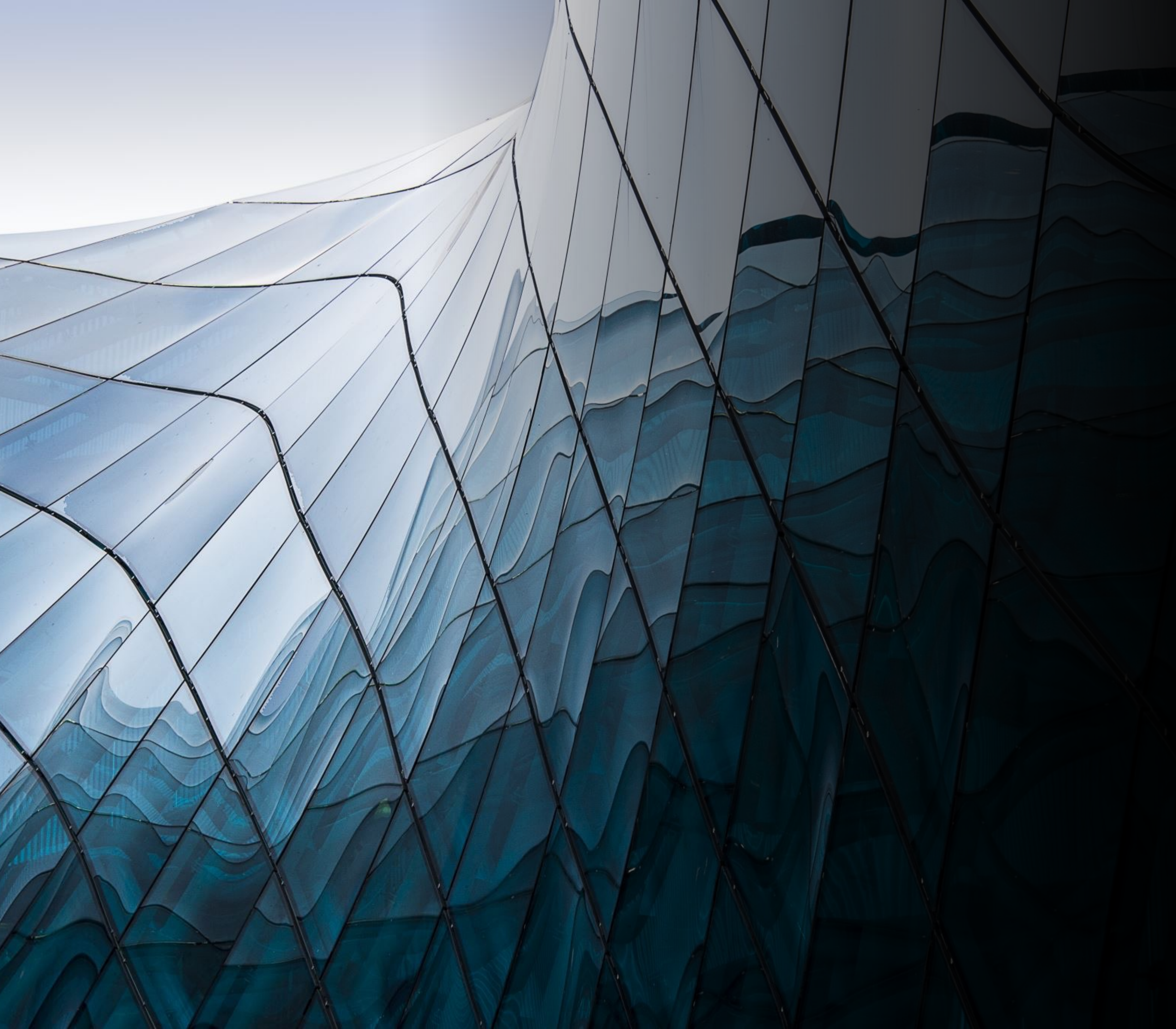


Text Classification of Misinformation with PySpark

William Simpson

DATA 603 – Platforms for Big Data Processing

Evidence Surfaces That The FBI Planned And Executed January 6 Capitol Riot

Desperate, angry, destructive: How Americans morphed into a mob

Research Question

Can factual news be distinguished from fake news by a machine learning model based purely on the words in a text and their frequency?

This is a Big Data Problem

- The volume and spread of content online (i.e., news sites and social media) enable misinformation and disinformation to proliferate on a massive scale.

- How do we begin to sift through this amount of information to determine what is factual and what is not?

- Need Big Data tools!

The original data

```
+-------------------+-------------------+---------------+------------------+------+
|              title|               text|        subject|              date|target|
+-------------------+-------------------+---------------+------------------+------+
|Trump to scrap pr...|WASHINGTON (Reute...|    politicsNews| September 4, 2017 |     1|
| BUSTED: Trump’s ...|It turns out that...|           News| December 18, 2016|     0|
|White House eyein...|WASHINGTON (Reute...|    politicsNews| November 14, 2017 |     1|
|Message to Presid...|21st Century Wire...|     Middle-east| February 10, 2017|     0|
|“GOOD-BYE SWEDEN”...|This blogger s pi...|Government News|      Nov 11, 2015|     0|
|MAXINE WATERS Gle...|Rep. Maxine Water...|        politics|       Aug 6, 2017|     0|
|[VIDEO] THEY BURN...|Barack and Michel...|       left-news|      Feb 20, 2016|     0|
|Obama vetoes Sept...|WASHINGTON (Reute...|    politicsNews|September 23, 2016 |     1|
|Return of defeate...|MOSCOW (Reuters) ...|       worldnews| December 12, 2017 |     1|
| Fox News Finally...|Donald Trump just...|            News|    March 20, 2017|     0|
|Republican Senato...|WASHINGTON (Reute...|    politicsNews|  December 1, 2017 |     1|
|LIST OF 22 TIMES ...|Oh the irony of a...|Government News|       Dec 3, 2015|     0|
|SUPREME COURT JUS...|What the heck is ...|        politics|      Jul 10, 2016|     0|
|Republican Collin...|WASHINGTON (Reute...|    politicsNews|    April 5, 2016 |     1|
|Senate leader McC...|WASHINGTON (Reute...|    politicsNews| December 22, 2017 |     1|
|Rival Tuaregs sig...|BAMAKO (Reuters) ...|       worldnews|September 21, 2017 |     1|
|Obama says he res...|WASHINGTON (Reute...|    politicsNews|    June 24, 2016 |     1|
|Trump signs execu...|WASHINGTON (Reute...|    politicsNews|  January 24, 2017 |     1|
|Korean 'comfort w...|TOKYO (Reuters) -...|       worldnews| December 19, 2017 |     1|
| WATCH: President...|President  Zero F...|            News|September 16, 2016|     0|
+-------------------+-------------------+---------------+------------------+------+
```

Source:
Kaggle – Fake and real news dataset

Shape:
38,729 rows x 4 columns

NOTE: After random shuffling, only 10,000 records were included in this analysis due to memory limitations that prevented the pipeline from running on the entire dataset.

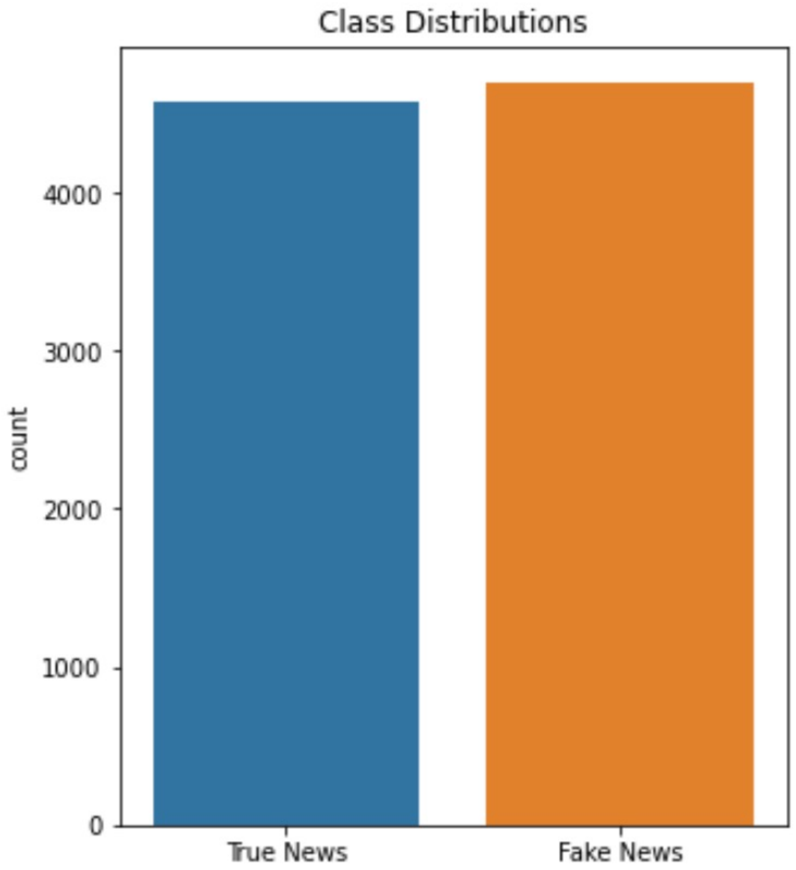
Steps in cleaning data

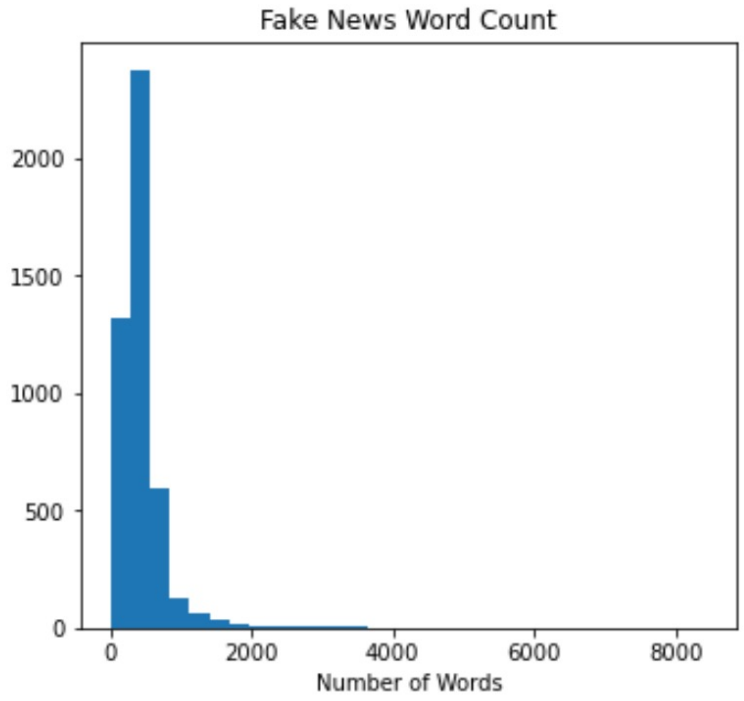
1) Drop missing values

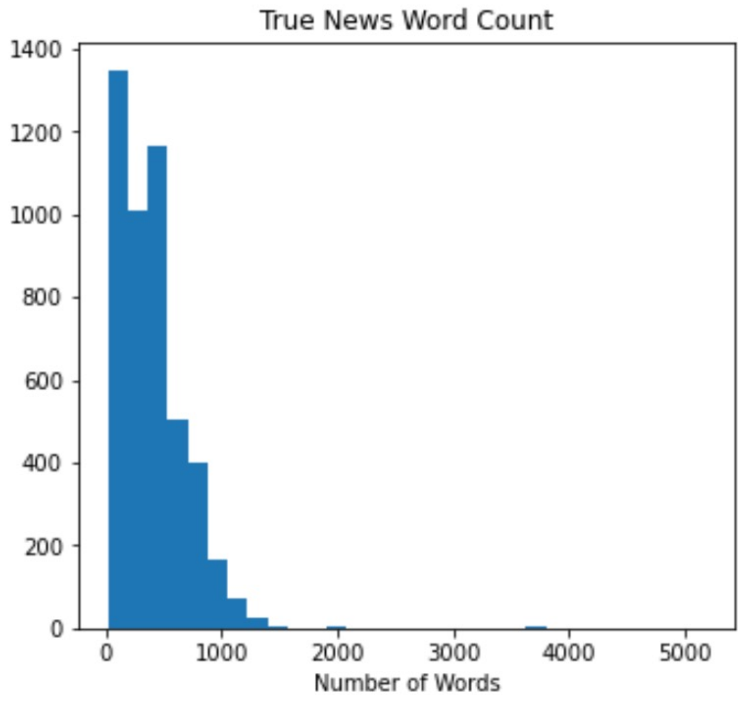
```
+-----+----+-------+----+------+
|title|text|subject|date|target|
+-----+----+-------+----+------+
|    0| 151|      0|   0|     0|
+-----+----+-------+----+------+
```

1) Remove duplicate records (*total of 298*)
2) Reformat the date to a datetime object

Exploratory Data Analysis








```
Fake maximum text length: 8435
Fake mean text length: 435.1392958670457

True maximum text length: 5174
True mean text length: 393.3552098870658
```

Clean the text

- Remove text that contained only Twitter metadata or website urls
- Remove "(Reuters)" that was present at beginning of all *true* documents

```
# remove punctuation, extra whitespace, special characters
def clean_text(doc):
    doc = re.sub(r'.*\(Reuters\)\s\-', '', doc)  # remove (Reuters) that exists at beginning of true news record
    doc = re.sub(r'[^\w\s]', '', doc.lower().strip())  # remove any character except word and space characters
    return doc

clean_text_udf = udf(clean_text, StringType())

all_news = all_news.withColumn('clean_text', clean_text_udf(col('text')))
all_news.show()
```

Meaningful word tokens

- Tokens obtained using *Tokenizer*
- Stop words removed using *StopWordsRemover*

clean_text	tokens	tokens_no_stopwords
donald trump is b...	[donald, trump, i...	[donald, trump, g...
according to an o...	[according, to, a...	[according, open,...
florida continues...	[florida, continu...	[florida, continu...
many americans ha...	[many, americans,...	[many, americans,...
while it s conven...	[while, it, s, co...	[convenient, ster...
north carolina go...	[north, carolina,...	[north, carolina,...
cops in america a...	[cops, in, americ...	[cops, america, c...
the united states...	[the, united, sta...	[united, states, ...
kneel before trum...	[kneel, before, t...	[kneel, trump, , ...
donald trump hate...	[donald, trump, h...	[donald, trump, h...
a trio of neonazi...	[a, trio, of, neo...	[trio, neonazi, c...
al franken return...	[al, franken, ret...	[al, franken, ret...
republicans in al...	[republicans, in,...	[republicans, ala...
sean spicer got o...	[sean, spicer, go...	[sean, spicer, go...
alec baldwin didn...	[alec, baldwin, d...	[alec, baldwin, d...
after this past w...	[after, this, pas...	[past, week, days...
the saga surround...	[the, saga, surro...	[saga, surroundin...
women especially ...	[women, especiall...	[women, especiall...
just when you tho...	[just, when, you,...	[thought, ann, co...
it s an epidemic ...	[it, s, an, epide...	[epidemic, inside...

Create the feature vector

- Convert clean text to a numerical representation to be used by the logistic regression classifier
  - *CountVectorizer*: to obtain term frequencies
  - *IDF (Inverse Document Frequency)*: to calculate the relative importance of tokens compared to their frequencies across all documents in our collection
- The resulting vectors contain **83,595 features**.

NOTE: text preprocessing and text featurization steps were encapsulated in a single pipeline for modeling

Run the Logistic Regression

- Default parameters were used for the model
- Training (80%), Test (20%)

```
# instantiate logistic regression model
lr = LogisticRegression(labelCol='target', featuresCol='features')
# fit model
lr_model = lr.fit(train_news)
# make predictions
target_pred = lr_model.transform(test_news)
```

Evaluate model performance

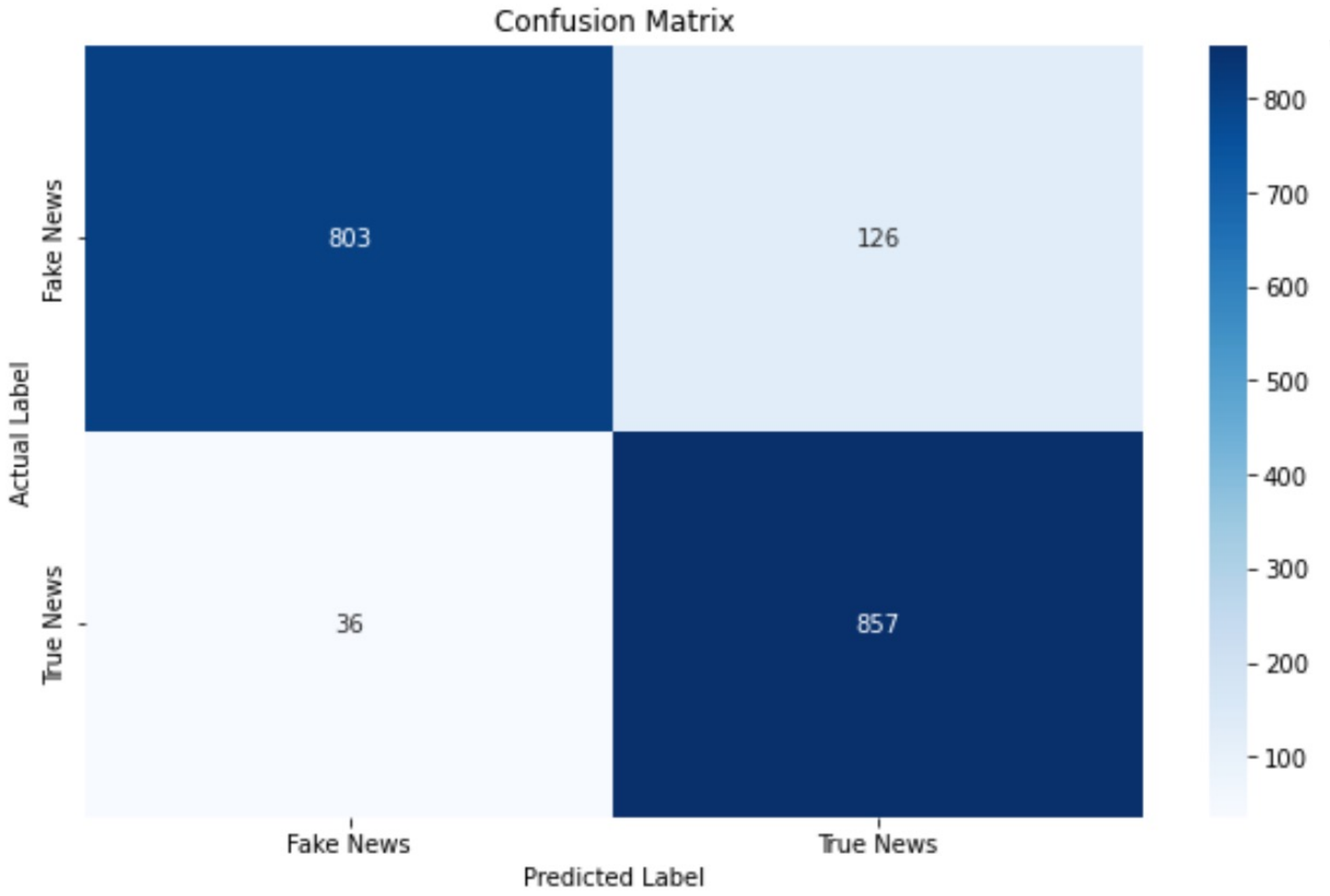
- Two methods of evaluating the model's performance
  - Built-in *AreaUnderROC* metric from *BinaryClassificationEvaluator*
  - Manual calculation of accuracy

```
# calculate accuracy
accuracy = target_pred.filter(col('target') == col('prediction')).count() / float(test_news.count())
print('Accuracy for Logistic Regression Model:', accuracy)
```

Results

Evaluation method	Value
Area Under Curve (AUC)	0.963
Accuracy	0.911

Results – Confusion Matrix



Results – Feature Coefficients (Top 5)

coefficients	featureName
{-12.127850097677...	yearbringing
{-12.127850097677...	expeditions
{-12.127850097677...	tock
{-12.127850097677...	inkling
{-12.127850097677...	watchseanhannity

Words associated with **fake** news

coefficients	featureName
{7.303274381425323}	leaner
{6.212181171375291}	1300
{6.138760758133321}	katya
{6.019484064105074}	abbate
{5.983340949032431}	hotheads

Words associated with **true** news

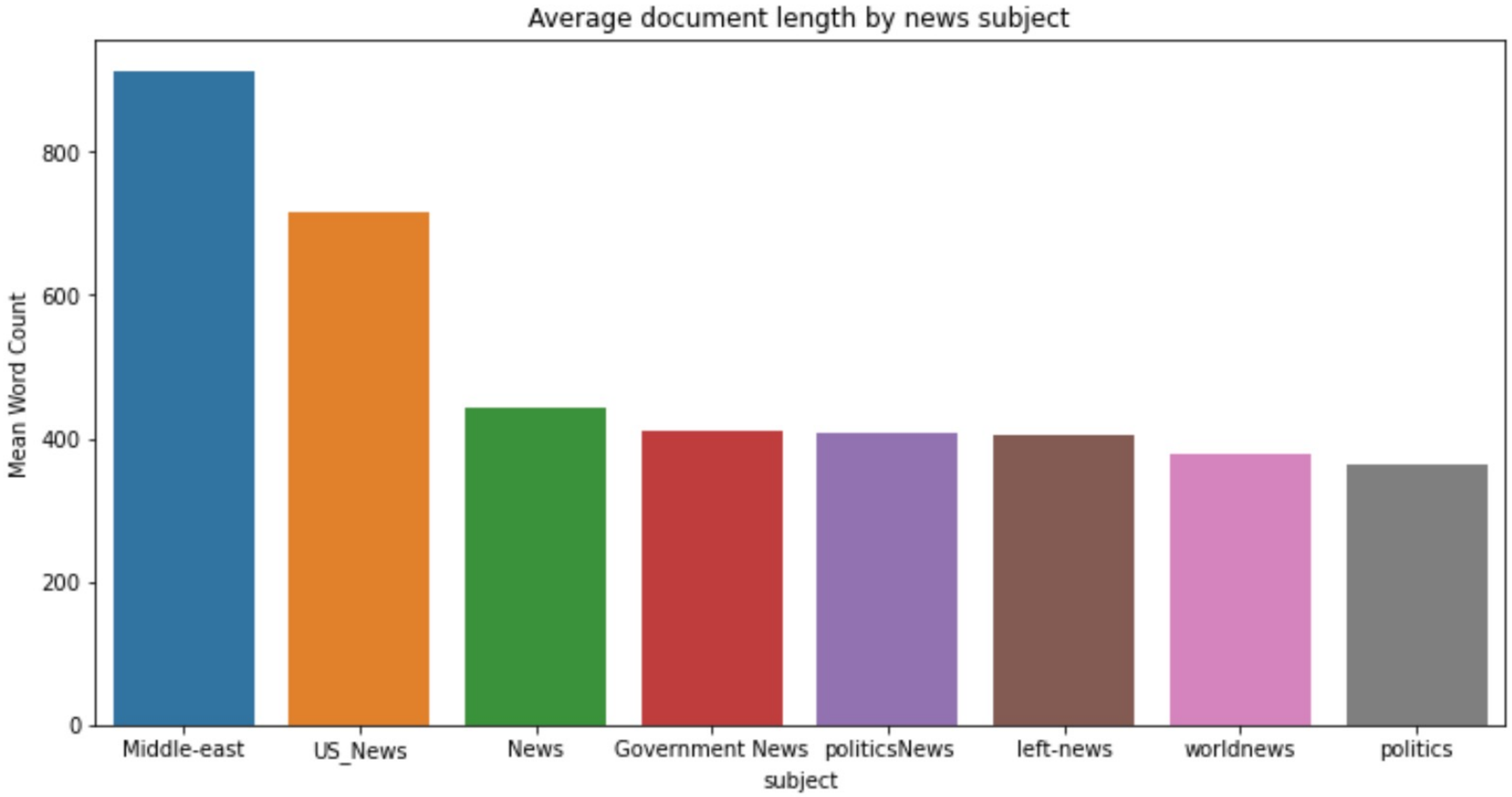
Conclusions

- **High accuracy (>90%) →** plausible to distinguish fake news based only on the words in the text and their frequency.
- **However**, coefficients of individual tokens does not reveal especially meaningful patterns in classifications.
- Therefore, question whether this specific model would truly be generalizable to novel datasets.

References

- https://tatumreport.com/evidence-surfaces-fbi-planned-executed-january-6-capitol-riot/
- https://www.washingtonpost.com/dc-md-va/2021/11/09/rioters-charges-arrests-jan-6-insurrection/
- https://www.scientificamerican.com/article/information-overload-helps-fake-news-spread-and-social-media-knows-it/
- https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset?select=Fake.csv

Additional EDA Figures



Additional EDA Figures