Fake News Detector

Team Techniators

Caesar Wong

Sarah Abdelazim

Lisa Sequeira

Mentor: Esthela Gallardo

Table of Contents

Impact and Social Threat

Project Steps

Use of Technology (AWS)

Conclusions and Demo

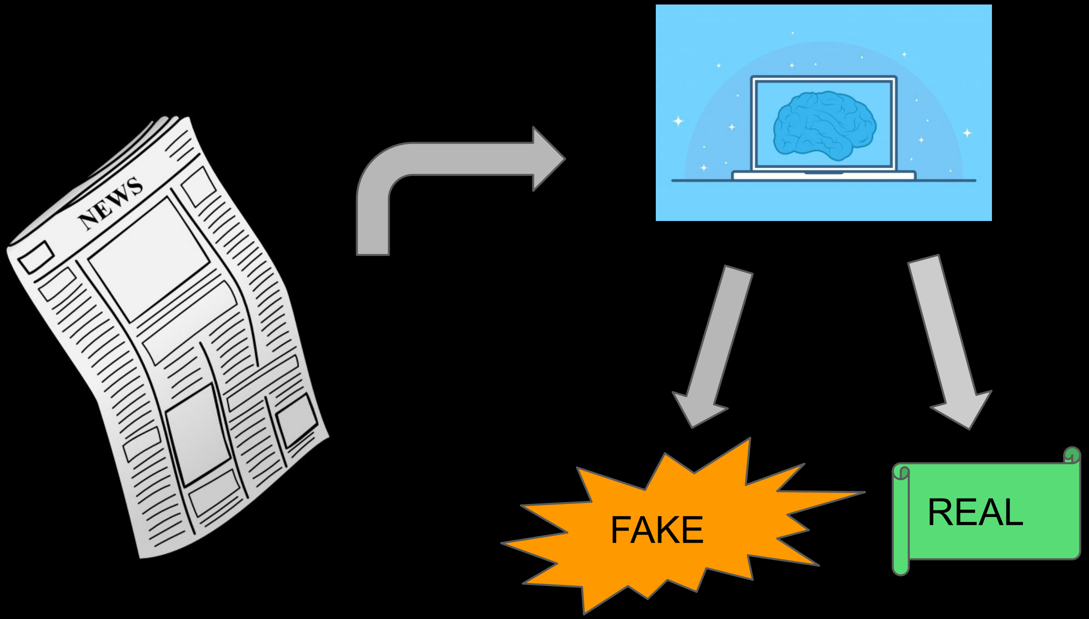
Future Work

Social Threat and Project Impact

Need a tool to build trust in media and open dialogue between conflicting viewpoints.

Our Solution: Fake News Detector

Create an app that makes fake news detection easier, more streamlined and consistent.



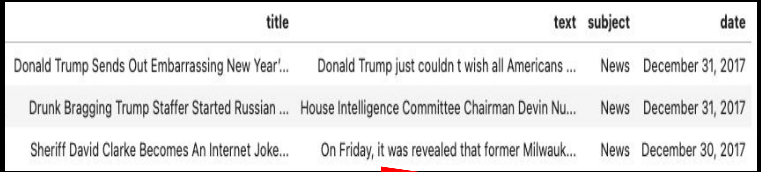
Fake News Detector – Project Plan

Our Steps:

1. Find the fake & real news dataset
2. Data cleaning
3. EDA
4. Feature Engineering
5. ML model training
6. App Development & Deployment

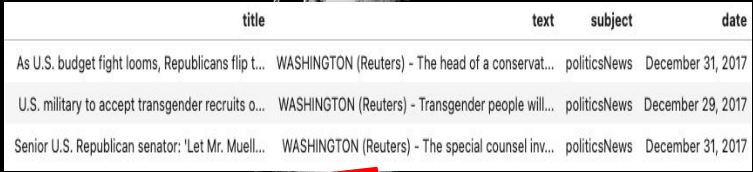
Data Cleaning

Fake



title	text	subject	date
Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017

Real



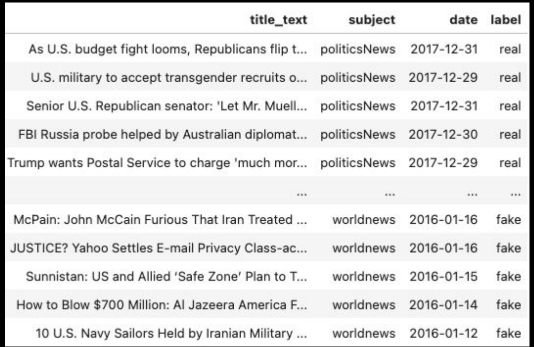
title	text	subject	date
As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017

Problem 1: Real df has extra characters in text column e.g. WASHINGTON (Reuters)

Label "real" or "fake" then concatenate, drop empty text rows, combine text and title columns

Problem 2: Multiple date formats, can't parse! e.g. December 31, 2017, 18-Feb-2018, Dec 12, 2016.

Problem 3: Multiple, infrequent subject categories e.g. News, PoliticNews, Middle-east, left-wing, etc...



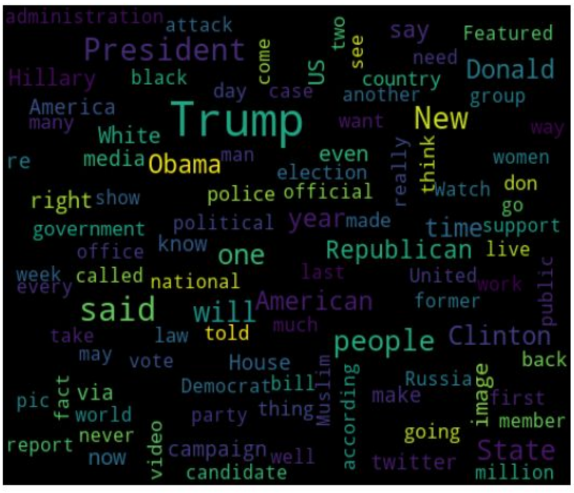
title_text	subject	date	label
As U.S. budget fight looms, Republicans flip t...	politicsNews	2017-12-31	real
U.S. military to accept transgender recruits o...	politicsNews	2017-12-29	real
Senior U.S. Republican senator: 'Let Mr. Muell...	politicsNews	2017-12-31	real
FBI Russia probe helped by Australian diplomat...	politicsNews	2017-12-30	real
Trump wants Postal Service to charge 'much mor...	politicsNews	2017-12-29	real
...	...	...	...
McPain: John McCain Furious That Iran Treated ...	worldnews	2016-01-16	fake
JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	worldnews	2016-01-16	fake
Sunnistan: US and Allied 'Safe Zone' Plan to T...	worldnews	2016-01-15	fake
How to Blow $700 Million: Al Jazeera America F...	worldnews	2016-01-14	fake
10 U.S. Navy Sailors Held by Iranian Military ...	worldnews	2016-01-12	fake

Processed dataframe (44267 x 4)

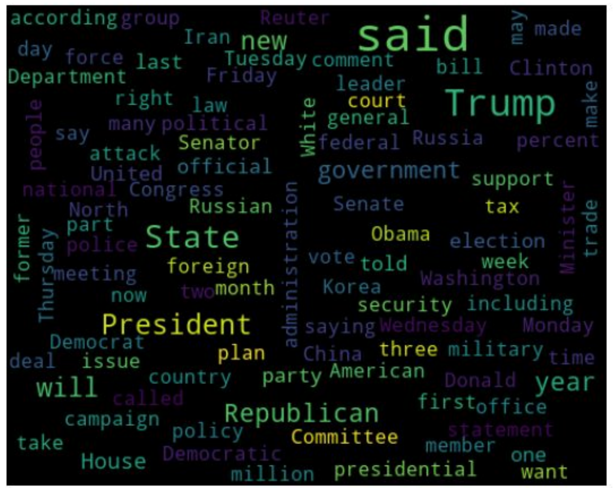
Exploratory Data Analysis

EDA Revealed limitations of Dataset

1. American & Politically focused news - limits diversity of news topic



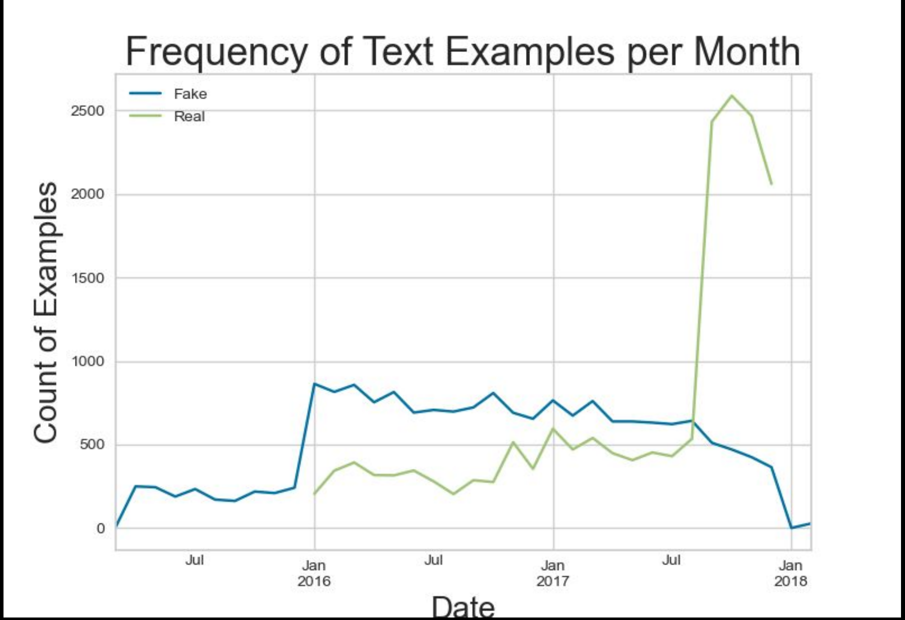
FAKE



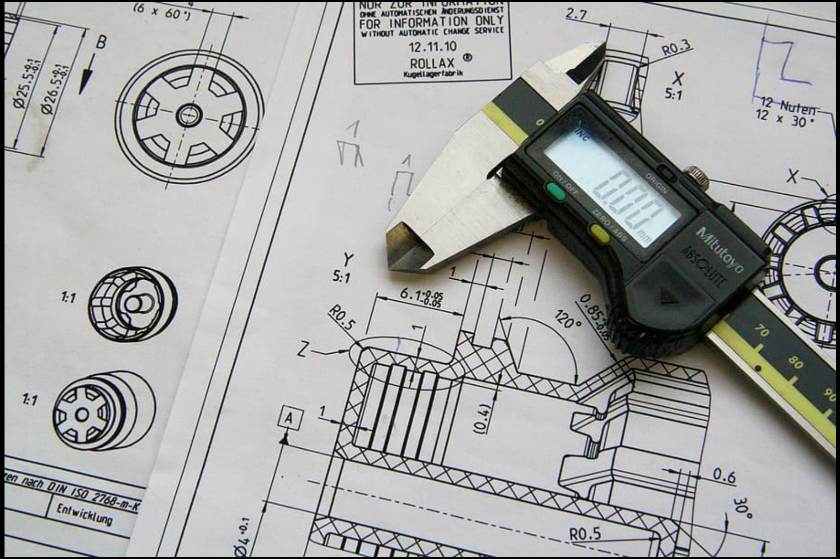
REAL

EDA Revealed limitations of Dataset

2. Data had a bias when reviewing dates

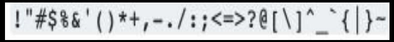


Feature Engineering



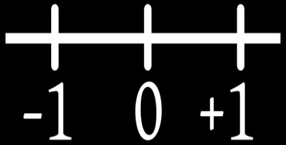
Three New Features

Special characters



`!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~`

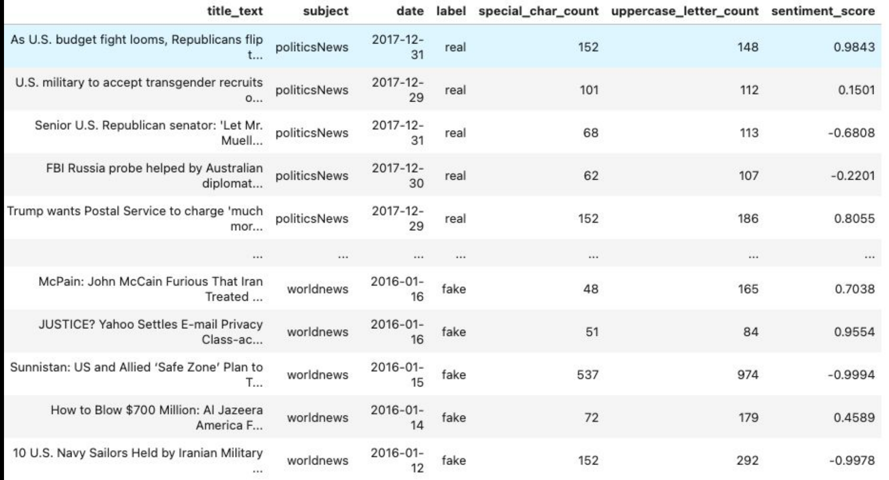
Sentiment Score



Uppercase letters

ABC DEF GHI
JKL MNO PQR
STU VWX YZ

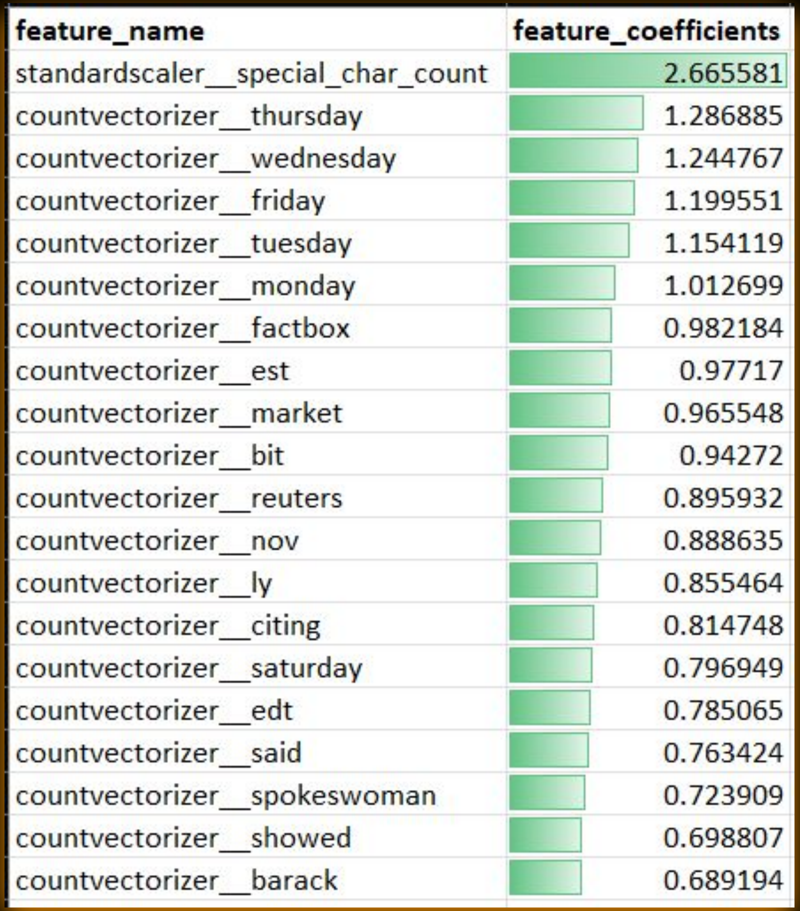
Final dataframe
(44267 x 7)



title_text	subject	date	label	special_char_count	uppercase_letter_count	sentiment_score
As U.S. budget fight looms, Republicans flip t...	politicsNews	2017-12-31	real	152	148	0.9843
U.S. military to accept transgender recruits o...	politicsNews	2017-12-29	real	101	112	0.1501
Senior U.S. Republican senator: 'Let Mr. Muell...	politicsNews	2017-12-31	real	68	113	-0.6808
FBI Russia probe helped by Australian diplomat...	politicsNews	2017-12-30	real	62	107	-0.2201
Trump wants Postal Service to charge 'much mor...	politicsNews	2017-12-29	real	152	186	0.8055
...	...	...	...	...	...	...
McPain: John McCain Furious That Iran Treated ...	worldnews	2016-01-16	fake	48	165	0.7038
JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	worldnews	2016-01-16	fake	51	84	0.9554
Sunnistan: US and Allied 'Safe Zone' Plan to T...	worldnews	2016-01-15	fake	537	974	-0.9994
How to Blow $700 Million: Al Jazeera America F...	worldnews	2016-01-14	fake	72	179	0.4589
10 U.S. Navy Sailors Held by Iranian Military ...	worldnews	2016-01-12	fake	152	292	-0.9978

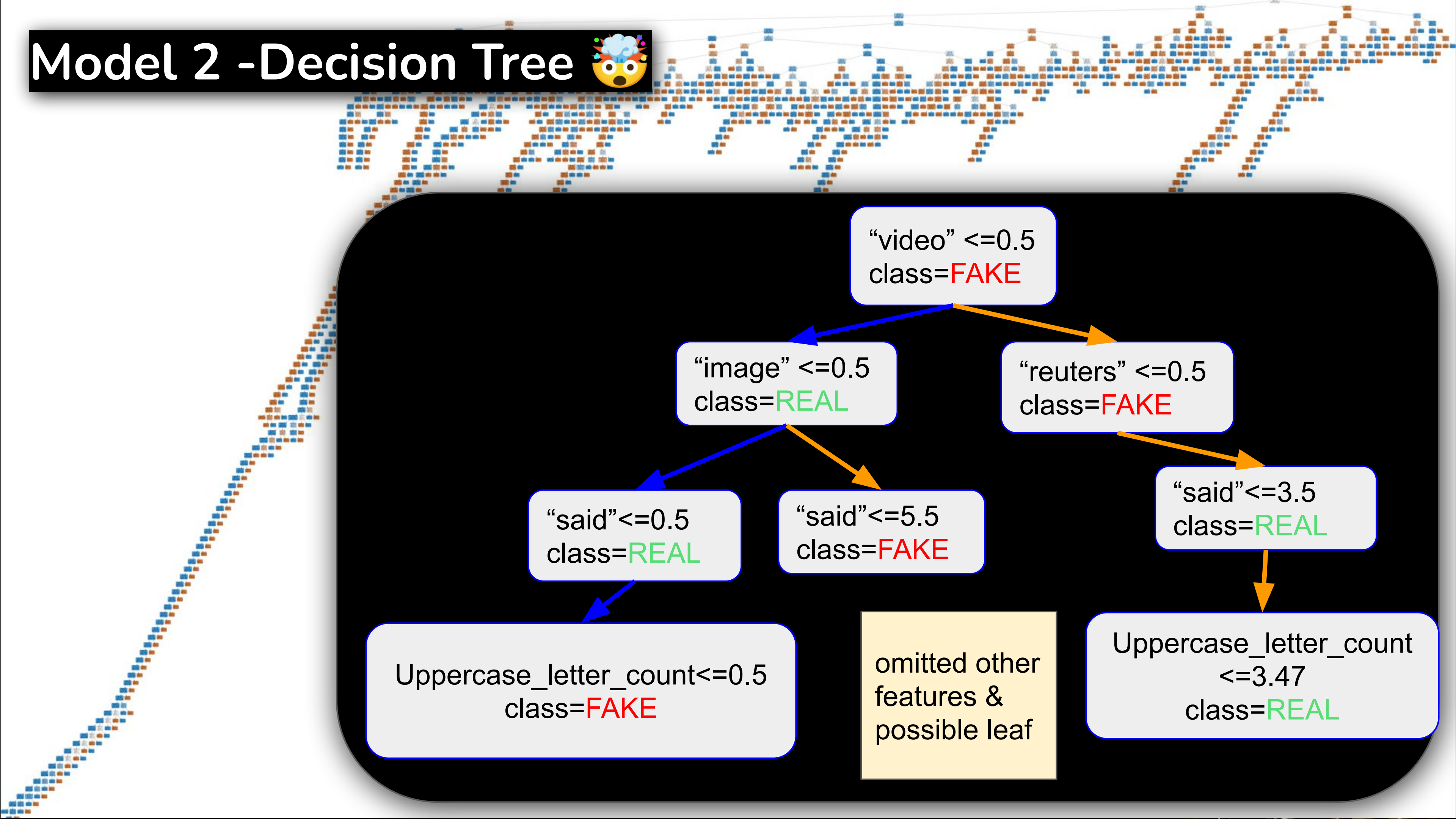
Model Building and Evaluation

Model 1 - Regression Model - Feature Coefficient



feature_name	feature_coefficients
standardscaler__special_char_count	2.665581
countvectorizer__thursday	1.286885
countvectorizer__wednesday	1.244767
countvectorizer__friday	1.199551
countvectorizer__tuesday	1.154119
countvectorizer__monday	1.012699
countvectorizer__factbox	0.982184
countvectorizer__est	0.97717
countvectorizer__market	0.965548
countvectorizer__bit	0.94272
countvectorizer__reuters	0.895932
countvectorizer__nov	0.888635
countvectorizer__ly	0.855464
countvectorizer__citing	0.814748
countvectorizer__saturday	0.796949
countvectorizer__edt	0.785065
countvectorizer__said	0.763424
countvectorizer__spokeswoman	0.723909
countvectorizer__showed	0.698807
countvectorizer__barack	0.689194

Model 2 -Decision Tree 🤯



"video" <=0.5
class=FAKE

"image" <=0.5
class=REAL

"reuters" <=0.5
class=FAKE

"said"<=0.5
class=REAL

"said"<=5.5
class=FAKE

"said"<=3.5
class=REAL

Uppercase_letter_count<=0.5
class=FAKE

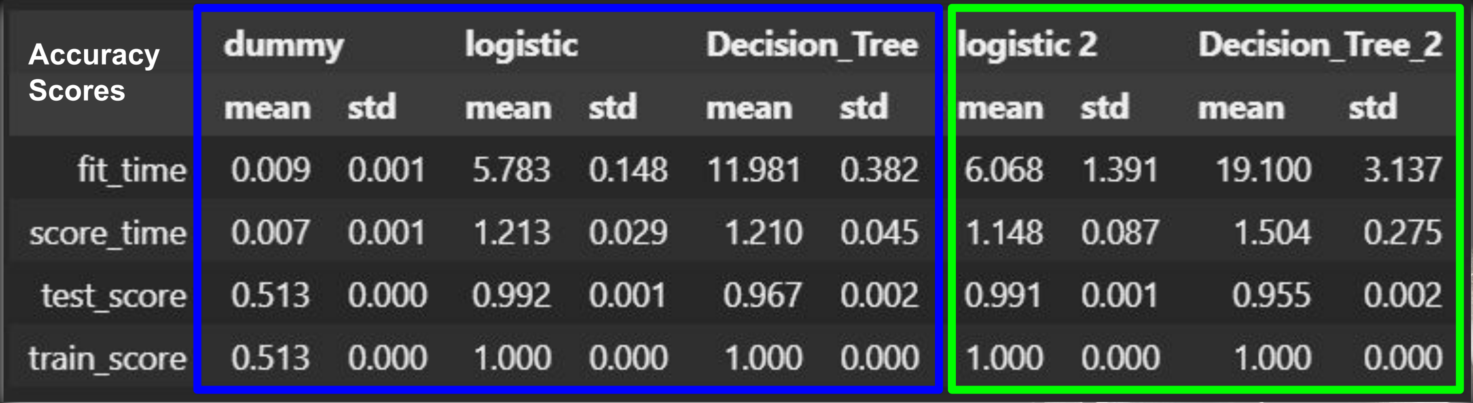
omitted other features & possible leaf

Uppercase_letter_count
<=3.47
class=REAL

Model Performance

With ‘subject’

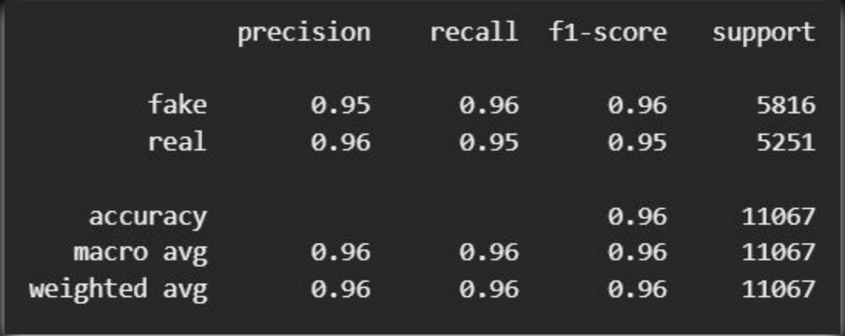
Without ‘subject’



Accuracy Scores	dummy		logistic		Decision_Tree		logistic 2		Decision_Tree_2	
	mean	std	mean	std	mean	std	mean	std	mean	std
fit_time	0.009	0.001	5.783	0.148	11.981	0.382	6.068	1.391	19.100	3.137
score_time	0.007	0.001	1.213	0.029	1.210	0.045	1.148	0.087	1.504	0.275
test_score	0.513	0.000	0.992	0.001	0.967	0.002	0.991	0.001	0.955	0.002
train_score	0.513	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000

Decision Tree Classification Report (Unseen Data)

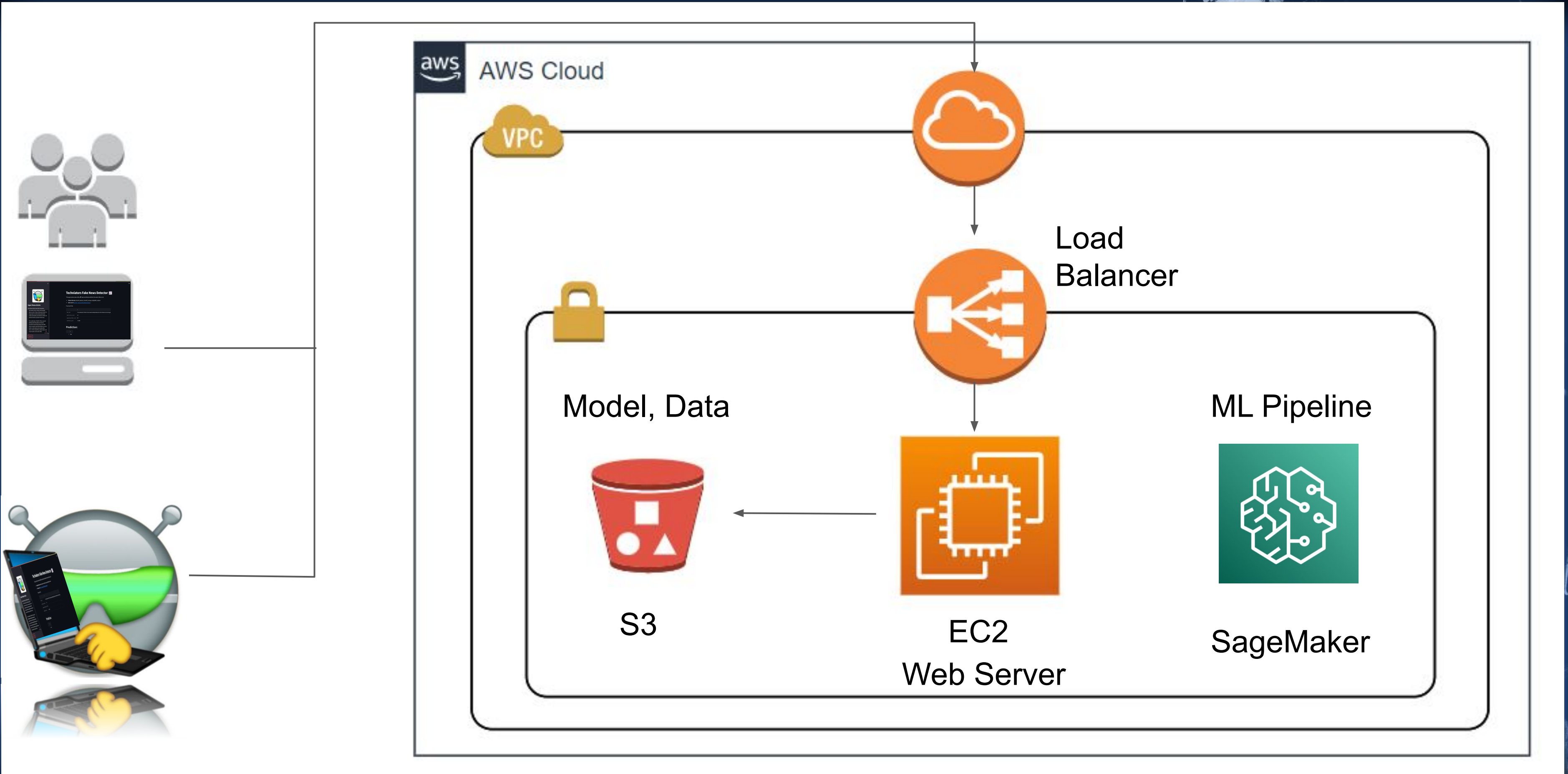
Without ‘subject’



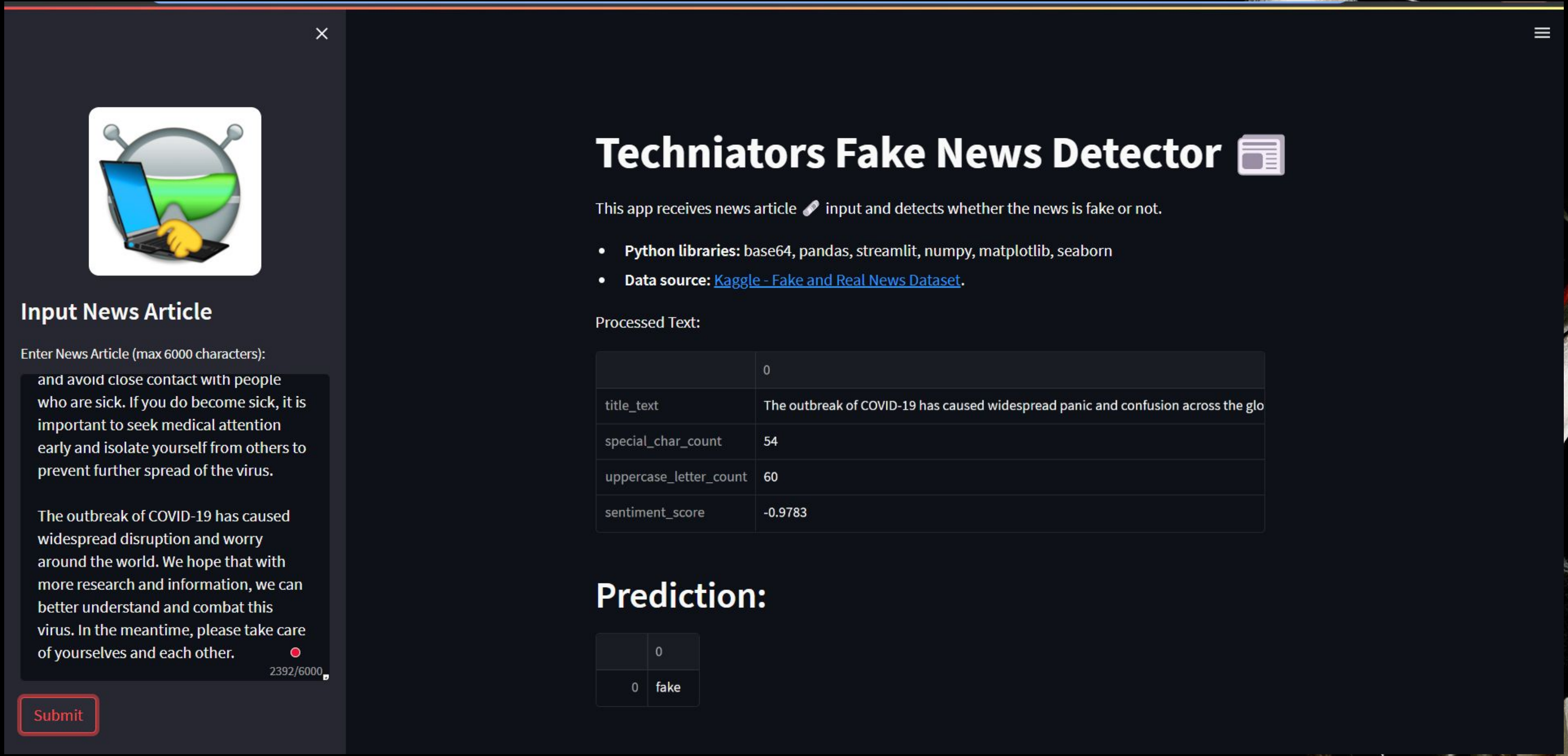
```
              precision    recall  f1-score   support

        fake       0.95      0.96      0.96      5816
        real       0.96      0.95      0.95      5251

    accuracy                           0.96     11067
   macro avg       0.96      0.96      0.96     11067
weighted avg       0.96      0.96      0.96     11067
```

Application and Demo

Architecture



Demo ([Local](#) / [Cloud](#))



Input News Article

Enter News Article (max 6000 characters):

and avoid close contact with people who are sick. If you do become sick, it is important to seek medical attention early and isolate yourself from others to prevent further spread of the virus.

The outbreak of COVID-19 has caused widespread disruption and worry around the world. We hope that with more research and information, we can better understand and combat this virus. In the meantime, please take care of yourselves and each other.

2392/6000

Techniators Fake News Detector 📰

This app receives news article 🩹 input and detects whether the news is fake or not.

- **Python libraries:** base64, pandas, streamlit, numpy, matplotlib, seaborn
- **Data source:** [Kaggle - Fake and Real News Dataset](#).

Processed Text:

	0
title_text	The outbreak of COVID-19 has caused widespread panic and confusion across the glo
special_char_count	54
uppercase_letter_count	60
sentiment_score	-0.9783

Prediction:

	0
0	fake

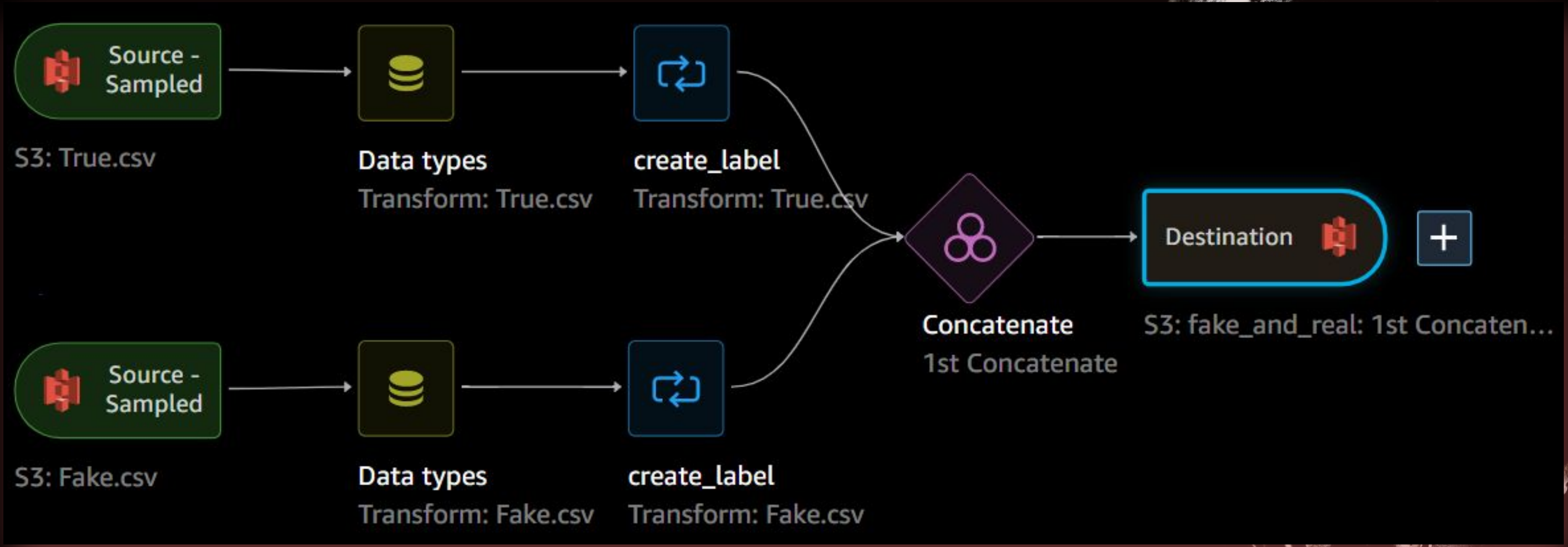
Future Work 💭🛠️🦾

- Use AI generated News dataset train the model and distinguish between AI and human written news
- Do additional feature engineering to improve model performance
- Automate ML pipeline to allow client upload a set of labelled fake and real news
- Adopt more ML models to compare the accuracy and performance
- Use RandomizedSearchCV to obtain the best parameters for our ML model

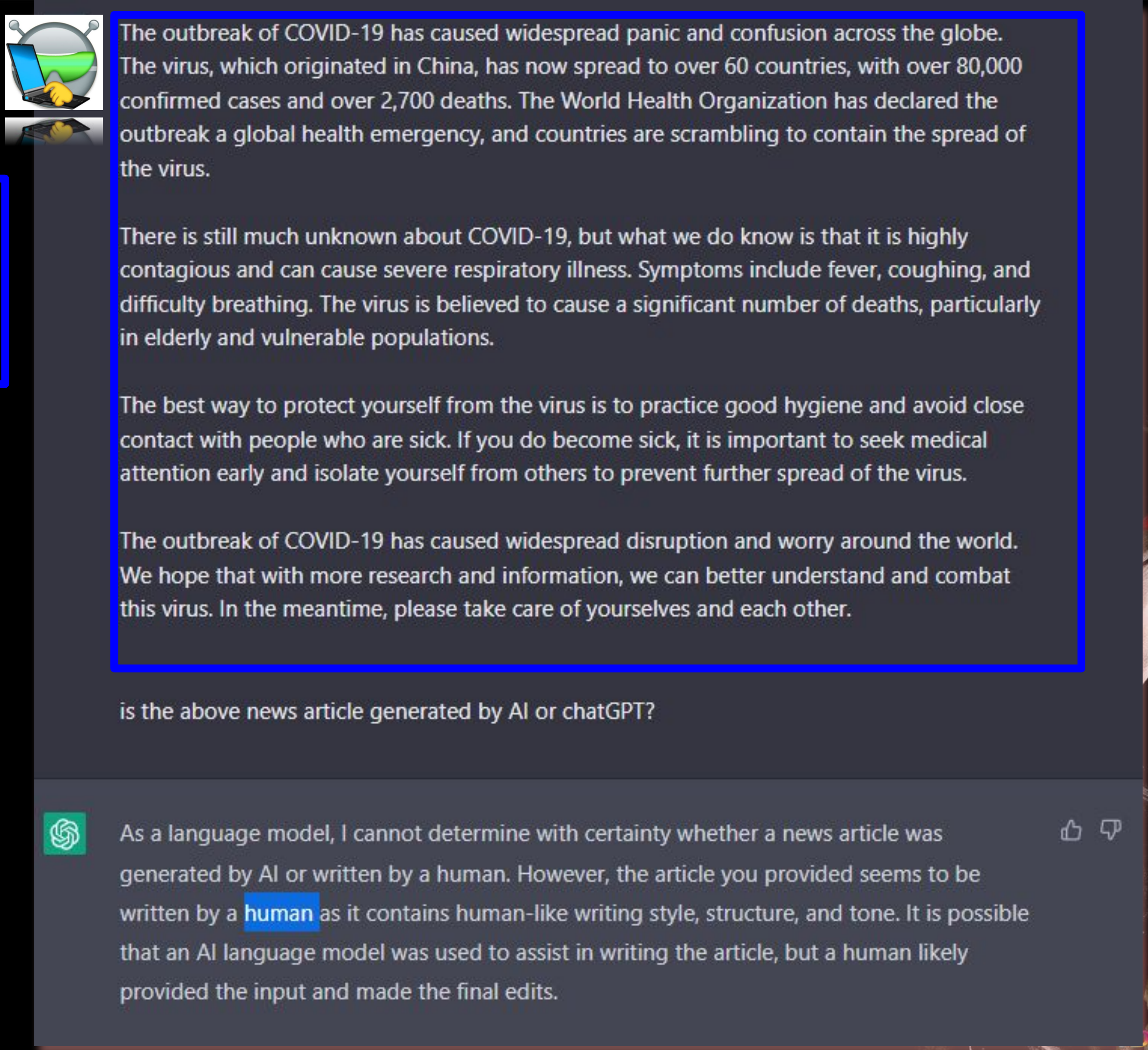
Thank you for your attention!

Appendix

Project Plan visual



AI generated news (previously collected)



The outbreak of COVID-19 has caused widespread panic and confusion across the globe. The virus, which originated in China, has now spread to over 60 countries, with over 80,000 confirmed cases and over 2,700 deaths. The World Health Organization has declared the outbreak a global health emergency, and countries are scrambling to contain the spread of the virus.

There is still much unknown about COVID-19, but what we do know is that it is highly contagious and can cause severe respiratory illness. Symptoms include fever, coughing, and difficulty breathing. The virus is believed to cause a significant number of deaths, particularly in elderly and vulnerable populations.

The best way to protect yourself from the virus is to practice good hygiene and avoid close contact with people who are sick. If you do become sick, it is important to seek medical attention early and isolate yourself from others to prevent further spread of the virus.

The outbreak of COVID-19 has caused widespread disruption and worry around the world. We hope that with more research and information, we can better understand and combat this virus. In the meantime, please take care of yourselves and each other.

is the above news article generated by AI or chatGPT?

As a language model, I cannot determine with certainty whether a news article was generated by AI or written by a human. However, the article you provided seems to be written by a human as it contains human-like writing style, structure, and tone. It is possible that an AI language model was used to assist in writing the article, but a human likely provided the input and made the final edits.