

FAKE NEWS DETECTION

ANASTASIIA HRYTSYNA
JANUARY, 2022



5
columns

44,9
thousand
entries

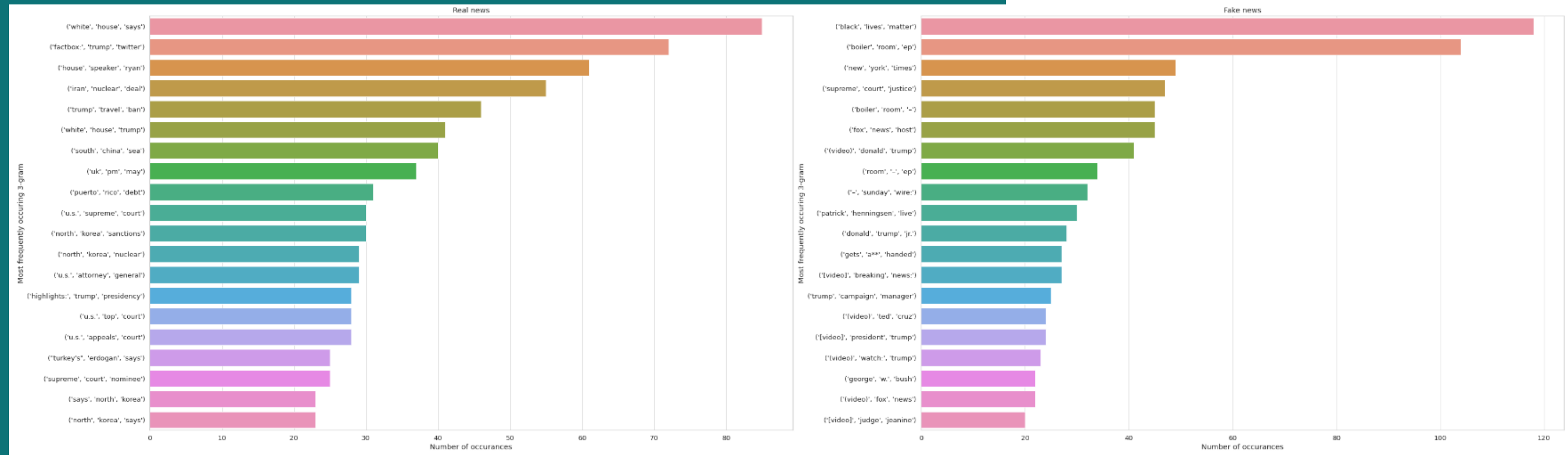
50/50
target
distribution

3
different
methods

Data

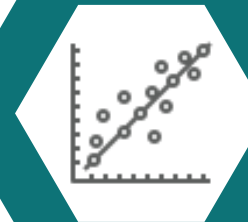
Data preprocessing: combined title and main text, removed stop words from the content, links and punctuation, changed some of the short word forms to the long ones, dates encoding.

EDA



In most cases fake news contains longer titles and text, using more complicated words. In addition, I checked the most popular common words for both real and fake classes (said, trump, us, would, president) and analyzed 3-grams for each class.

1 method



Logistic Regression

49% accuracy 46% precision



Naive Bayes Classifier

48% accuracy 47% precision



Random Forest

51% accuracy 67% precision

2 method

LSTM

98% accuracy

97% precision

98% recall

98% F1-score

0.98

0.03

0.016

0.97

Confusion matrix

3 method

BERT

52% accuracy 52% precision
100% recall 68% F1-score

EPOCH 1/10	train loss 0.702	val loss 0.701
EPOCH 2/10	train loss 0.698	val loss 0.701
EPOCH 3/10	train loss 0.703	val loss 0.701
EPOCH 4/10	train loss 0.703	val loss 0.701
EPOCH 5/10	train loss 0.705	val loss 0.701
EPOCH 6/10	train loss 0.702	val loss 0.701
EPOCH 7/10	train loss 0.703	val loss 0.701
EPOCH 8/10	train loss 0.707	val loss 0.701
EPOCH 9/10	train loss 0.701	val loss 0.701
EPOCH 10/10	train loss 0.700	val loss 0.701

The image features a solid teal background. In the center is a white hexagon with a thick teal border. The word "QUESTIONS" is written in a bold, dark grey, sans-serif font across the middle of the hexagon. The corners of the image are decorated with geometric shapes: teal triangles pointing inwards and grey triangles pointing outwards, creating a layered, architectural feel.

QUESTIONS