
Classification

Fake and True News

The slide features a light blue background with a central title in orange. The title is framed by two horizontal teal bars at the top and bottom. Below the top teal bar is a thin light blue line. Below the bottom teal bar is a thin light blue line. Two short, thick olive-green horizontal bars are positioned on the left and right sides of the slide, below the main title.

Dataset

Dataset downloaded from kaggle

The procedure of collecting is unknown

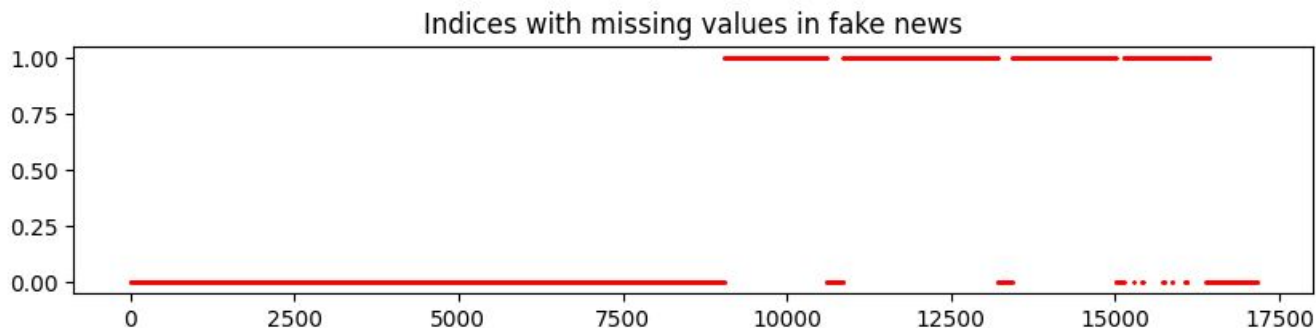
25% of duplicates in fake.csv, 5% in true.csv

99% of true data comes from reuters, 25% of fake data comes from twitter (comments to the post are included)

Scraping of fake dataset replaces apostrophes with white space (removal of lone letters)

Dataset

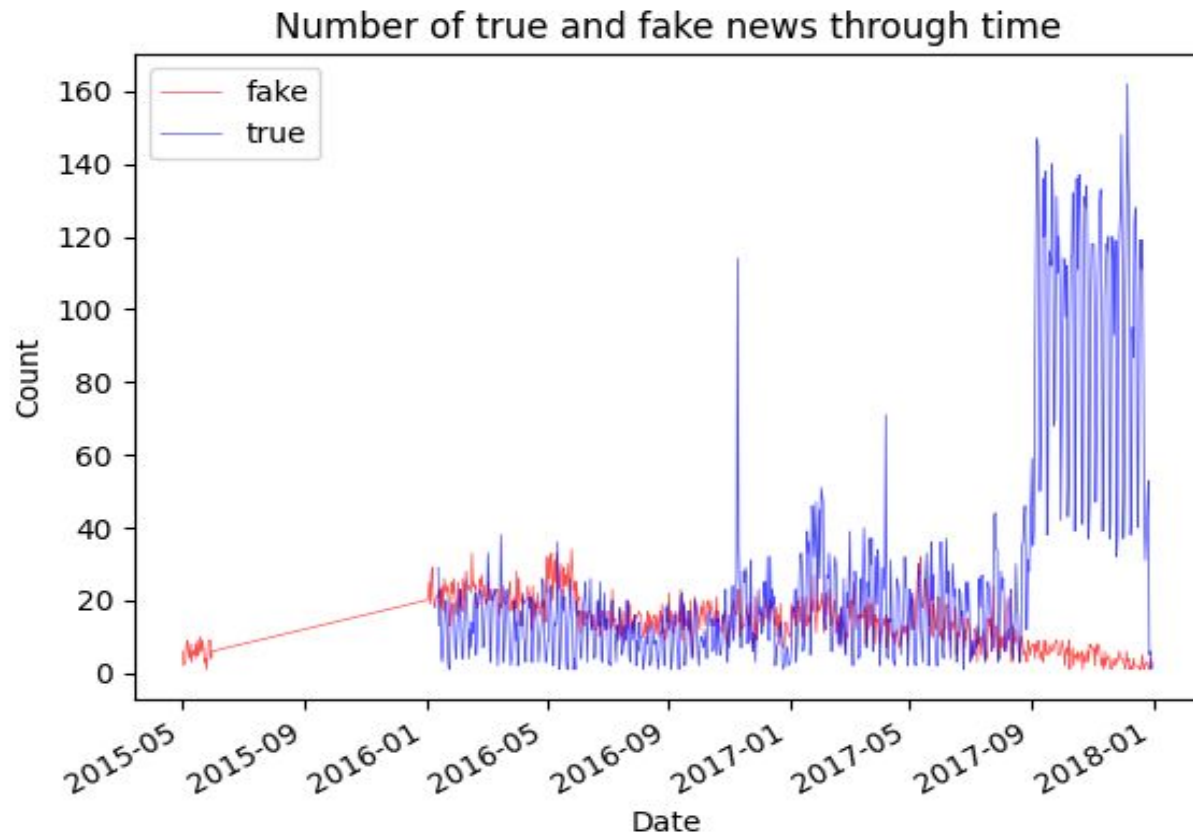
40% of dates in fake news are missing, we can't confidently fill in with standard techniques



First value in fake news for date:
date 2017-12-31 00:00:00
Name: 0, dtype: object

Last value in fake news for date:
date 2016-01-02 00:00:00
Name: 17165, dtype: object

Dataset

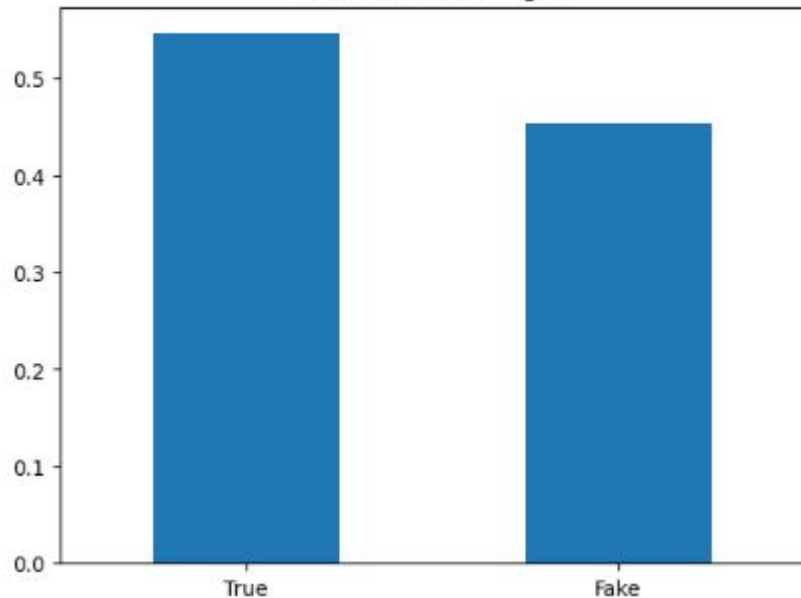


Data Split

Stratified

Rows of training set: 29601
Rows of validation set: 3700
Rows of test set: 3701

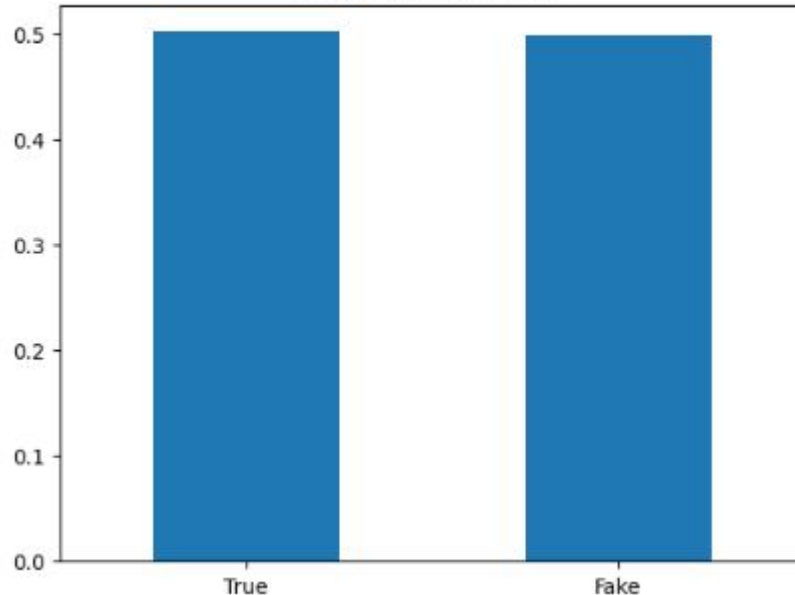
Distribution of Target



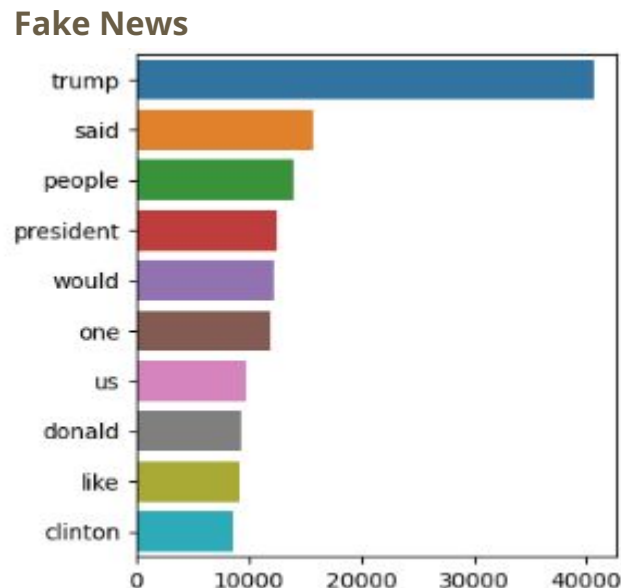
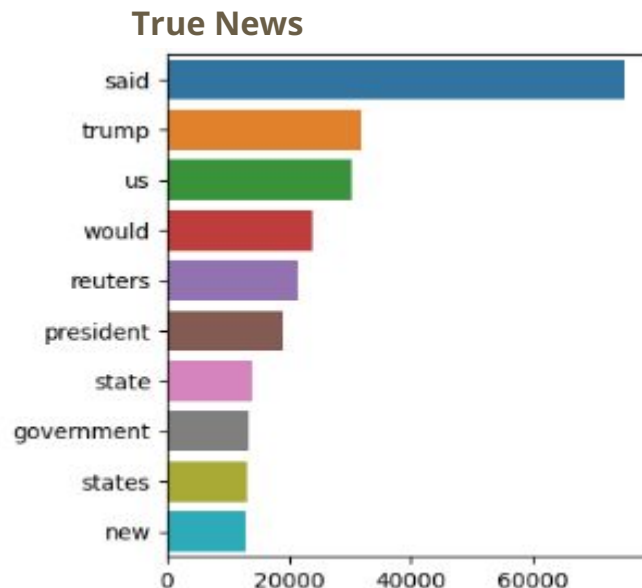
By Date

Rows of training set: 19690
Rows of validation set: 494
Rows of test set: 494

Distribution of Target



Most Common Unigrams in True and Fake News



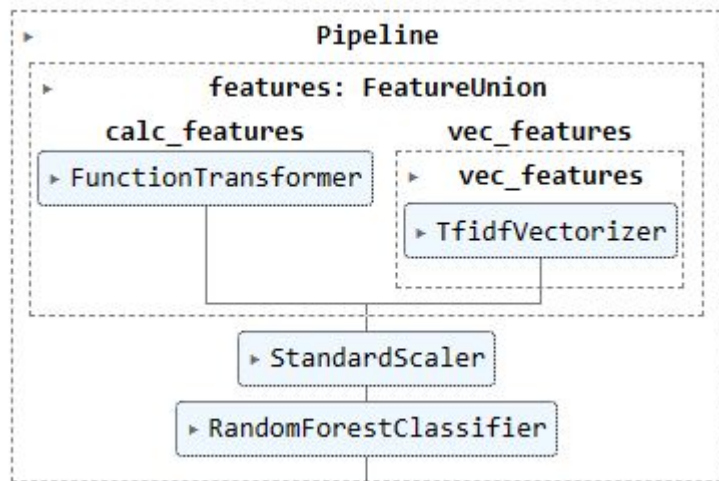
Most common words in both corpuses:

said, trump, us, would, president, people, reuters, one, state, new, also, states, house, government, donald, republican, could, united, told, clinton, obama, white, campaign, last, election, two, like, party, time, year

Analysis of Text Structure

Fake News	True News
Higher proportion of uppercase letters to grab attention	Consistent average word length
Higher variance in length	Slightly longer
Higher frequency of exclamation marks	Longer words - more sophisticated language
More outliers	Higher frequency of dots

Feature Engineering



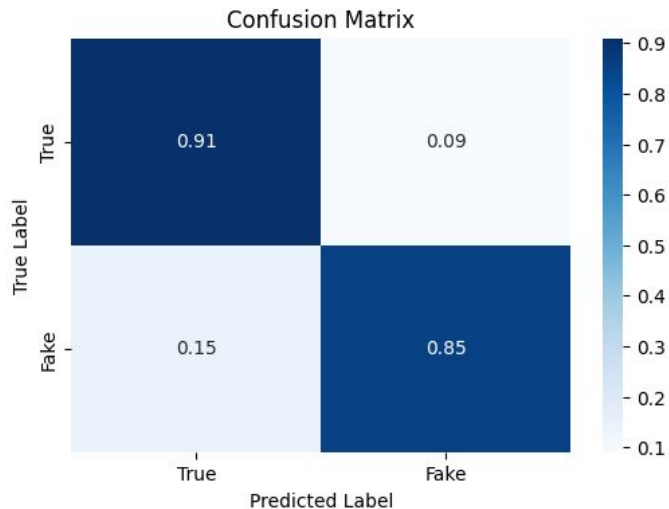
calculated features:

- Proportion of upper case letters
- Average word length
- Number of characters in news
- Count of words
- Proportion of punctuation
- Proportion of exclamation marks
- Proportion of dots

vectorized words:

- Said
- Would
- People
- One
- Also
- Could
- Last
- Two
- Like
- Time
- Year

Modeling with Algorithms from Sklearn

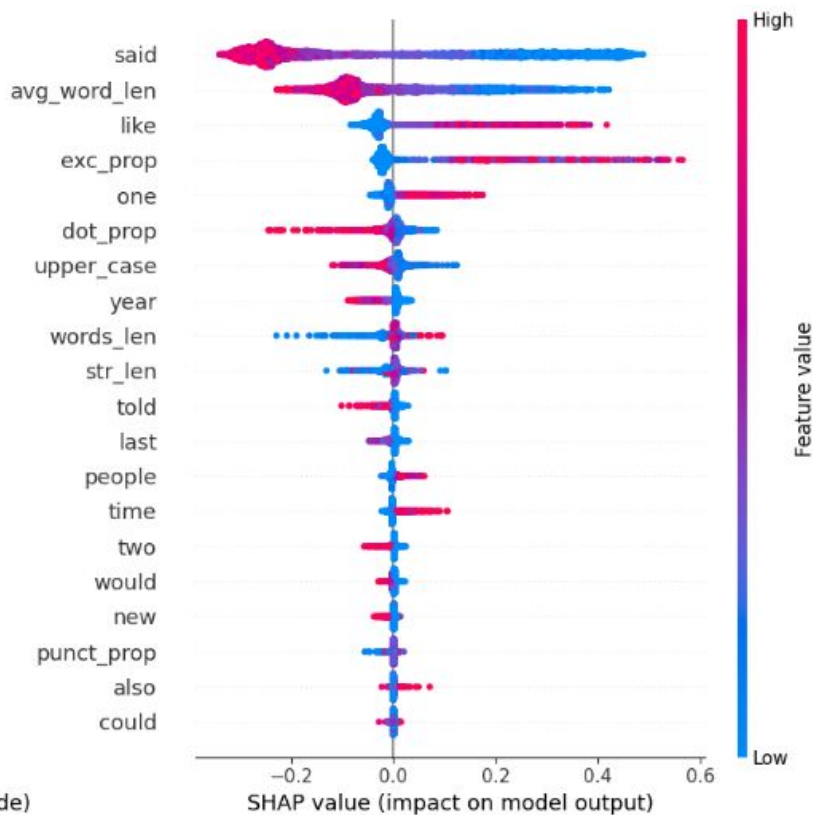
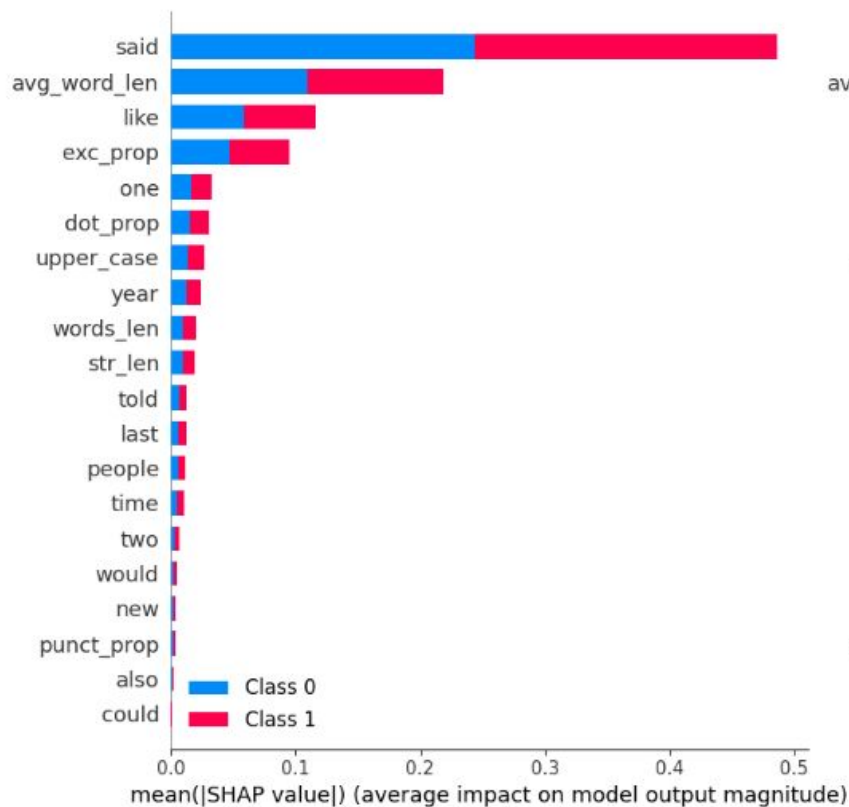


Classification Report:

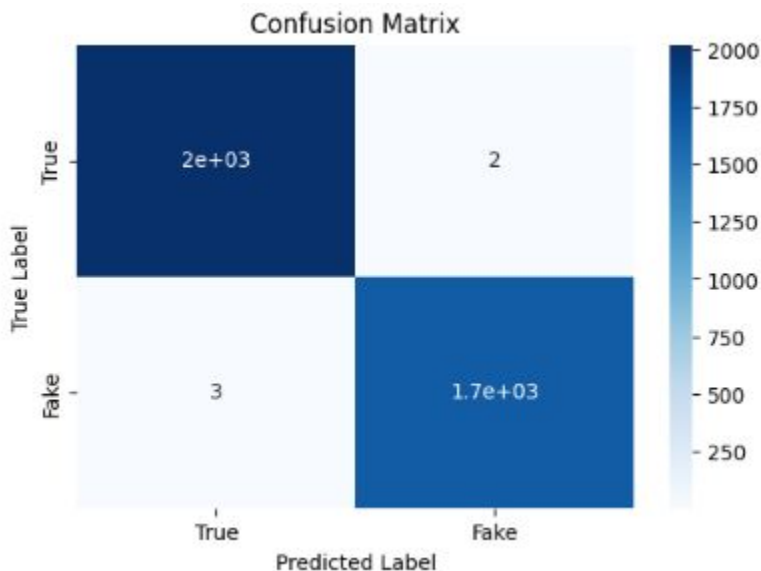
	precision	recall	f1-score	support
True	0.88	0.91	0.90	2023
Fake	0.89	0.85	0.87	1678
accuracy			0.88	3701
macro avg	0.88	0.88	0.88	3701
weighted avg	0.88	0.88	0.88	3701

- Best performing model - RandomForestClassifier
- Weighted f1_score on calculated features 0.77
- Weighted f1_score using all features on test set 0.88
- Adding more vectorized words didn't improve the outcome

Interpretation of Random Forest Classifier Predictions



Modeling *title* with Pretrained distilbert-base-cased



Classification Report:

	precision	recall	f1-score	support
True	1.00	1.00	1.00	2023
Fake	1.00	1.00	1.00	1678
accuracy			1.00	3701
macro avg	1.00	1.00	1.00	3701
weighted avg	1.00	1.00	1.00	3701

- The model was mistaken only 0.1% of the time.
- Training reached accuracy 1.00 on validation set after 1 epoch.

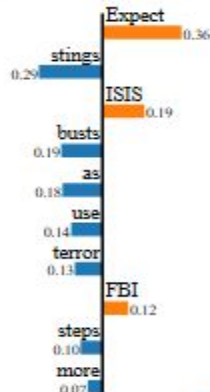
Examples of Wrongly Predicted Titles

Prediction probabilities



True

False



Text with highlighted words

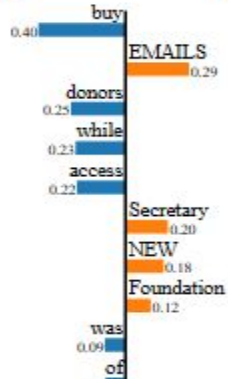
Expect more 'terror busts' as FBI steps up its use of "ISIS stings"

Prediction probabilities



True

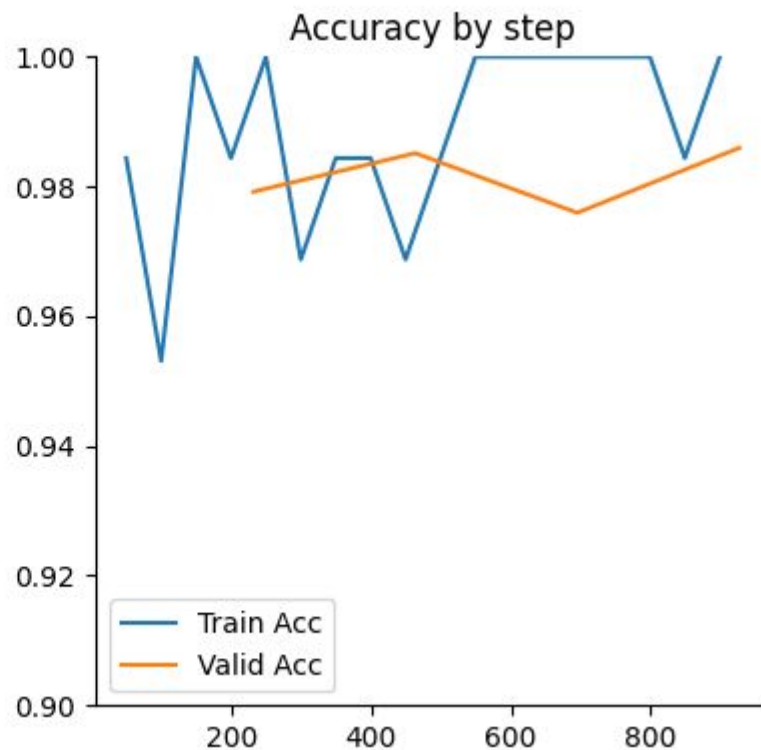
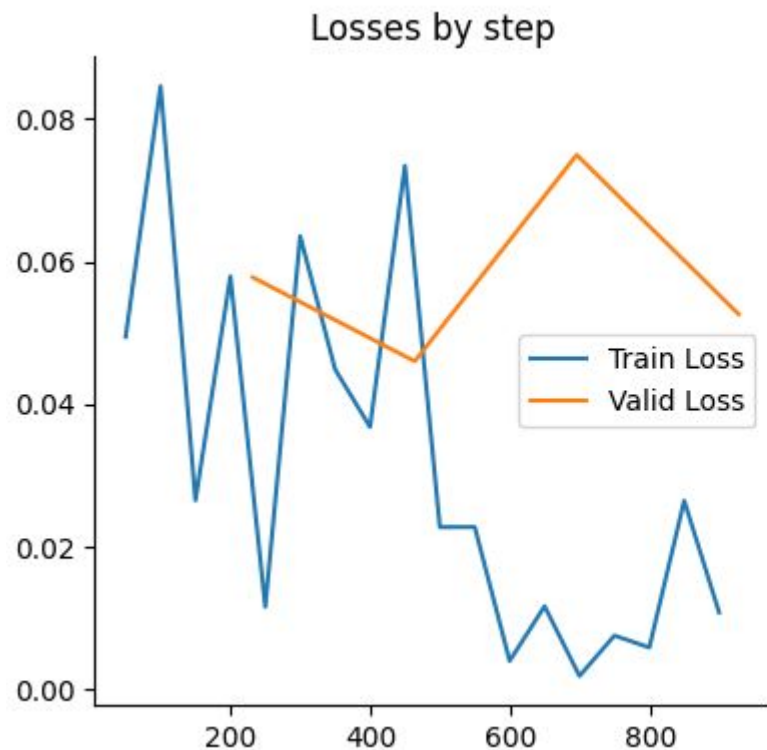
False



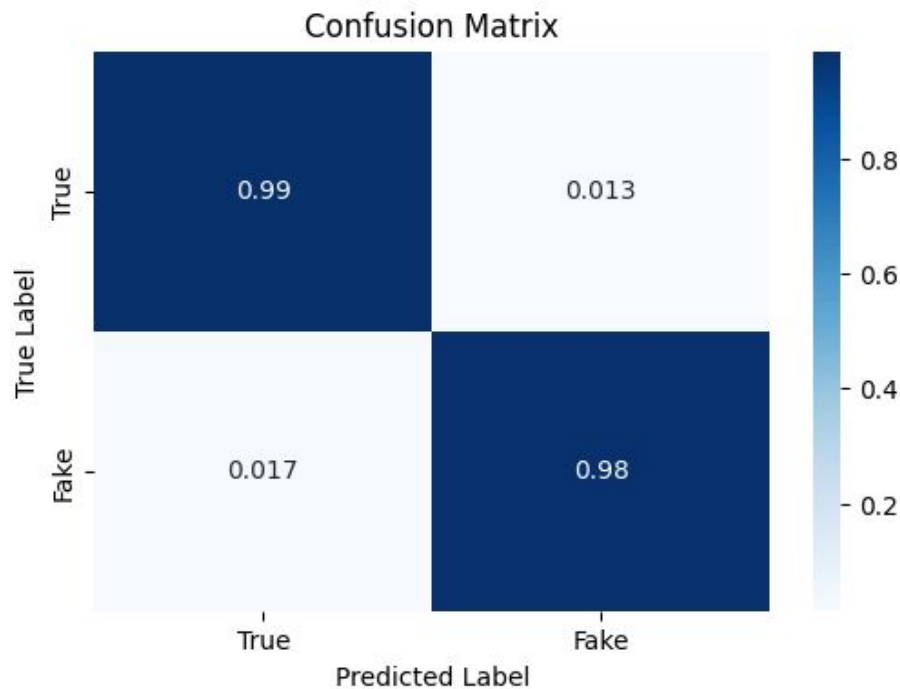
Text with highlighted words

NEW EMAILS: Clinton Foundation VIP donors buy access – while Hillary was Secretary of State

Modeling *text* with Pretrained distilbert-base-cased



Performance of the model



Classification Report:

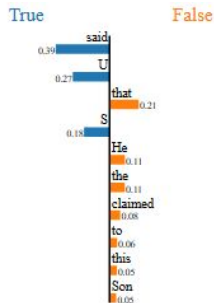
	precision	recall	f1-score	support
True	0.99	0.99	0.99	2023
Fake	0.98	0.98	0.98	1678
accuracy			0.98	3701
macro avg	0.98	0.98	0.98	3701
weighted avg	0.98	0.98	0.98	3701

The model is wrong every 3 in 100 news.

Problems with correctly classifying fake news

Examples of Wrong Predictions

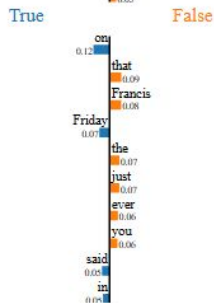
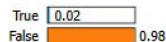
Prediction probabilities



Text with highlighted words

Donald Trump **said** Monday **that** Japan SoftBank has agreed to invest **billion** in **the** **U.S.** aiming to create **jobs**. **He** announced **the** deal after meeting with SoftBank CEO Masayoshi Son Japanese billionaire and technology investor at Trump Tower in New York. On Twitter Trump **claimed** **that** Son **said** he would never do this if Trump had not won **the** presidential election. **But** **the** billion will come from previously announced billion international investment fund set up with Saudi Arabia according to Dow Jones. Via cnbc SO MUCH WINNING! Trump secures million investment from Japan SoftBank Group in US! Wahoo

Prediction probabilities



Text with highlighted words

VATICAN CITY Reuters If **you** **ever** wondered how to say fake news **in** Latin it nuntii fallaces and Pope **Francis** is writing document **on** **just** **that** **Francis** announced it himself **in** tweet to his nearly million followers **on** **Friday** saying **the** theme of his message for **the** Roman Catholic Church next World Day of Social Communications will be The truth will set you free. Fake news and journalism for peace. In Latin one of **the** nine languages **the** pope uses to tweet **that** would be Veritas liberavit vos. Nuntii fallaces et diurniariorum opus ad pacem. Vatican statement **said** **the** issue was important enough for **the** pope to address because fake news contributes to generating and nurturing strong polarisation of opinions. distortion of facts it **said** can have repercussions at **the** level of individual and collective behaviour. **Francis** **said** **in** an interview last year **that** media **that** focus **on** scandals and spread fake news to smear politicians risk becoming like people who have morbid fascination with excrement. The Vatican statement **said** **the** leader of **the** billion member Church wanted to offer reflection **on** **the** causes **the** logic and **the** consequences of disinformation **in** **the** media and helping to promote professional journalism which always seeks **the** truth The Church next World Day of Social Communications will be celebrated **on** May. The pope message will be released **on** January feast of St **Francis** de Sales **the** patron of journalists. This version of **the** story corrected day of Day of Social Communications to May from Jan.

Conclusions

- Models rely on writing style of the writer
- The structure of title seems to be the most important factor of predictions
- True news have more reporting style of writing
- The size and performance of classic ML and NN models differs by 0.1 in weighted f1 score
- The source of dataset is not known.
- Estimating model performance on examples that are consistent with business are crucial
- Deployment of the model could depend on business integrity, ethics and financial resources

Improvements

- Checking for features that can leak data about the predictions.
- Further feature engineering and optimization of model hyperparameters.
- Estimating the performance of different sets of data to establish if models are performing as good as in the notebook.
- Experimenting with uncased models.
- Building LSTM model which would be smaller but hopefully performed similar.
- Estimating time of getting results and cost of deployment.