Mini Project 3

Can we use Natural Language Processing and Machine Learning Models to Detect ‘Fake’ vs ‘Real’ News?

Gayle Ferguson

The Data

- Fake.csv 23482 rows
- True.csv 21417 rows
- 4 columns: ‘Title’, ‘Text’, ‘Date’, ‘Subject’
- Target = ‘True’ (1) or ‘Fake’ (0)

- Dataset from Kaggle: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

- Clément Bisaillon (Owner)

Publications:
Ahmed H, Traore I, Saad S. (2018) “Detecting opinion spams and fake news using text classification”, *Journal of Security and Privacy*, Volume 1(1).

Ahmed H, Traore I, Saad S. (2017) “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments.* ISDDC 2017. Lecture Notes in Computer Science, Vol 10618. Springer, Cham (pp. 127-138).

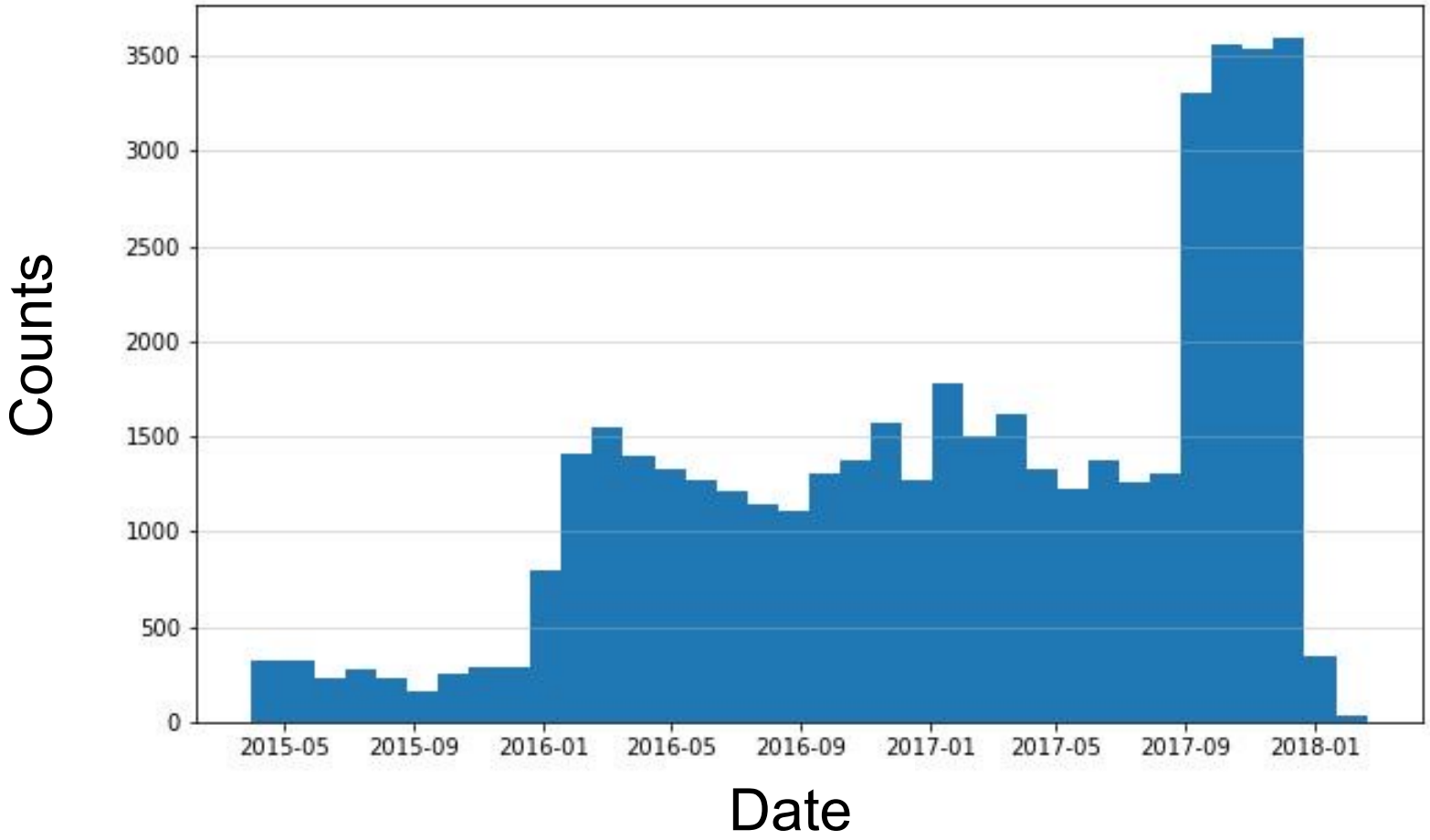
The Data

	title	text	subject	date	target
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0

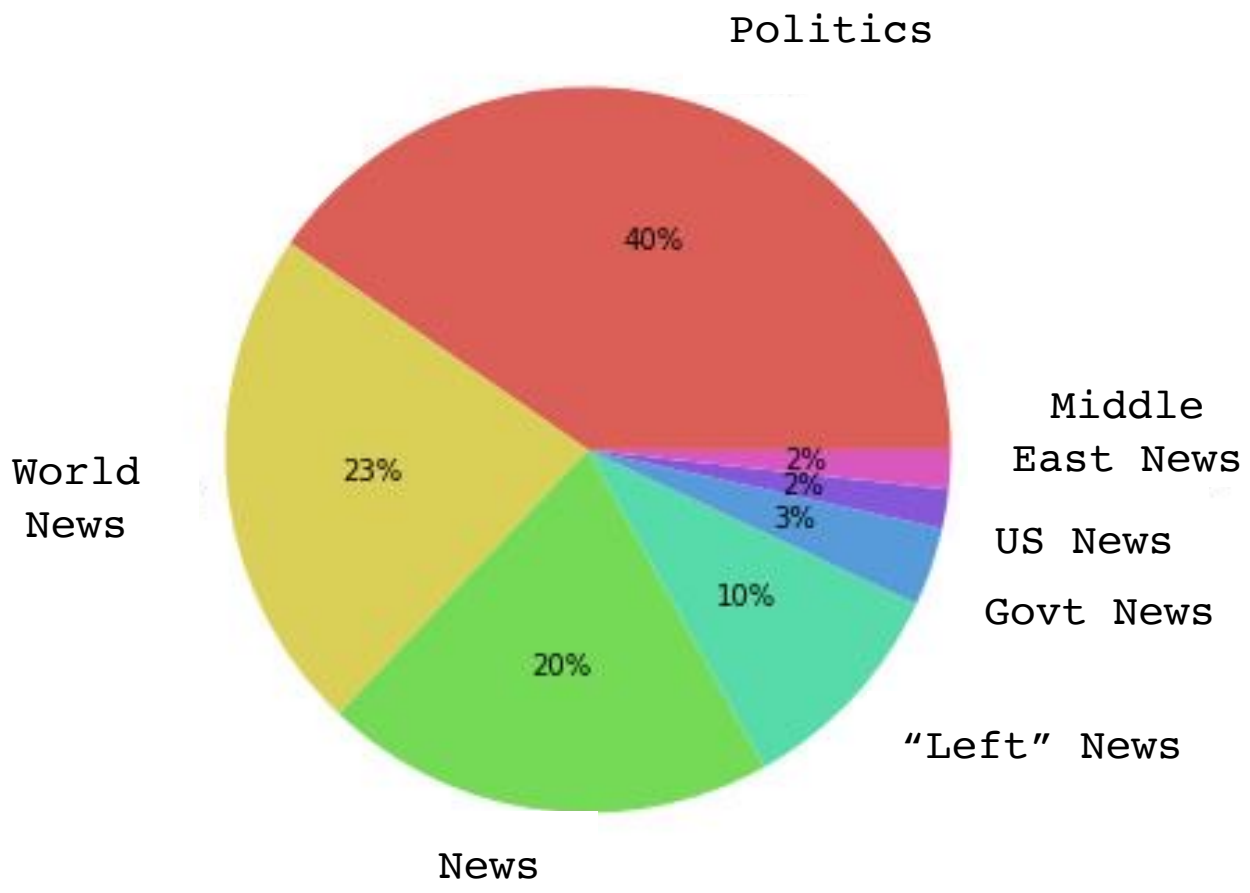
	title	text	subject	date	target
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1

Article Counts over Time

The articles span May 2015 – Jan 2018,
with the majority from late 2017



Corpus Classification by Subject

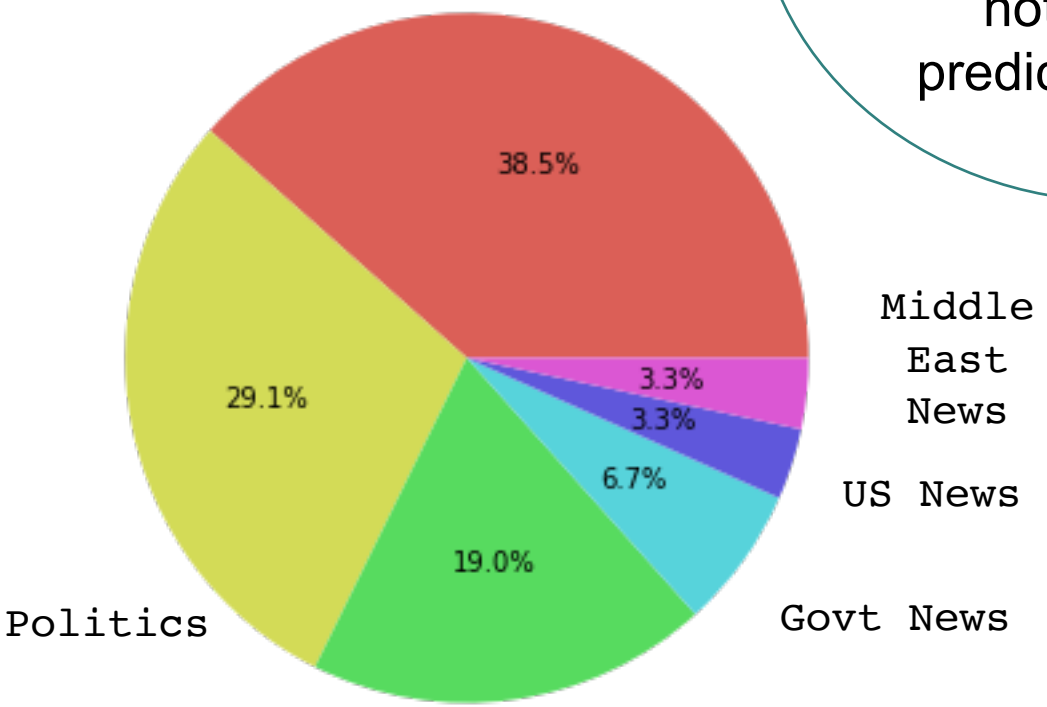


Could the ‘subject’ labels be a useful predictive feature?

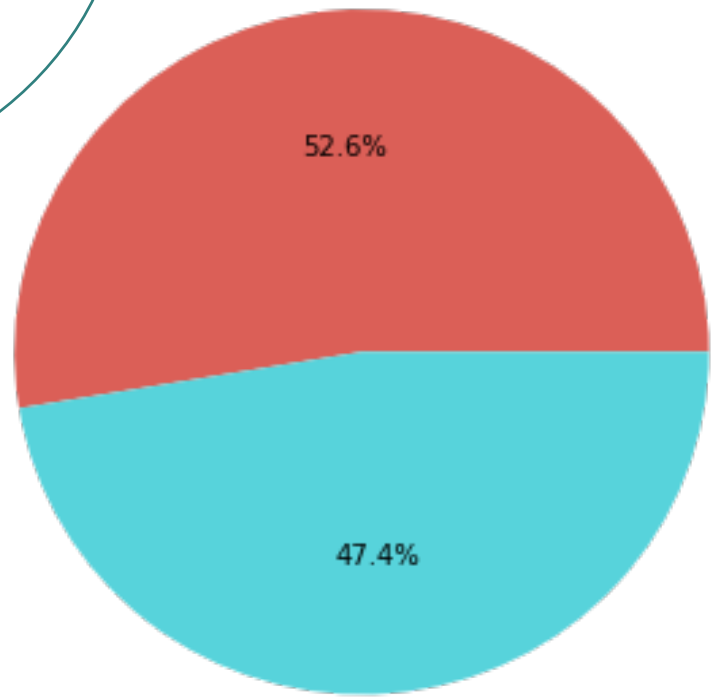
Corpus Classification by Subject

Subject labelling is inconsistent between the ‘fake’ and ‘true’ datasets. Therefore, this is not a useful predictive feature

Fake News



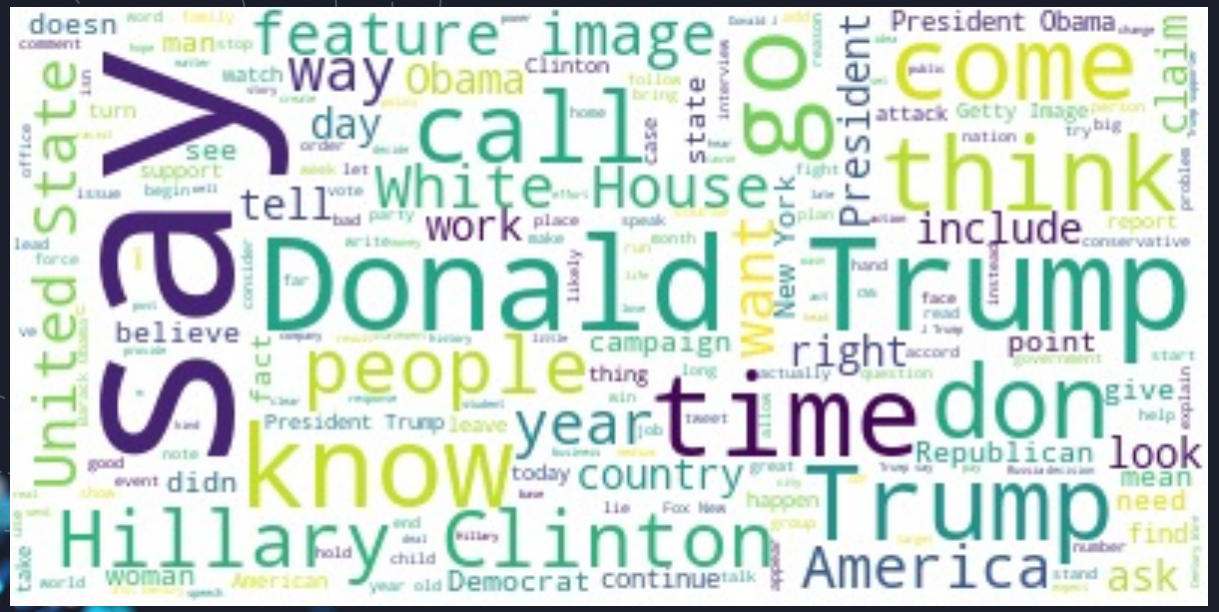
True News



Do the two classes differ in word content?

Fake News WordCloud!

38850618 words in the combination of all documents.



True News WordCloud!

34881474 words in the combination of all documents.

They differ surprisingly little!

Workflow

Libraries:

Pandas and NumPy
Regex
SpaCy
SciKitLearn
Matplotlib / Seaborn

1. Data Cleaning

- Removal of STOPWORDS & punctuation
- Stemming
- Lemmatization

test/train split (20:80)

2. Feature Engineering

- Word count, character count, word density, punctuation count, Title Case count, UPPER CASE count on both **'title'** and **'text'**
- Count Vectorization ('**title**' and '**text**')
- TF-IDF Vectorizer ('**title**' and '**text**')
  - bi-grams, tri-grams on both words and characters
- Topic modelling (LDA) ('**title**' only)

Dataset 1

Dimension reduction *TruncatedSVD*

Concatenation

Dataset 2

Results of preliminary modelling (non-optimized)

Dataset 1
(35910, 26)

Dataset 2
(35910, 200)

	Dataset 1 (X_train)	Dataset 2 (X_train_combined_2)
Naïve Bayes	0.849744	0.853976
Logistic Regression	0.922143	0.973936
LinearSVC	0.715972	0.972934
Random Forest	0.963021	0.980062
Gradient Boosting	0.946536	0.917465

Accuracy Scores for Cross-Validated Models on Dataset 1

Model	Set	Accuracy Score	Meta Classifier
Linear SVC	Train	0.81 +/- 0.01	
	Test	0.70 +/- 0.09	
Naïve Bayes (Gaussian)	Train	0.81 +/- 0.01	
	Test	0.83 +/- 0.00	
Logistic Regression	Train	0.89 +/- 0.01	
	Test	0.90 +/- 0.01	
Gradient Boosting	Train	0.95 +/- 0.00	
	Test	0.94 +/- 0.01	
Random Forest	Train	0.96 +/- 0.00	
	Test	0.94 +/- 0.01	
Stacking Classifier	Train	0.96 +/- 0.00	Random Forest
	Test	0.94 +/- 0.01	

SCORES (LOWER TO HIGHER)

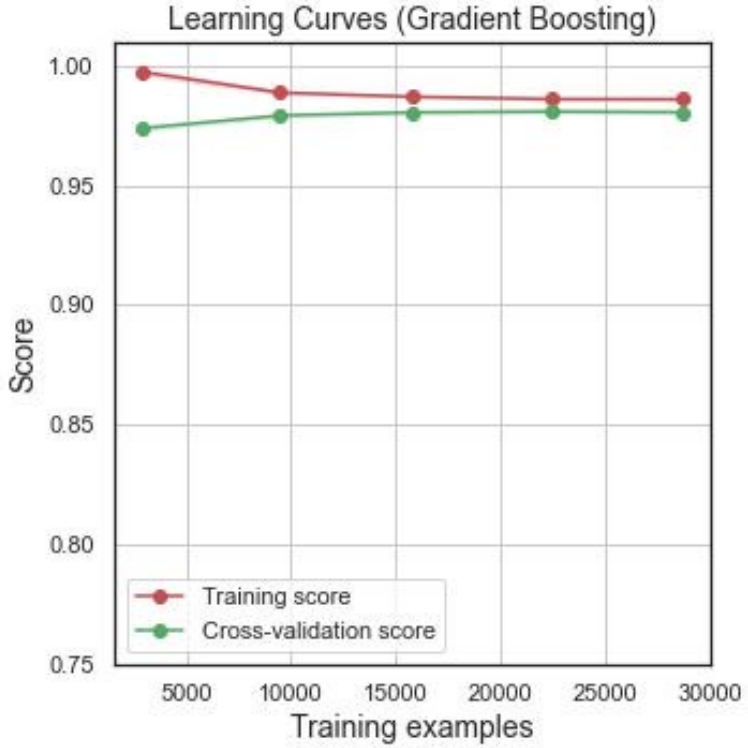
Accuracy scores for cross-validated models on Dataset 2

Model	Set	Accuracy Score	Meta Classifier
Naïve Bayes (Gaussian)	Train	0.85 +/- 0.00	
	Test	0.85 +/- 0.01	
Logistic Regression	Train	0.97 +/- 0.00	
	Test	0.97 +/- 0.00	
Random Forest	Train	0.98 +/- 0.00	
	Test	0.97 +/- 0.00	
Stacking Classifier	Train	0.98 +/- 0.00	Random Forest
	Test	0.97 +/- 0.00	
Linear SVC	Train	0.98 +/- 0.00	
	Test	0.98 +/- 0.00	
Gradient Boosting	Train	0.98 +/- 0.00	
	Test	0.98 +/- 0.00	

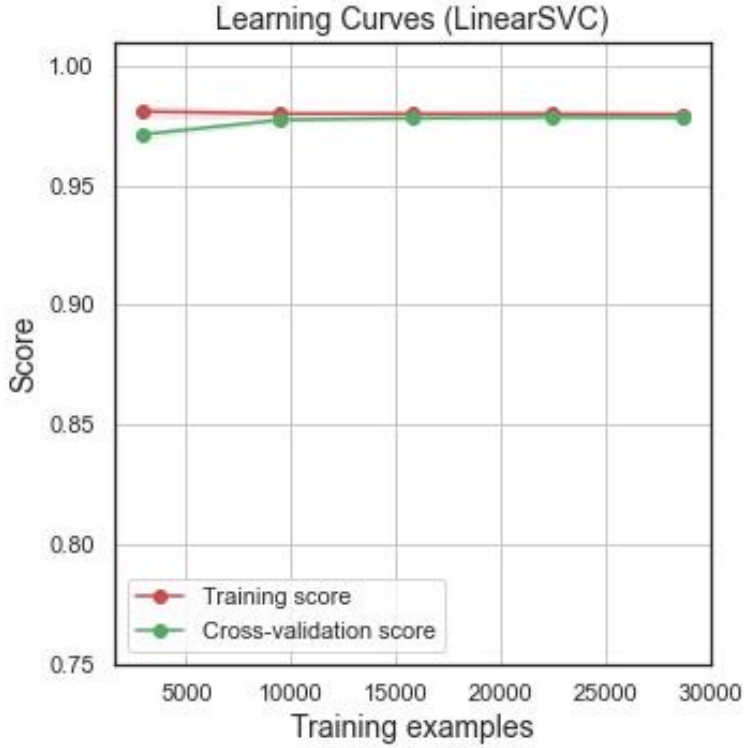
SCORES (LOWER TO HIGHER)

Model Evaluation
Dataset 2

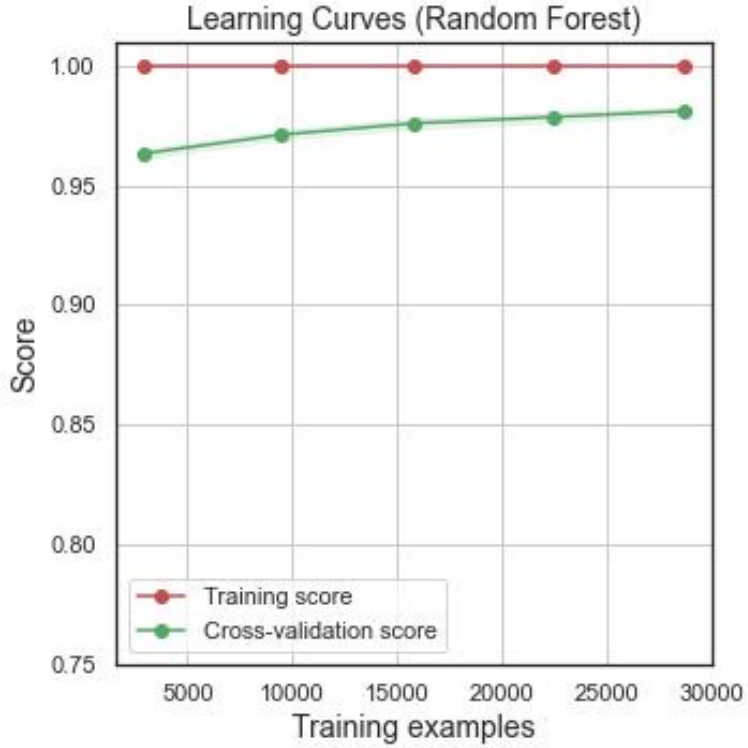
Gradient Boosting



Linear SVC

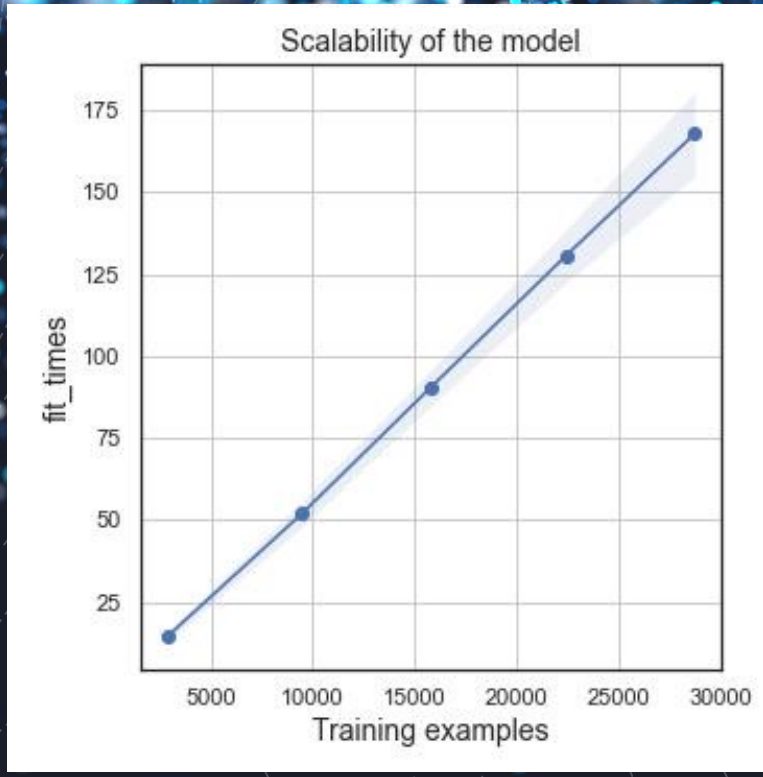


Random Forest

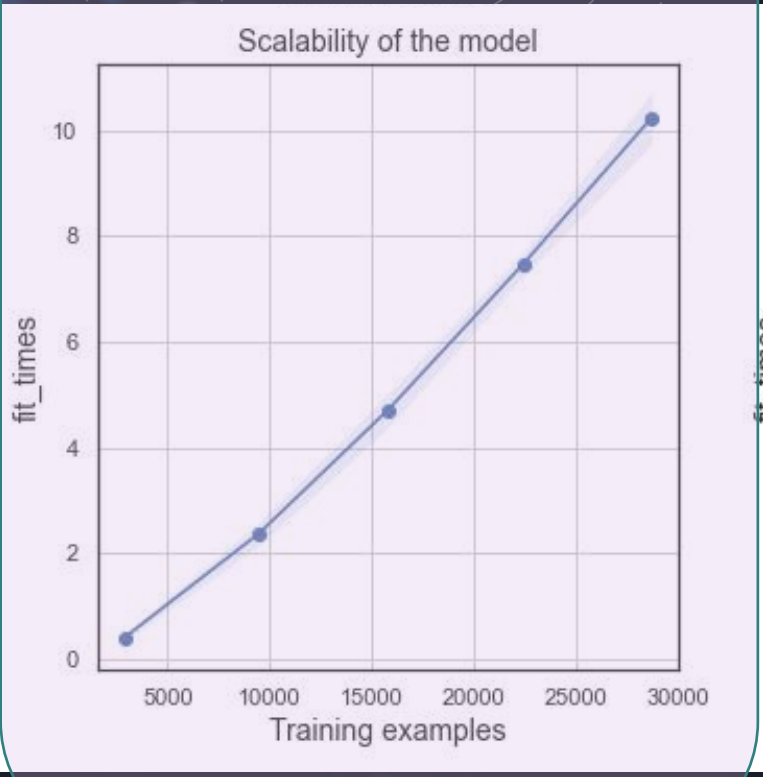


Model Evaluation
Dataset 2

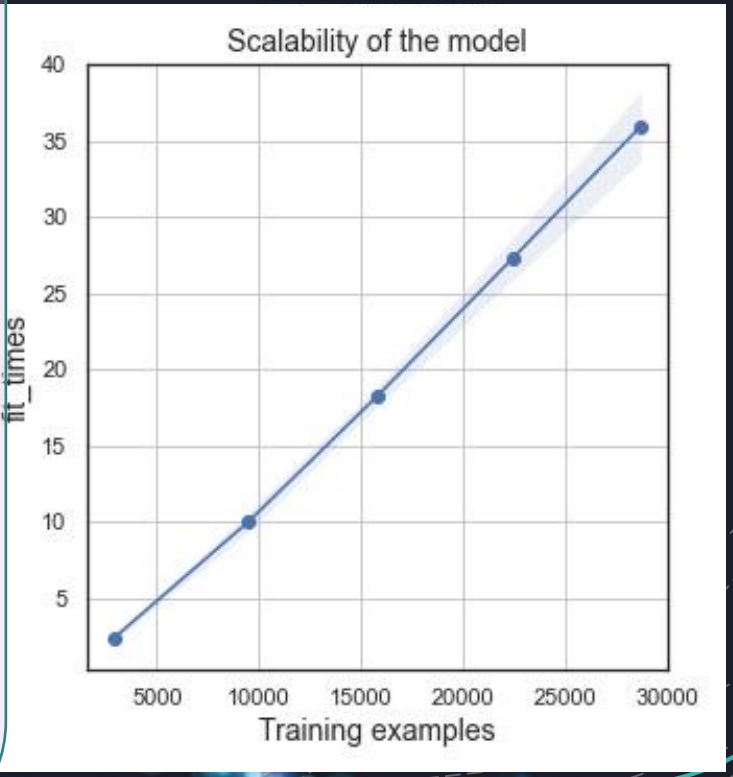
Gradient Boosting



Linear SVC



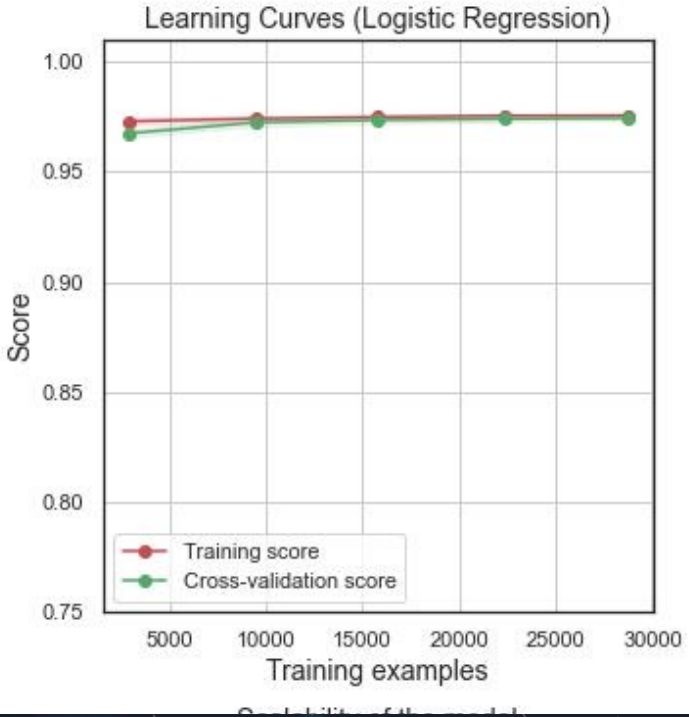
Random Forest



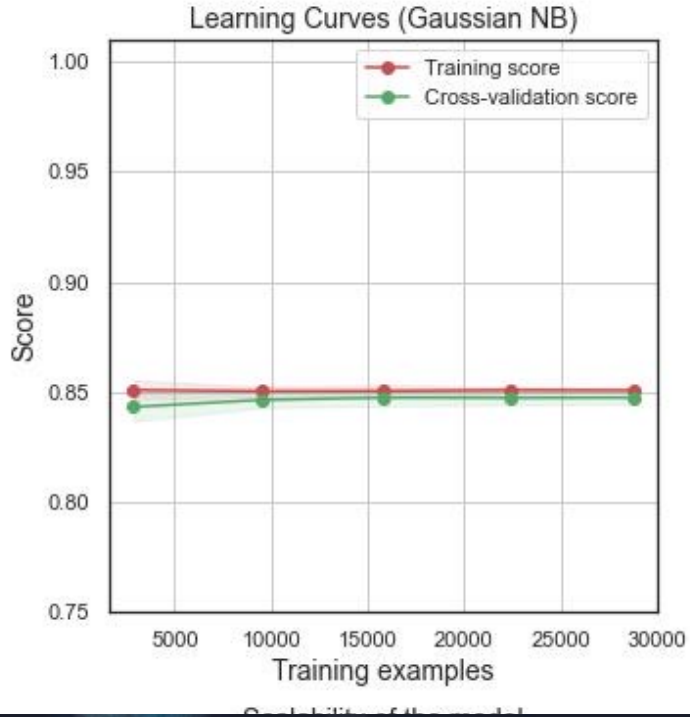
All things being equal, linear SVC wins on speed!

Model Evaluation
Dataset 2

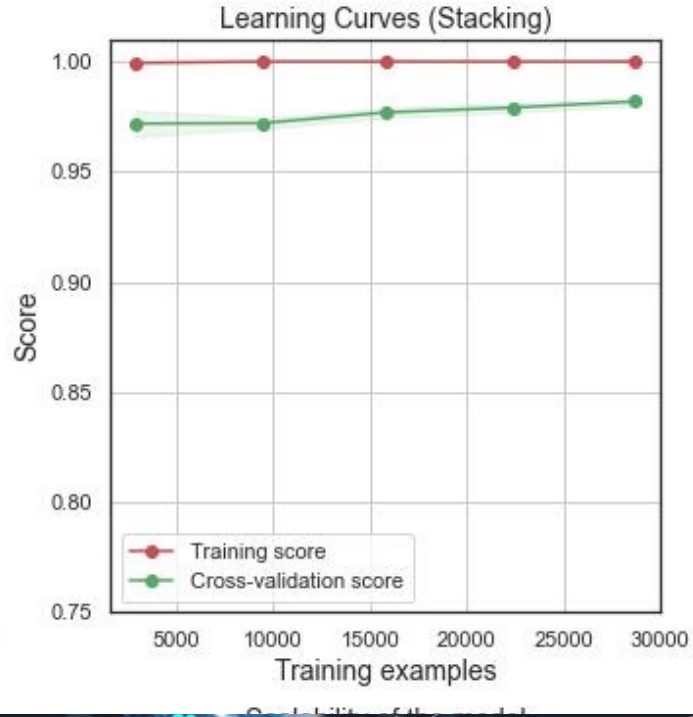
Logistic Regression



Gaussian NB

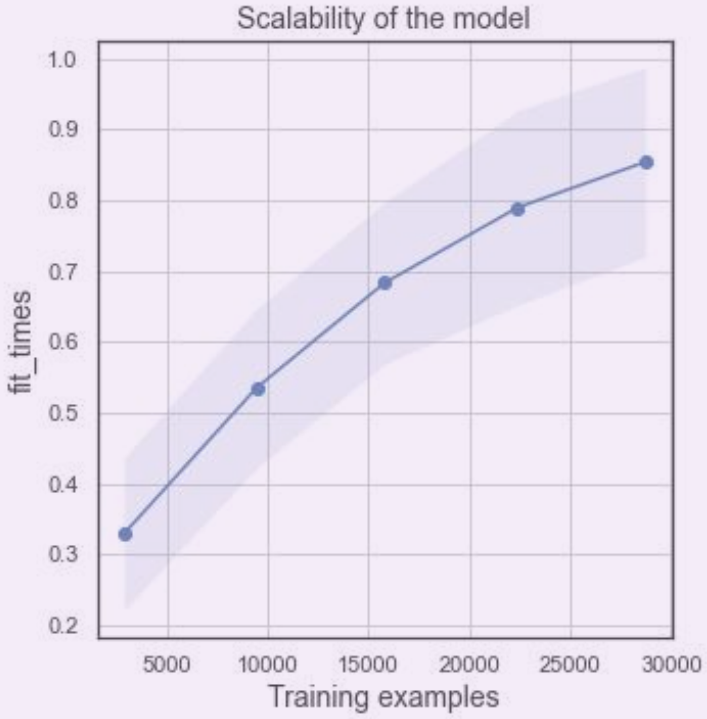


Stacking



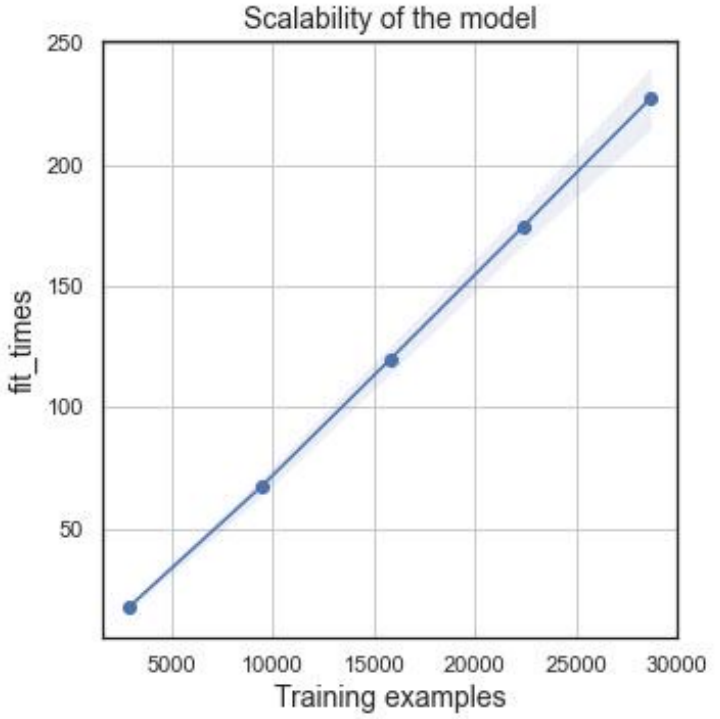
Model Evaluation
Dataset 2

Logistic Regression



Gaussian NB

Stacking



Logistic Regression wins on speed!

Model Evaluation
Dataset 2

Model	Accuracy (test dataset)	Precision (the extent to which we are not classifying fake news as true)	Recall / Sensitivity (the extent to which we are finding the true news)	AUC
Logistic Regression	0.9732	0.9696	0.9744	0.9955
Linear SVC	0.9747	0.9759	0.9711	0.9964
Random Forest	0.9810	0.9784	0.9818	0.9979
Gradient Boosting	0.9738	0.9719	0.9735	0.9969
Stacking Classifier	0.9792	0.9779	0.9786	0.9928

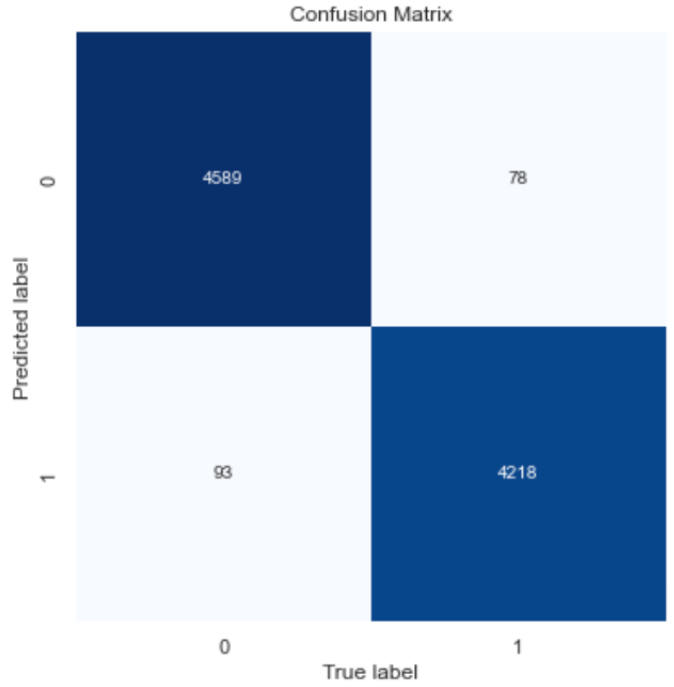
SPEED (FAST TO SLOW)

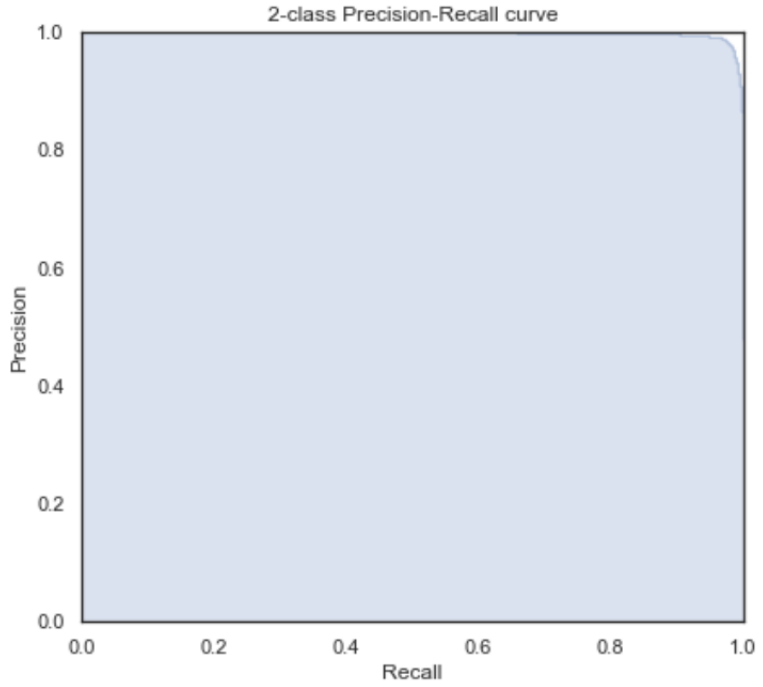
Model Evaluation (Random Forest) Dataset 2

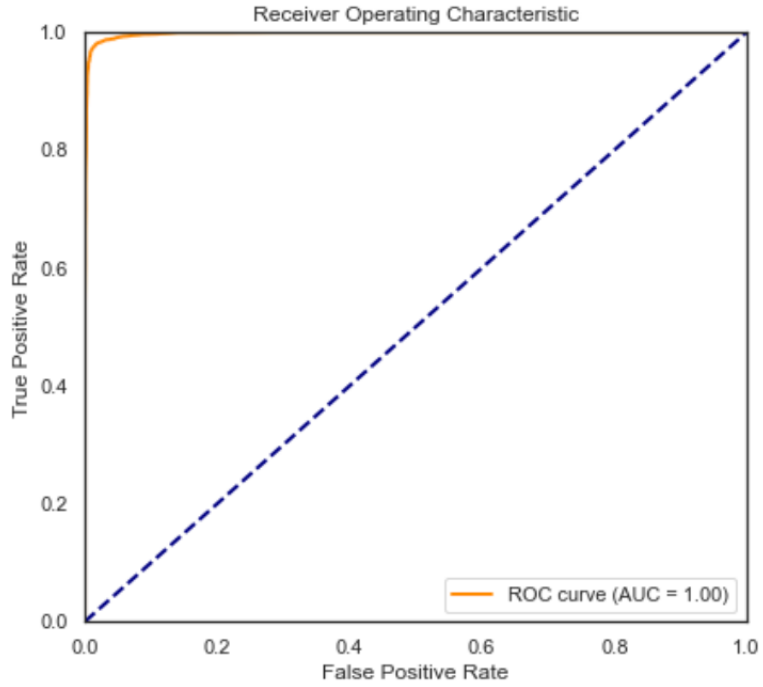
Random Forest

```
Accuracy  : 0.9810 [TP / N] Proportion of predicted labels that match the true labels. Best: 1, Worst: 0
Precision: 0.9784 [TP / (TP + FP)] Not to label a negative sample as positive.        Best: 1, Worst: 0
Recall    : 0.9818 [TP / (TP + FN)] Find all the positive samples.                    Best: 1, Worst: 0
ROC AUC   : 0.9979                                                                    Best: 1, Worst: < 0.5
```

TP: True Positives, FP: False Positives, TN: True Negatives, FN: False Negatives, N: Number of samples







Conclusions

- All models (with the exception of Naïve Bayes) were highly accurate, even without optimization; very little trade-off between precision and recall.
- If speed and computational demands are of primary concern, we'd recommend a Linear SVC model or Logistic Regression model
- For the highest accuracy, Random Forest is recommended: this model is a good compromise between speed/computational demands and accuracy but is recommended only for very large volumes of training data.

NEXT STEPS:

- Optimization of the hyperparameters for the Random Forest (and other models);
- Topic modelling on the full text may improve model accuracy, but this is unlikely to be worth the effort.
- Exploration of different meta-classifiers in a Stacking Model to see if accuracy can be further improved.