NLP WITH REDDIT

Maithili Joshi

SUBREDDIT PAGES

'Pros And Cons Of Making Birth Control Available Over The Counter'

'elliott abrams defends war crimes as happening back in the ’80s when everyone was doing it'

'Plant-Based Meat Sales Rise, Fueled by Carnivores'

'cast of cats film to be size of actual cats'

THE DATA SCIENCE QUESTION

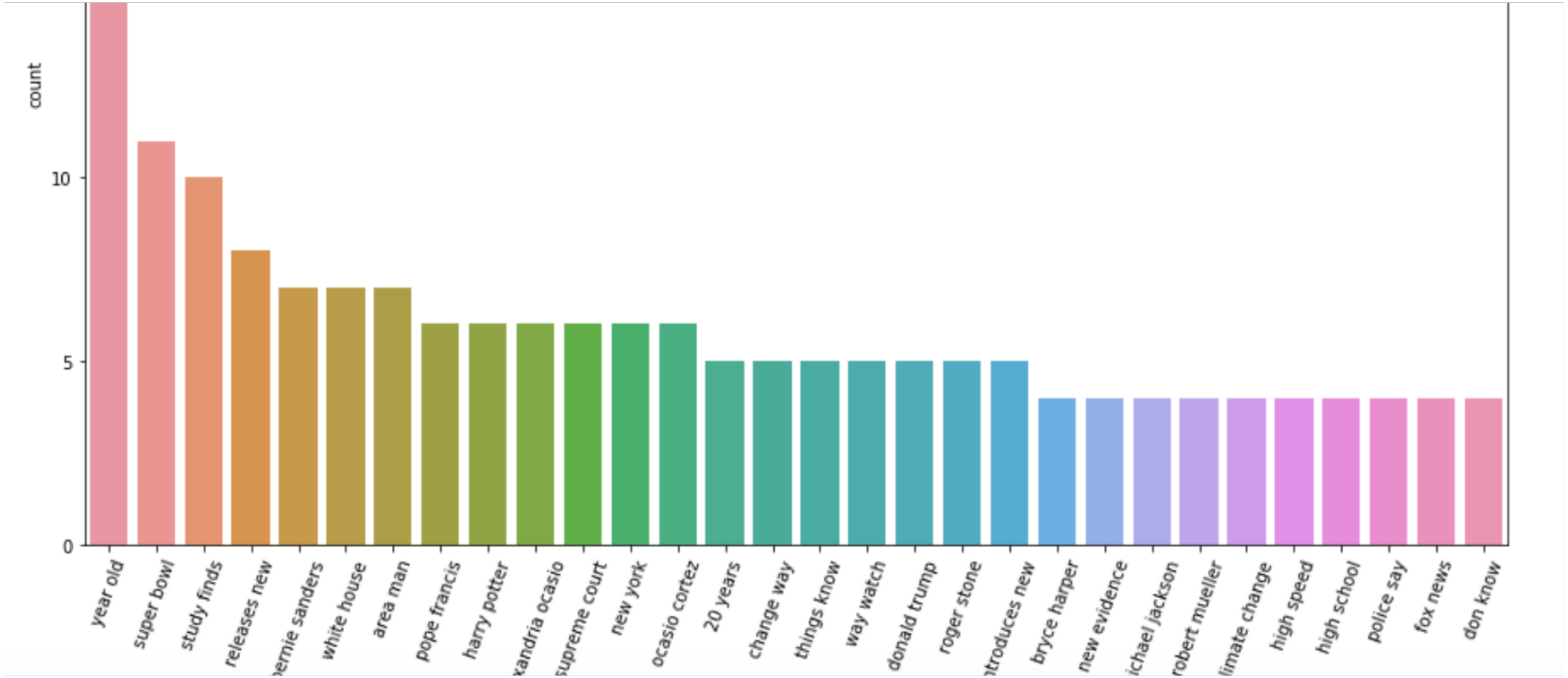
- I want to know if there is any validity to the subreddit "Not the Onion"

THE DATA SCIENCE QUESTION

- I want to know if there is any validity to the subreddit "Not the Onion"
- Are these articles REALLY all so bad that they sound fake?

THE DATA SCIENCE QUESTION

- I want to know if there is any validity to the subreddit "Not the Onion"
- Are these articles REALLY all so bad that they sound fake?
- 0 = The Onion
- 1 = Not the Onion


count
10
5
0
year old
super bowl
study finds
releases new
white house
area man
pope francis
harry potter
supreme court
new york
ocasio cortez
20 years
change way
things know
way watch
donald trump
roger stone
bryce harper
new evidence
high speed
high school
police say
fox news
don know

975
400

0	0.709091
1	0.290909

THE DATA

400 points from Not the Onion versus 975

975
400

0	0.709091
1	0.290909

THE DATA

400 points from Not the Onion versus 975

Unbalanced model

```
975
400
```

```
0    0.709091
1    0.290909
```

THE DATA

400 points from Not the Onion versus 975

- Unbalanced model

We have a baseline score of 70.9% that our models have to beat.

MODELING

Logistic Regression with Ridge

Lasso

Naïve Bayes w/ CV (Mult)

Naïve Bayes w/ TF-IDF

Random Forest w/ CV

Random Forest with TF-IDF

```
CV score LogReg Lasso: 0.7982
Training accuracy LogReg Lasso: 0.9874
Testing accuracy LogReg Lasso: 0.811
-------------------------------------------------
CV score LogReg Ridge: 0.7982
Training accuracy LogReg Ridge: 0.9874
Testing accuracy LogReg Ridge: 0.811
-------------------------------------------------
CV score Naive Bayes & Count Vectorizer(Mult): 0.3928286665728625
Training accuracy Naive Bayes & Count Vectorizer(Mult): 0.9990300678952473
Testing accuracy Naive Bayes & Count Vectorizer (Mult): 0.7790697674418605
-------------------------------------------------
CV score Naive Bayes & TF-IDF (Mult): 0.7031893438394071
Training accuracy Naive Bayes & TF-IDF (Mult): 0.8816682832201745
Testing accuracy Naive Bayes & TF-IDF (Mult): 0.7441860465116279
-------------------------------------------------
CV score Random Forest & Count Vectorizer: 0.773
Training accuracy Random Forest & Count Vectorizer: 0.9941804073714839
Testing accuracy Random Forest & Count Vectorizer: 0.7733
-------------------------------------------------
CV score Random Forest & TF-IDF: 0.7575
Training accuracy Random Forest & TF-IDF: 0.988360814742968
Testing accuracy Random Forest & TF-IDF: 0.7674418604651163
```

MODELING

Logistic Regression with Ridge

Lasso

Naïve Bayes w/ CV (Mult)

Naïve Bayes w/ TF-IDF

Random Forest w/ CV

Random Forest with TF-IDF

```
CV score LogReg Lasso: 0.7982
Training accuracy LogReg Lasso: 0.9874
Testing accuracy LogReg Lasso: 0.811
-----------------------------------------------
CV score LogReg Ridge: 0.7982
Training accuracy LogReg Ridge: 0.9874
Testing accuracy LogReg Ridge: 0.811
-----------------------------------------------
CV score Naive Bayes & Count Vectorizer(Mult): 0.3928286665728625
Training accuracy Naive Bayes & Count Vectorizer(Mult): 0.9990300678952473
Testing accuracy Naive Bayes & Count Vectorizer (Mult): 0.7790697674418605
-----------------------------------------------
CV score Naive Bayes & TF-IDF (Mult): 0.7031893438394071
Training accuracy Naive Bayes & TF-IDF (Mult): 0.8816682832201745
Testing accuracy Naive Bayes & TF-IDF (Mult): 0.7441860465116279
-----------------------------------------------
CV score Random Forest & Count Vectorizer: 0.773
Training accuracy Random Forest & Count Vectorizer: 0.9941804073714839
Testing accuracy Random Forest & Count Vectorizer: 0.7733
-----------------------------------------------
CV score Random Forest & TF-IDF: 0.7575
Training accuracy Random Forest & TF-IDF: 0.988360814742968
Testing accuracy Random Forest & TF-IDF: 0.7674418604651163
```

```
--------------------------------------------------
CV score Naive Bayes & TF-IDF (Mult): 0.7031893438394071
Training accuracy Naive Bayes & TF-IDF (Mult): 0.8816682832201745
Testing accuracy Naive Bayes & TF-IDF (Mult): 0.7441860465116279
```

BEST PERFORMING MODEL

mexico border	-5.102845
spicy chips	-5.102845
ar 15	-5.102845
supreme court	-5.030111
police say	-4.999865
russian man	-4.952176
year old	-4.355409

words	coef
placing new	-8.147367
powerpoint company	-8.147367
powerpoint opened	-8.147367
pr nightmare	-8.147367
pr stunt	-8.147367
precious time	-8.147367

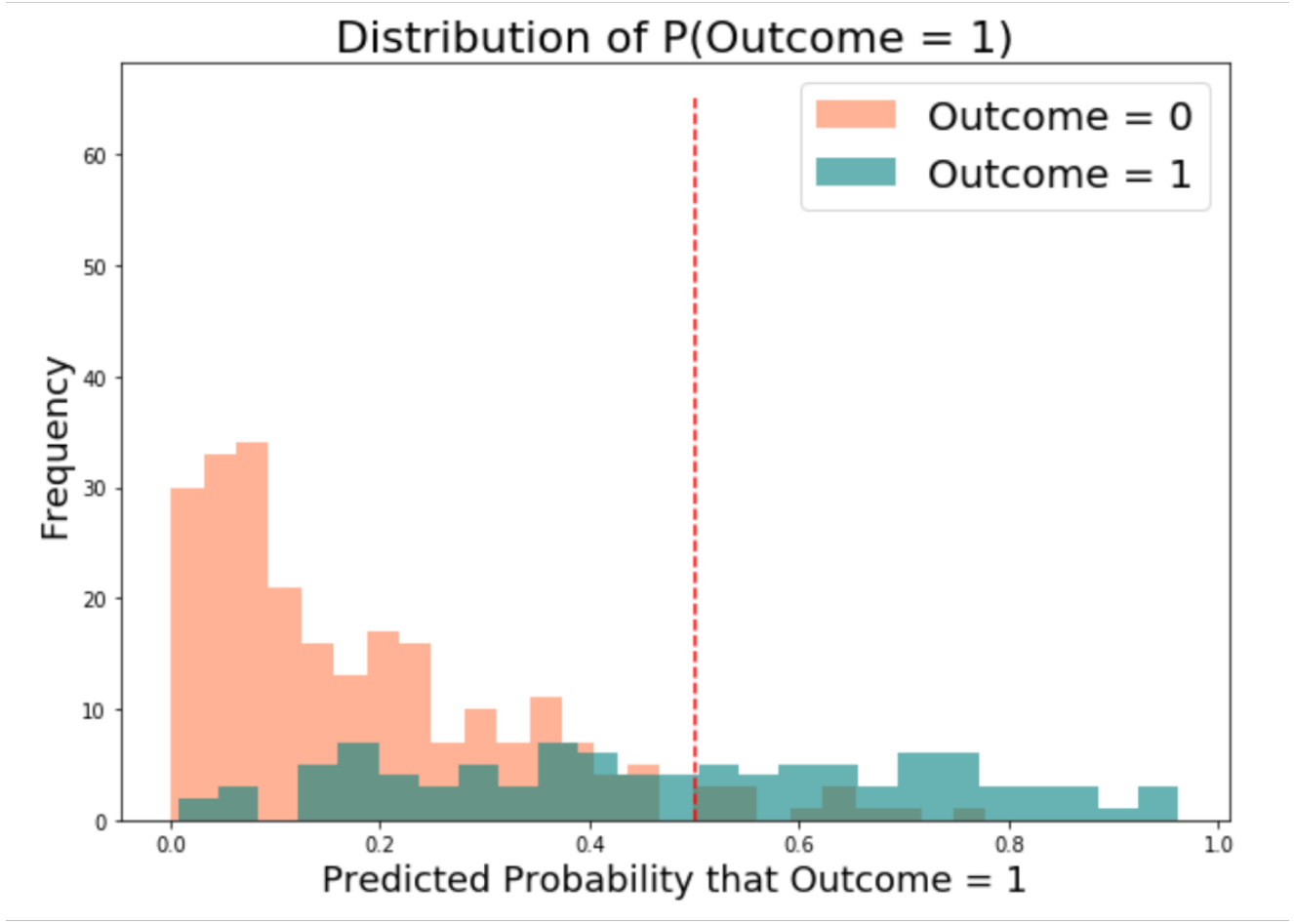
COEFFICIENTS

CONFUSION MATRIX

Not great, but not the worst.

```
True Negatives: 243
False Positives: 1
False Negatives: 87
True Positives: 13
----------------------
Accuracy: 0.7442
Specificity: 2.0
Sensitivity: 88.0
Misclassification: 0.2558
Precision: 0.9286
```

VISUALIZATION



CONCLUSIONS

- My model did not do a great job predicting weather or not a title came from The Onion or Not the Onion.

CONCLUSIONS

- My model did not do a great job predicting weather or not a title came from The Onion or Not the Onion.
  - Accuracy = 74.7%

CONCLUSIONS

- My model did not do a great job predicting weather or not a title came from The Onion or Not the Onion.
  - Accuracy = 74.7%
  - Specificity = 14%

CONCLUSIONS

- My model did not do a great job predicting weather or not a title came from The Onion or Not the Onion.
  - Accuracy = 74.7%
  - Specificity = 14%
  - Sensitivity = 75%

CONCLUSIONS

- The computer has a hard time predicting the difference

CONCLUSIONS

- The computer has a hard time predicting the difference
- Unbalanced model likely had a lot to do with my predictions.

CONCLUSIONS

- The computer has a hard time predicting
- Unbalanced model likely had a lot to do with my predictions.
- Also likely that it just is hard to tell the difference and my data is pretty much meaningless

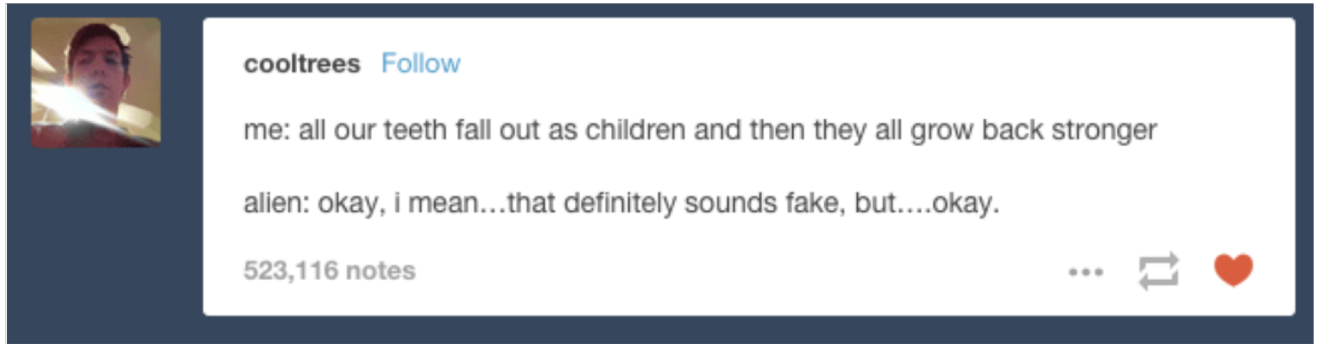
SUBREDDIT PAGES

'Pros And Cons Of Making Birth Control Available Over The Counter'

'elliott abrams defends war crimes as happening back in the ’80s when everyone was doing it'

'Plant-Based Meat Sales Rise, Fueled by Carnivores'

'cast of cats film to be size of actual cats'



FURTHER RESEARCH

- Randomly selecting out The Onion article baseline score is 50%
- Adding another subreddit
  - r/FloridaMan
  - r/fakenews
- Trying to bootstrap/bag?
- Lemmatizing/Stemming?