Is this news real or fake?

Using text classification to detect fake news

Agenda

- Introduction of problem
- Introduction of data
- Exploratory data analysis
- Model explanation
- Summary
- Questions

With the quickness in which articles can be shared on social media and the rise of “clickbait” headlines, how do we know if the news is true?

Dataset - Fake News Detection

The dataset used in this project is Fake News Detection from Bhavik Jikadara via Kaggle. It contains two CSVs, true.csv for true news stories and fake.csv for fake news stories.

Each contain four columns, title, text, subject, and date.

Exploratory Data Analysis

Fake

- 23481 rows
- Subjects
  - News
  - Politics
  - Government News
  - Left-news
  - US_News
  - Middle-east

True

- 21417 rows
- Subjects
  - Politics News
  - Worldnews

Exploratory Data Analysis

Combined data with additional column ‘true’

- 44898 rows, 5 columns
- Data types:
    - title   object
    - text   object
    - subject   object
    - date   object
    - true   int64

Model Explanation

For the models used in this project, I used the Spacy Python package for natural language processing.

Models Used

- Bag of Words
- TF-IDF
- Naïve Bayes

Model Explanation - Text

CountVectorizer aka Bag of Words

- Accuracy: 0.95
- Precision: 0.9506632499246308
- Recall: 0.95
- F1 Score: 0.9499498746867167

Naïve Bayes with CountVectorizer

- Accuracy: 0.905
- Precision: 0.9050120192307693
- Recall: 0.905
- F1 Score: 0.9049833546093967

TF-IDF (Term Frequency-Inverse Document Frequency)

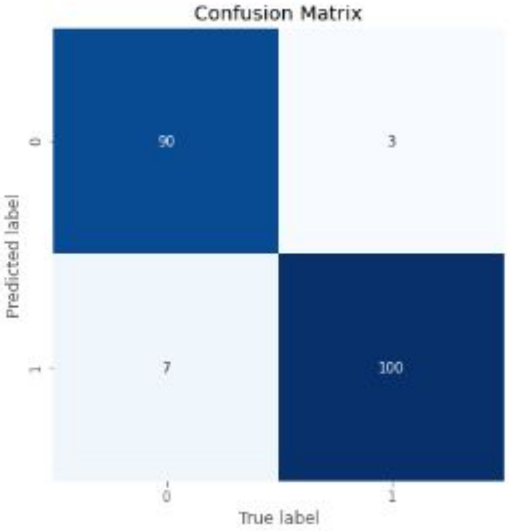
- Accuracy: 0.94
- Precision: 0.9402120212021202
- Recall: 0.94
- F1 Score: 0.9400120048019207

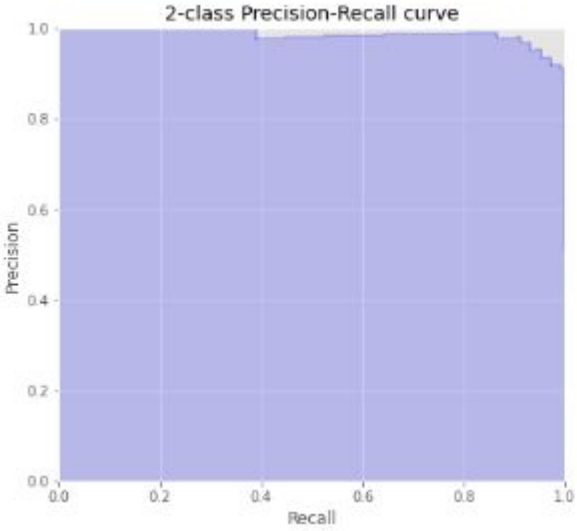
Naïve Bayes with TF-IDF

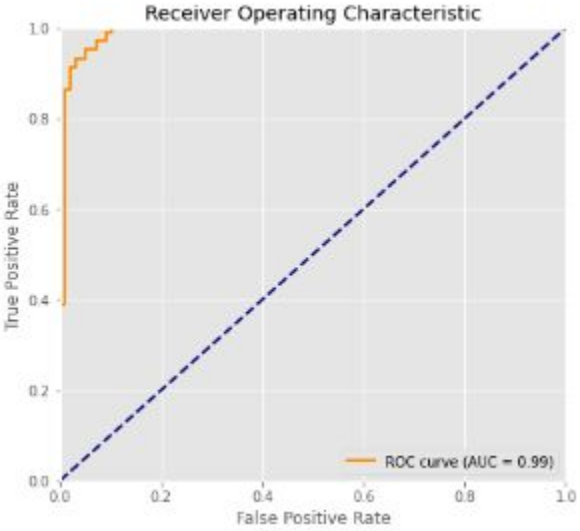
- Accuracy: 0.895
- Precision: 0.895008012820513
- Recall: 0.895
- F1 Score: 0.8949816024630173

Model Explanation - Text

CountVectorizer aka Bag of Words

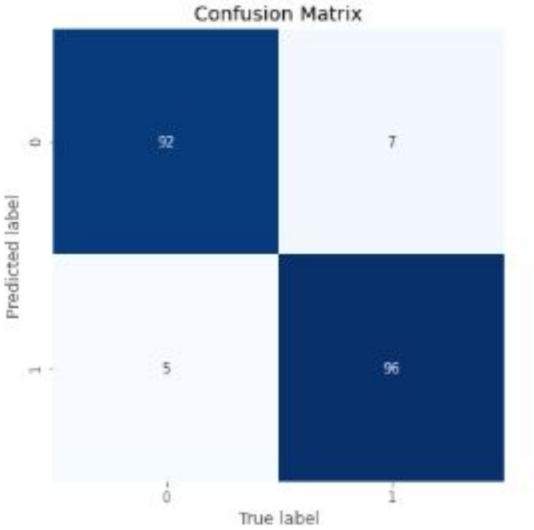


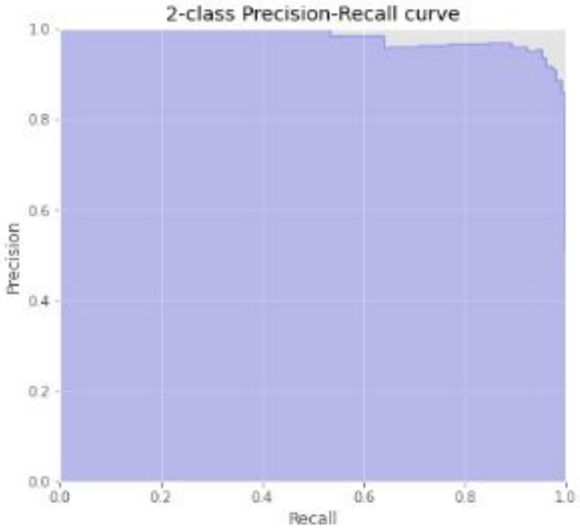


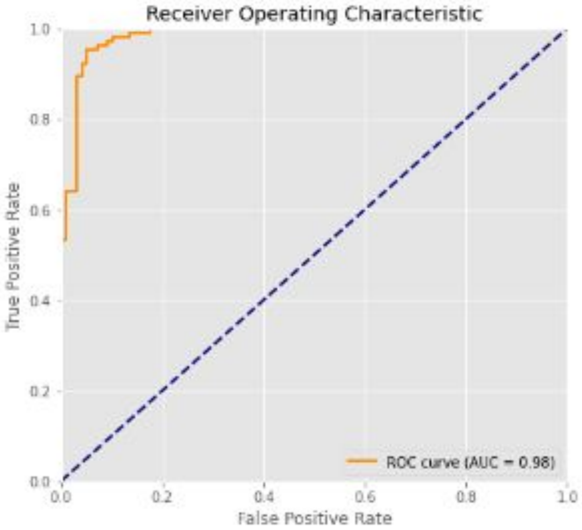


Model Explanation - Text

TF-IDF







Model Explanation - Title

CountVectorizer aka Bag of Words

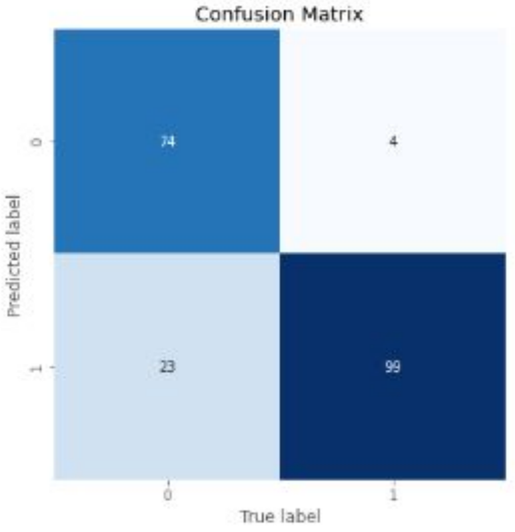
- Accuracy : 0.8650
- Precision: 0.8115
- Recall   : 0.9612
- ROC AUC : 0.9514

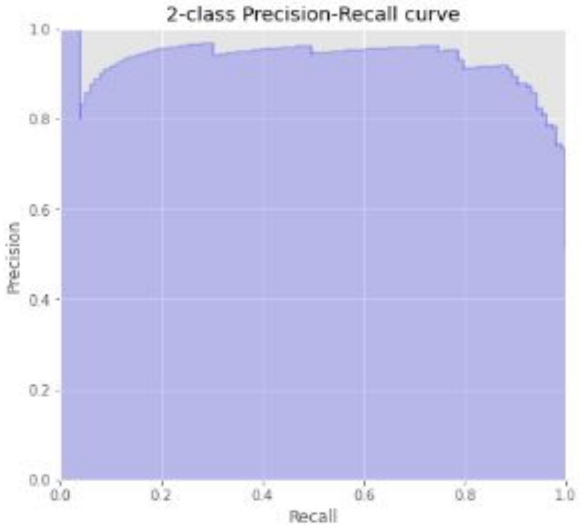
TF-IDF

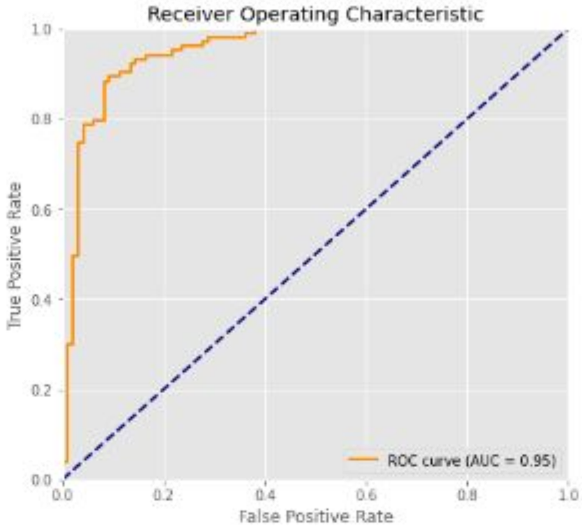
- Accuracy : 0.8900
- Precision: 0.8932
- Recall   : 0.8932
- ROC AUC : 0.9554

Model Explanation - Title

CountVectorizer aka Bag of Words

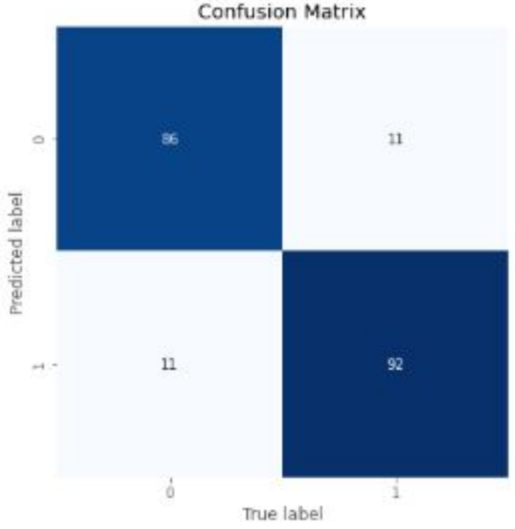


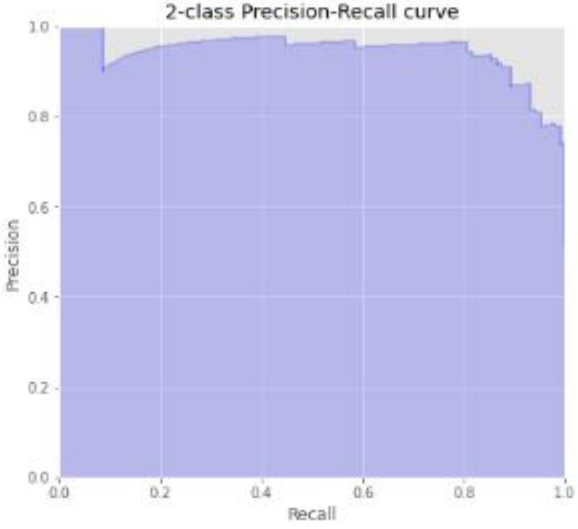


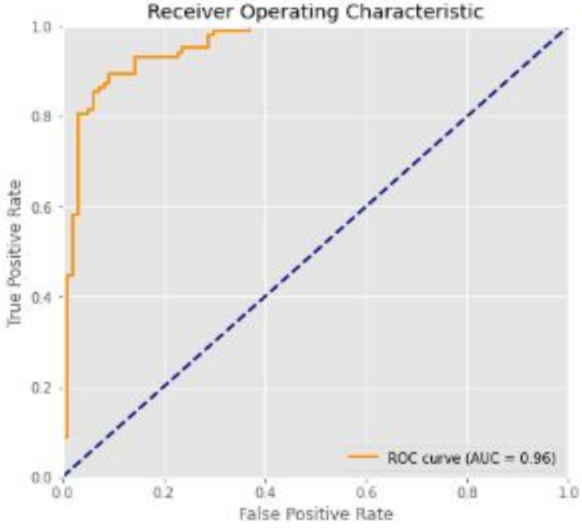


Model Explanation - Title

TF-IDF







Summary

Of the models tested, Bag of Words had the best scores using a cleaned up version of the text body from the news articles as the feature column with a .95 accuracy score.

Questions?