Subreddit Classification Using Web APIs & NLP

Danielle Medellin

Problem Statement

NBC is looking to see how people engage with some of their most famous sitcoms on social media. An intern at NBC was tasked with gathering as many posts as he could for NBC sitcoms like *Will & Grace*, *The Office*, *Parks and Recreation*, *Brooklyn 99*, and *The Good Place*. Unfortunately, this intern was terrible at organization and dumped all of the posts he found into one general folder. NBC needs to identify which show these posts belong to and have asked for our help, specifically with their difficulty classifying two shows: *The Office* and *Parks and Recreation*.

r/DunderMifflin

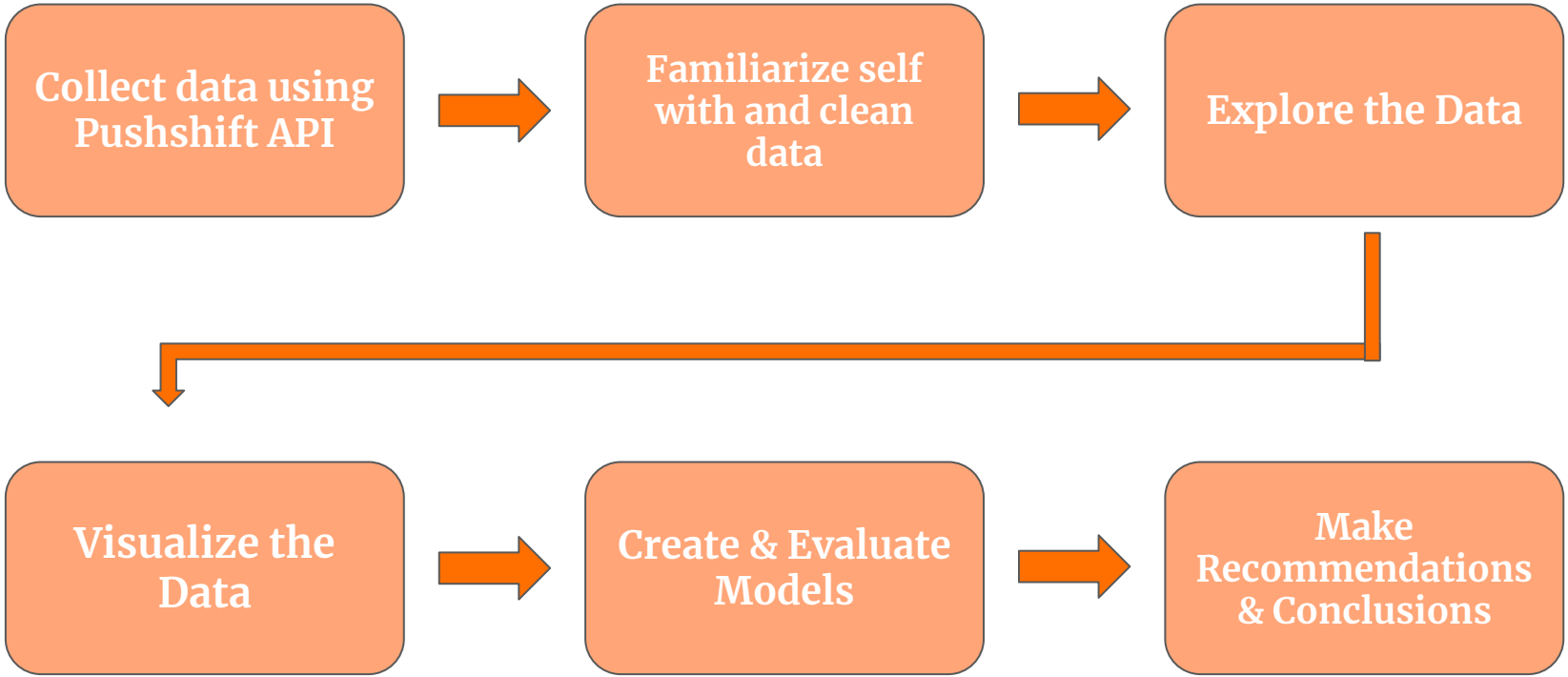
r/PandR

Goal: build a classification model that will help to sort each post to its appropriate subreddit. Our model's success will be measured with an accuracy score.

Possible Issues

- Same creators: Greg Daniels & Michael Schur
- Workplace sitcoms on NBC
- Both have characters named Andy
- Characters & actors share names:
  Michael Scott & Adam Scott
- Rashida Jones

Methodology & Workflow



Data Gathering

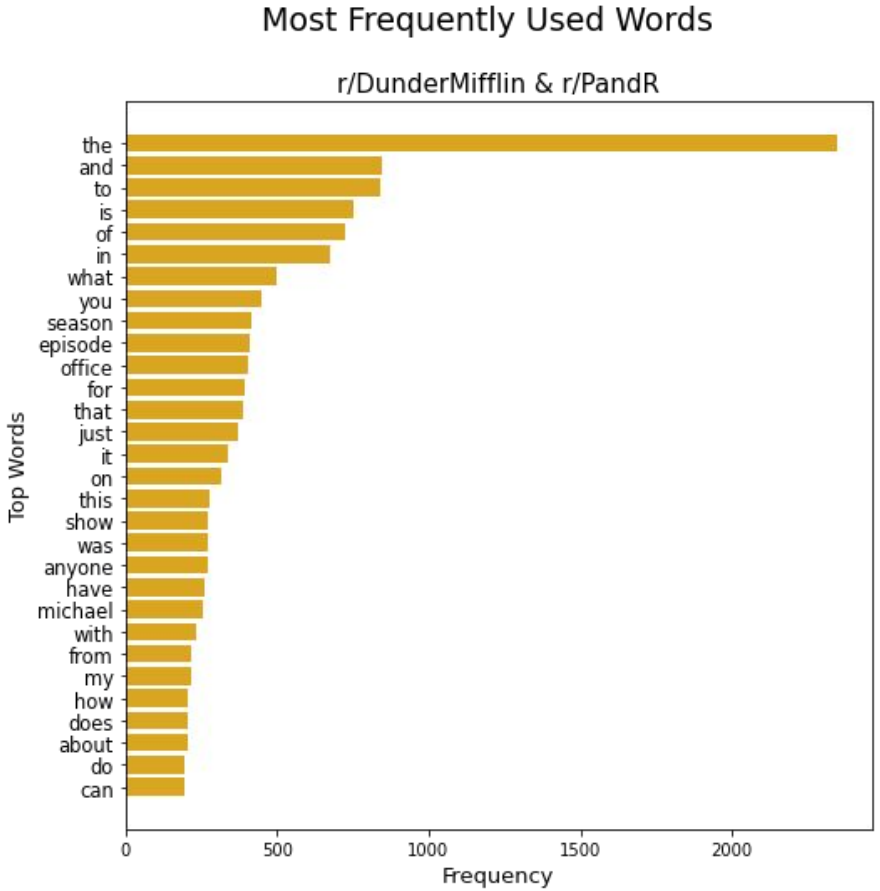
Gathered Data:

- Pushshift Reddit API
- Submissions from r/DunderMifflin & r/PandR
- Pulled data over 6 month intervals for the last 10 years
- Over 4,000 submissions

Data:

- **Title of submission***
- Self text
- Author
- Number of comments
- Score: number of up votes minus down votes
- Date of submission
- Subreddit (0 → r/DunderMifflin; 1 → r/PandR)

Exploratory Data Analysis

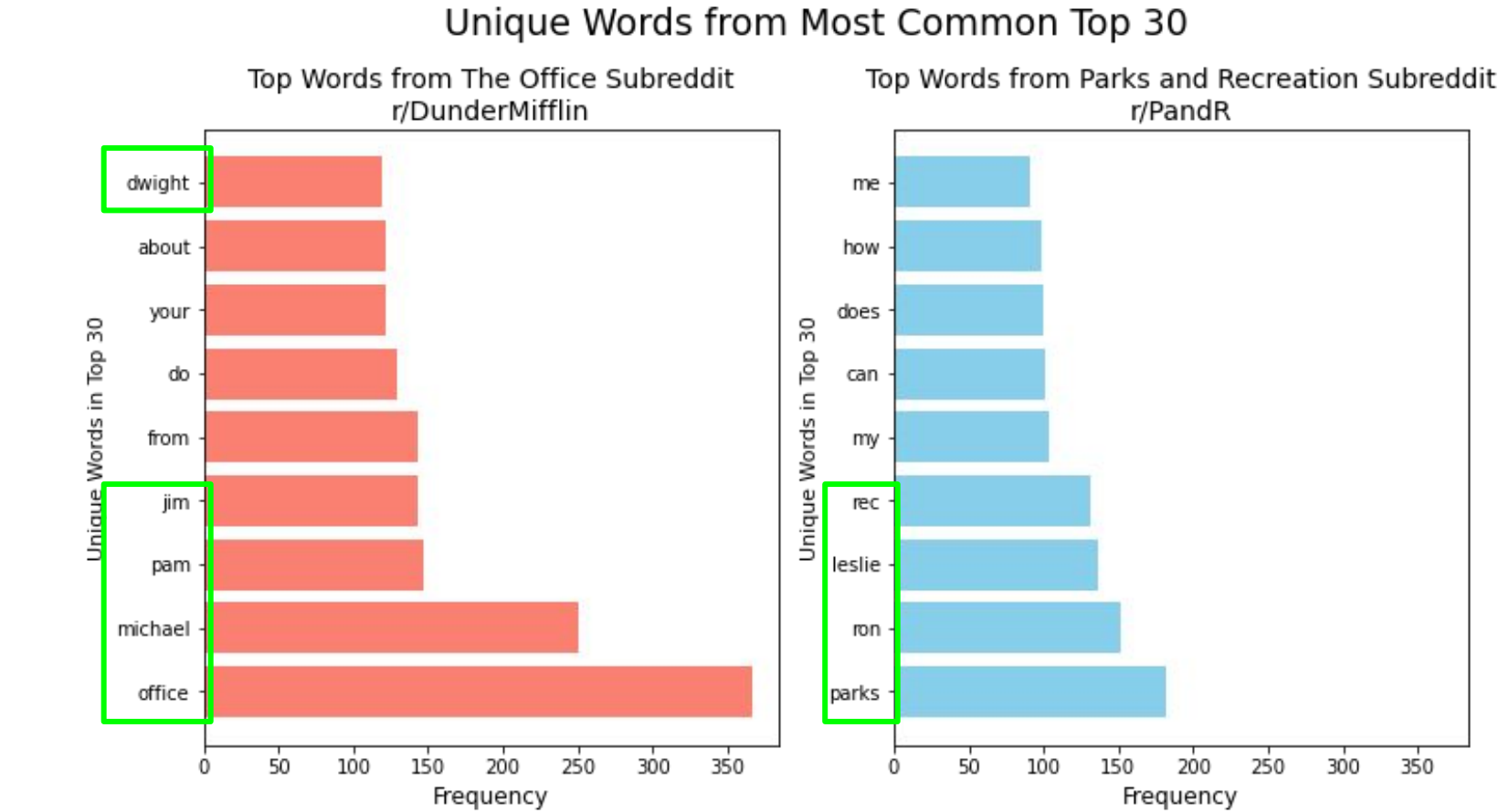


Words of Note:

- season, episode, show
- michael, office

WORDS

Exploratory Data Analysis



Modeling

Count Vectorizer: turns text data to numeric data -- takes frequencies of words in each submission title

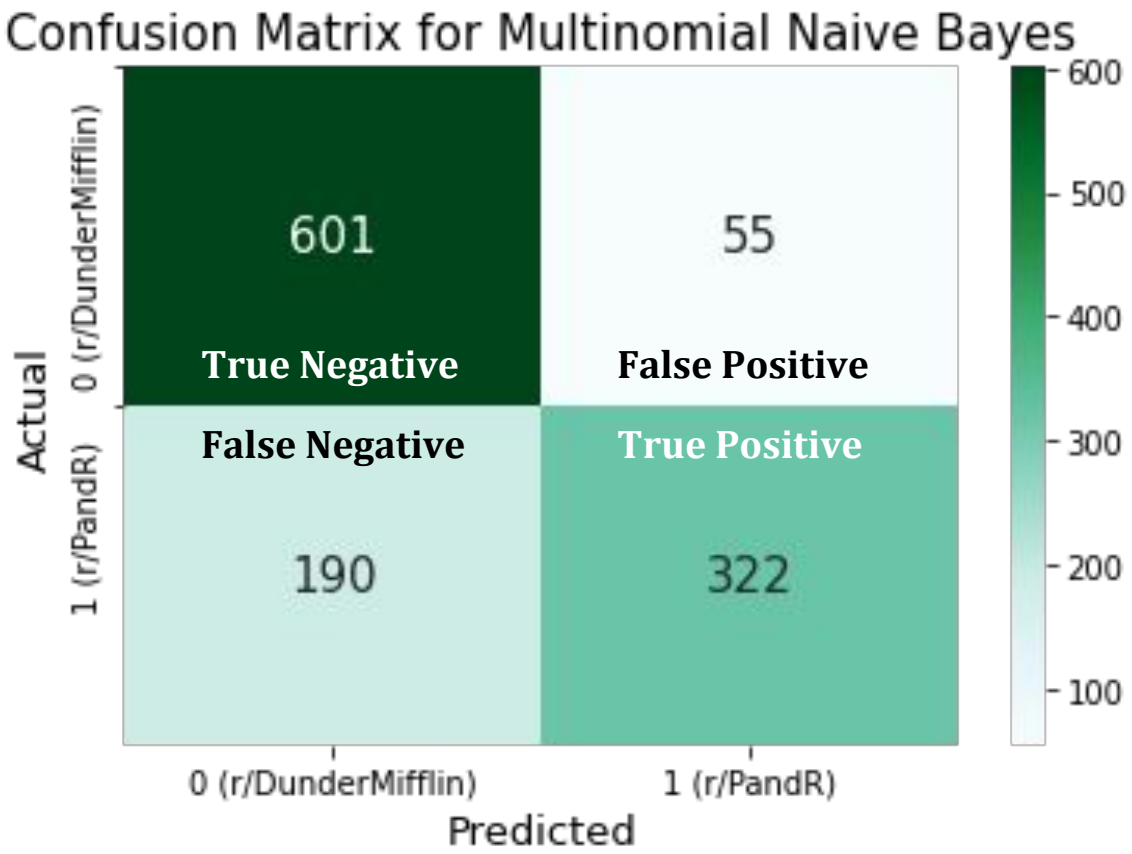
Classification Metric: Accuracy (what percentage did we get right?)

Baseline Model: simplest model, performing at 56% accuracy

Interpretable →
Predictive →

Model	Training Accuracy	Testing Accuracy	Difference (Train - Test)
Logistic Regression	89.7%	79.9%	9.79%
Multinomial Naive Bayes	85.5%	79%	6.45%
Support Vector Machine	85.5%	78.4%	7.11%
Decision Tree	85.9%	77.2%	8.68%
Random Forest	81.9%	75.1%	6.85%
Bagging	85.7%	74.9%	10.76%
k-Nearest Neighbors	78.6%	71.6%	7.0%
Baseline	56.1%	56.2%	- .1%

Model Evaluation - Confusion Matrix



Accuracy:

79%

Misclassification Rate:

21%

True Positive/ Parks & Rec rate:

63%

True Negative/ The Office Rate:

92%

Model Evaluation - Coefficients

Word	Coef
ron	7.455006
leslie	6.825718
jerry	6.185149
april	5.992854
parks	5.930971
ben	5.597879
ann	5.587543
amp	5.076798
pawnee	5.028507
pandr	4.879771
chris	4.611710
mark	4.119129
donna	3.917979
tom	3.606681
tammy	3.317171

$e^{5.028507} = 152.7$

In our model, for every 1 unit increase in the frequency of the word ***pawnee*** in a submission title, that submission was **152.7 times** as likely to belong to the *r/PandR* subreddit, all else held equal.

Word	Coef
best episode	-3.070436
darryl	-3.158133
dundermifflin	-3.240538
room	-3.349333
erin	-3.596560
scranton	-3.662051
angela	-3.673531
ryan	-3.685674
creed	-4.368170
toby	-4.639908
jim	-4.838117
pam	-5.775606
dwight	-6.262191
office	-6.937631
michael	-7.531737

$(e^{-7.531737} - 1) \times 100 = -99.946$

In our model, for every 1 unit increase in the frequency of the word ***michael*** in a submission title, that submission was **99.9% less likely** to belong to the *r/PandR* subreddit, all else held equal.

Conclusions & Recommendations

- Our models are not perfect, but they are accurately classifying almost 80% of the data
- Words that reference people and places from the TV shows are the best indicators for what subreddit a submission belongs in
- Not a foolproof rule:
  Because these shows are so similar, they have an overlapping fan base that often compare the two and mention each other in posts
- Our models classified the r/DunderMifflin subreddit better, so we can look deeper for better indicators of the r/PandR subreddit and look for more stop words
- Apply this model to other social media platforms

References

“The Office.” IMDb, IMDb.com, 24 Mar. 2005, www.imdb.com/title/tt0386676/.

“Parks and Recreation.” IMDb, IMDb.com, 9 Apr. 2009, www.imdb.com/title/tt1266020/.

Pushshift.io. “Pushshift Reddit API Documentation.” GitHub, 1 Oct. 2019, github.com/pushshift/api.

“r/DunderMifflin: People Person's Paper People.” Reddit, www.reddit.com/r/DunderMifflin/.

“r/PandR: Tommy's Place.” Reddit, www.reddit.com/r/PandR/.

“r/Reddit.com.” Reddit, www.reddit.com/wiki/faq#wiki_how_is_a_submission.27s_score_determined.3F.