Subreddit Classification

r/LifeProTips vs. r/UnethicalLifeProTips

Jocelyn Lutes
DSI-12
July 2020

Background

Life Pro Tips

vs.

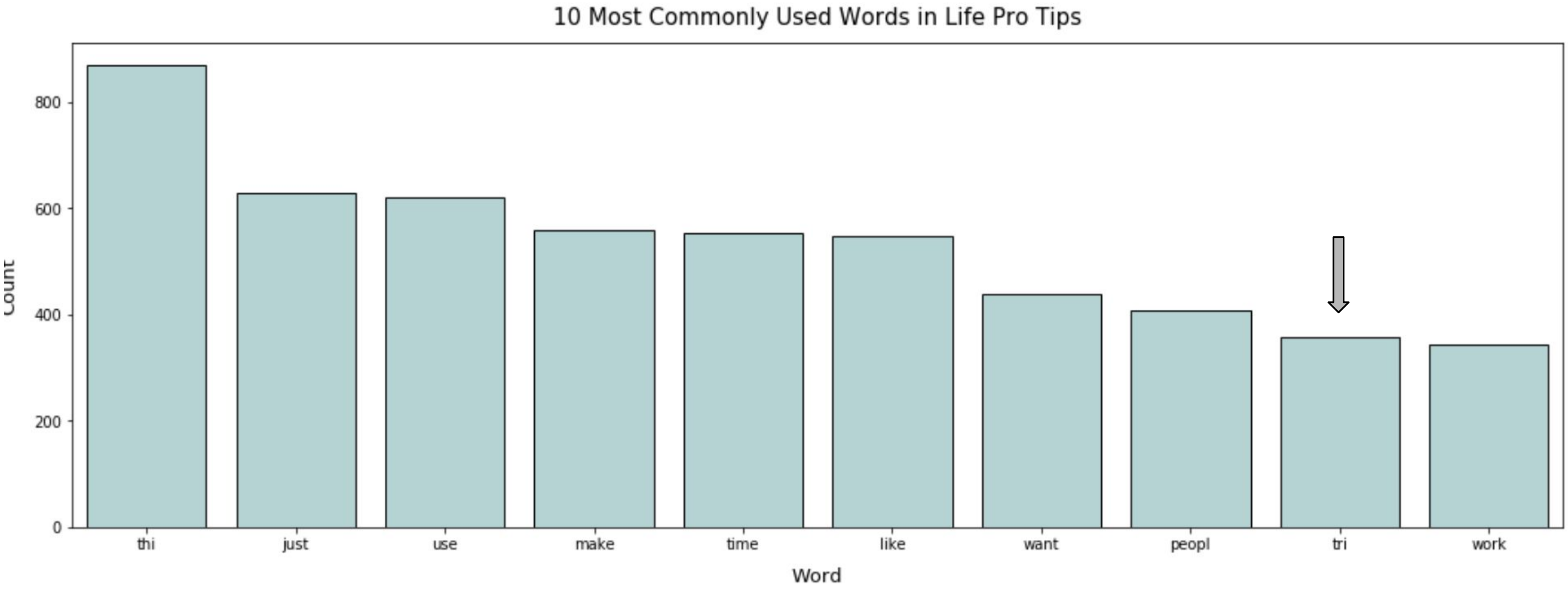
Unethical Life Pro Tips

Can a model be built to accurately classify posts from each of the subreddits?

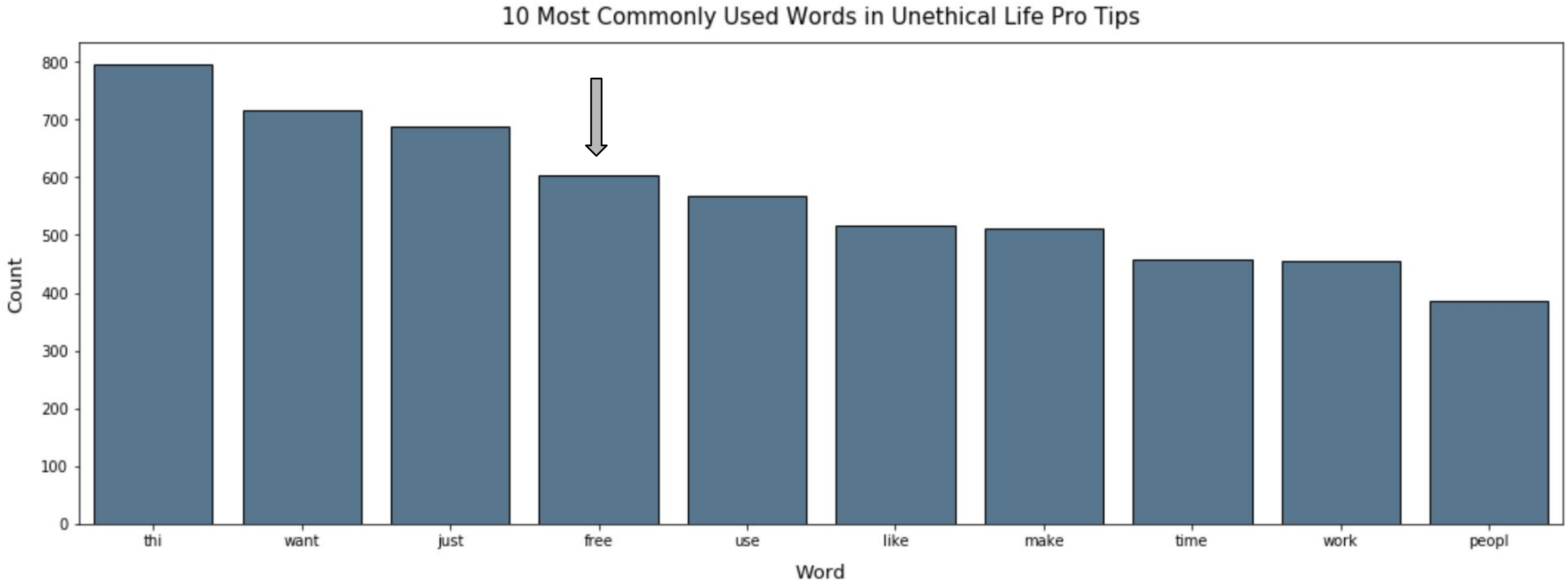
Work Flow

Step	Description
Obtain Data	Gathered using Pushift API
Clean Data	Removed subreddit tags, unuseful text
Text Preprocessing	Expanded contractions, stemmed words, sentiment analysis
Exploratory Data Analysis	Visualized distributions, word frequencies, sentiment scores
Model Preparation	Created a training and testing set
Modeling and Model Selection	Built 9 different classification models
Model Evaluation	Determined how well the model did for each subreddit
Model Interpretation	Investigated which features seemed to be most important

Top Stemmed Words on Life Pro Tips



Top Stemmed Words on Unethical Life Pro Tips



Performance Metrics for Best Models

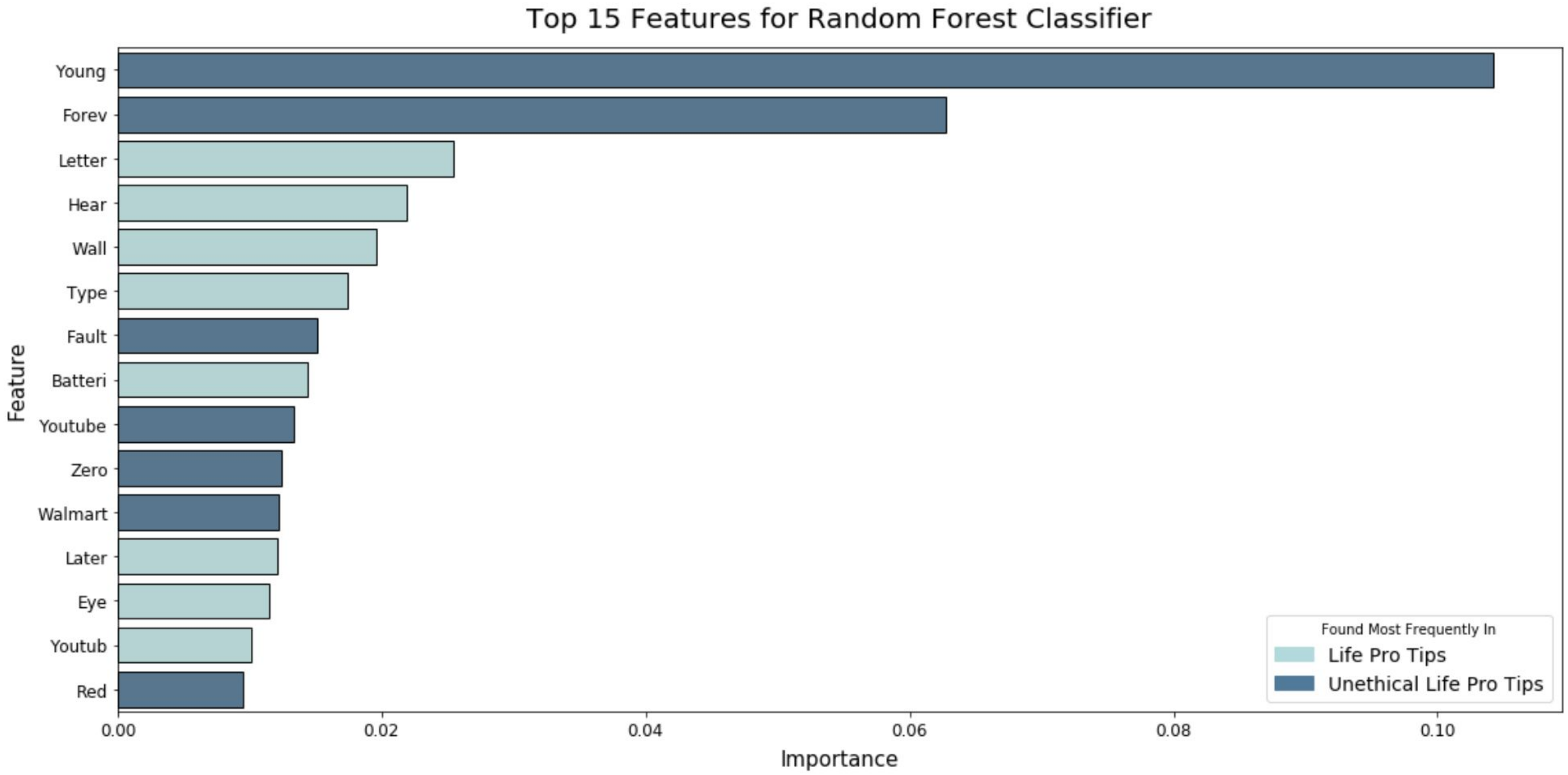
Model	Training Accuracy	Testing Accuracy	Testing Sensitivity
Baseline/Null	50.5%	50.1%	54%
Multinomial Naive Bayes	90.2%	79.1%	88.5%
Random Forest Classifier	69.9%	78.0%	90.6%

Confusion Matrix for Best Models

	Predicted LPT	Predicted ULPT
True LPT	45%	55%
True ULPT	9.4%	91%

This model correctly classifies Unethical Life Pro Tips with approximately 91% accuracy, but performs under baseline for Life Pro Tips.

Most Important Features



Conclusions

- The best-performing model was a Random Forest Classifier with an accuracy of 78% and a sensitivity of 90.6%.

- This model is good at accurately predicting posts from the r/UnethicalLifeProTips Subreddit, but it performs almost at Baseline for posts from r/LifeProTips.

- To begin the next round of improvements, we will learn from our models and manually prune features from the vector of words.

- Once the dimensionality of the models has been reduced, we will be able to tune our models in hopes of improving the overall accuracy.