reddit -April Fools' 2021!

Ryan McDonald – Reddit Data Scientist

https://git.generalassemb.ly/spyder6146/project_3

Background

Reddit has a rich history of 'April Fools' hijinks! Lighthearted 'experiments' deployed onto the site to wow and amaze, or frustrate users!

From its start in 2011 releasing reddit 'mold' (a play on 'gold') that would make your user experience worse, each year has brought a surprise!

Previous April Fools' Deployments (not all are listed)	
2013	Video game 'Team Fortress 2' was released and randomly assigned users onto 1 of 2 teams to battle it out
2015	'The Red Button' social experiment were users would press a button and reset a 60s global timer. Who's time is longest?
2016	'Robin' chat widget that gave users a short time to decide if their chat would join, split, or remain in another chat
2017	Social experiment 'place'. Collaborative pixel art canvas where each user could place one pixel every 5 minutes.
2020	'Imposter' release. Users needed to choose which response from many was machine-generated.

HTTPS://EN.WIKIPEDIA.ORG/WIKI/REDDIT
https://www.reddit.com/r/TheoryOfReddit/comments/2gdv9z/back_of_the_napkin_calculation_reddits_server/

Problem Statement

For 2021, we have proposed another excellent hijinks based on 'limiting similar subreddits'!

- r/sinkorswim bot will (temporarily) disable subreddits that appear too similar to other subreddits
- We are tasked with two objectives:

1. For two subreddits, use Natural Language Processing to train a classifier Model that can determine which subreddit a given post came from.

   Decide on metrics that determine the cut-off for subreddit disablements

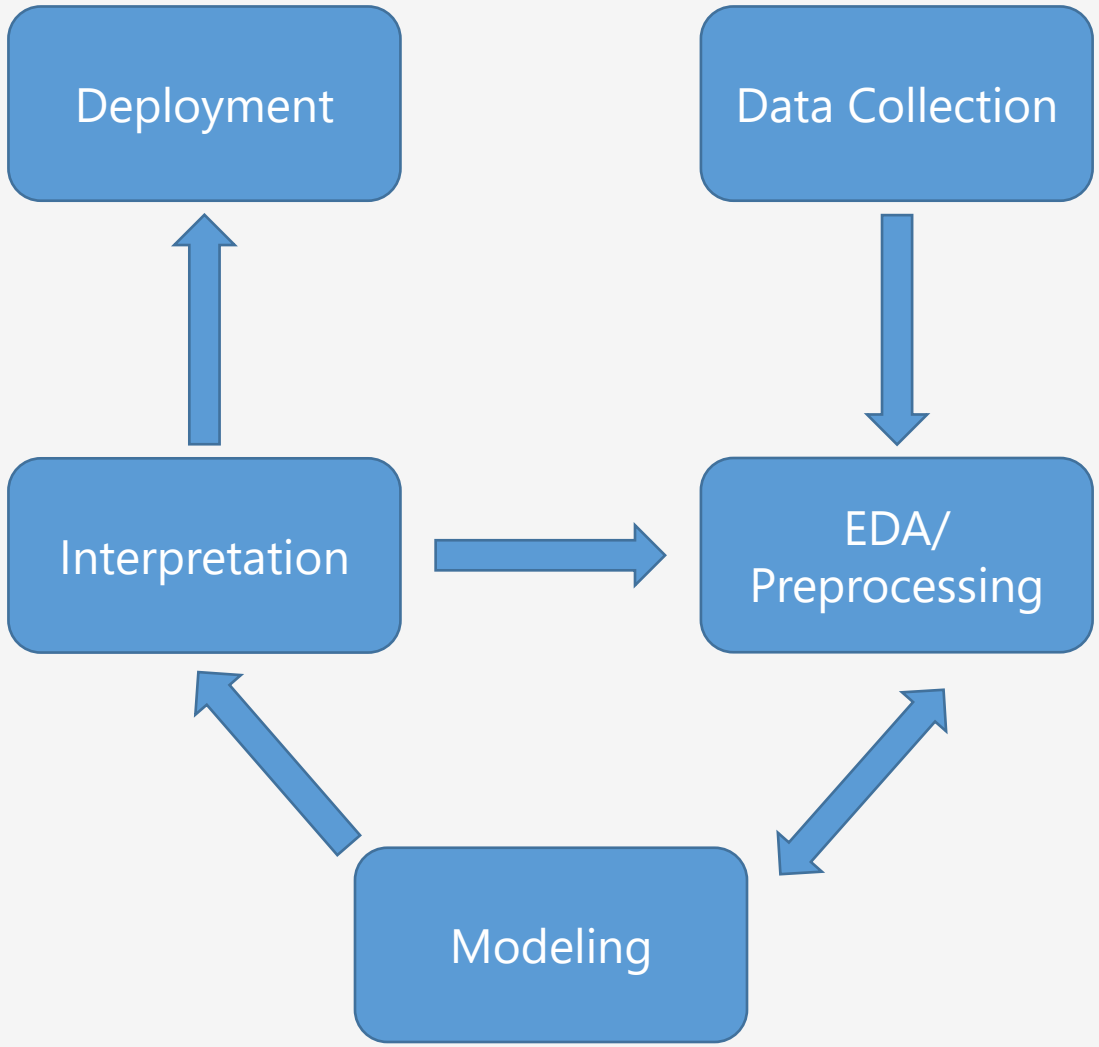
2.
   - If the model outperforms the metric, there is enough unique substance in each subreddit for them both to remain.
   - Otherwise, the subreddit is disabled!

HTTPS://EN.WIKIPEDIA.ORG/WIKI/REDDIT

Methodologies

Although, meant to be lighthearted, our modeling process will be prescriptive.

- Automate subreddit submission title retrieval with a function
- Perform necessary EDA/Preprocessing
- Impute data into models and iterate through hyperparameters to determine best fit
- Interpret results and justify the deployment of a Production Model
- Watch as reddit users go crazy on April 1st!



Data Analysis

Data was extensively cleaned and interpreted with the problem statement in mind.

- A function was developed with use of the pushshift API to pull 4000 posts from each of two subreddits.

- Our analysis was conducted on two active, seemingly similar subreddits (but varying popularity):
  - r/VanLife (109k members) (7 mo. of posts)
  - r/camping (1.9M members) (3 mo. of posts)

pushshift.io

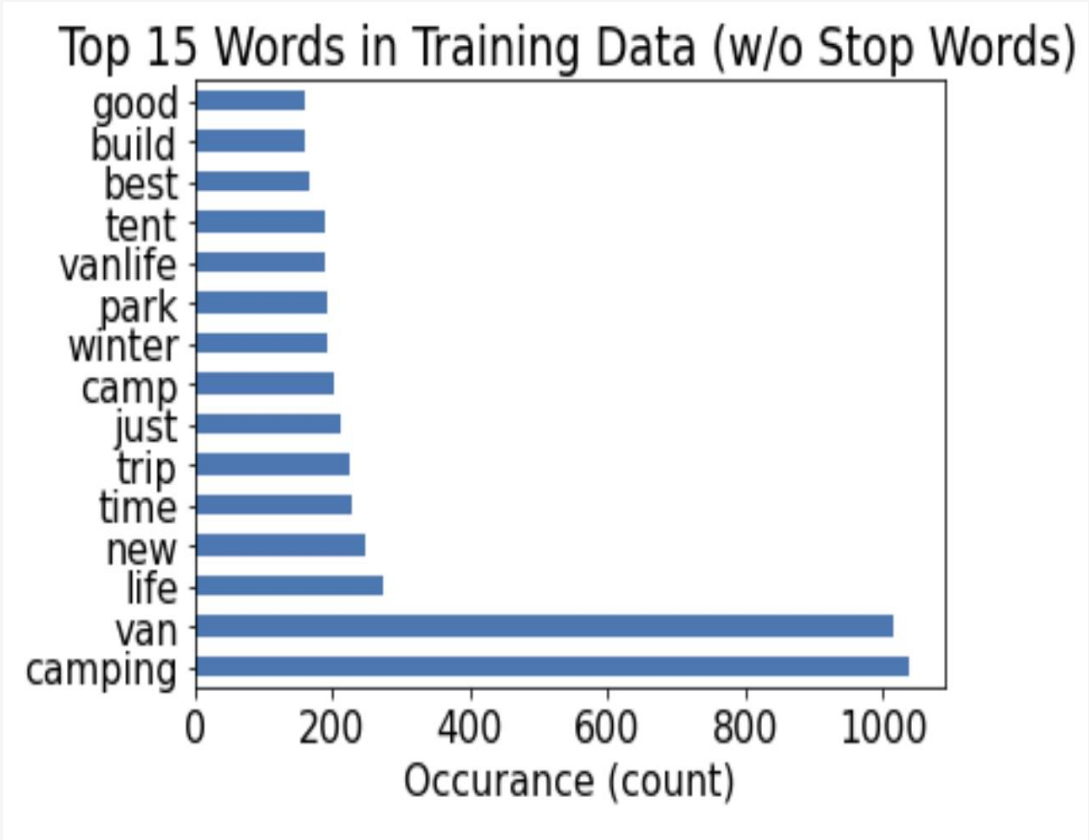
Two Data Frames were developed:

- Raw title data, unprocessed
  - Baseline score of 0.50
- Sentence-Tokenized title data
  - Baseline score of 0.54
- Our ‘not to anger too many users’
  - Baseline score of 0.80

Data Analysis (cont.)

- Pre-Processing Findings of Interest

Mean Sentiment Analysis				
	neg	**neu**	**pos**	**c.pound**
VanLife	0.03	0.86	0.11	0.146
Camping	0.02	0.86	0.11	0.154



Modeling Overview

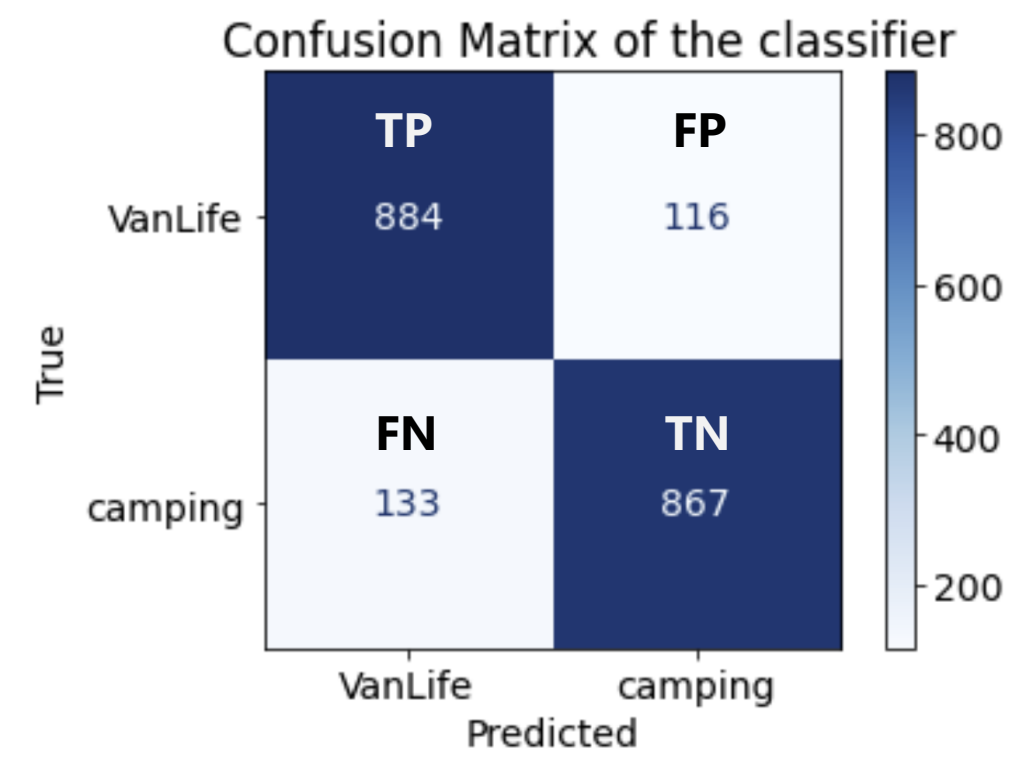
Number of Models Developed	6
Transformers Utilized	CountVectorizer, TfidfVectorizer
Classifiers Utilized	Bernoulli NaïveBayes, Logistic Regression, DecisionTree, LinearSCV
Ensemble Methods Utilized	RandomForest, AdaBoost, XGBoost
Workflow Automation Methods	Pipline
Hyperparameter Tuning Technique	GridSearchCV

Production Model	
Transformer	TF-IDF Vectorizer (with bigrams)
Classifier	Bernoulli Naïve Bayes
Automation/HP Tuning	Pipeline/ GridSearchCV , cv=8

Production Model Performance / Evaluation

So, how well does it work?

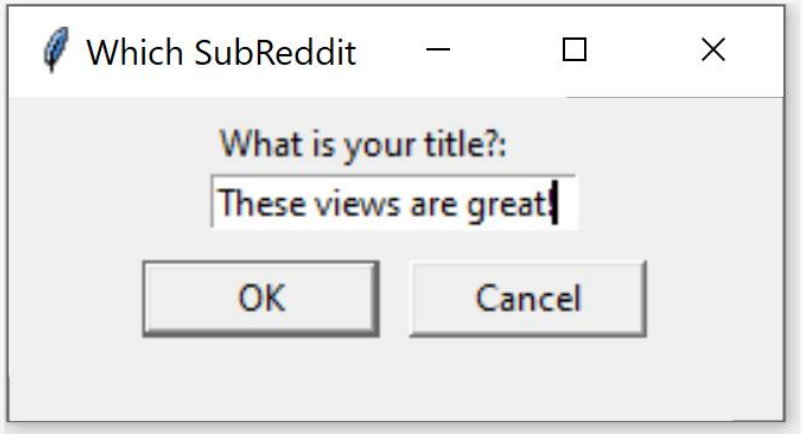
- Production Model Testing Accuracy = **87.6%**
- Models 2-6 range between 79%-84%



IDF-rated Importance	Word	Predicted Class
1	die	VanLife
2	spring camping	Camping
3	saving	VanLife
4	scary	VanLife
5	searching	VanLife
6	does experience	VanLife
7	secure	VanLife
8	services	VanLife
9	sf	VanLife
10	shed	Camping

Production Model Evaluation

Fun User Tool for evaluating proposed title!



Random Subreddit Model Performance	
Subreddits Ran Through Model	**Testing Accuracy Score**
Spacex/Politics (different content)	97.2%
IndoorGardening/GardeningIndoors (similar content)	56%
Houseplants/ Babies (somewhere in the middle)	85.8%

Conclusion

- Our Production Model is a success! We, at Reddit, are looking forward to releasing this hijinks to unsuspecting Reddit users on April 1st. The model performed as well as we could have hoped, weeding out similar subreddits and promoting others as 'unique'

- We will be monitoring users that day to see how everyone reacts, and are looking forward to putting together our 2022 ideas to work soon!