A Smart Tool for Classifying Food Posts from Reddit Threads

DATA SCIENCE FOR FOOD LOVERS

Jonna Pander

PROBLEM STATEMENT

How can you take a bank of subreddit submissions collected from multiple subreddits and decipher which one it came from?

I want Indian Food but I want it Whole30!

SUBREDDITS

🌶️ /r/IndianFood 🍚 76.5k Members 🍛

14,055 Submission Titles Scraped

🥑 /r/whole30 🥗 55.4k Members 🥓

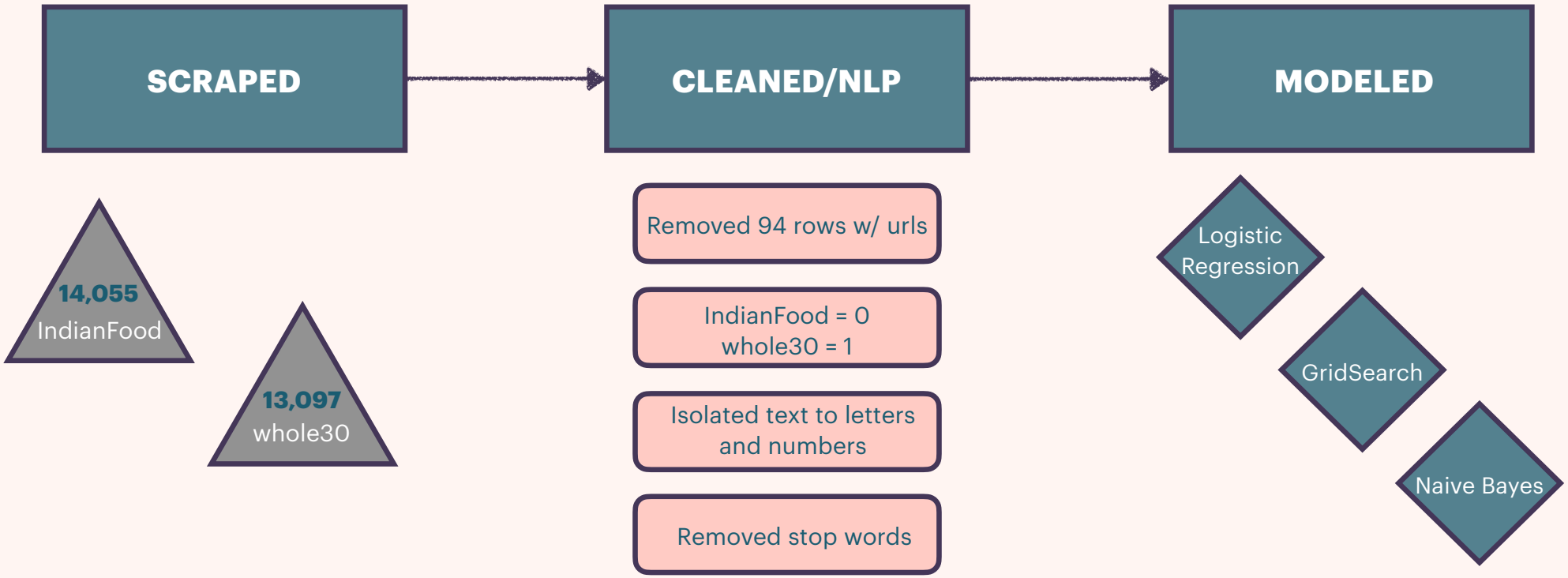
13,097 Submissions Titles Scraped

BASELINE

Baseline accuracy is 48.23%

Subreddit	Percent	Binary
IndianFood	51.77%	0
whole30	48.23%	1

DATA PROCESSING



WHY INCLUDE DIGITS?

- Whole30
  - Day 1
    - Day3

```
text_ltrs_nums[:50]
```

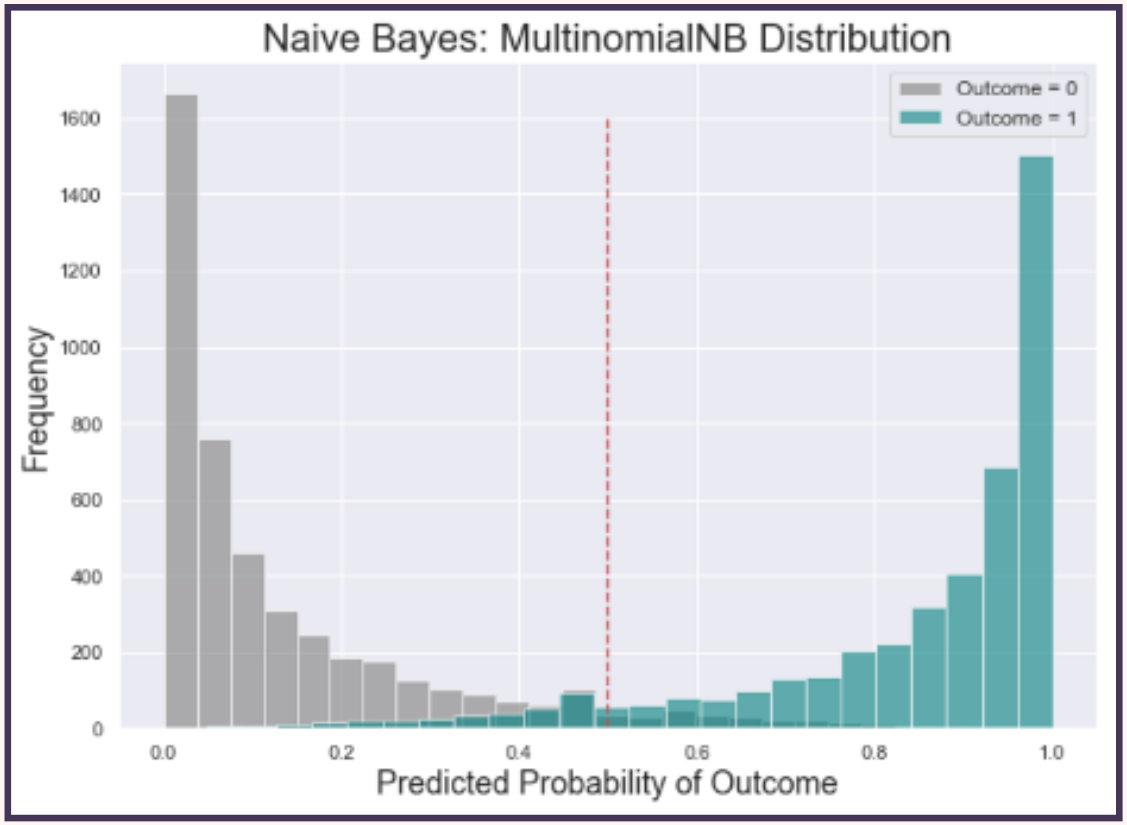
```
['day',
 '17',
 'and',
 'i',
 'would',
 'still',
 'steal',
 'cake',
 'from',
 'a',
 'hungry',
 'homeless',
 'man.']
```

For whole30, they matter.....

MODEL SUMMARY

Model	Training	Testing
Logistic Regression - CountVectorizer	96.83%	93.28%
GridSearch - CountVectorizer	96.46%	93.45%
Naive Bayes - Multinomial - TFIDFVectorizer	96.15%	93.52%

DISTRIBUTION OF OUTCOMES



CONFUSION MATRIX

Sensitivity

(True Positive Rate):

91.97%

	Predicted whole30	**Predicted IndianFood**
Actual whole30	4,368 (TP)	232 (FN)
Actual IndianFood	345 (FP)	3,954 (TN)

Specificity

(True Negative Rate):

94.96%

CONCLUSION

Whittles down the selection

Secondary check

QUESTIONS?