NEWS HEADLINE IS REAL OR FAKE NEWS

NLP CHALLENGE

Tiago Santos

DATA SET

Contains News headline

To create a classifier to predict labels

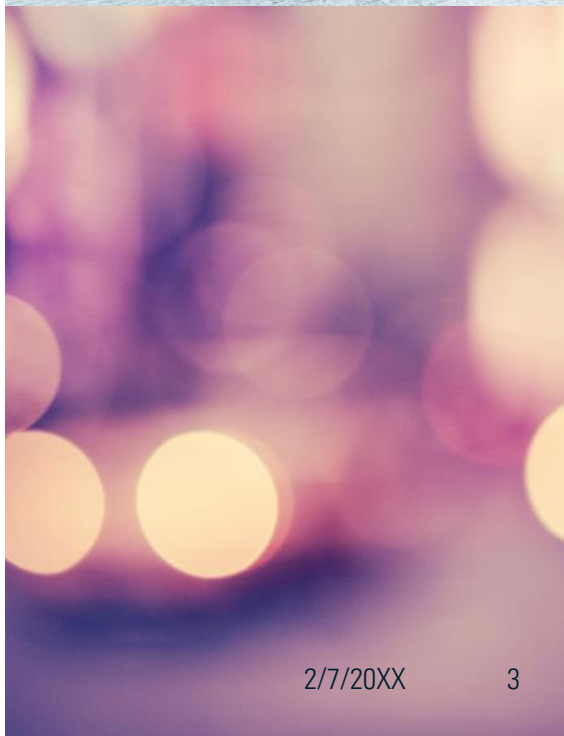
- is the headline real (1)

or

- fake news (0)

TEXT TREATMENT

- ✓ Eliminate “\t”
- ✓ Only use Labels scores “0” and “1”
- ✓ Standard columns
- ✓ Create columns names
  - ✓ Classification
  - ✓ News
- ✓ Cleaning (words)
- ✓ Cleaning Bad words



EDA

```
classification     int64
news              object
dtype: object
```

```
classification
0    17571
1    16580
```

```
classification
0    0.514509
1    0.485491
```

	classification	news
0	0	drunk bragging trump staffer started russian c...
1	0	sheriff david clarke becomes an internet joke ...
2	0	trump is so obsessed he even has obama,s name ...
3	0	pope francis just called out donald trump duri...
4	0	racist alabama cops brutalize black boy while ...
...	...	...
34146	1	tears in rain as thais gather for late king's ...
34147	1	pyongyang university needs non-u.s. teachers a...
34148	1	philippine president duterte to visit japan ah...
34149	1	japan's abe may have won election\tbut many do...
34150	1	demoralized and divided: inside catalonia's po...

34151 rows × 2 columns

TEXT PROCESSING

- Tokenization With punkt_tab
  - create a new column
- From Tokenization create Pos-tags
  - create a new column
- From Tokenization create stemmed
  - create a new columnFrom
- From stemmed create lemmatized
  - create a new columnFrom
- Apply StopWords
- Bag-of-Words vectorization (feature extraction)

MODEL

```
Training data class

classification
0    14054
1    13266
```

```
Test data class

classification
0    3517
1    3314
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.91      0.90      3517
           1       0.90      0.89      0.90      3314

    accuracy                           0.90      6831
   macro avg       0.90      0.90      0.90      6831
weighted avg       0.90      0.90      0.90      6831
```

- Use Train Test Split
  - 80%, 20%
- Model = RandomForestClassifier

TRY ANOTHER MODEL

```
Training data class

classification
0    14054
1    13266
```

```
Test data class

classification
0    3517
1    3314
```

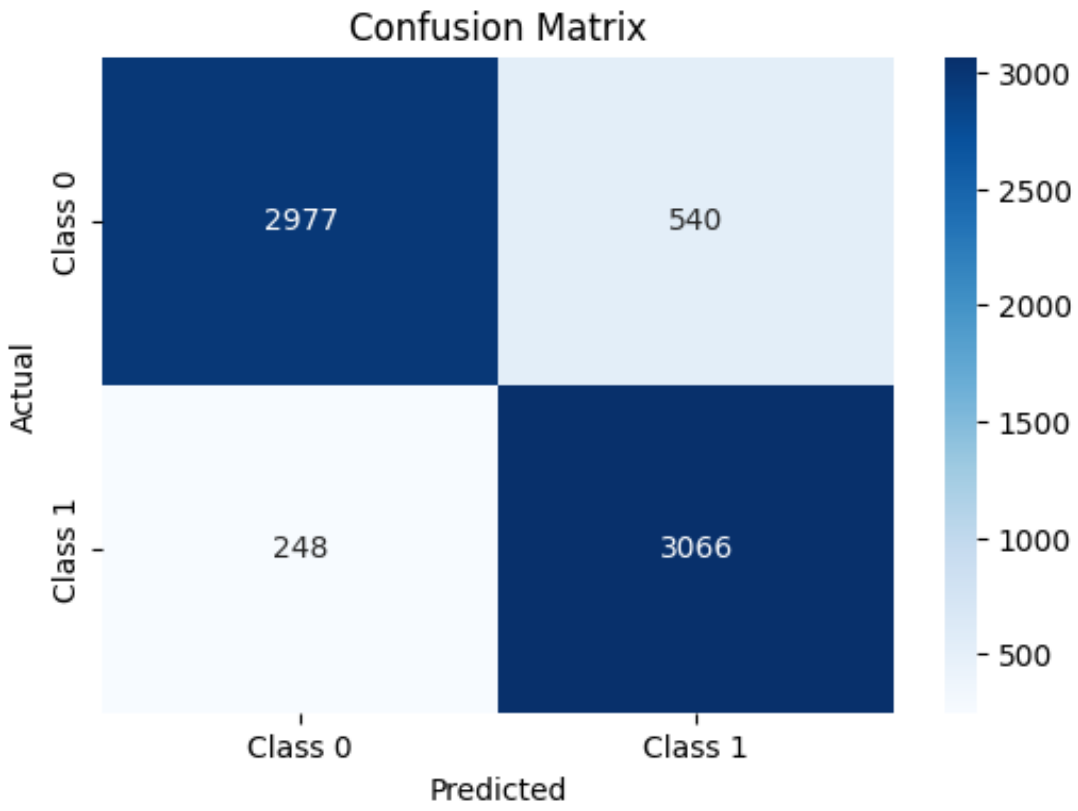
```
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.73      0.81      3517
           1       0.76      0.93      0.84      3314

    accuracy                           0.82      6831
   macro avg       0.84      0.83      0.82      6831
weighted avg       0.84      0.82      0.82      6831
```

- Use Train Test Split
  - 80%, 20%
- Model = GradientBoostingClassifier

CONFUSION MATRIX



Right Predictions

- Fake News 84.6% (-15.3%)
- Real News 92.5% (-7.05%)

THANK YOU