Subreddit Binary Classification

Modeling to differentiate between two subreddits

Will Hanley, General Assembly

Project Overview

- Web-scraping and requesting Reddit API
- Data Cleaning
- Natural Language Processing and Modeling

Differentiate posts between the two different subreddits

Two Subreddits

r/TheOnion

the ONION

AMERICA'S FINEST NEWS SOURCE · ONION.COM

'ONION' PRINT REVENUES UP 5,000%

Experts: Digital Media Revolution Still Another 70 Or 80 Years Away

Google CEO Fondly Recalls 'Onion' Print Ad That Put Company On Map

'Onion' Reporter Glad He's Not In The Business Of Chasing Web Traffic

r/NotTheOnion

API Requests

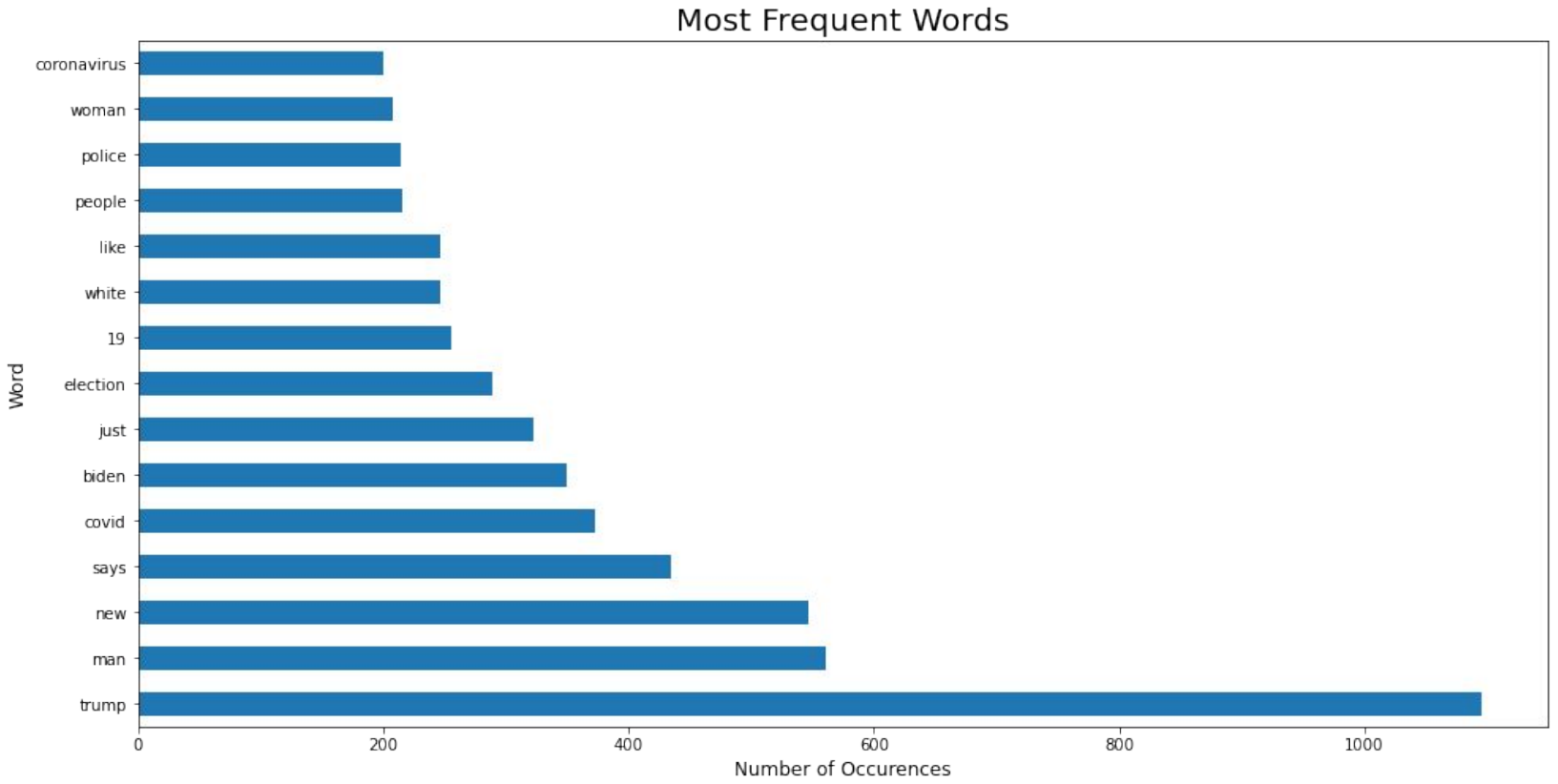
- Retrieving the data from the Reddit API
- Epoch Time
- Creating Dataframe

```
# Function takes the subreddit name and the number of posts wanted (in returns of 100 posts)

def reddit_scannerZ(subreddit, number_posts):

    # Empty list to append the scrapes
    mega = []

    # URL from the reddit api
    url = 'https://api.pushshift.io/reddit/search/submission'

    # Starting point for date of scrape
    new = 1606338449

    # Loop that runs for the number of posts wanted
    for i in range(number_posts):

        # params for the scape (name, number of posts (max: 100), and date)
        params = {'subreddit' : subreddit,
             'size' : 100,
             'before': new}

        # Use requests to scrape API
        pull = requests.get(url, params).json()

        # Create new date from which the posts will come
        new = pull['data'][-1]['created_utc']

        # Append dataframe of requests taking only the subreddit and post
        mega.append(pd.DataFrame(pull['data'])[['subreddit','title']])

        # Wait 2 seconds before looping again
        time.sleep(2)

    # Return concatinated dataframe
    return pd.concat(mega).reset_index()
```

The Data

	subreddit	title
241	1	Disney World On Lockdown After Mickey Escapes Enclosure, Rampages Through Park
5879	0	Andrew Cuomo to get Emmy for use of TV during COVID-19 pandemic
5462	0	Veteran who founded veterans charity in Connecticut sentenced to prison for scamming donors
8893	0	GOP Rep. Dan Crenshaw to Marjorie Taylor Greene: 'Start acting' like a member of Congress
146	1	Motorcyclist Who Identifies As Bicyclist Sets Cycling World Record

Most Common Words



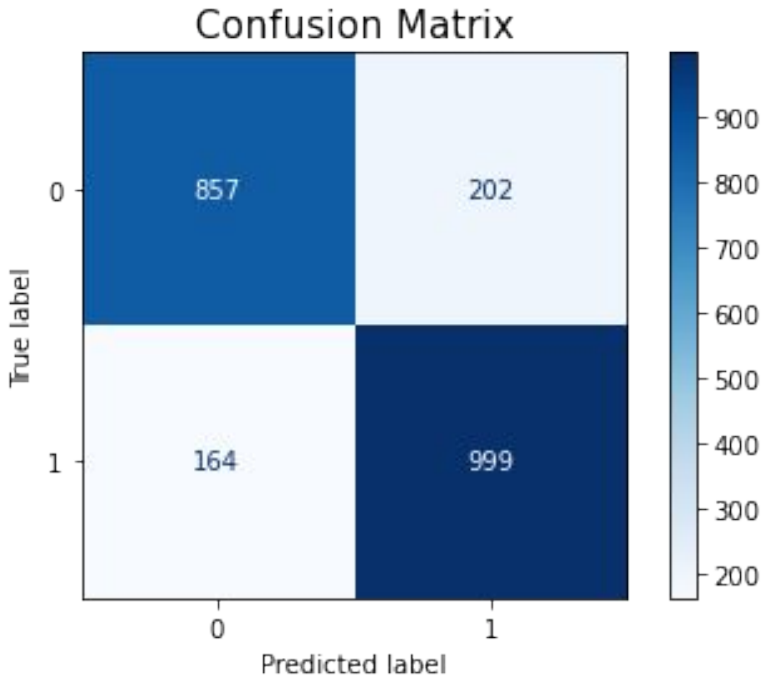
Models

	Transformer	Estimator
Model 1	Countvectorizer	Logistic Regression
Model 2	TFIDF	Multinomial Naive Bayes
Model 3	Countvectorizer	Random Forest
Model 4	TFIDF	Support Vector Machine

Logistic Regression w/ Countvectorizer

Model 1

Data	Accuracy Score
Training Split	0.966
Testing Split	0.835

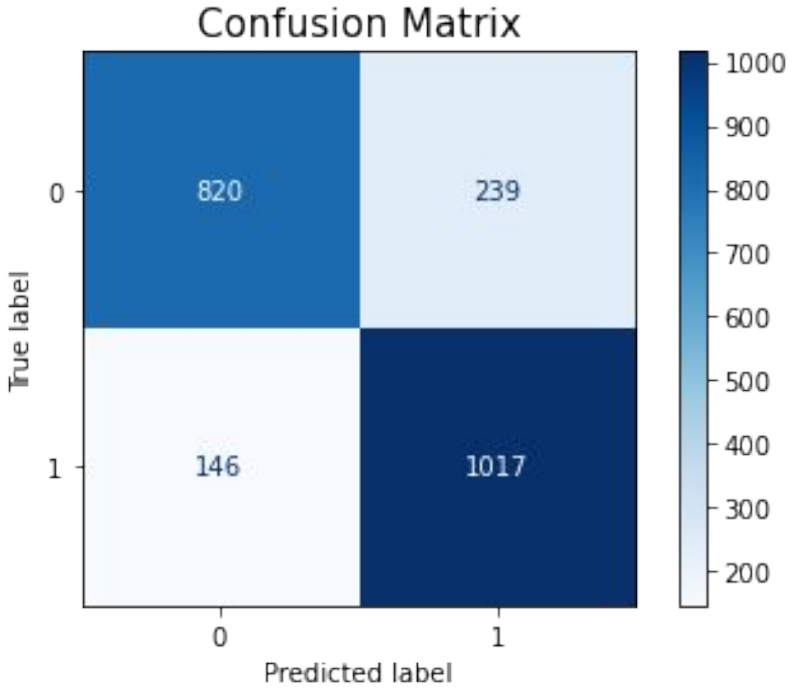


Baseline Accuracy: 52.3%

Multinomial Naive Bayes w/ TFIDF

Model 2

Data	Accuracy Score
Training Split	0.920
Testing Split	0.826

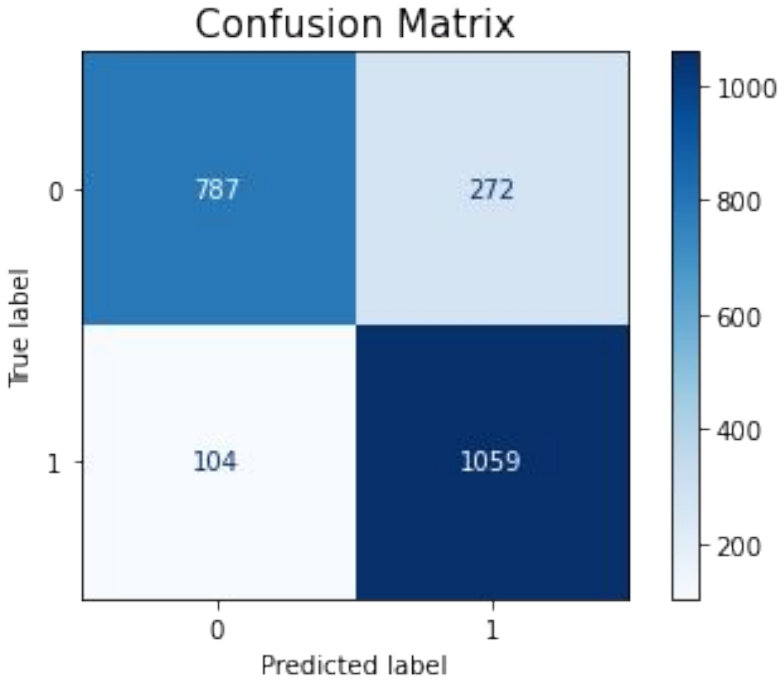


Baseline Accuracy: 52.3%

Random Forest w/ CountVectorizer

Model 3

Data	Accuracy Score
Training Split	0.999
Testing Split	0.825

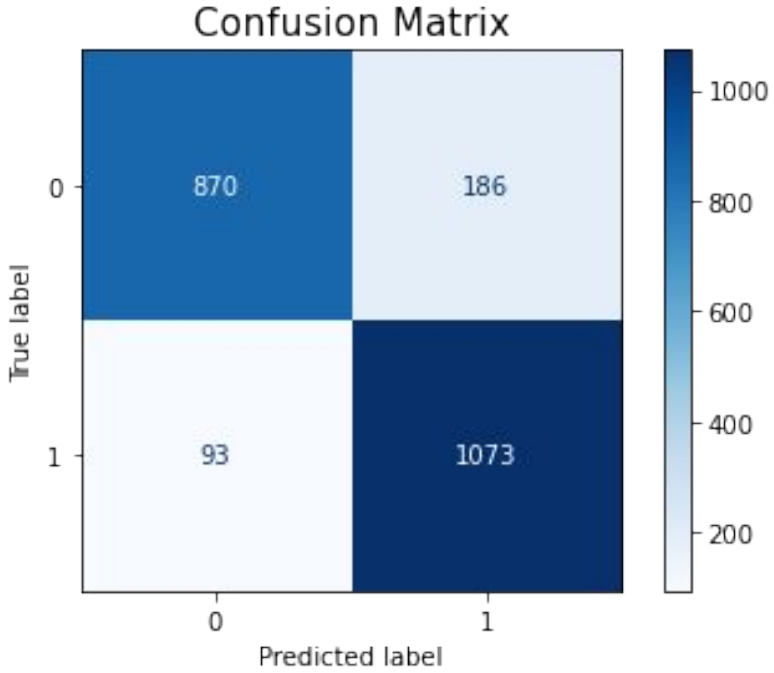


Baseline Accuracy: 52.3%

Support Vector Machine w/ TFIDF

Model 4

Data	Accuracy Score
Training Split	0.998
Testing Split	0.874



Baseline Accuracy: 52.3%

87.4% Accuracy!

On test data from the Support Vector Machine w/ TFIDF

Correctly predicted 1,943 out of 2,222 submissions

Can you beat the model?

- “America Has Just Found A Way To Deep Fry Water”

- “Earwax Analysis Shows WWII Was Stressful For Whales”

Both are from r/NotTheOnion!!!

(Both are real headlines,
If you can believe it)

Thank You for Listening!

Questions?

Images:
https://www.oratoryprepomega.org/2018/04/18/exploration-of-not-the-onion-subreddit/
www.theonion.com