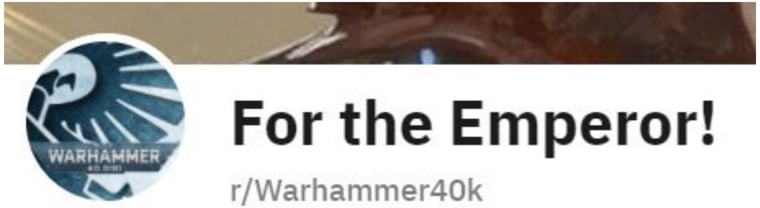
# Classifying posts from two role playing game subreddits:

Dungeons and Dragons & Warhammer 40k

Dungeons & Dragons

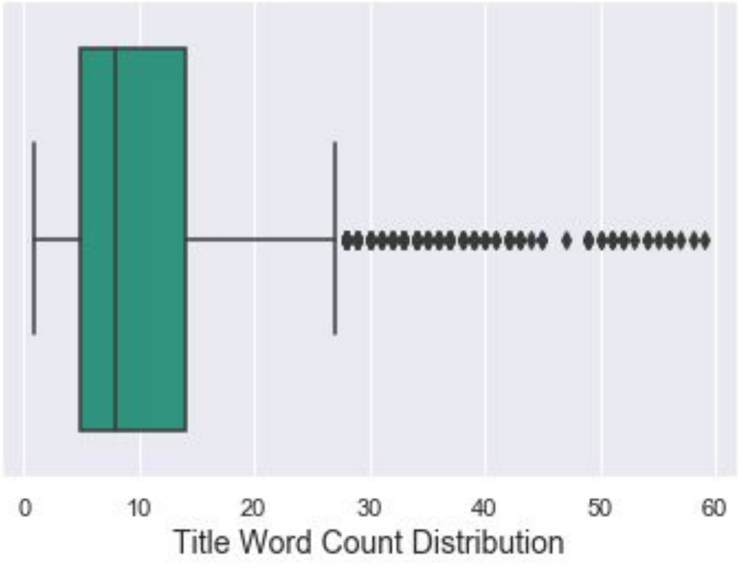
r/DungeonsAndDragons


For the Emperor!

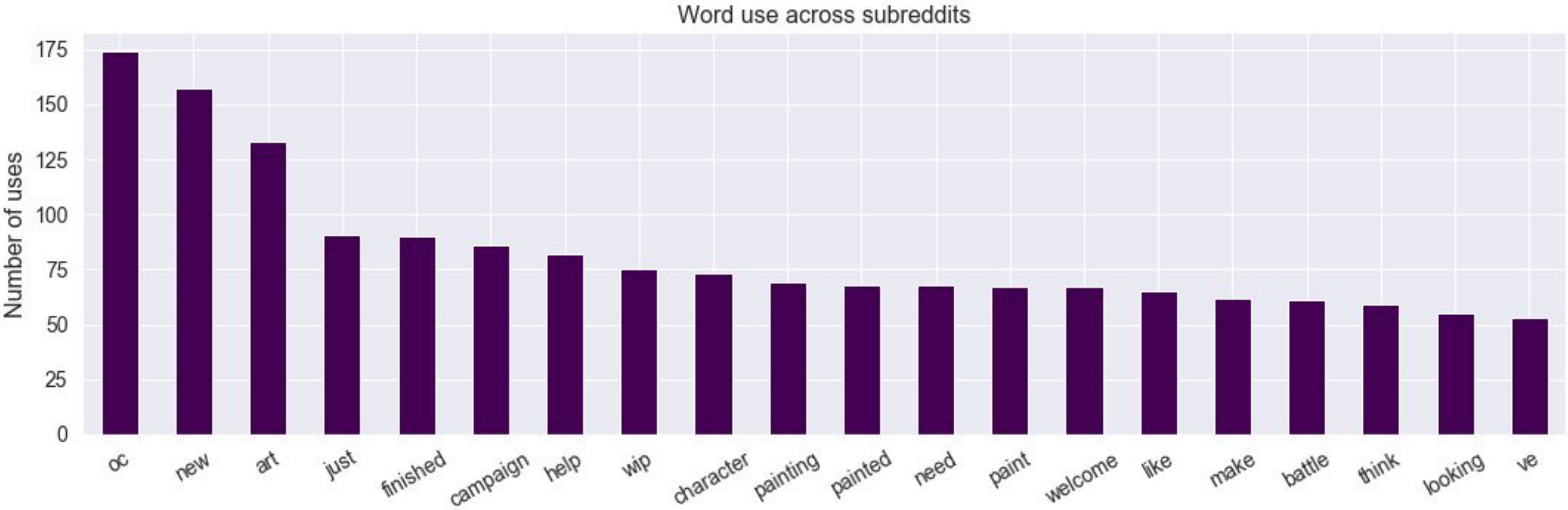
r/Warhammer40k

# Summary: classifying titles

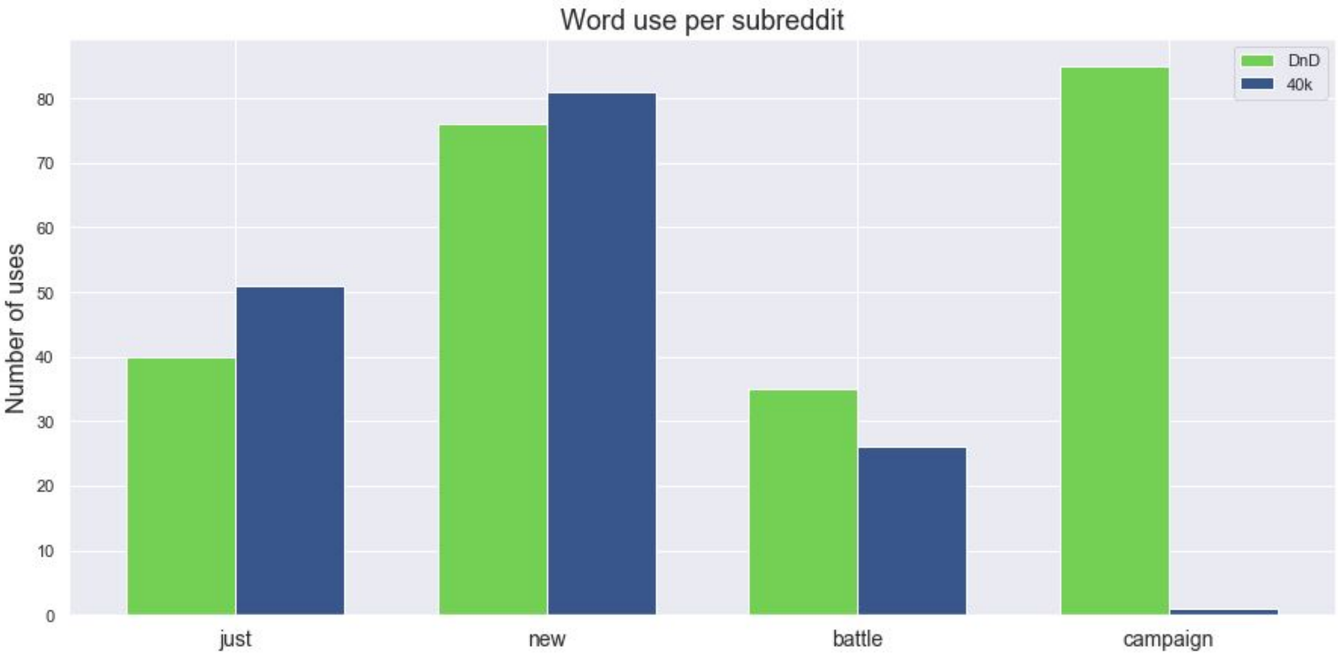
- Focus on subreddit titles
- Data cleaning and analysis of corpus
- Feature inclusion/exclusion
- Naïve Bayes & Random Forest
- Conclusions and Recommendations



# Data Cleaning

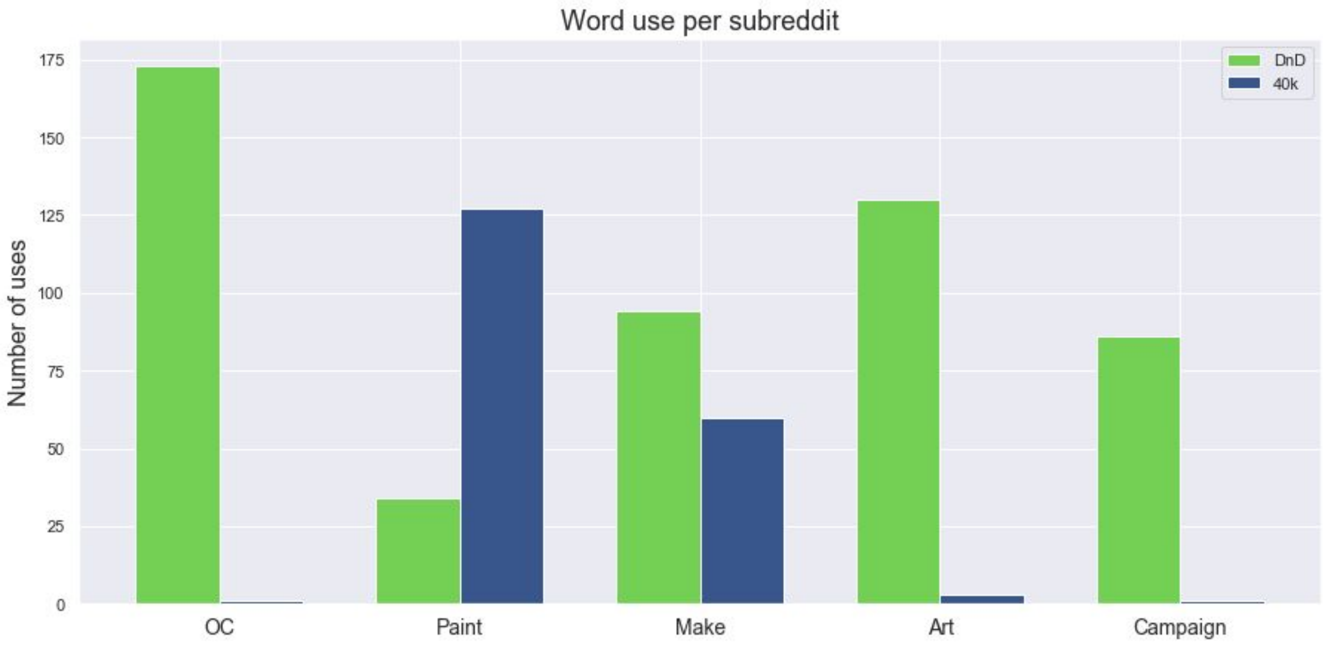


# Data cleaning and EDA



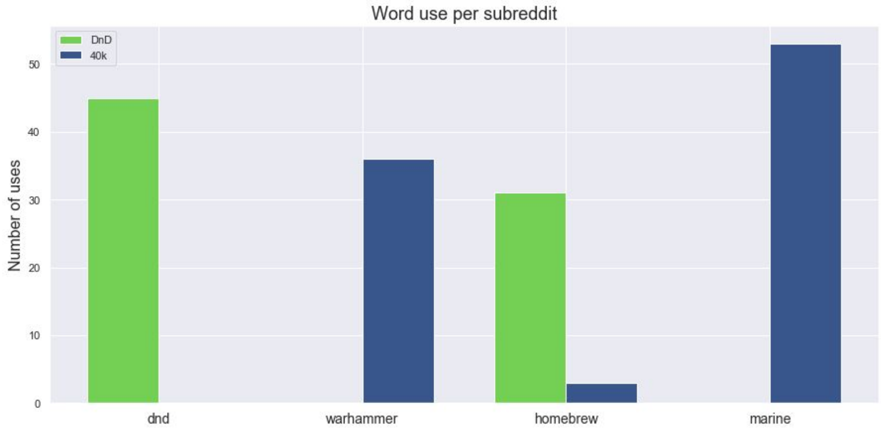
Some examples of popular words in corpus that may or may not be useful.

# Experimental Data Analysis

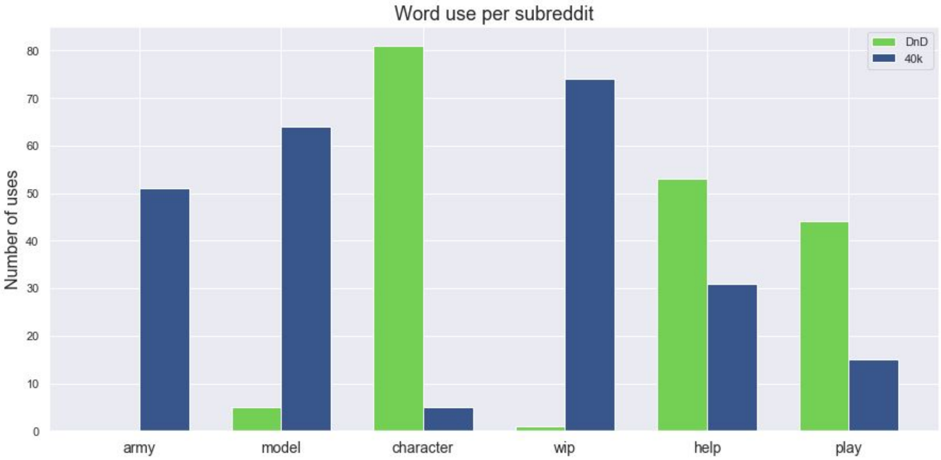


Frequency of most common words in each group indicates differences between subreddits.

# More fun words!



Words that will have an obvious improvement on model accuracy



Other fun words commonly used in role playing games

# Modeling: Naïve Bayes & Random Forest

Final data set:

- 2574 rows
  - 1283 from r/DungeonsAndDragons
  - 1291 from r/Warhammer40k
- 954 columns

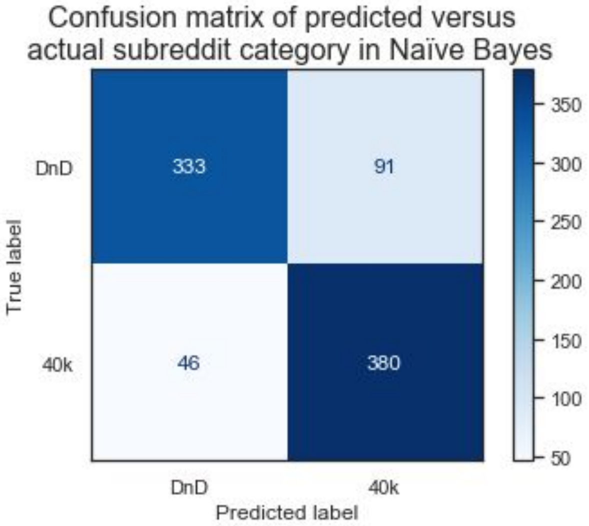
| **Naïve Bayes** | Model training data | Model testing data |
|---|---|---|
| Accuracy | 0.9176 | 0.8388 |

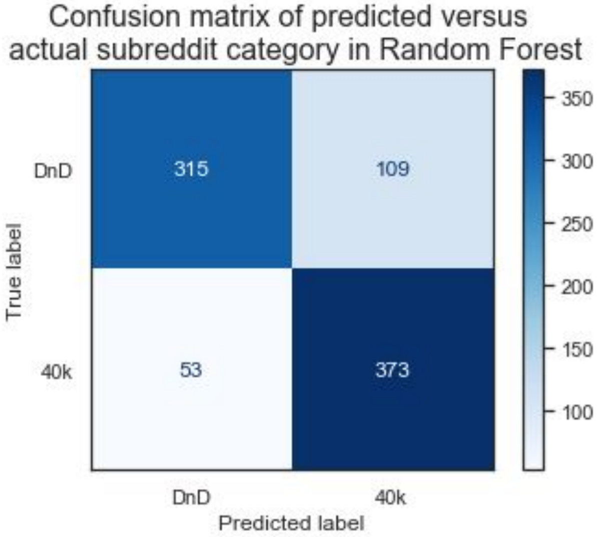
| **Random Forest** | Model training data | Model testing data |
|---|---|---|
| Accuracy | 0.9751 | 0.8035 |

# Model evaluation

*Terminology relative to predicting posts from r/Warhammer40k

| | **Naïve Bayes** | **Random Forest** |
|---|---|---|
| **specificity** | 0.7854 | 0.7358 |
| **sensitivity** | 0.8920 | 0.8709 |
| **precision** | 0.8068 | 0.7681 |


Confusion matrix of predicted versus actual subreddit category in Naïve Bayes


Confusion matrix of predicted versus actual subreddit category in Random Forest

# Conclusions and recommendations

- Models have high variance but Naïve Bayes was slightly better
- Naïve Bayes model has good accuracy given data available

- Collect more data
- Analysis on the body of posts
- Try other models
- Image analysis