PROJECT 3: GOT VS LOTR SUBREDDIT NLP CHALLENGE

GAME OF THRONES

BY: MELISSA MCMILLAN

JAN 22, 2021

LORD OF THE RINGS

OUTLINE

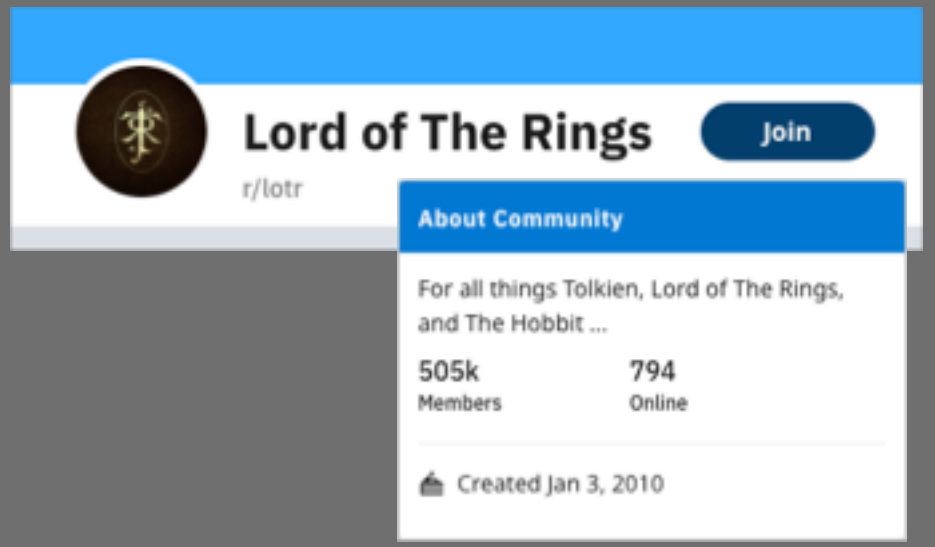
- Problem statement
- Data Collection
- EDA & Cleaning
- Naïve bayes modelling results
- Random forest modelling results
- Conclusions & Recommendations

PROBLEM STATEMENT

- MAIN GOT SUBREDDIT MODERATOR NEEDS TO FILTER LOTR POSTS OUT OF HER BEAUTIFUL, PRISTINE SUBREDDIT PAGE
- CAN I DEVELOP A MODEL TO MAKE HER LIFE EASIER?
- MODELS: NAÏVE BAYES, RANDOM FOREST CLASSIFIER
- KEY METRICS: ACCURACY & SENSITIVITY

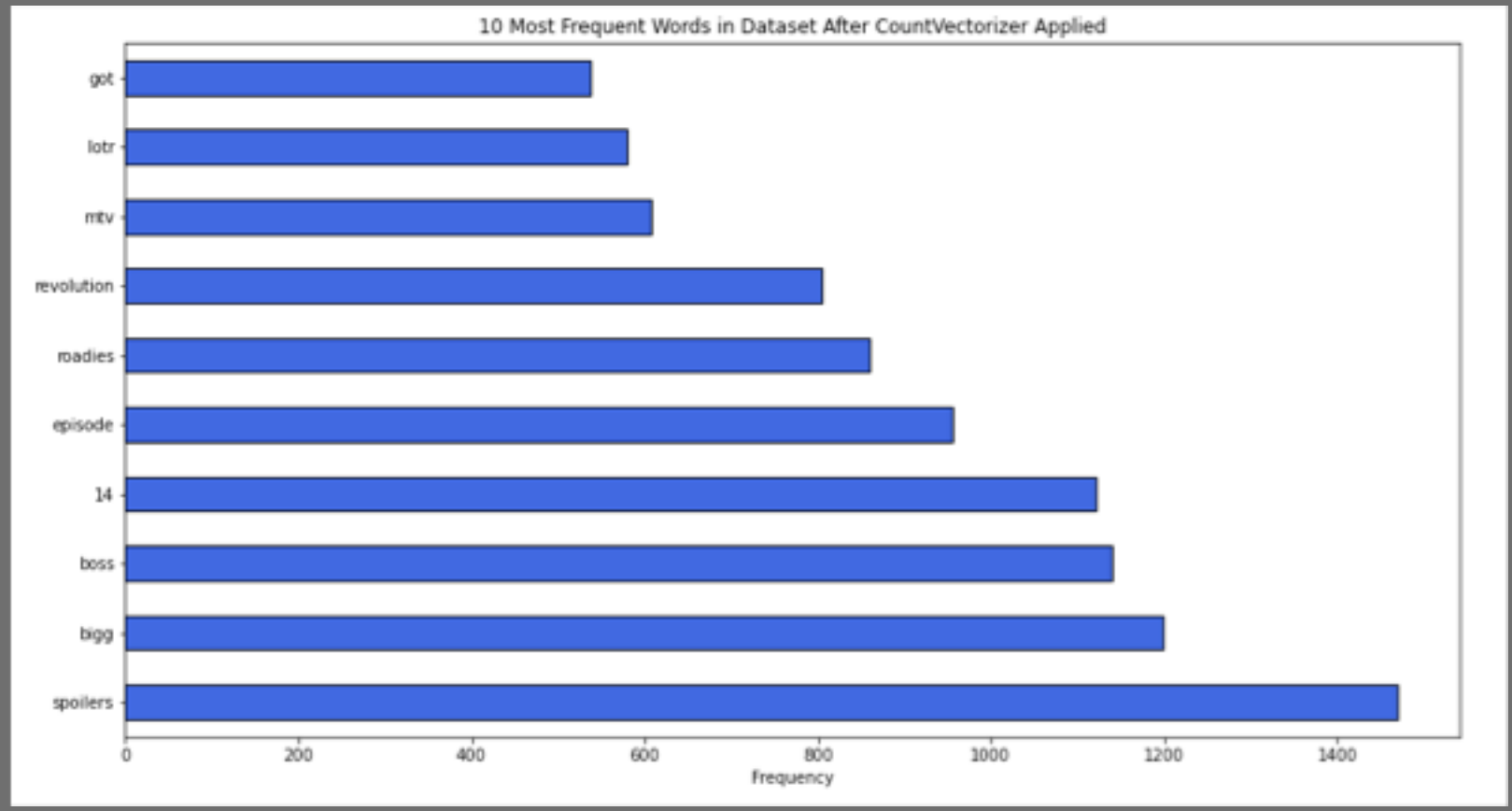
DATA COLLECTION: R/GOT AND R/LOTR

- USED PUSHSHIFT'S API
- 5,100 TITLES WITH DESCRIPTIONS FROM BOTH SUBREDDITS
- 10,200 TOTAL SUBREDDIT POSTS
- ONLY USED TITLES (NOT DESCRIPTIONS) IN PROJECT



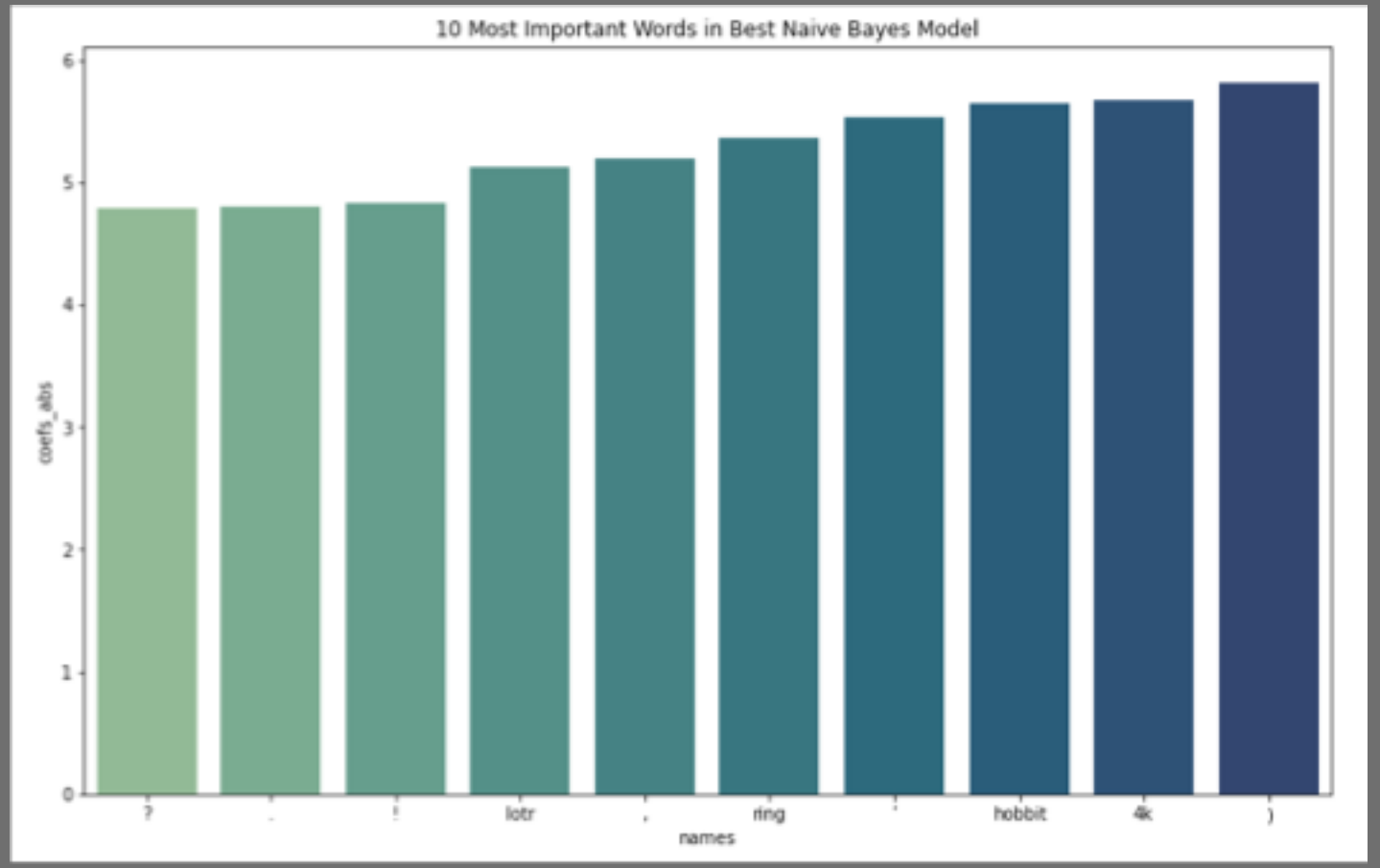
EDA & CLEANING

- Dropped post descriptions
- Dropped nulls
- Utilized regex tokenizer:
  - Removed emojis
  - Removed special characters
  - Removed symbols



NAÏVE BAYES MODEL RESULTS

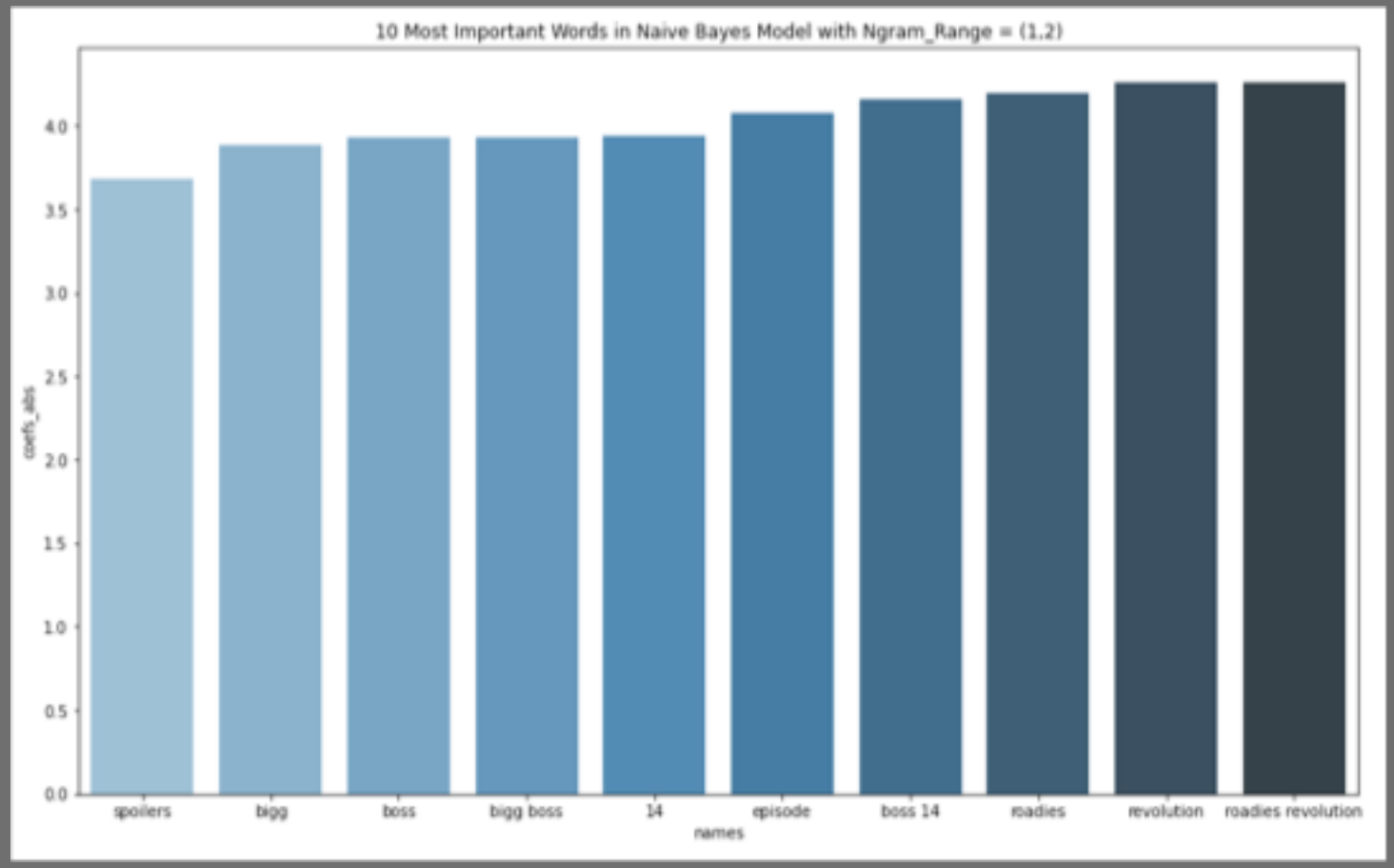
- BEST MODEL:
  - TFIDF VECTORIZER W/ WORDNET LEMMATIZER
  - TEST SCORE: 93%
  - RECALL SCORE = 0.89
  - NORM FN = 11% (LOWEST OF NB MODELS)
  - STOP WORDS = ENGLISH
  - NO MAX FEATURES
  - NB ALPHA = 1.0
  - DIRTY DATA



NOTE: ABS(IMPORTANCE)

NAÏVE BAYES MODEL OBSERVATION: WINTER IS NOT COMING

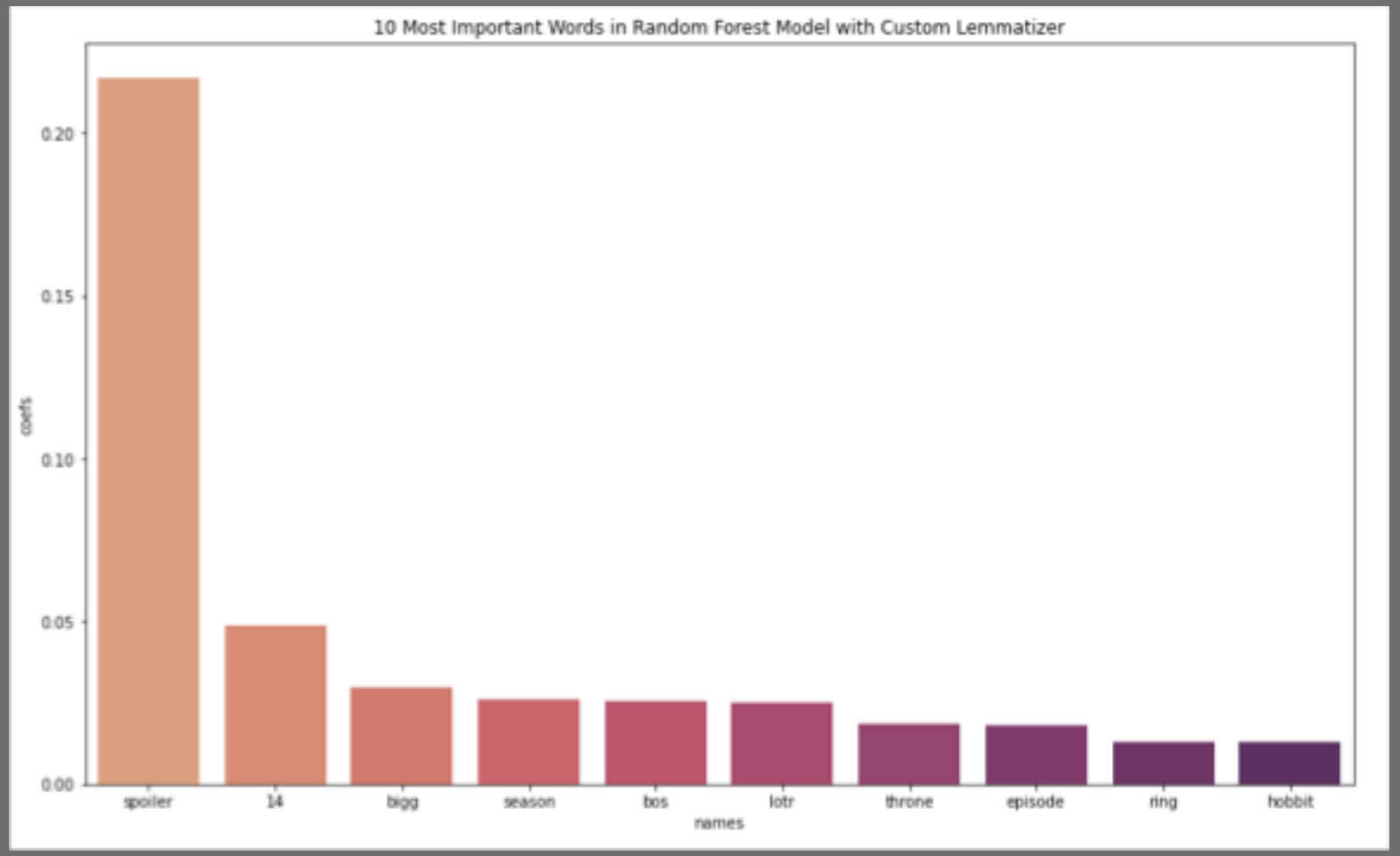
- EXPERIMENTED WITH NGRAM RANGES (1, 2), (2,2), (1,3), (1,4), AND (3,3)
- DIDN'T HELP MY MODEL MUCH
- BEST NGRAM MODEL WAS (1, 2)
- TWO-WORD FEATURES WERE BARELY IN THE TOP IMPORTANT FEATURES
- WORDNET LEMMATIZER USED HERE



NOTE: ABS(IMPORTANCE)

RANDOM FOREST MODEL RESULTS

- Less overfitting
- Used clean data
- Best model:
  - Tfidfvectorizer w/ wordnet lemmatizer
  - Test score: 93%
  - Recall = .91
  - Norm FN = 9% (lowest of rf models)
  - Stop words = English
  - max features = 1000
  - Default random forest params



CONCLUSIONS/RECOMMENDATIONS

- Can delineate between subreddits with ~93-94% accuracy
- Will need to do some data cleaning
- Random forest is the preferred model (less overfitting)
- Lemmatizers matter
- The tfidf vectorizer with wordnet lemmatizer did best
  - Utilize stop words, max features
- Future Work:
  - Incorporate more types of models
  - Utilize the Roc-Auc score to evaluate sensitivity
  - Sentiment Analysis on each type of subreddit

IMAGE REFERENCES
(ORDERED BY WHEN THEY APPEAR)

1. https://do816.com/events/2018/4/13/lord-of-the-rings-vs-game-of-thrones
2. https://www.reddit.com/r/gameofthrones/
3. https://imgur.com/gallery/D4NcyvD/comment/189944230?nc=1
4. https://www.reddit.com/r/lor/
5. https://catvills.com/game-of-thrones-cat-names/
6. https://www.pinterest.com/pin/179229260144682544/
7. https://www.reddit.com/r/lotr/comments/jzglqz/meet_my_cat_the_white_wizard
8. https://www.popsugar.com/family/Game-Thrones-Cats-46082964