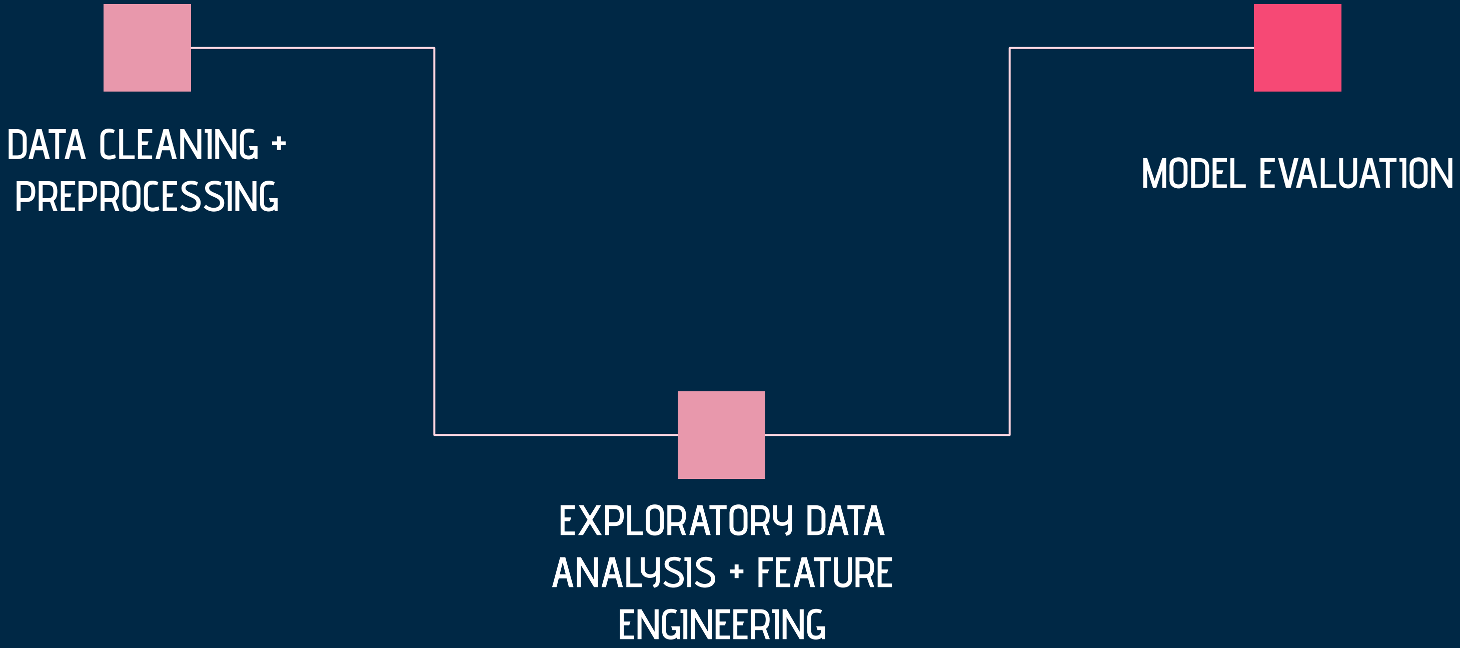
Fake News Project

By: Michael Murphy and Rohan Kathuria

STEPS TO BUILD CLASSIFIER



DATA CLEANING + PREPROCESSING

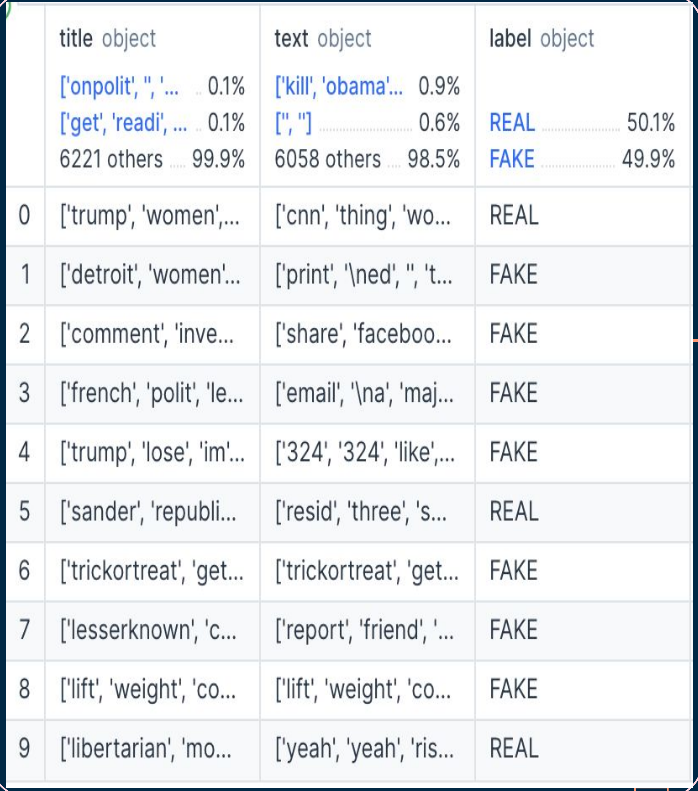
EXPLORATORY DATA ANALYSIS + FEATURE ENGINEERING

MODEL EVALUATION

Data Cleaning

- Removal of duplicates
- Removal of missing data (we chose to drop)
- Removal of punctuation
- Tokenization:
  strings → word lists
- Removal of stop words:
  “and”,”the”,”this”,”of”
- Stemming:
  “stemming” → “stem”

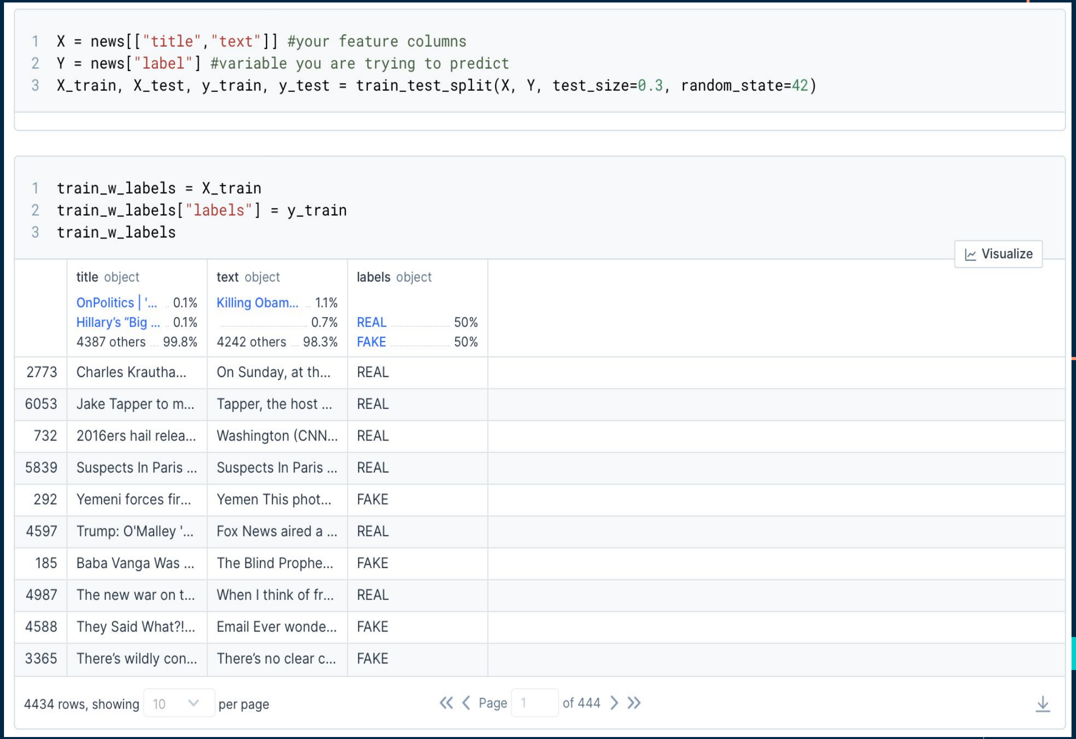
Final Data Frame:



	title object	text object	label object
	['onpolit', '', '... 0.1% ['get', 'readi', ... 0.1% 6221 others 99.9%	['kill', 'obama'... 0.9% ['', ''] 0.6% 6058 others 98.5%	REAL 50.1% FAKE 49.9%
0	['trump', 'women',...	['cnn', 'thing', 'wo...	REAL
1	['detroit', 'women'...	['print', '\ned', '', 't...	FAKE
2	['comment', 'inve...	['share', 'faceboo...	FAKE
3	['french', 'polit', 'le...	['email', '\na', 'maj...	FAKE
4	['trump', 'lose', 'im'...	['324', '324', 'like',...	FAKE
5	['sander', 'republi...	['resid', 'three', 's...	REAL
6	['trickortreat', 'get...	['trickortreat', 'get...	FAKE
7	['lesserknown', 'c...	['report', 'friend', '...	FAKE
8	['lift', 'weight', 'co...	['lift', 'weight', 'co...	FAKE
9	['libertarian', 'mo...	['yeah', 'yeah', 'ris...	REAL

Splitting the data

- To work on model development, we split our data into training and testing datasets.
- This is to check the performance of our model on unseen data.



```
X = news[["title","text"]] #your feature columns
Y = news["label"] #variable you are trying to predict
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.3, random_state=42)
```

```
train_w_labels = X_train
train_w_labels["labels"] = y_train
train_w_labels
```

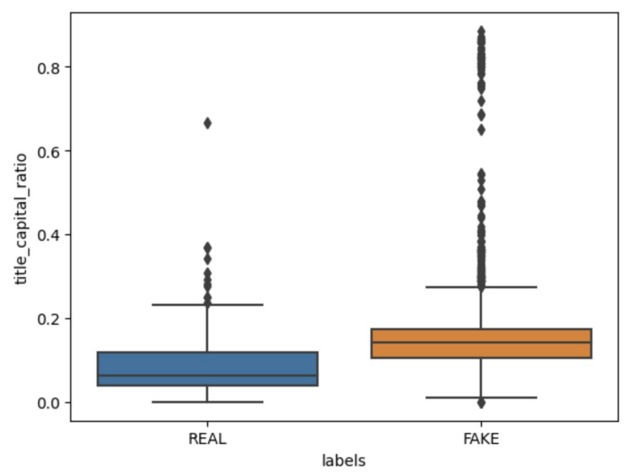
	title object	text object	labels object
	OnPolitics | '... 0.1% Hillary's "Big ... 0.1% 4387 others 99.8%	Killing Obam... 1.1% 0.7% 4242 others 98.3%	REAL 50% FAKE 50%
2773	Charles Krautha...	On Sunday, at th...	REAL
6053	Jake Tapper to m...	Tapper, the host ...	REAL
732	2016ers hail relea...	Washington (CNN...	REAL
5839	Suspects In Paris ...	Suspects In Paris ...	REAL
292	Yemeni forces fir...	Yemen This phot...	FAKE
4597	Trump: O'Malley '...	Fox News aired a ...	REAL
185	Baba Vanga Was ...	The Blind Prophe...	FAKE
4987	The new war on t...	When I think of fr...	REAL
4588	They Said What?!...	Email Ever wonde...	FAKE
3365	There's wildly con...	There's no clear c...	FAKE

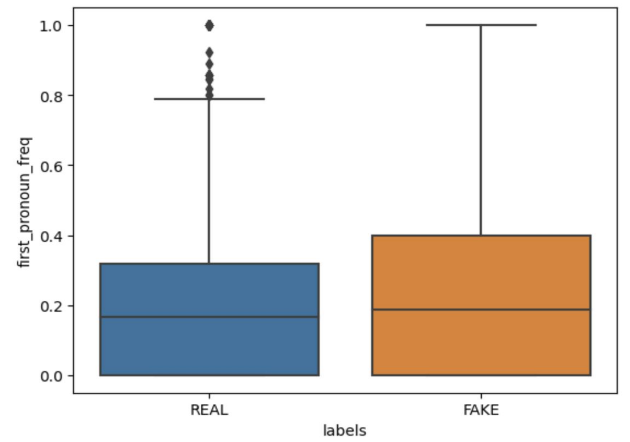
4434 rows, showing 10 per page

EDA + Feature Engineering

- Visualizations: box plots (numerical features), bar plots (word proportions)
- Feature engineering with word map:
  a. Fake articles: “!”, “?”, “best”, “worst”, capitalized words/sentences
  b. Real articles: “according”, “Dr”, “report”, “claim”

Good Feature (capital ratio)

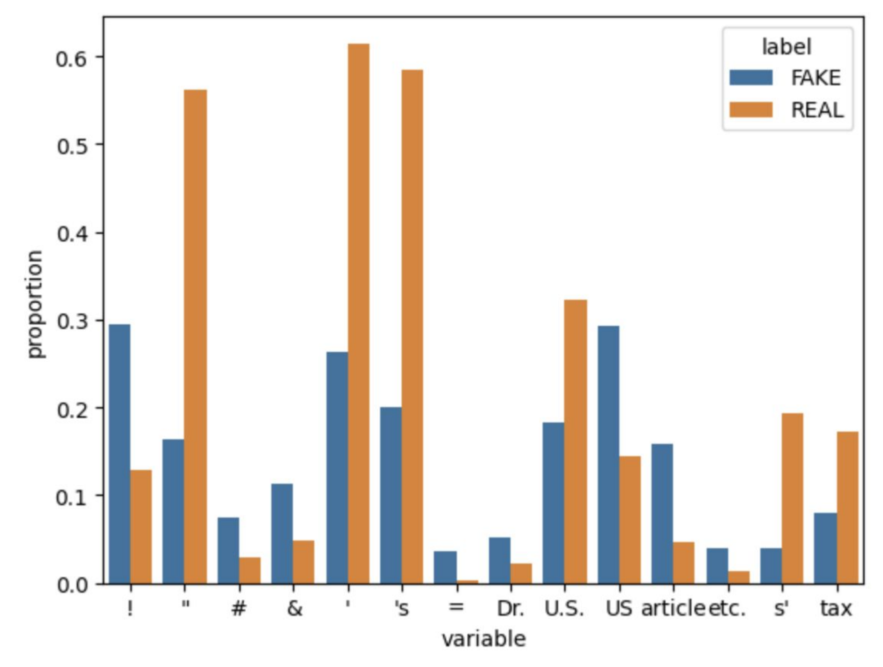




Not as good feature (first person pronoun frequency)

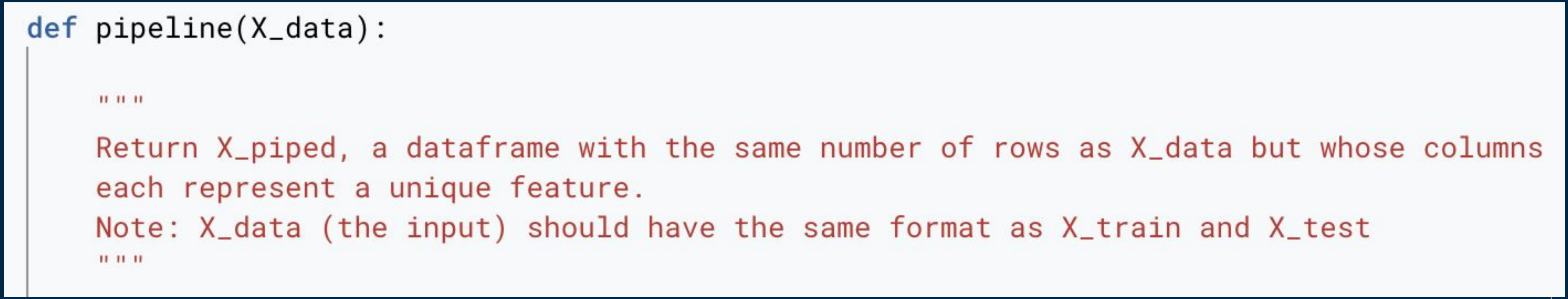
EDA: Word Frequency Bar Plots

- Words w/ bars that are higher with a greater difference between classes = better features



Model Evaluation

- Defining our pipeline function to bring together our earlier developed features
- Our goal is use our specific features to fit a logistic regression model with relatively high accuracy



```
def pipeline(X_data):

    """
    Return X_piped, a dataframe with the same number of rows as X_data but whose columns
    each represent a unique feature.
    Note: X_data (the input) should have the same format as X_train and X_test
    """
```

Accuracy

81.2%

Training Accuracy

80.9%

Testing Accuracy

Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF is a text vectorization technique that takes your text and transforms it into a matrix wherein each word is corresponding to a decimal value that indicates the significance of that particular world. We can then use this matrix to train our logistic regression model to get the following accuracies:

95.3%

Training Accuracy

90.4%

Testing Accuracy

Reflection

Overall

- Initial learning curve
- Importance of the notebooks provided to us
- Understanding the importance of thought behind our code, especially when choosing features

Model specific

- With more time, we could have tried out more features and evaluated their impact on the accuracy of our model
- The TF-IDF model was more accurate than our initial model, which could be because of the vectorization techniques utilized
- Overall, satisfied with our model accuracy

THANKS FOR LISTENING!