PROJECT TRUTHMINER

Mining through headlines to uncover real vs fake

Sergio Eguakun
Natalia Dominguez
Kaan Tirpanci
Zahory Vasquez

Table of Contents

Project Overview

Goal:

- Classifying news headlines as either real (1) or fake (0)

Approaches:

- Data preprocessing
- Feature Engineering: TF-IDF
- Models and Evaluation
- Transformers (distilBert)
- Test data
- Transfer learning

Data Preprocessing

- Steps taken here to clean the data involved :
- Lowercase conversions
- Removing stopwords
- Removing unnecessary punctuations
- Checking missing values and missing text
- Ensuring Data Types were consistent throughout (str)
- Fixing whitespace

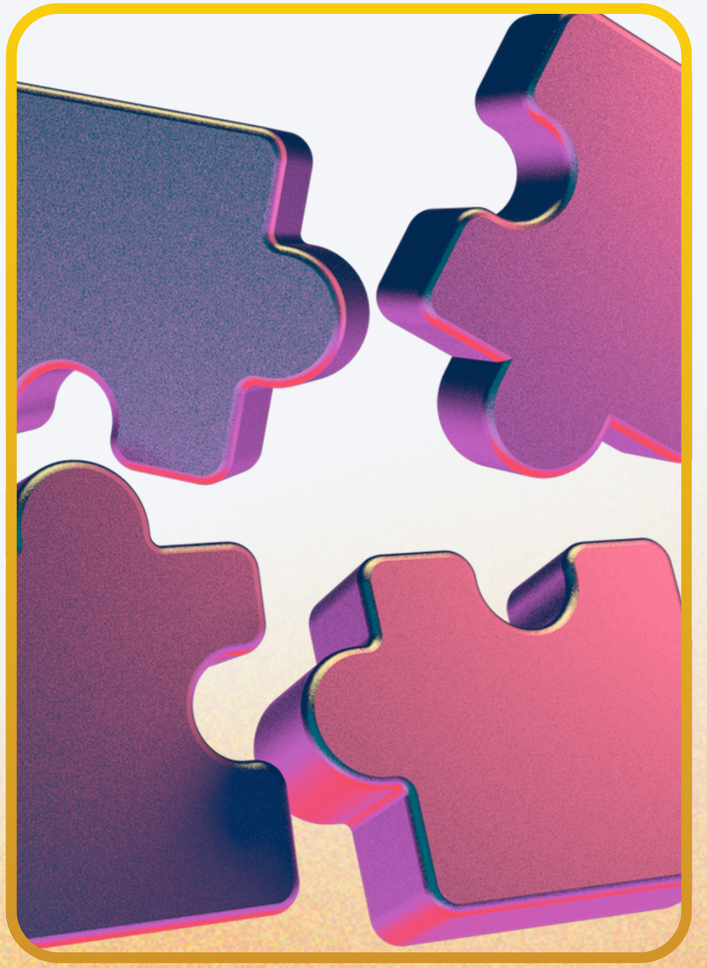
Conclusion: The headlines show very little differences after implementing the text cleaning.

Feature Engineering: Term Frequency—Inverse Document Frequency (TF-IDF)

- Create TF-IDF features

```
Training features: (23906, 5000)
Validation features: (10246, 5000)
```

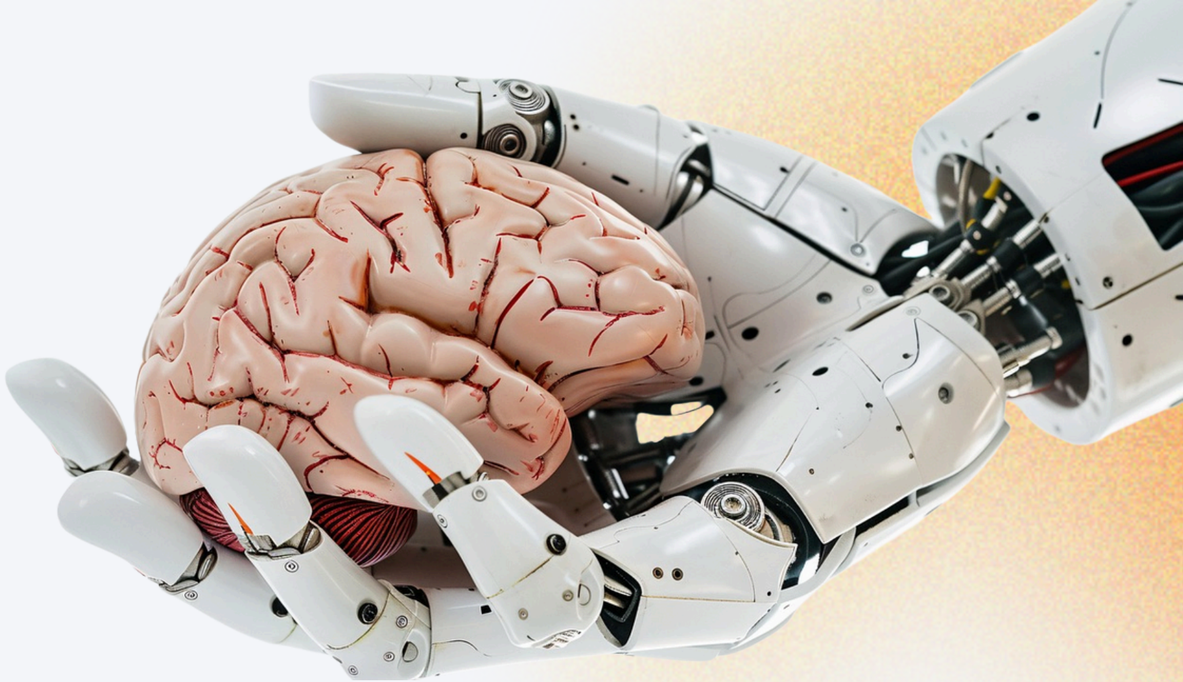
- Parameters used for vectorizer here:
- max features=5000
- stop_words
- ngram_range=(1,2)
- max_df=0.95



Models and Evaluation

Model	Accuracy
Logical Regression	0.9277
Naive Bayes	0.9233
Random Forest Classifier	0.9133
Linear SVM	**0.9319**
XGBoost Classifier	0.8751

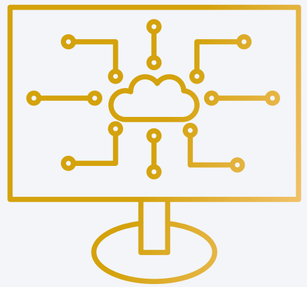
Transformer: distil Bidirectional Encoder Representations (distilBert)



- Bert Transfer Model used without fitting on the data
- Reason for this was to compare performance on Transfer models before and after fitting

- Accuracy: 0.5493

Test Data



Data Preprocessing

Term Frequency—Inverse Document Frequency (TF-IDF)

Predictions

Transfer learning

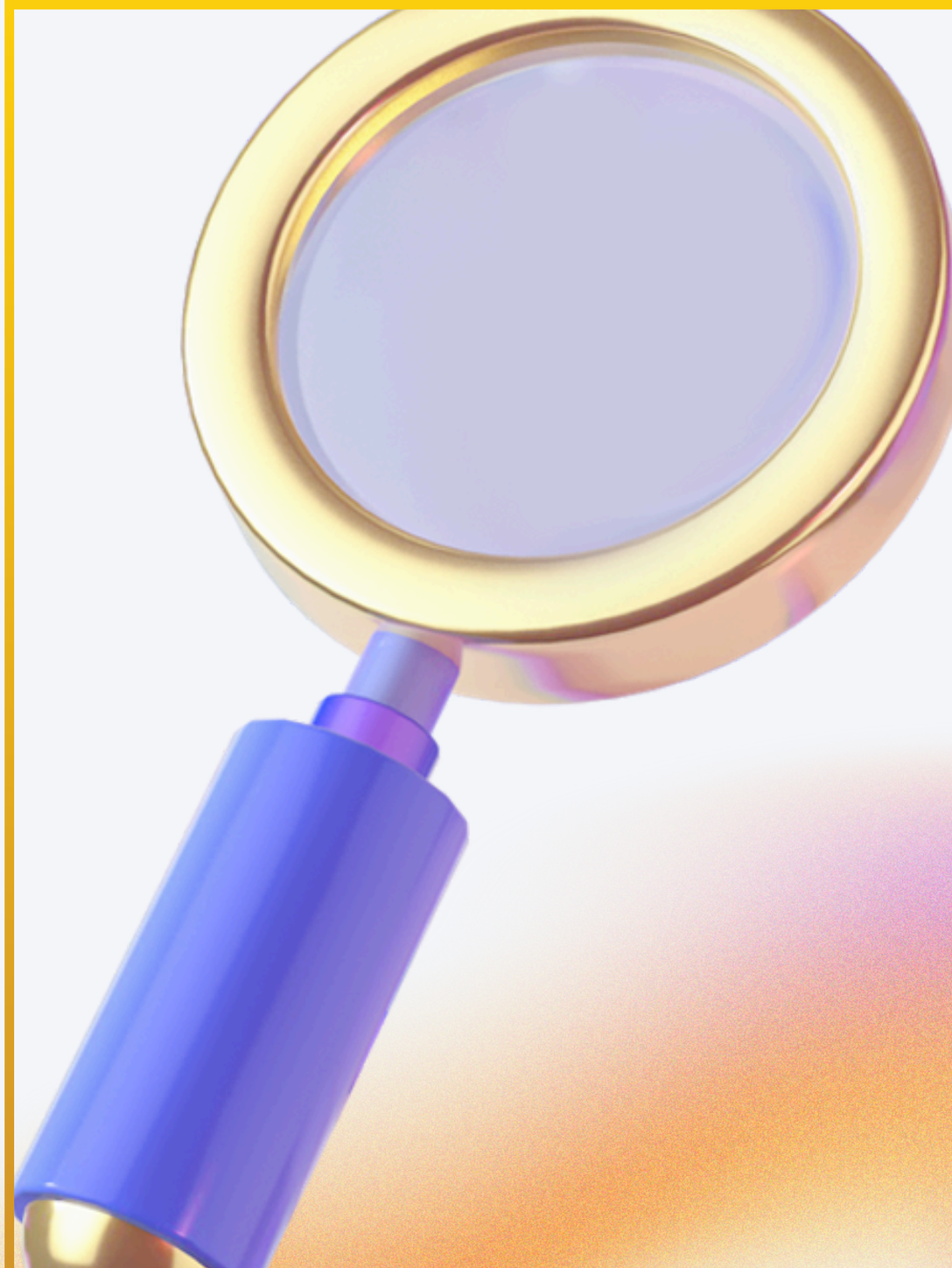
- Train the Pre-trained Model distilBERT and fit it with our data in order to generate better results

Transformers	Accuracy
distilBert	0.5493
Training PTD	**0.9868**

Conclusion

01 The best model was Linear SVM (0.9319)

02 After the transfer learning the accuracy improve (0.9868)



Thank you!