Real and Fake News

Sara Satti
Data Science Career Track

Objective

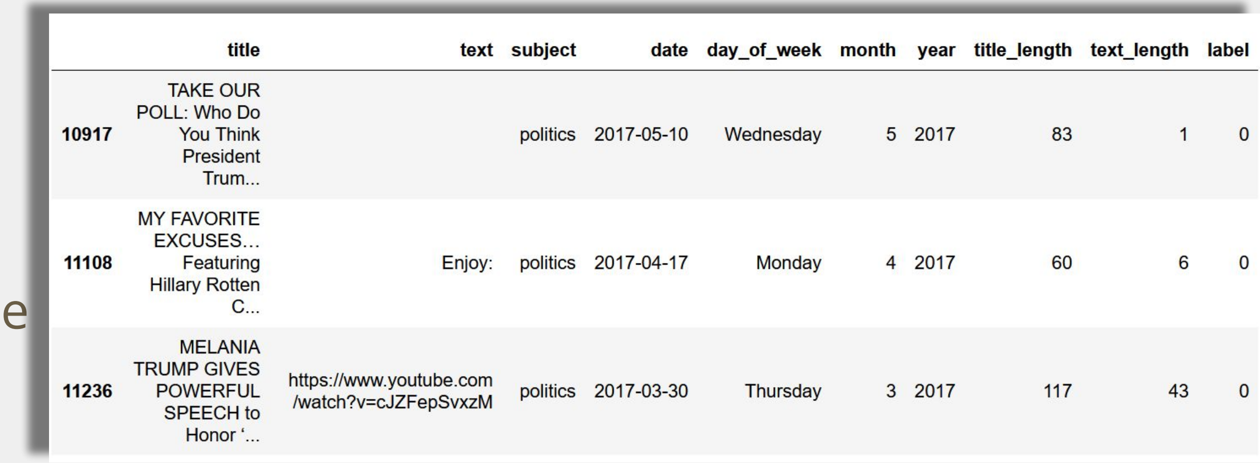
Build a model to distinguish whether an article is Real or Fake

Outline

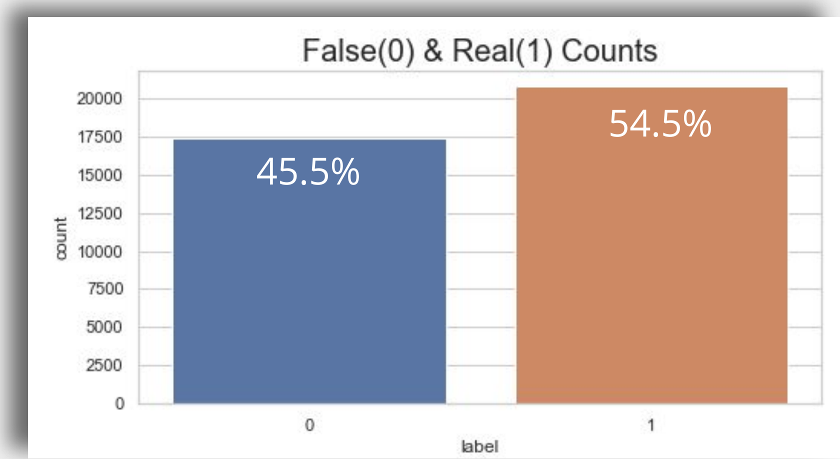
- Data Wrangling
- Exploring the Data
- WordClouds
- Statistical Analysis
- Machine Learning
- Model Comparisons
- Stress tests
- Conclusions
- Recommendations

Data Wrangling

- Load Fake / Real data
- Remove ‘http’ entries
- Drop rows with duplicate entries
- Feature engineering:
  - date,
  - title_length
  - text_length
- Label and combine Fake and Real datasets

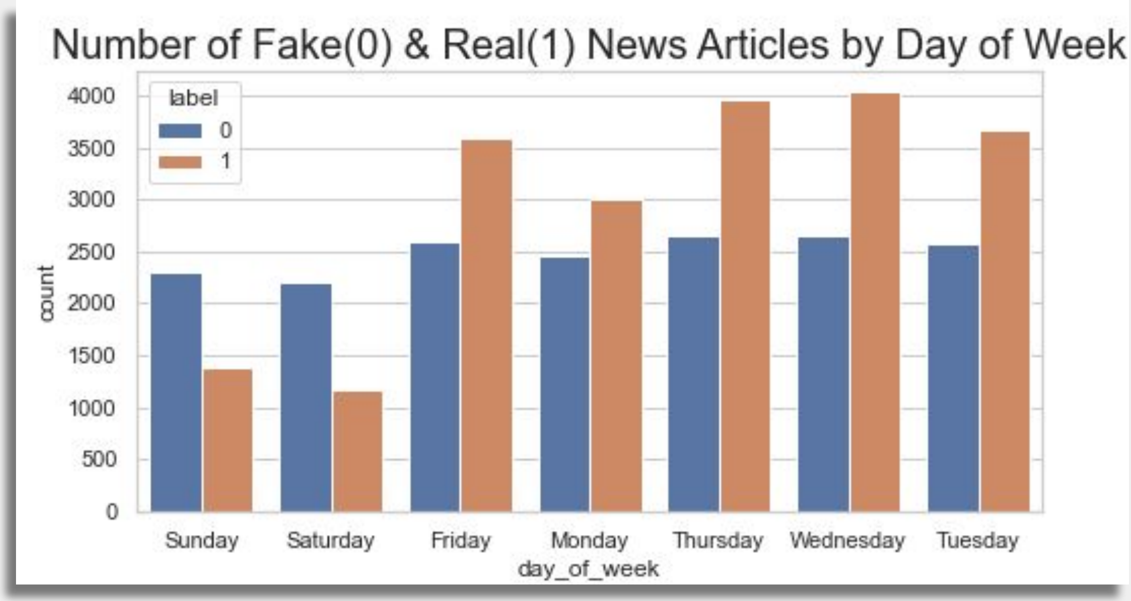


	title	text	subject	date	day_of_week	month	year	title_length	text_length	label
10917	TAKE OUR POLL: Who Do You Think President Trum...		politics	2017-05-10	Wednesday	5	2017	83	1	0
11108	MY FAVORITE EXCUSES… Featuring Hillary Rotten C...	Enjoy:	politics	2017-04-17	Monday	4	2017	60	6	0
11236	MELANIA TRUMP GIVES POWERFUL SPEECH to Honor ‘...	https://www.youtube.com /watch?v=cJZFepSvxzM	politics	2017-03-30	Thursday	3	2017	117	43	0



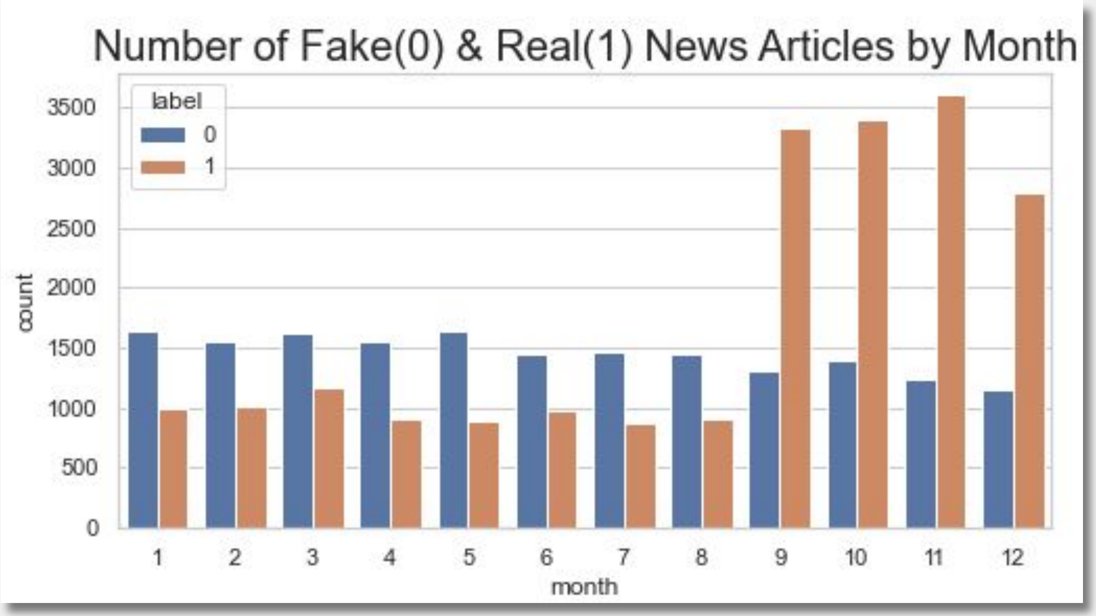
Exploring the Data - Fake news knows no rest!

- Fake news - Uniform distribution
- Real news -
  - Peaks mid-week
  - 60% drop in number of articles over the weekend.



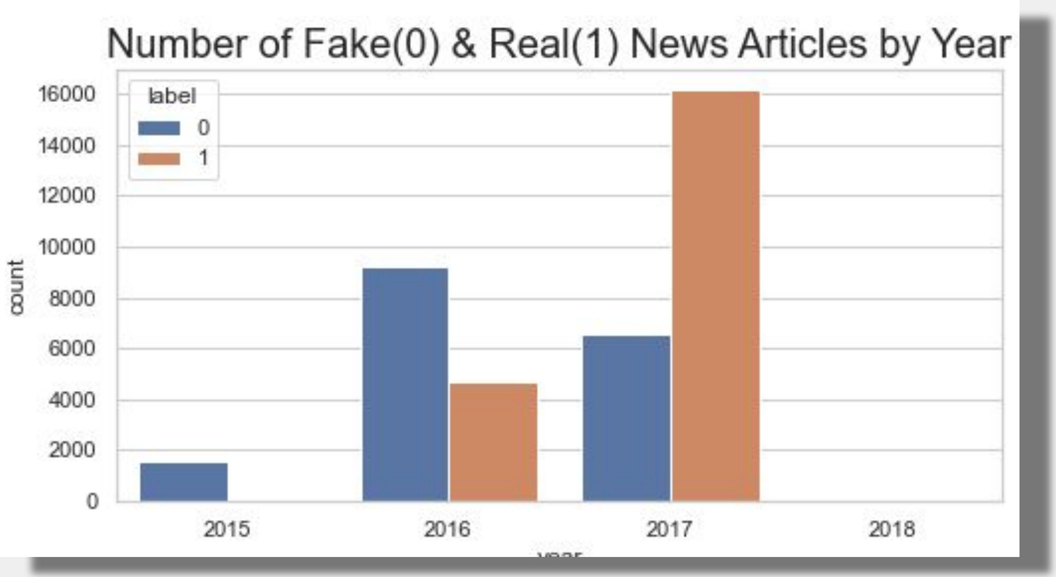
Exploring the Data

- Fake news -
  - Uniform distribution
  - Jan-August at higher count than real news
- Real news -
  - At 1000 articles Jan-August
  - Jump to >3X Sept to Dec



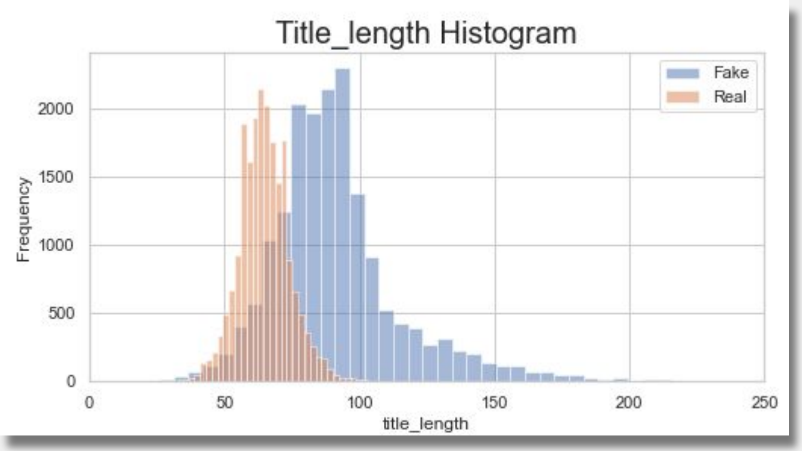
Exploring the Data - Election year:

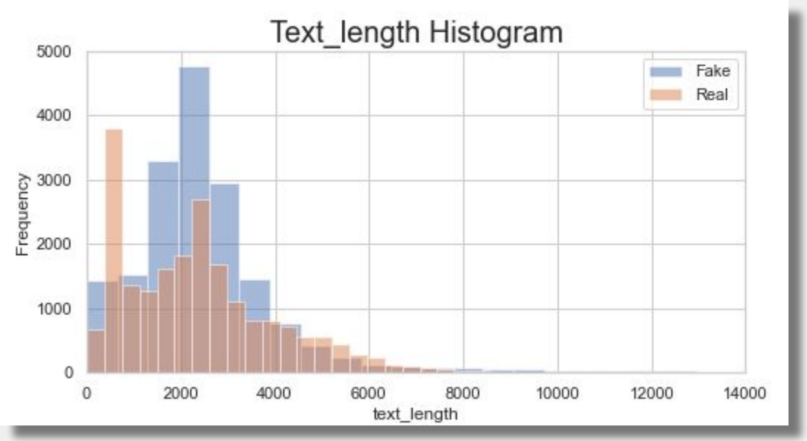
- Fake news - 2015 - 2017
- Real news - 2016 - 2017



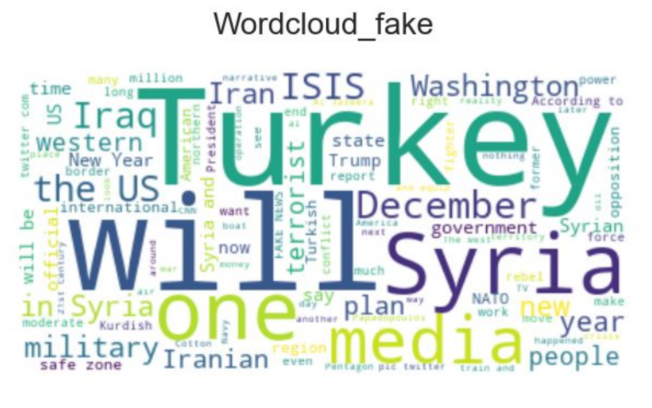
Exploring the Data - Text / Title-length

- Title-length:
  - Fake news -
    - Mean: 91
    - Few titles >150 characters
  - Real news
    - Mean: 65
    - Normally distributed
- Text-length: fake and real news overlap

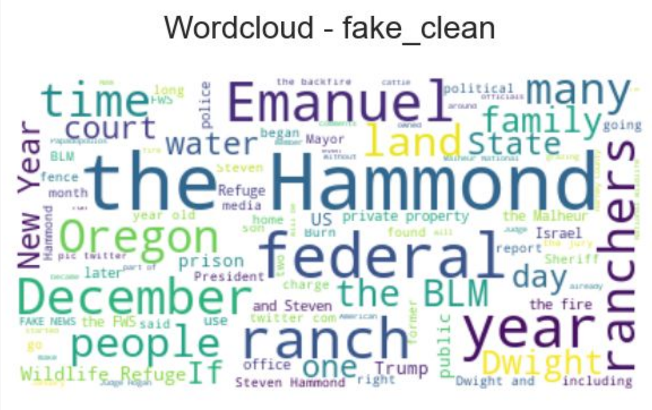


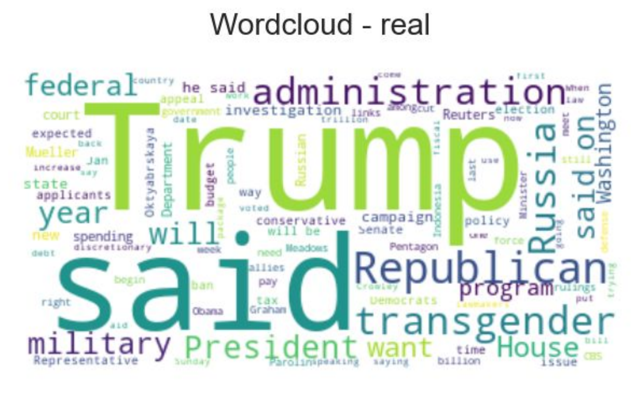


Exploring the Data

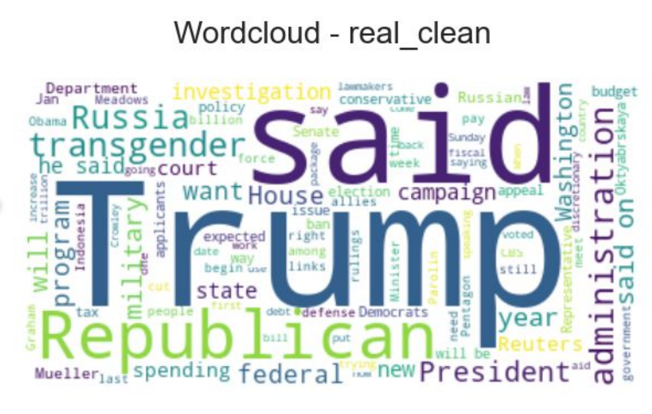

Wordcloud_fake

Delete urls
Drop duplicate rows
Drop row (text < 45)

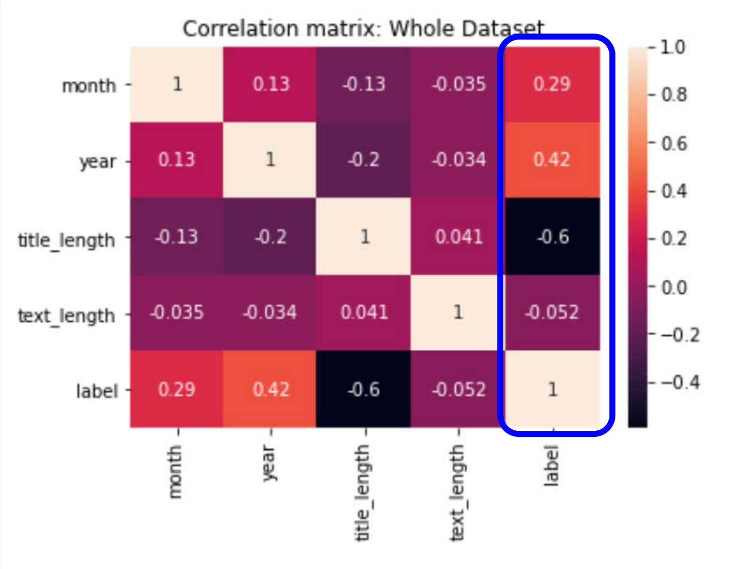

Wordcloud - fake_clean


Wordcloud - real

Drop duplicate rows


Wordcloud - real_clean

Statistical Analysis

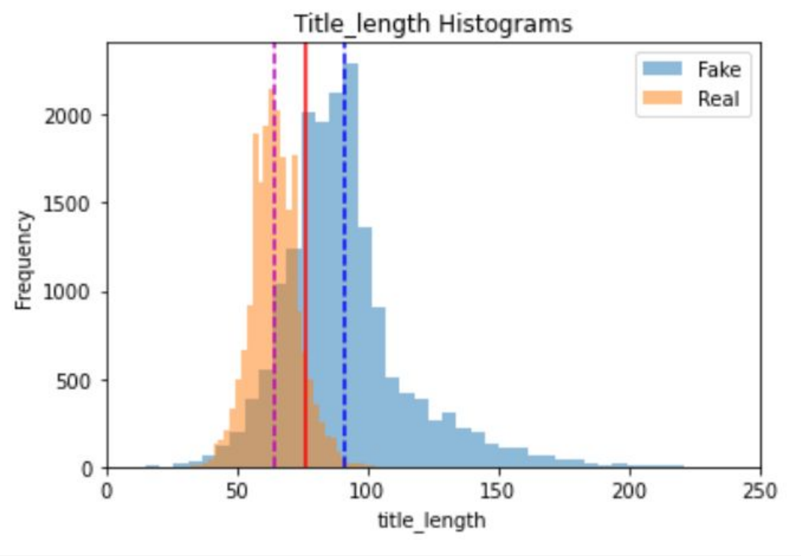


Correlation matrix: Whole Dataset

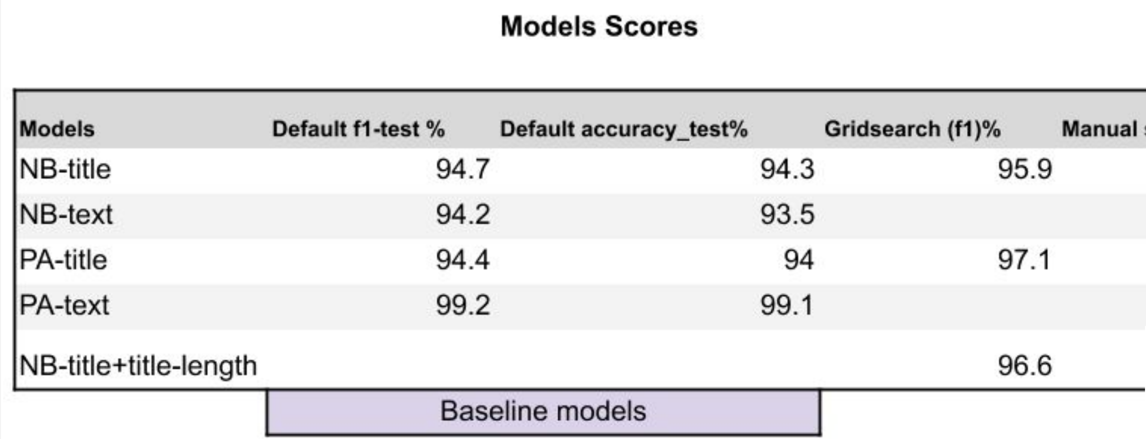
	month	year	title_length	text_length	label
month	1	0.13	-0.13	-0.035	0.29
year	0.13	1	-0.2	-0.034	0.42
title_length	-0.13	-0.2	1	0.041	-0.6
text_length	-0.035	-0.034	0.041	1	-0.052
label	0.29	0.42	-0.6	-0.052	1

- Highest correlation - title_length/label
- Label has lowest correlation to text_length
- Positive corr with month/year - reflected in barcharts

- Samples are independent (pearson’s cor coef = 0.006)
- T-statistic of 133, p-value=0 => probably different distributions.



Machine Learning

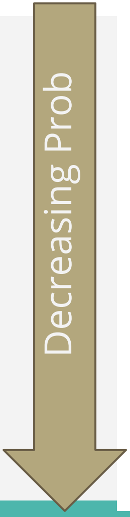


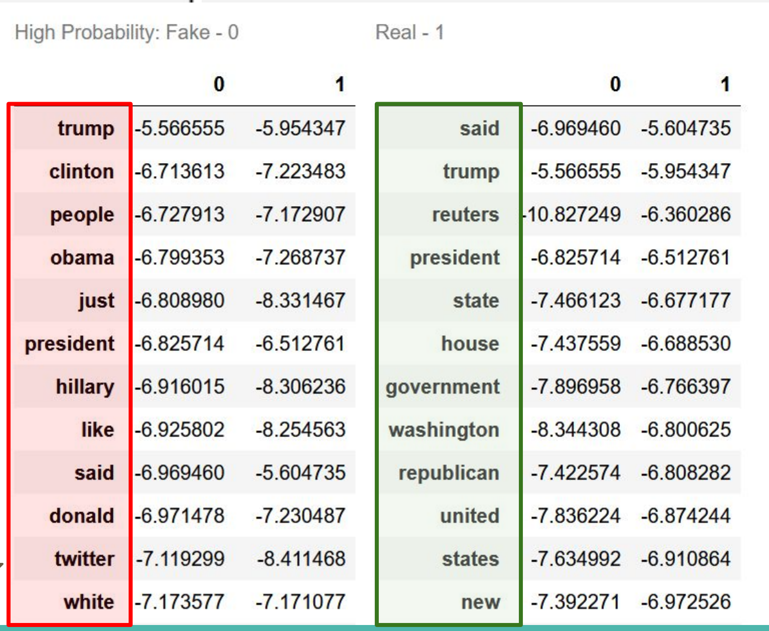
Models Scores

Models	Default f1-test %	Default accuracy_test%	Gridsearch (f1)%	Manual search (f1)%
NB-title	94.7	94.3	95.9	
NB-text	94.2	93.5		99
PA-title	94.4	94	97.1	
PA-text	99.2	99.1		99.9
NB-title+title-length			96.6	

Baseline models

- Models ('text') - better scores.
- Model using title *and* title_length - slightly improved f1 score (95.9 to 96.6%).





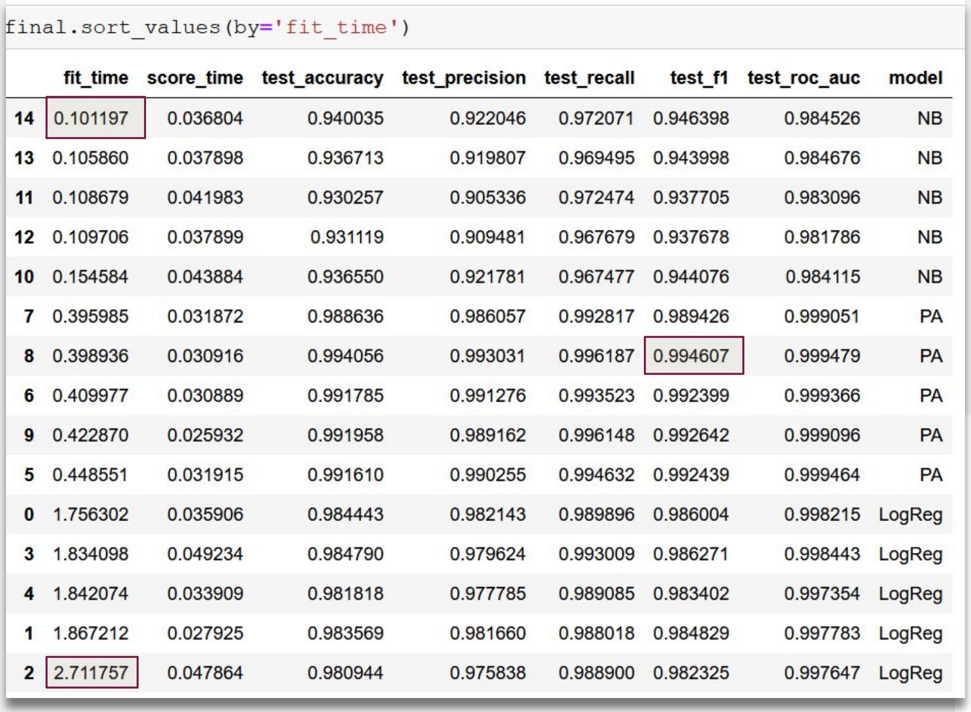
High Probability: Fake - 0

	0	1
trump	-5.566555	-5.954347
clinton	-6.713613	-7.223483
people	-6.727913	-7.172907
obama	-6.799353	-7.268737
just	-6.808980	-8.331467
president	-6.825714	-6.512761
hillary	-6.916015	-8.306236
like	-6.925802	-8.254563
said	-6.969460	-5.604735
donald	-6.971478	-7.230487
twitter	-7.119299	-8.411468
white	-7.173577	-7.171077

Real - 1

	0	1
said	-6.969460	-5.604735
trump	-5.566555	-5.954347
reuters	-10.827249	-6.360286
president	-6.825714	-6.512761
state	-7.466123	-6.677177
house	-7.437559	-6.688530
government	-7.896958	-6.766397
washington	-8.344308	-6.800625
republican	-7.422574	-6.808282
united	-7.836224	-6.874244
states	-7.634992	-6.910864
new	-7.392271	-6.972526

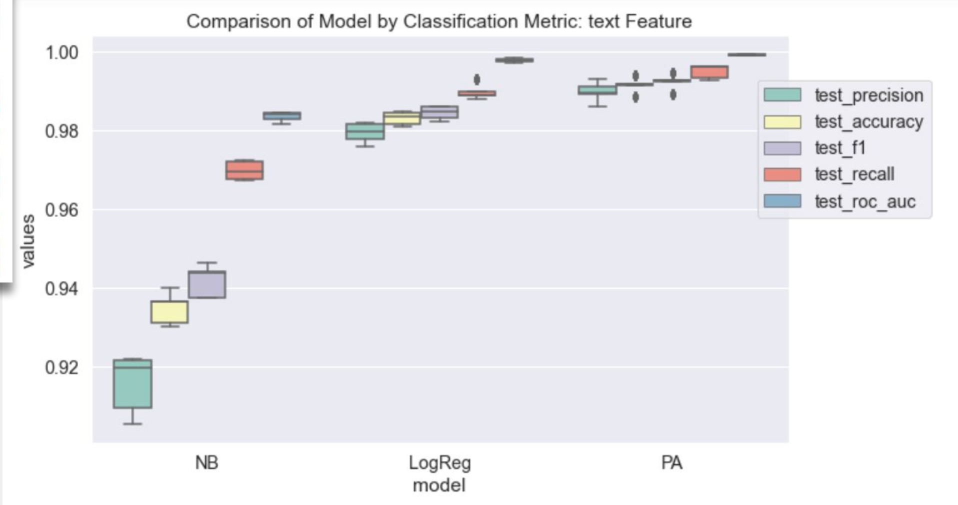
Model Comparisons



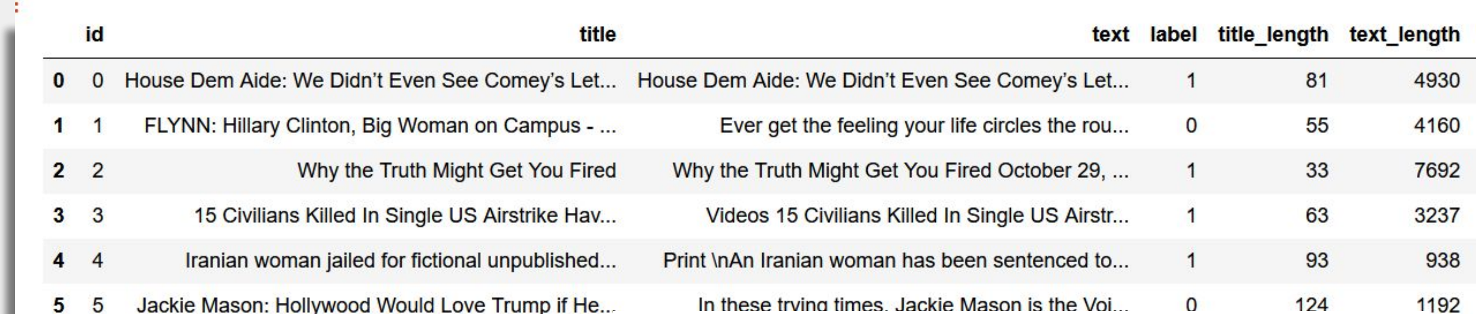
```
final.sort_values(by='fit_time')
```

	fit_time	score_time	test_accuracy	test_precision	test_recall	test_f1	test_roc_auc	model
14	0.101197	0.036804	0.940035	0.922046	0.972071	0.946398	0.984526	NB
13	0.105860	0.037898	0.936713	0.919807	0.969495	0.943998	0.984676	NB
11	0.108679	0.041983	0.930257	0.905336	0.972474	0.937705	0.983096	NB
12	0.109706	0.037899	0.931119	0.909481	0.967679	0.937678	0.981786	NB
10	0.154584	0.043884	0.936550	0.921781	0.967477	0.944076	0.984115	NB
7	0.395985	0.031872	0.988636	0.986057	0.992817	0.989426	0.999051	PA
8	0.398936	0.030916	0.994056	0.993031	0.996187	0.994607	0.999479	PA
6	0.409977	0.030889	0.991785	0.991276	0.993523	0.992399	0.999366	PA
9	0.422870	0.025932	0.991958	0.989162	0.996148	0.992642	0.999096	PA
5	0.448551	0.031915	0.991610	0.990255	0.994632	0.992439	0.999464	PA
0	1.756302	0.035906	0.984443	0.982143	0.989896	0.986004	0.998215	LogReg
3	1.834098	0.049234	0.984790	0.979624	0.993009	0.986271	0.998443	LogReg
4	1.842074	0.033909	0.981818	0.977785	0.989085	0.983402	0.997354	LogReg
1	1.867212	0.027925	0.983569	0.981660	0.988018	0.984829	0.997783	LogReg
2	2.711757	0.047864	0.980944	0.975838	0.988900	0.982325	0.997647	LogReg

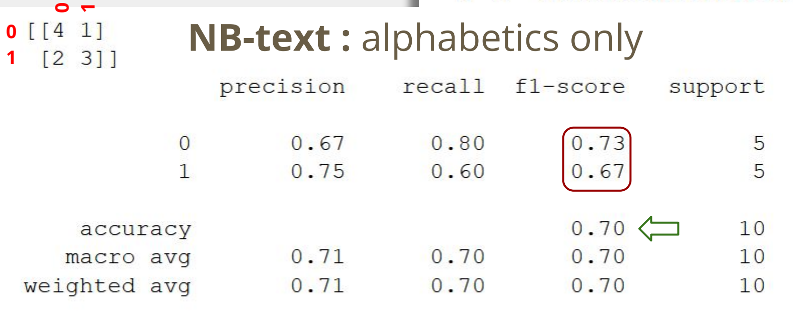
- NB - lowest training time.
- PA - highest f1 score.
- Logistic Regression - slowest in training.



Stress Tests



	id	title	text	label	title_length	text_length
0	0	House Dem Aide: We Didn't Even See Comey's Let...	House Dem Aide: We Didn't Even See Comey's Let...	1	81	4930
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Ever get the feeling your life circles the rou...	0	55	4160
2	2	Why the Truth Might Get You Fired	Why the Truth Might Get You Fired October 29, ...	1	33	7692
3	3	15 Civilians Killed In Single US Airstrike Hav...	Videos 15 Civilians Killed In Single US Airstr...	1	63	3237
4	4	Iranian woman jailed for fictional unpublished...	Print \nAn Iranian woman has been sentenced to...	1	93	938
5	5	Jackie Mason: Hollywood Would Love Trump if He...	In these trying times, Jackie Mason is the Voi...	0	124	1192
		Elton John's 6 Favorite ...				729
		ch Socialist Party's Pre...				923
		cript for Donald Trump'...				177
		Jkraine and Russia, Co...				261

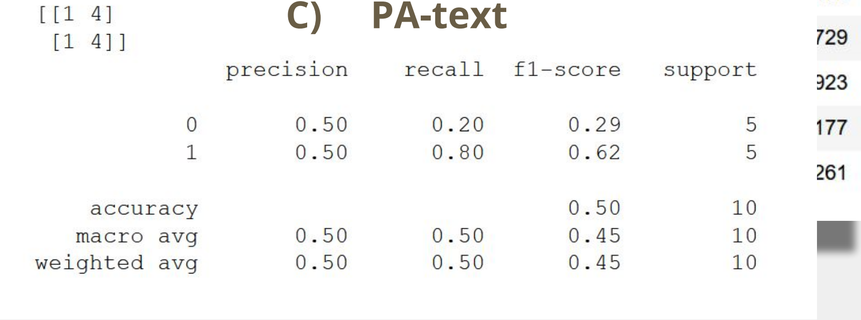


NB-text : alphabetics only

```
   0 1
0 [[4 1]
1  [2 3]]
              precision    recall  f1-score   support

           0       0.67      0.80      0.73         5
           1       0.75      0.60      0.67         5

    accuracy                           0.70        10
   macro avg       0.71      0.70      0.70        10
weighted avg       0.71      0.70      0.70        10
```

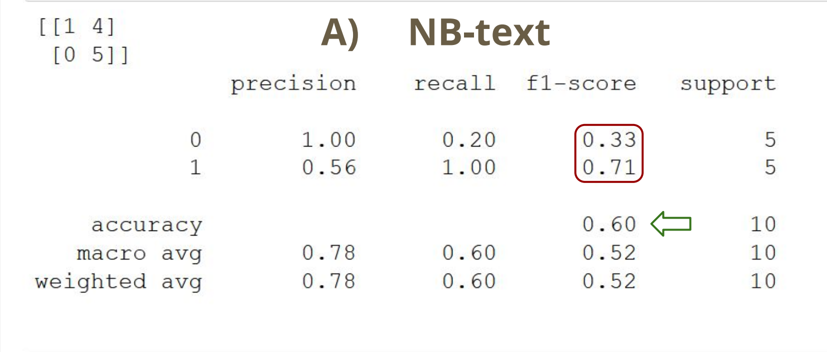


C) PA-text

```
[[1 4]
 [1 4]]
              precision    recall  f1-score   support

           0       0.50      0.20      0.29         5
           1       0.50      0.80      0.62         5

    accuracy                           0.50        10
   macro avg       0.50      0.50      0.45        10
weighted avg       0.50      0.50      0.45        10
```

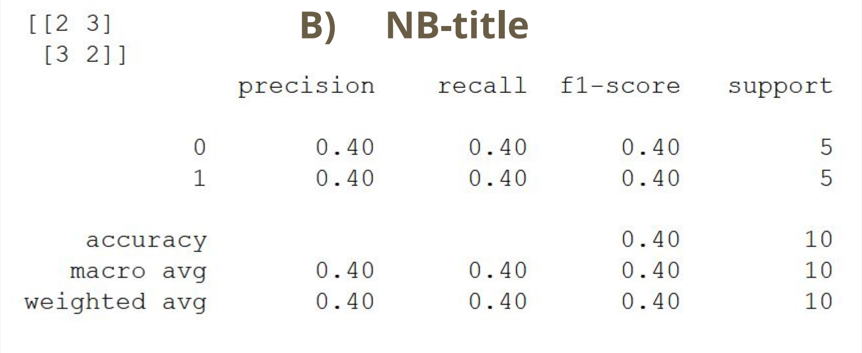


A) NB-text

```
[[1 4]
 [0 5]]
              precision    recall  f1-score   support

           0       1.00      0.20      0.33         5
           1       0.56      1.00      0.71         5

    accuracy                           0.60        10
   macro avg       0.78      0.60      0.52        10
weighted avg       0.78      0.60      0.52        10
```



B) NB-title

```
[[2 3]
 [3 2]]
              precision    recall  f1-score   support

           0       0.40      0.40      0.40         5
           1       0.40      0.40      0.40         5

    accuracy                           0.40        10
   macro avg       0.40      0.40      0.40        10
weighted avg       0.40      0.40      0.40        10
```

Conclusions

Supervised learning worked well in predicting whether a news article is Real or Fake from within the **same kaggle dataset**. However, when introducing articles from other datasets, the performance was considerably lower. Removal of ***characters, numbers and symbols*** from the body of the news article ***considerably improved*** the performance of the models. Maximum improvement was seen in the **Naive Bayes model where accuracy increased by 10% (from 60% to 70%), Fake news f1-score increased by 40%(from 33-73%), and f1-score for Real news dropped by 4% (71-67%)**.

Numerical, categorical and datetime features were skipped in order to focus on Natural Language features.

Title_length of the Fake and Real data showed a separation between their means. A multi-feature model using 'title' and 'title_length' was tested. This model led to a slight improvement in the performance (from f1:95.9 to 96.6%) when compared to the 'title' only model.

Recommendations

Combine features from title, and text and perhaps title_length to produce a model containing these and any other numerical features extracted from the data.

More complex methods such as neural networks to be tested in the future in order to further generalize the model.