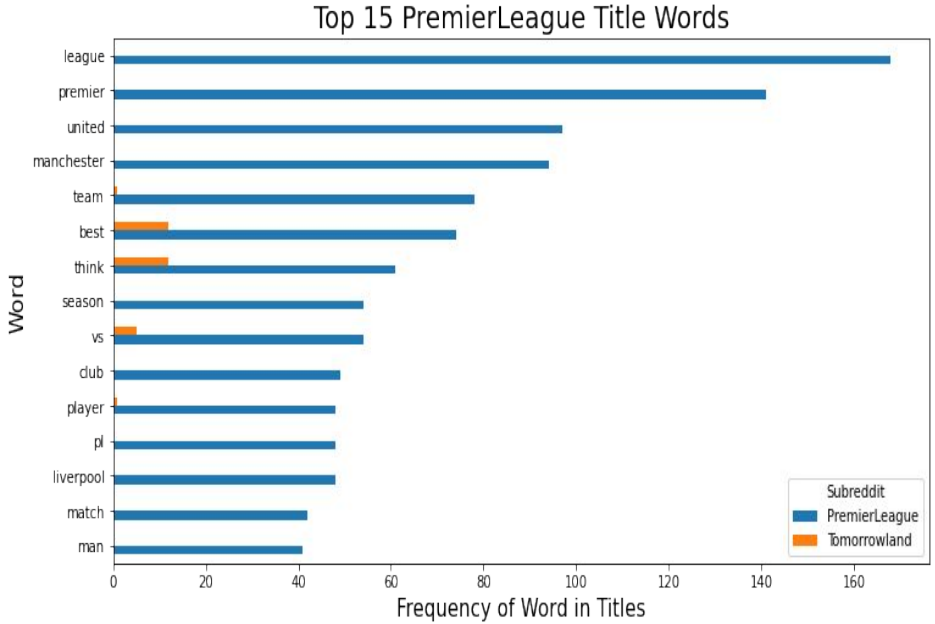
Predicting Subreddits - Premier League vs. Tomorrowland

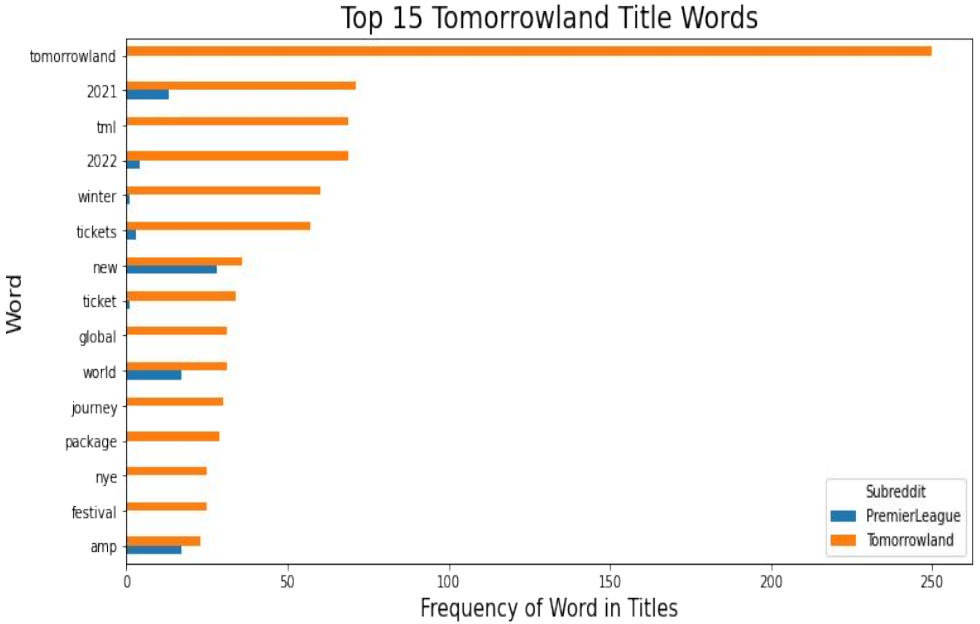
Presentation by Aaron O’Neal

Project Overview:

The goal of this project was to take two subreddits and create a model to accurately determine which subreddit a specific post came from. I sought out to determine which part of the post (the title, the selftext, or a combination of the two) would be the best information to give a model in order to help it predict which subreddit the post came from. For this project, I chose the Tomorrowland and Premier League subreddits: two passionate fan bases that were fairly active in each subreddit.

Cleaning and Exploratory Analysis

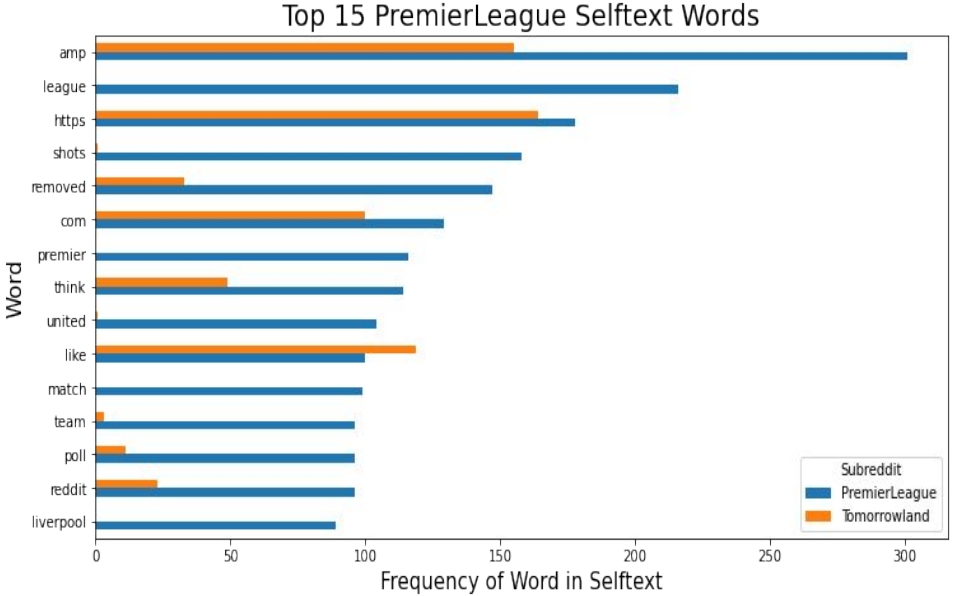


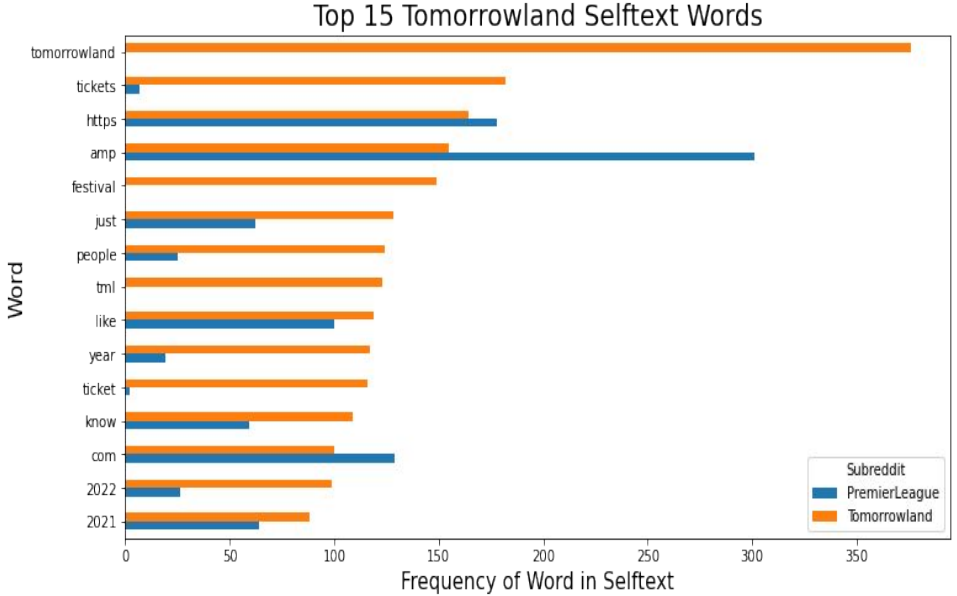


Average number of characters in title: 61
Average number of words in title: 11

Average number of characters in title: 48
Average number of words in title: 8

Cleaning and Exploratory Analysis

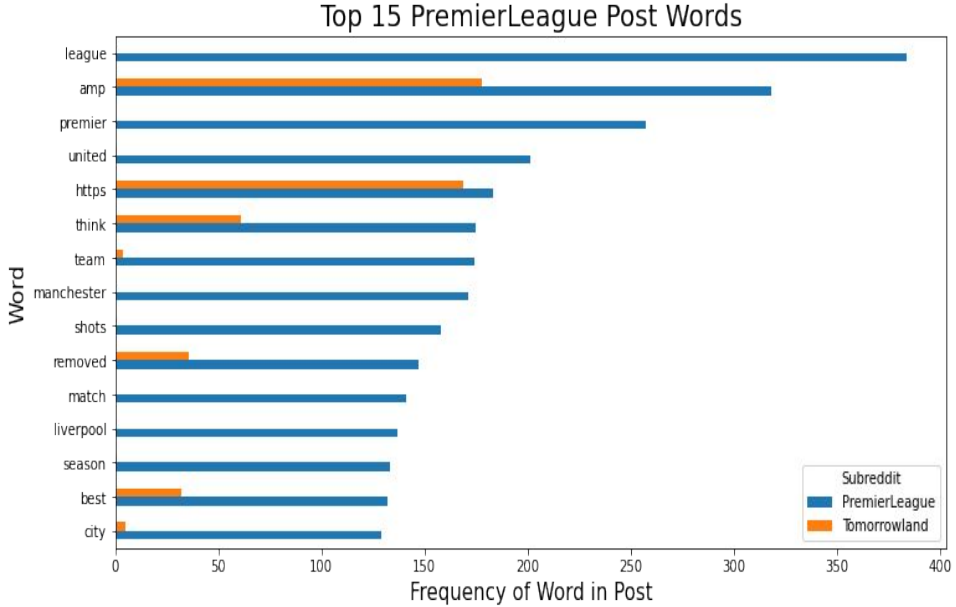


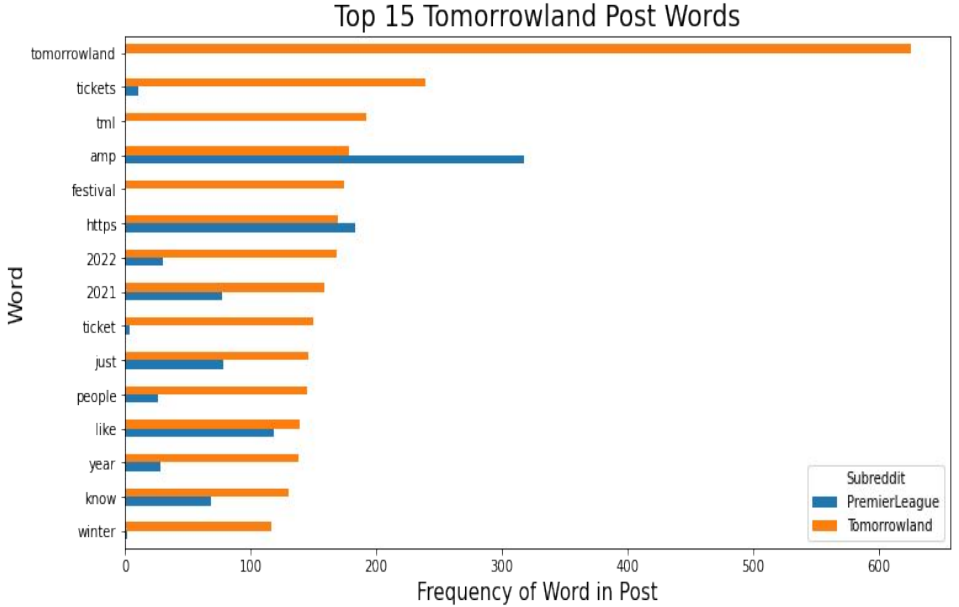


Average number of characters in selftext: 182
Average number of words in selftext: 28

Average number of characters in selftext: 197
Average number of words in selftext: 33

Cleaning and Exploratory Analysis





Average number of characters in post: 244
Average number of words in post: 39

Average number of characters in post: 245
Average number of words in post: 42

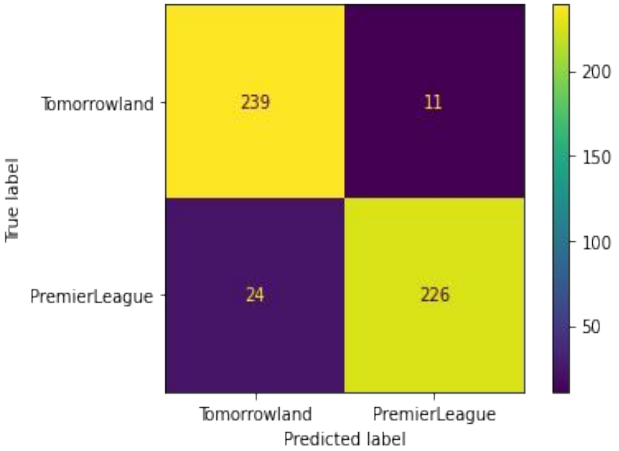
Models and Accuracies

		Title	Selftext	Entire Post
KNN	CountVectorizer	83.4%	66.9%	83.6%
	TFIDFVectorizer	84.5%	68.7%	84.6%
Logistic Regression	CountVectorizer	89.8%	73.6%	92.9%
	TFIDFVectorizer	91.1%	74.0%	93.9%
Decision Tree	CountVectorizer	87.3%	71.3%	90.3%
	TFIDFVectorizer	87.1%	70.3%	89.9%
Random Forest	CountVectorizer	88.5%	73.7%	92.0%
	TFIDFVectorizer	88.8%	73.5%	92.3%

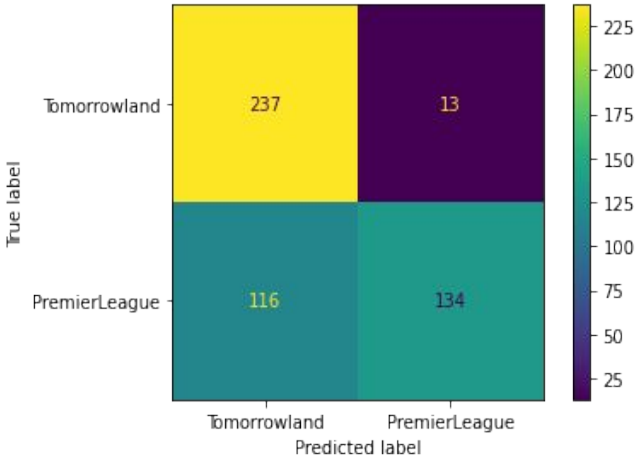
Examining the Errors - Best Models

Logistic Regression and TFIDFVectorizer Model Confusion Matrices

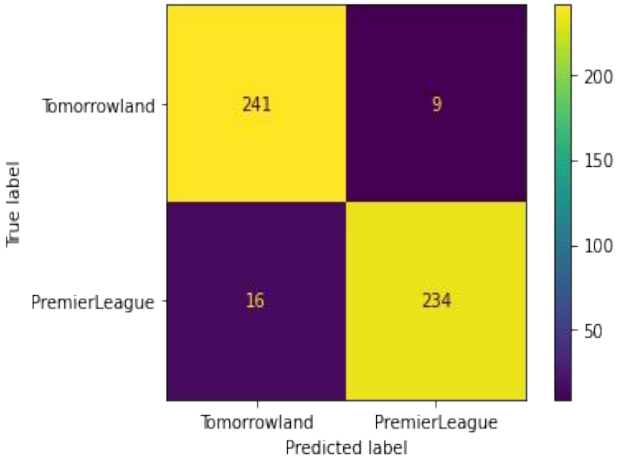
Title



Selftext



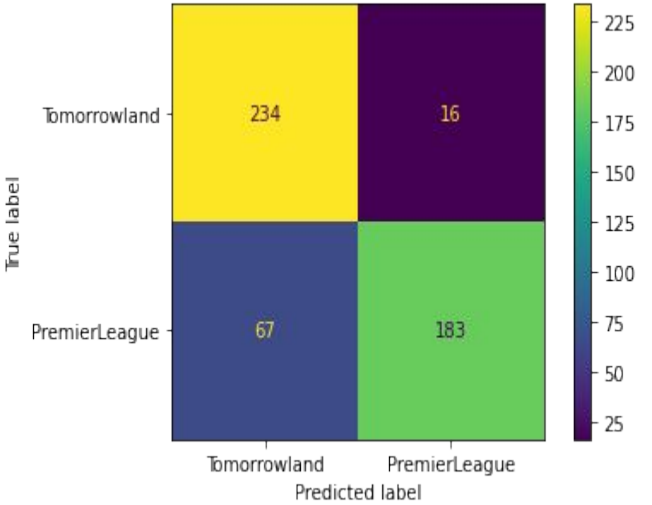
Entire Post



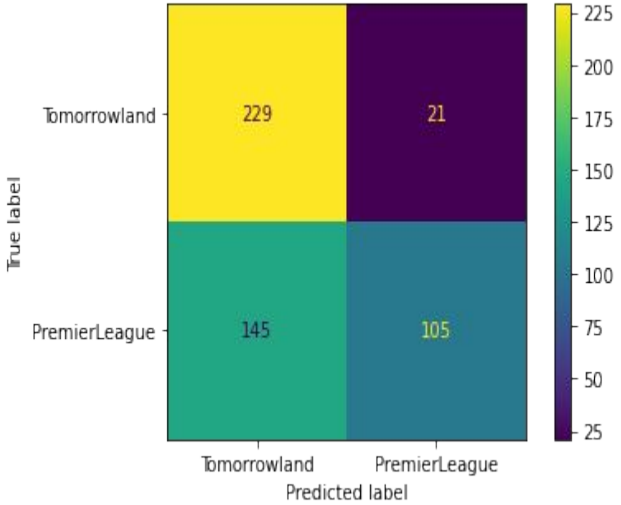
Examining the Errors - Worst Models

KNN and CountVectorizer Model Confusion Matrices

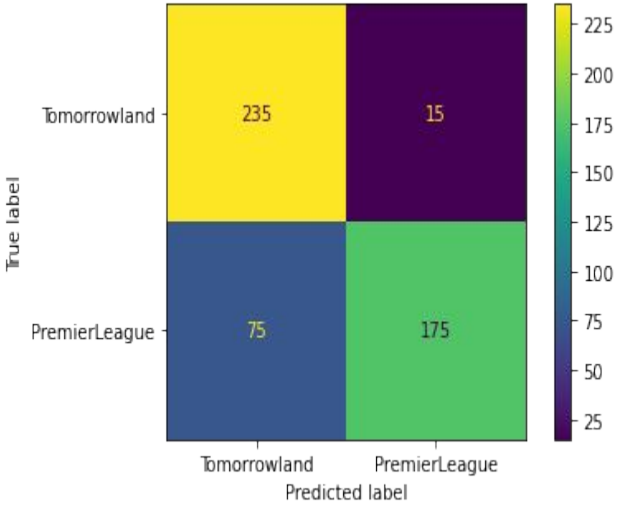
Title



Selftext



Entire Post

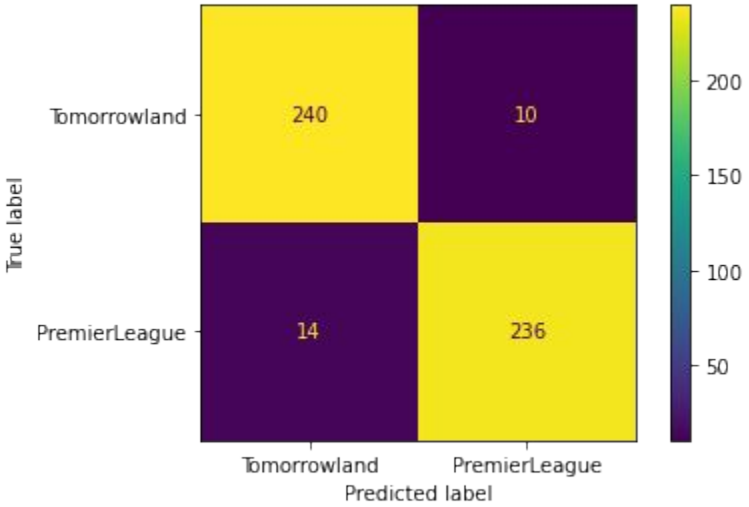


Conclusion and Recommendation

Given two subreddits, the best way to create a model for predicting which subreddit a post comes from would be to give it both the title and selftext information to help it predict which subreddit that post came from. Some subreddits do not have much in the selftext of a post but at least some information can be gained from the title. Therefore I would recommend creating a model based on the entire post and not just the title or selftext and using a logistic regression model with a tfidfvectorizer to get the most accurate predictions.

BONUS

I created a stacked model with the 3 most accurate models using the entire post with a logistic regression model as the 2nd level model that was able to predict the correct subreddit with 95.2% accuracy.



QUESTIONS???