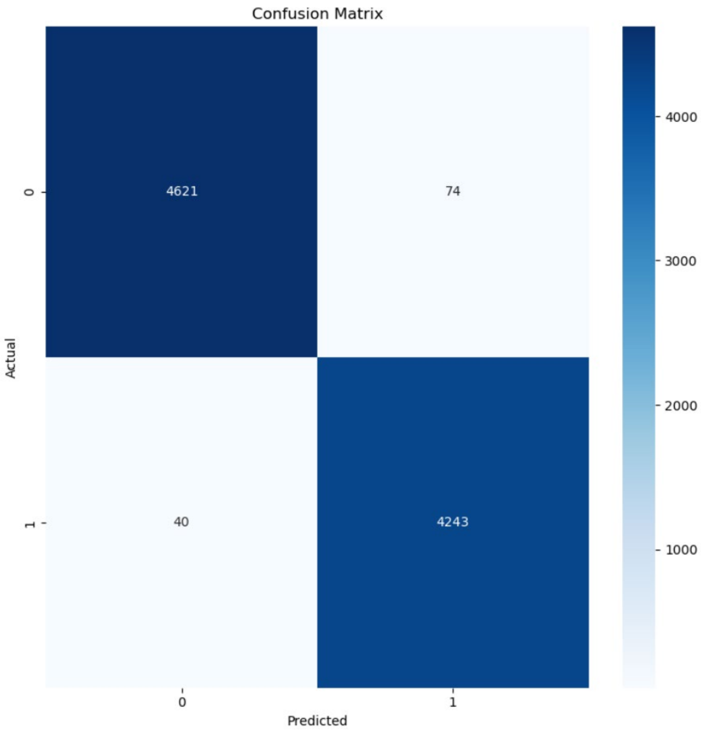
- Logistic regression
- Naïve Bayes

**For non-classical models, we tried out,**

- RoBERTa
- DistilBERTa

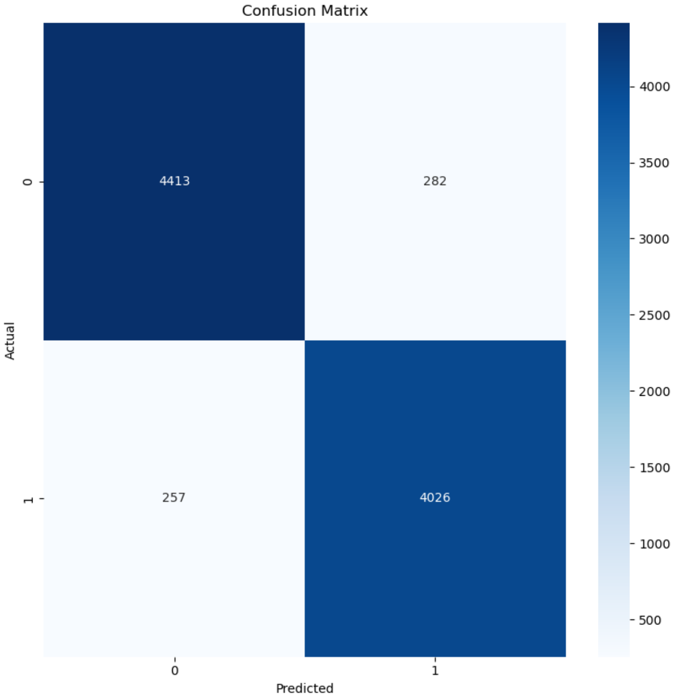
# Logistic Regression

```
====== Logistic regression results ======
Accuracy: 0.9873
f1-score: 0.9867
==========================
Confusion matrix
[[4621   74]
 [  40   4243]]
==========================
Classification report
             precision    recall  f1-score   support

          0       0.99      0.98      0.99      4695
          1       0.98      0.99      0.99      4283

    accuracy                          0.99      8978
   macro avg       0.99      0.99      0.99      8978
weighted avg       0.99      0.99      0.99      8978
```



# Naïve Bayes

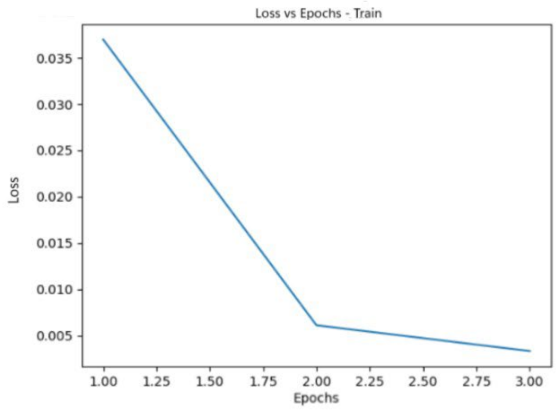
```
====== Naive Bayes results ======
Accuracy: 0.9400
f1-score: 0.9373
=========================
Confusion matrix
[[4413  282]
 [ 257 4026]]
=========================
Classification report
              precision    recall  f1-score   support

           0       0.94      0.94      0.94      4695
           1       0.93      0.94      0.94      4283

    accuracy                           0.94      8978
   macro avg       0.94      0.94      0.94      8978
weighted avg       0.94      0.94      0.94      8978
```

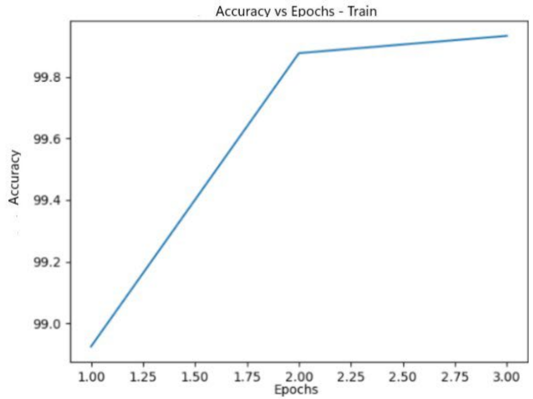


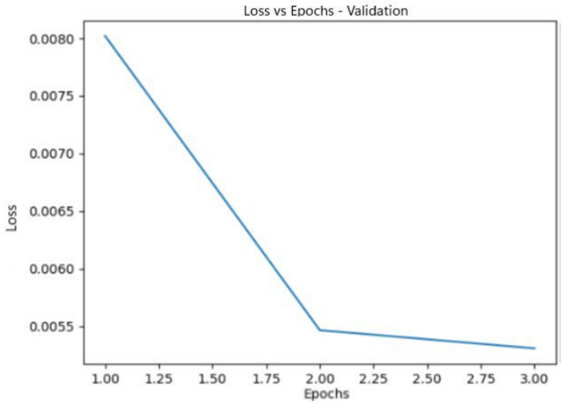
# DistilBERTa

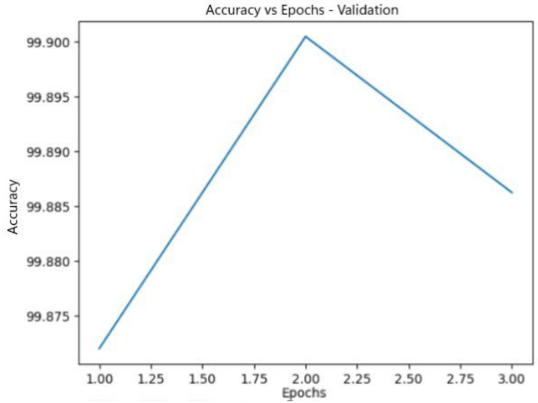
| Epoch | Batch_size | Learning_rate | Max_len | Training_accuracy | Validation_accuracy |
|---|---|---|---|---|---|
| 3 | 32 | 0.00001 | 256 | 99.91 | 99.87 |
| 2 | 32 | 0.01 | 256 | 53.47 | 53.45 |

# DistilBERTa





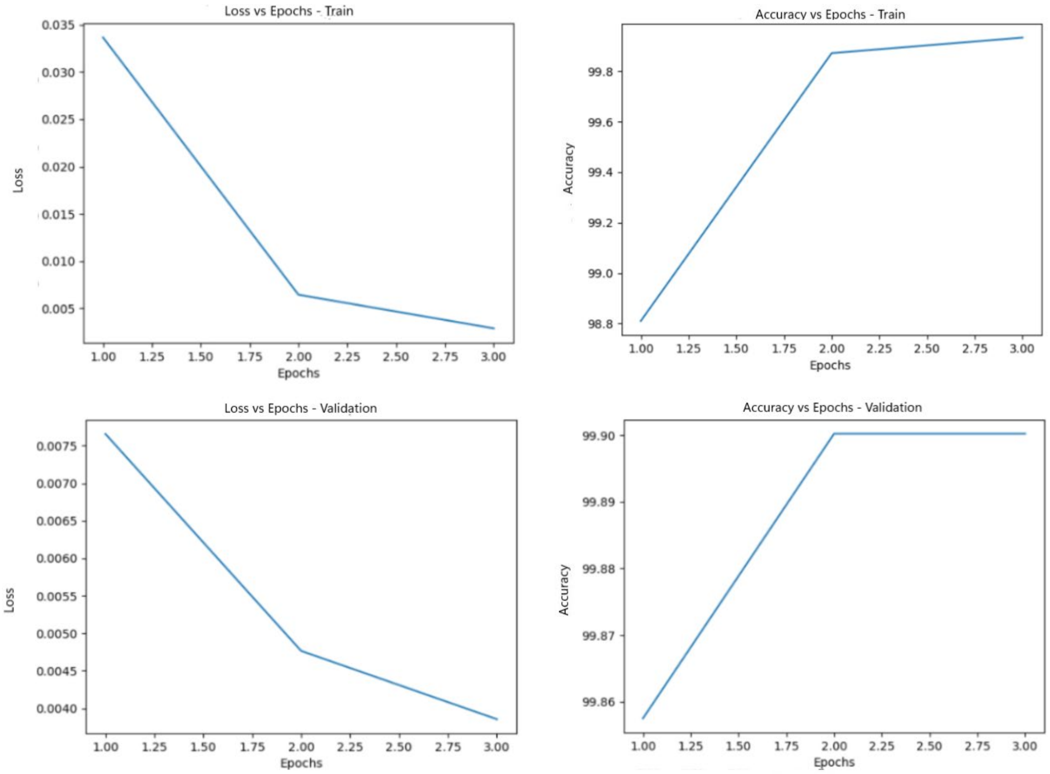




# RoBERTa

| Epoch | Batch_size | Learning_rate | Max_len | Training_accuracy | Validation_accuracy |
|---|---|---|---|---|---|
| 3 | 32 | 0.00001 | 256 | 99.87 | 99.89 |
| 3 | 16 | 0.00001 | 256 | 99.94 | 99.9 |

# RoBERTa


Loss vs Epochs - Train
Loss
0.035
0.030
0.025
0.020
0.015
0.010
0.005
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
Epochs
Accuracy vs Epochs - Train
Accuracy
99.8
99.6
99.4
99.2
99.0
98.8
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
Epochs
Loss vs Epochs - Validation
Loss
0.0075
0.0070
0.0065
0.0060
0.0055
0.0050
0.0045
0.0040
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
Epochs
Accuracy vs Epochs - Validation
Accuracy
99.90
99.89
99.88
99.87
99.86
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
Epochs

# Conclusion

After the analysis, we concluded that,

- Classical models have performed surprisingly well for this data
- Non-classical models however outperformed the classical models by a slim margin.
- RoBERTa has the highest accuracy in non-classical models.

# Thank you!