Classifying Posts in Two Popular Gaming Subreddits

Dominic Martorano

DSB-EC-1211

Problem Statement

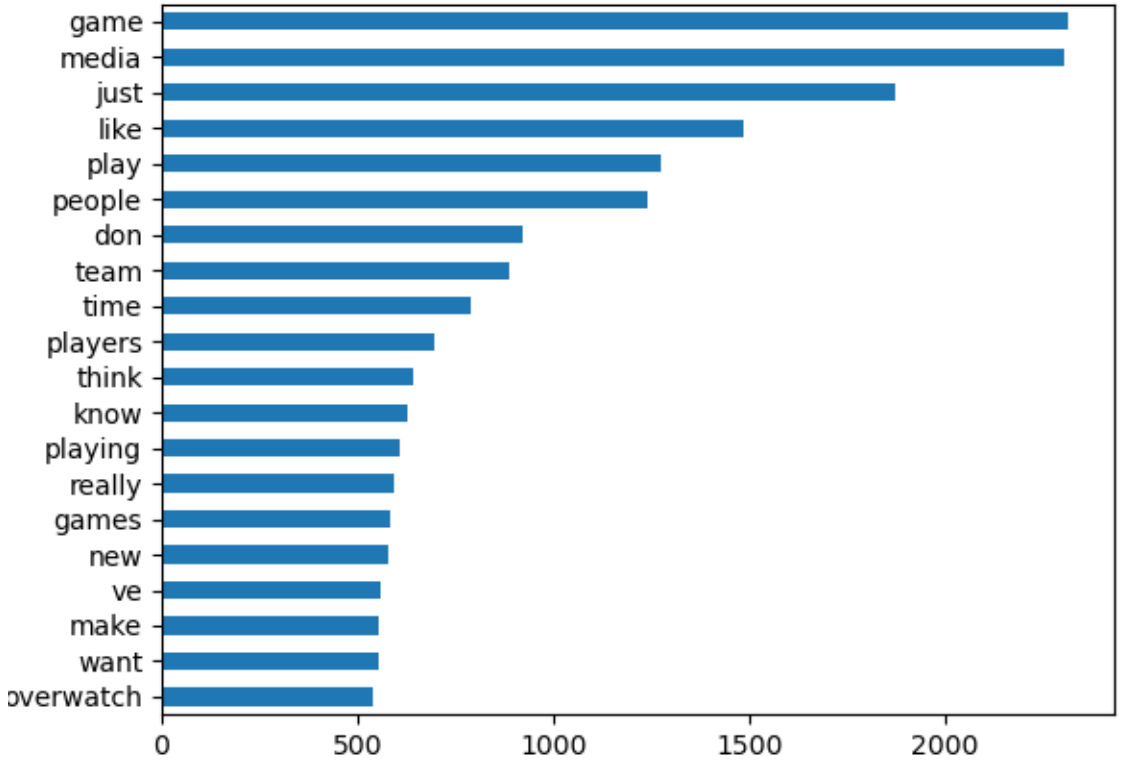
Data scientist asked by group of gamer friends to create a model that predicts what subreddit a post has been taken from

Data Collection and Cleaning

- Gathered posts from Overwatch and Apex Legends subreddits over three days
- Took data from the ‘new’, ‘top’, and ‘controversial’ streams
- Ensured that no duplicate posts came from any of the streams
- Made any blank “selftext” posts into the string “Media”

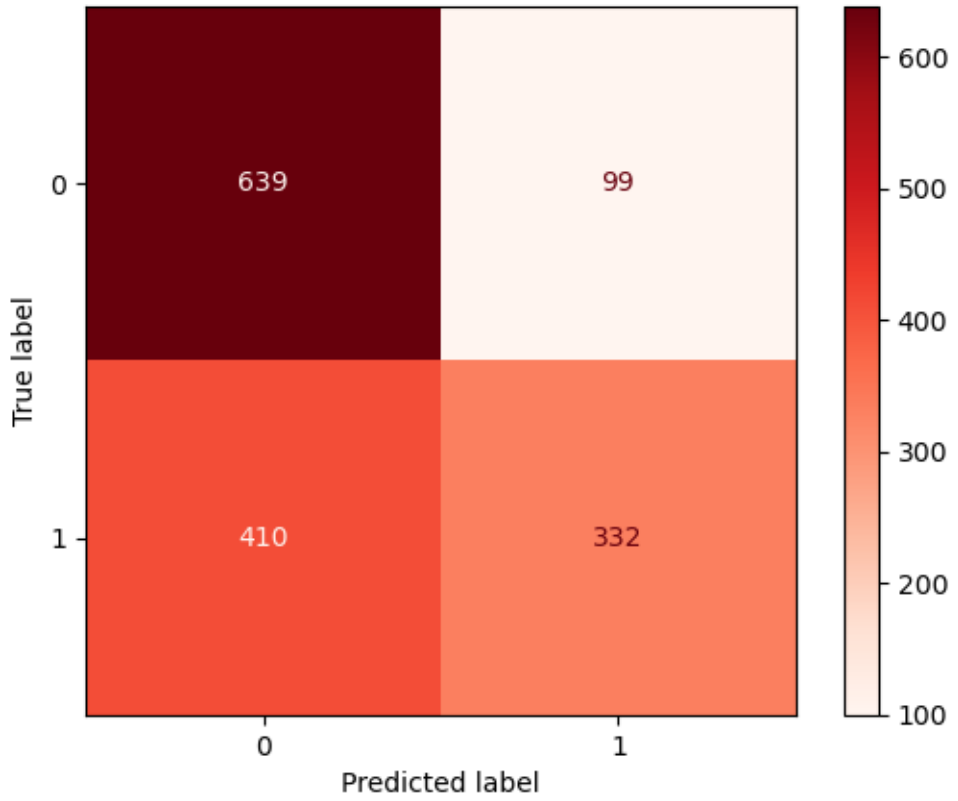
Observations about Words

- Some outliers present in word counts
- Ran models and found that with or without outliers, the difference in accuracy was not significant



Model Selection

- Fit and tested four different models
- Logistic Regression with TF-IDF Vectorizer was best model
- Model produced significantly more Type II error



Conclusion

- Model was able to predict better than baseline accuracy
- Relatively small improvement over baseline could be indicative of communities being fairly similar
- Gamers slightly disappointed with predictive strength