
Capstone:
Fake News
Classification
By Andrew Chia

Table of Contents

01 Background & Problem Statement

02 Exploratory Data Analysis

03 Modelling & Evaluation

04 Conclusion & Recommendations

01

Background & Problem Statement

Fake News

What?

Fake News: false or misleading information presented as news.

Why?

Damage the reputation of a person or entity or making money through advertising revenue.

How Severe is the Problem?

2.8B

Of social media engagement among top 100 news sources are fake

USD $78B

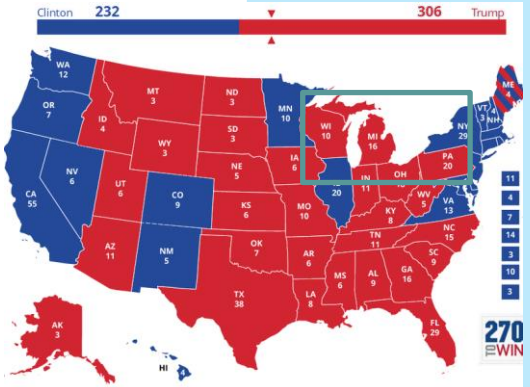
Estimated global economic cost of fake news in year 2019

Fake News in US

Fake News in US are mainly political in nature.

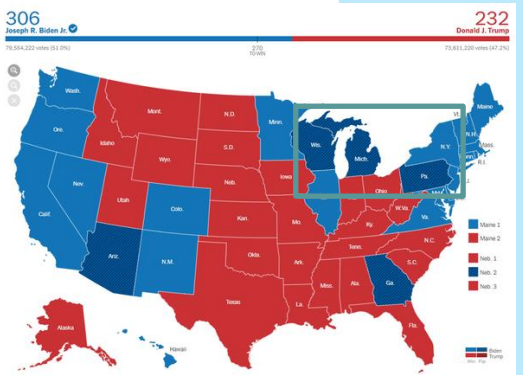
<u>2016 Presidential Election</u>

- Fake News became a “buzzword”.
- Studies suggest that fake news play an important role in Donald Trump’s win.
- Some examples of fake news:
  - Hillary Clinton approved weapons sales to ISIS
  - Hillary Clinton was in “very poor health condition”.



<u>2020 Presidential Election</u>

- Multiple fake news being generated to portray that the election was fraud and was stolen from Republicans and Donald Trump.
- Some examples of fake news:
  - Burning of ballot boxes with votes belonging to Trump.
  - Claims that dead people went voting.

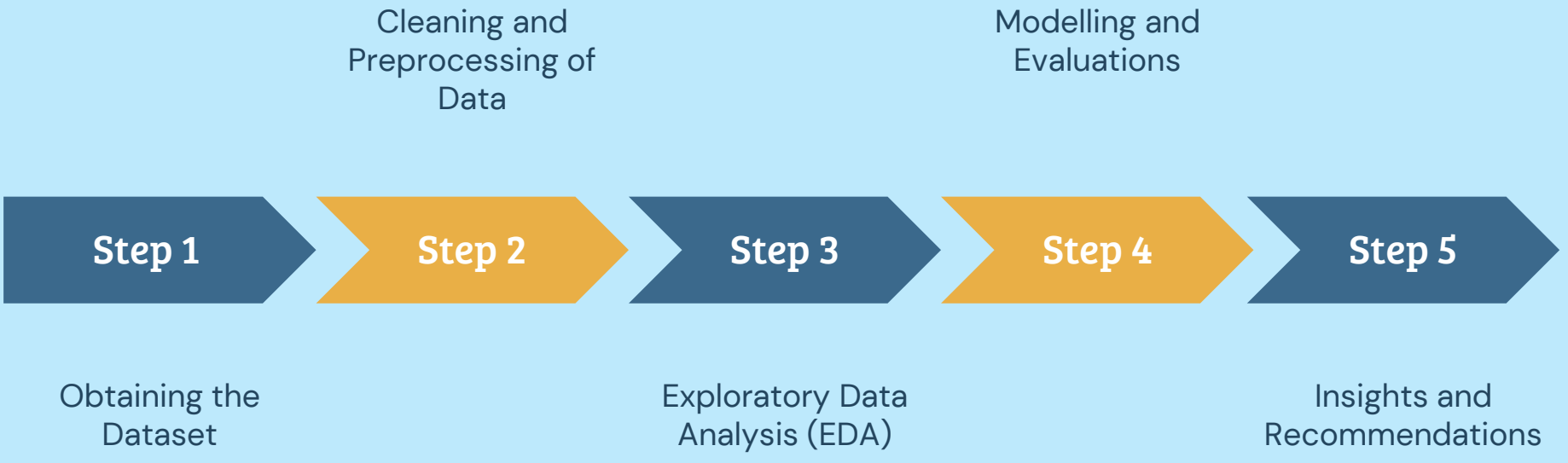


Effects of Fake News in US

Problem Statement

“ This projects aim to use data science methods (classification modellings) to predict whether a news article is fake or real, with prediction having the higher the accuracy, F1 and AUC Score the better (as close to 1 as possible), so as to enable US citizens to have a better capability to differentiate between fake or real news, by allowing them to do it themselves“.

Workflow Outline



Dataset Used

Main Dataset

- Records the title and text of the news articles in US
- Around 25,000 entries of articles
- Time range: Around the US 2016 Presidential Election

Data Cleaning / Data Pre-processing

Data Cleaning

- Imputation of null values with “-” to keep the other columns info
- Remove duplicate entries

Natural Language Processing Techniques

- Removing punctuations
- Tokenizing
- Remove stopwords
- Done lemmatization (more accurate compared to stemming)

Vectorization

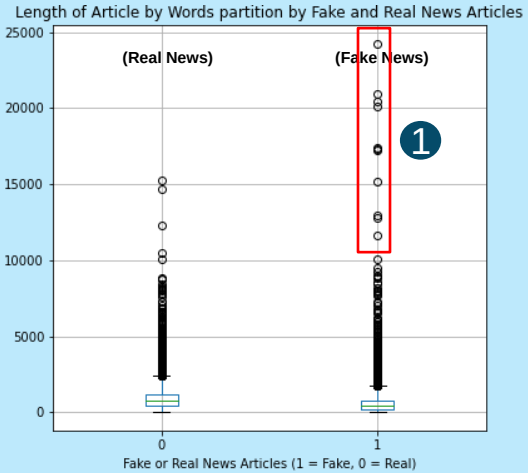
- Count Vectorization
- TF-IDF Vectorization

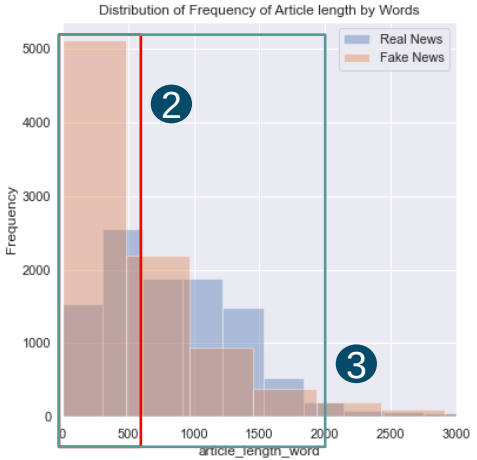
02

Exploratory Data Analysis

Length of News Articles By Words

1. There are a few more fake news articles having > 10,000 words.
2. However, based on distribution, a lot more fake news articles have 0-500 words compared to real news.
3. Most of real/fake news articles have length of < 2,000 words.



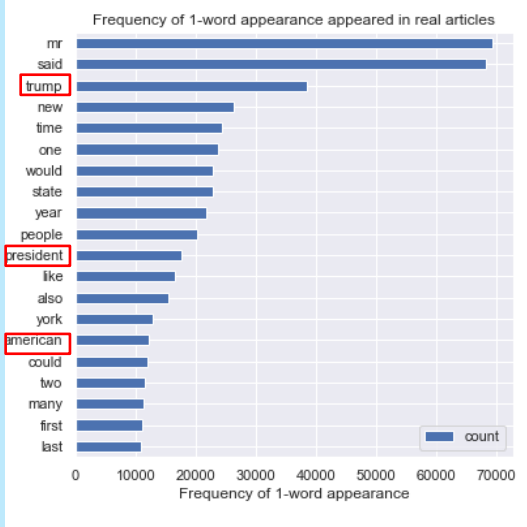


Frequently Appeared 1-Word

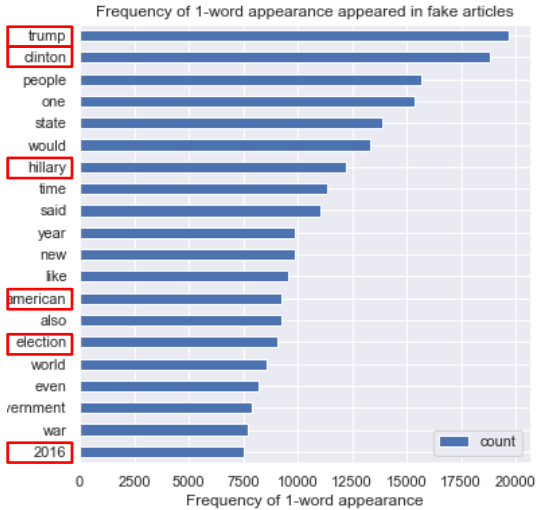
Using N-Gram, below are top-20 1-word that appears frequently in both real and fake news articles.

1. Not much insight could be gathered here except that the words appeared in both shown that these articles are very related to 2016 US Presidential Election with following words:

Real Articles



Fake Articles



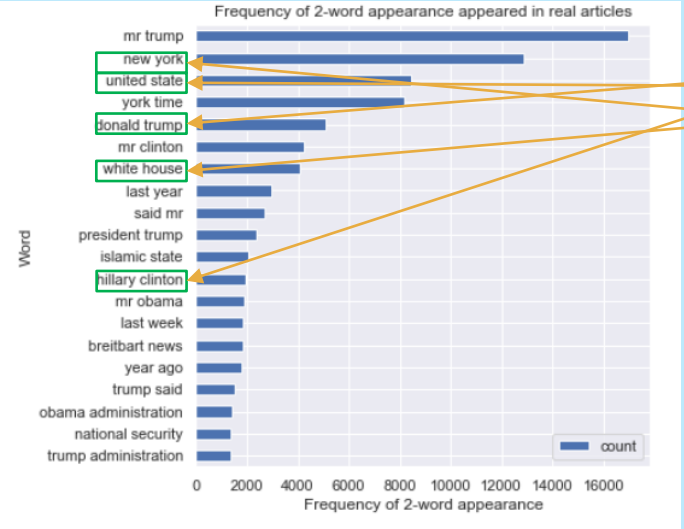
Frequently Appeared 2-Words

Using N-Gram, below are top-20 2-words that appear frequently in both real and fake news articles.

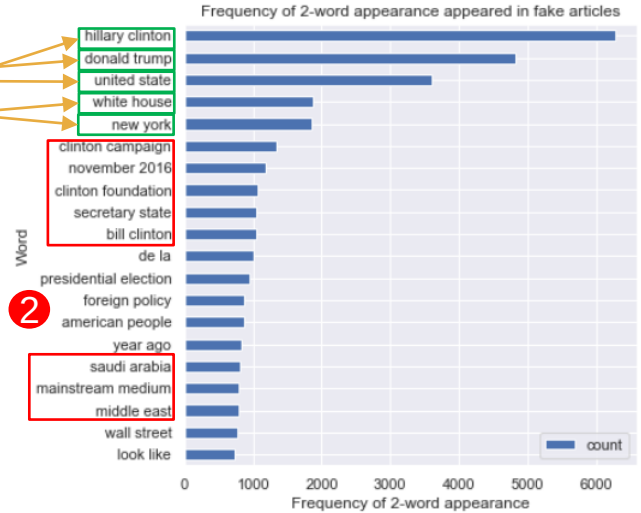
1. A lot of the 2-words combination appear in both the real and fake news articles. Hence, unable to use the word frequency to properly classify fake or real news.

2. However, can observe that a lot of fake articles are targeting on Hillary Clinton or Middle East foreign policy.

Real Articles



Fake Articles



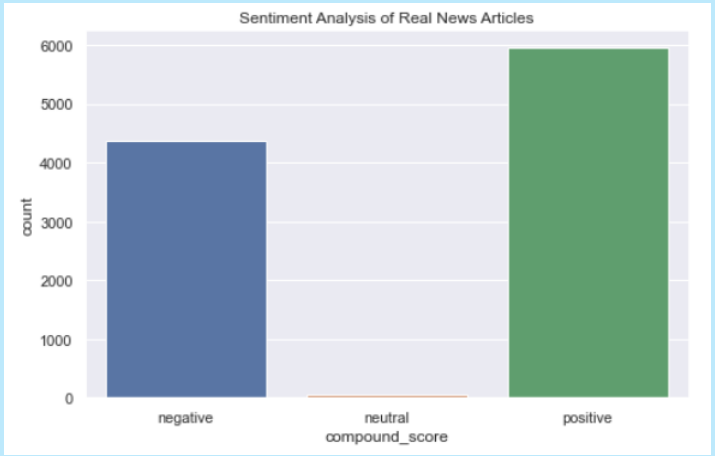
Sentiment Analysis of News Articles

Using Sentiment Analysis Compound Score, below are sentiment of both real and fake articles.
Compound Score < (-0.05): Negative Sentiment, -0.05 to 0.05: Neutral, > 0.05: Positive Sentiment

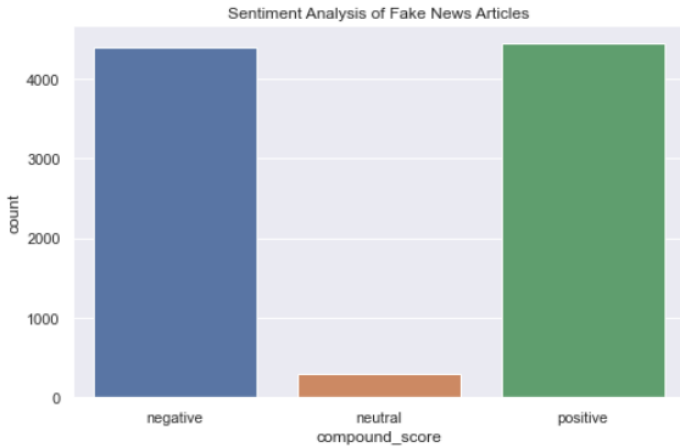
1. Within the real news articles, the proportion of positive to negative is around 60:40, which is more positive than the fake news articles with proportion of around 50:50.

2. However, the sentiment is not too binary (real == +ve, fake == -ve), hence sentiment analysis is not a good way to classify fake/real news.

Real Articles



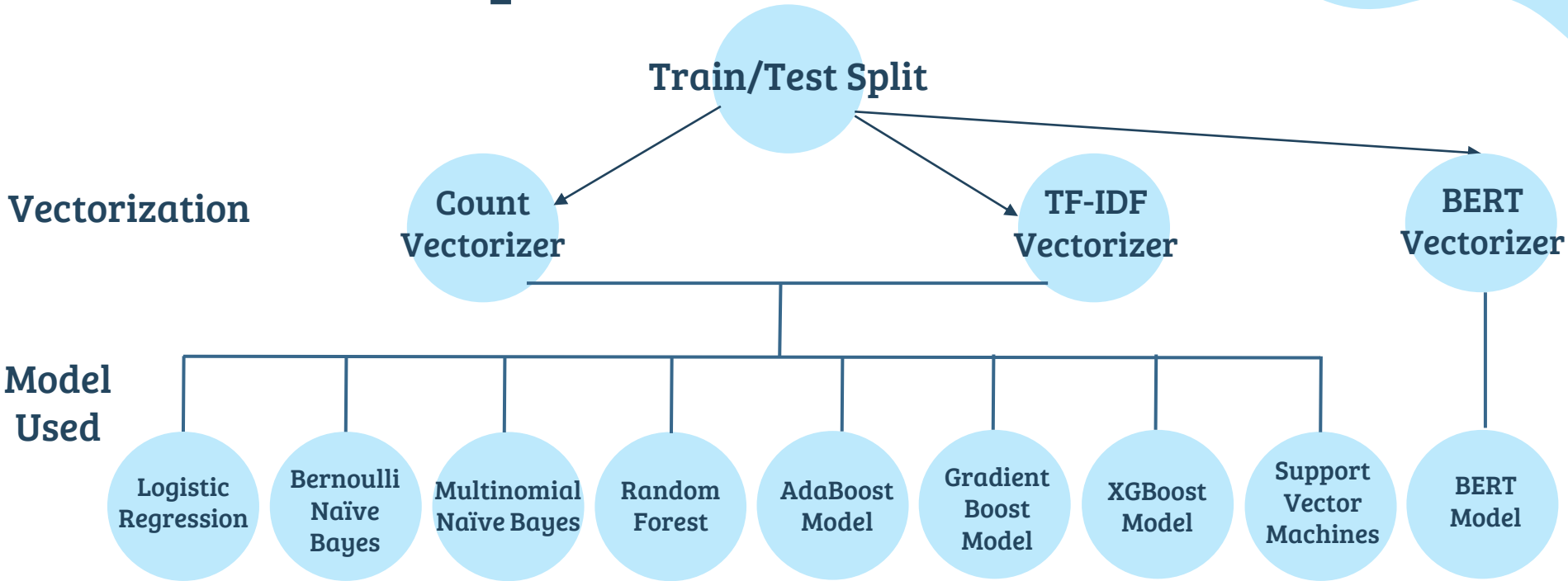
Fake Articles



03

Modelling & Evaluation

Model Preparation



Evaluation Metrics

Test Accuracy

Indicator of how accurate is the prediction

Better to be as close to 100% as possible

F1 Score

Balance perspective on precision-recall performance

Better to be as close to 100% as possible

AUC Score

Indicator of performance at differentiating the +ve and –ve classes

Better to be as close to 1 as possible

Train – Test Accuracy

Indicator of overfitting or underfitting

Better to be as close to 0% as possible

Model Evaluation

1. Test Accuracy

All models using various vectorization have test accuracy higher than baseline, better than baseline.

Classification Model	Vectorization	Test Accuracy	F1 Score	AUC Score	Train – Test Accuracy
Baseline	-	**53.0%**	-	-	-
Logistic Regression	**Count**	95.0%	96.0%	0.990	1.0%
	TF-IDF	94.0%	96.0%	0.990	1.0%
Bernoulli Naïve Bayes	**Count**	76.0%	77.0%	0.850	0%
	TF-IDF	76.0%	77.0%	0.850	0%
Multinomial Naïve Bayes	**Count**	91.0%	91.0%	0.965	0%
	TF-IDF	90.0%	91.0%	0.975	1.0%
Random Forest	**Count**	93.0%	94.0%	0.989	1.0%
	TF-IDF	91.0%	93.0%	0.984	2.0%

Classification Model	Vectorization	Test Accuracy	F1 Score	AUC Score	Train – Test Accuracy
AdaBoost Model	**Count**	95.0%	96.0%	0.992	1.0%
	TF-IDF	95.0%	95.0%	0.991	1.0%
Gradient Boosting	**Count**	96.0%	96.0%	0.994	0%
	TF-IDF	95.0%	96.0%	0.993	0%
XGBoost Model	**Count**	97.0%	98.0%	0.998	0%
	TF-IDF	96.0%	97.0%	0.997	1.0%
Support Vector Machine	**Count**	82.0%	69.0%	0.868	1.0%
	TF-IDF	85.0%	86.0%	0.933	2.0%
BERT Model	**BERT**	95.0%	95.0%	0.948	1.0%

Model Evaluation

1. Test Accuracy

XGBoost Model using Count Vectorization have the highest test accuracy at 97%.

Classification Model	Vectorization	Test Accuracy	F1 Score	AUC Score	Train – Test Accuracy
Baseline	-	53.0%	-	-	-
Logistic Regression	Count	95.0%	96.0%	0.990	1.0%
	TF-IDF	94.0%	96.0%	0.990	1.0%
Bernoulli Naïve Bayes	Count	76.0%	77.0%	0.850	0%
	TF-IDF	76.0%	77.0%	0.850	0%
Multinomial Naïve Bayes	Count	91.0%	91.0%	0.965	0%
	TF-IDF	90.0%	91.0%	0.975	1.0%
Random Forest	Count	93.0%	94.0%	0.989	1.0%
	TF-IDF	91.0%	93.0%	0.984	2.0%
AdaBoost Model	Count	95.0%	96.0%	0.992	1.0%
	TF-IDF	95.0%	95.0%	0.991	1.0%
Gradient Boosting	Count	96.0%	96.0%	0.994	0%
	TF-IDF	95.0%	96.0%	0.993	0%
XGBoost Model	Count	97.0%	98.0%	0.998	0%
	TF-IDF	96.0%	97.0%	0.997	1.0%
Support Vector Machine	Count	82.0%	69.0%	0.868	1.0%
	TF-IDF	85.0%	86.0%	0.933	2.0%
BERT Model	BERT	95.0%	95.0%	0.948	1.0%

Model Evaluation

2. F1 Score

XGBoost Model using Count Vectorization have the highest F1 score at 98%.

Classification Model	Vectorization	Test Accuracy	F1 Score	AUC Score	Train – Test Accuracy
Baseline	-	53.0%	-	-	-
Logistic Regression	Count	95.0%	96.0%	0.990	1.0%
	TF-IDF	94.0%	96.0%	0.990	1.0%
Bernoulli Naïve Bayes	Count	76.0%	77.0%	0.850	0%
	TF-IDF	76.0%	77.0%	0.850	0%
Multinomial Naïve Bayes	Count	91.0%	91.0%	0.965	0%
	TF-IDF	90.0%	91.0%	0.975	1.0%
Random Forest	Count	93.0%	94.0%	0.989	1.0%
	TF-IDF	91.0%	93.0%	0.984	2.0%

Classification Model	Vectorization	Test Accuracy	F1 Score	AUC Score	Train – Test Accuracy
AdaBoost Model	Count	95.0%	96.0%	0.992	1.0%
	TF-IDF	95.0%	95.0%	0.991	1.0%
Gradient Boosting	Count	96.0%	96.0%	0.994	0%
	TF-IDF	95.0%	96.0%	0.993	0%
XGBoost Model	Count	97.0%	98.0%	0.998	0%
	TF-IDF	96.0%	97.0%	0.997	1.0%
Support Vector Machine	Count	82.0%	69.0%	0.868	1.0%
	TF-IDF	85.0%	86.0%	0.933	2.0%
BERT Model	BERT	95.0%	95.0%	0.948	1.0%

Model Evaluation

3. AUC Score

XGBoost Model using Count Vectorization have the highest AUC score at 0.998, very close to 1.

Classification Model	Vectorization	Test Accuracy	F1 Score	AUC Score	Train – Test Accuracy
Baseline	-	53.0%	-	-	-
Logistic Regression	Count	95.0%	96.0%	0.990	1.0%
	TF-IDF	94.0%	96.0%	0.990	1.0%
Bernoulli Naïve Bayes	Count	76.0%	77.0%	0.850	0%
	TF-IDF	76.0%	77.0%	0.850	0%
Multinomial Naïve Bayes	Count	91.0%	91.0%	0.965	0%
	TF-IDF	90.0%	91.0%	0.975	1.0%
Random Forest	Count	93.0%	94.0%	0.989	1.0%
	TF-IDF	91.0%	93.0%	0.984	2.0%

Classification Model	Vectorization	Test Accuracy	F1 Score	AUC Score	Train – Test Accuracy
AdaBoost Model	Count	95.0%	96.0%	0.992	1.0%
	TF-IDF	95.0%	95.0%	0.991	1.0%
Gradient Boosting	Count	96.0%	96.0%	0.994	0%
	TF-IDF	95.0%	96.0%	0.993	0%
XGBoost Model	Count	97.0%	98.0%	0.998	0%
	TF-IDF	96.0%	97.0%	0.997	1.0%
Support Vector Machine	Count	82.0%	69.0%	0.868	1.0%
	TF-IDF	85.0%	86.0%	0.933	2.0%
BERT Model	BERT	95.0%	95.0%	0.948	1.0%

Model Evaluation

4. Train – Test Accuracy

A lot of models using various vectorizations have train – test accuracy = 0, no overfitting or underfitting.

Classification Model	Vectorization	Test Accuracy	F1 Score	AUC Score	Train – Test Accuracy
Baseline	-	53.0%	-	-	-
Logistic Regression	Count	95.0%	96.0%	0.990	1.0%
	TF-IDF	94.0%	96.0%	0.990	1.0%
Bernoulli Naïve Bayes	Count	76.0%	77.0%	0.850	0%
	TF-IDF	76.0%	77.0%	0.850	0%
Multinomial Naïve Bayes	Count	91.0%	91.0%	0.965	0%
	TF-IDF	90.0%	91.0%	0.975	1.0%
Random Forest	Count	93.0%	94.0%	0.989	1.0%
	TF-IDF	91.0%	93.0%	0.984	2.0%

Classification Model	Vectorization	Test Accuracy	F1 Score	AUC Score	Train – Test Accuracy
AdaBoost Model	Count	95.0%	96.0%	0.992	1.0%
	TF-IDF	95.0%	95.0%	0.991	1.0%
Gradient Boosting	Count	96.0%	96.0%	0.994	0%
	TF-IDF	95.0%	96.0%	0.993	0%
XGBoost Model	Count	97.0%	98.0%	0.998	0%
	TF-IDF	96.0%	97.0%	0.997	1.0%
Support Vector Machine	Count	82.0%	69.0%	0.868	1.0%
	TF-IDF	85.0%	86.0%	0.933	2.0%
BERT Model	BERT	95.0%	95.0%	0.948	1.0%

Model Evaluation

Overall Evaluation

XG Boost Model using Count Vectorizer have the best performance.

Classification Model	Vectorization	Test Accuracy	F1 Score	AUC Score	Train – Test Accuracy
Baseline	-	53.0%	-	-	-
Logistic Regression	Count	95.0%	96.0%	0.990	1.0%
	TF-IDF	94.0%	96.0%	0.990	1.0%
Bernoulli Naïve Bayes	Count	76.0%	77.0%	0.850	0%
	TF-IDF	76.0%	77.0%	0.850	0%
Multinomial Naïve Bayes	Count	91.0%	91.0%	0.965	0%
	TF-IDF	90.0%	91.0%	0.975	1.0%
Random Forest	Count	93.0%	94.0%	0.989	1.0%
	TF-IDF	91.0%	93.0%	0.984	2.0%

Classification Model	Vectorization	Test Accuracy	F1 Score	AUC Score	Train – Test Accuracy
AdaBoost Model	Count	95.0%	96.0%	0.992	1.0%
	TF-IDF	95.0%	95.0%	0.991	1.0%
Gradient Boosting	Count	96.0%	96.0%	0.994	0%
	TF-IDF	95.0%	96.0%	0.993	0%
XGBoost Model	Count	97.0%	98.0%	0.998	0%
	TF-IDF	96.0%	97.0%	0.997	1.0%
Support Vector Machine	Count	82.0%	69.0%	0.868	1.0%
	TF-IDF	85.0%	86.0%	0.933	2.0%
BERT Model	BERT	95.0%	95.0%	0.948	1.0%

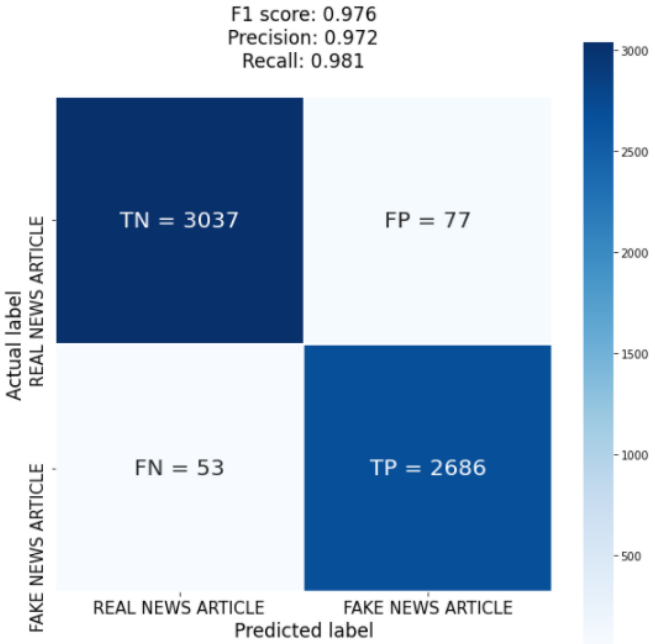
04

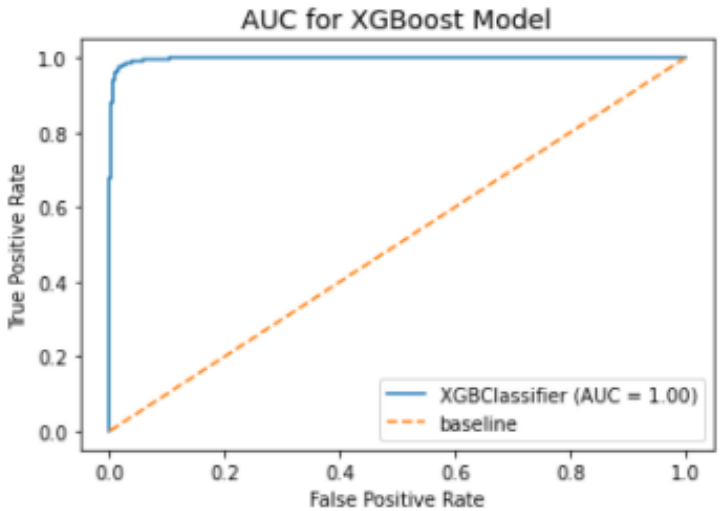
Conclusion & Recommendations

Conclusion

XG Boost Model using Count Vectorizer have the best performance.

- Accurate classification ~ 97% of the time; 5,723 out of 5,853 predictions
- AUC score ~ 1%

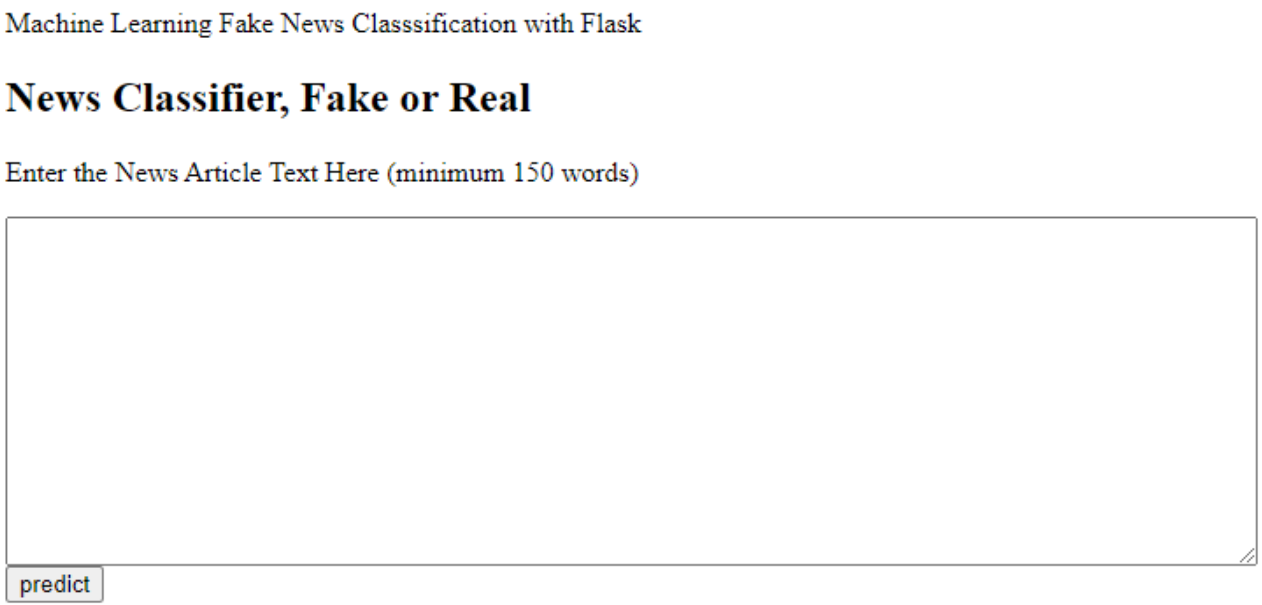




Model Demonstration

A simple flask app is created to demonstrate on how US citizens can use the model to differentiate between real or fake news, all by themselves.

Flask App Link



Recommendations

01

Deploy the model for use as a self fact-check system in testing with actual articles

Given the strong performance of model (97% accuracy score, 98% F1 score, ~1 AUC Score), US citizens can benefit from deploying the model for purpose of self fact-check of political news.

02

Retrain the model periodically with new political news articles.

As time passes, there would be a need to retrain the model with new and more relevant political news articles. This is to continuing to ensure that the model can continually maintained a high performance.

03

Explore the expansion of the model to cover other news area.

Currently, the model only includes news articles from politics. In the near future, it would be better if the model can expand to include fact-check of other areas: healthcare, education, sports, etc.

Thank You!

Do you have any questions?