LYRICS BASED GENRE CLASSIFICATION

EMILY KRUEGER

OCTOBER 2023

INTRODUCTION

- **Krueger Consulting** has been contracted to explore the following question:

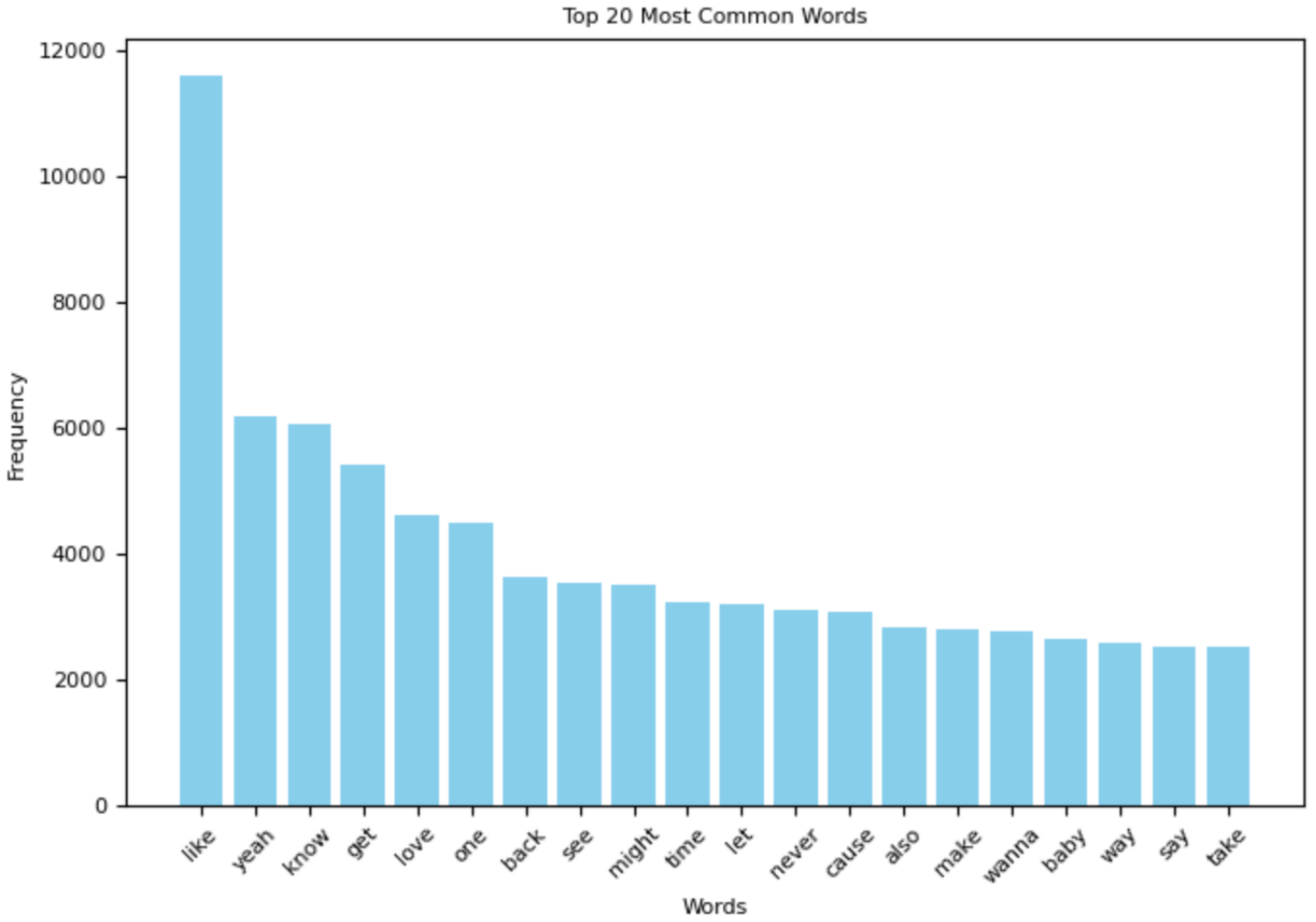
Can song lyrics be used to predict genre?

- We've developed multiple classifier models utilizing NLP analysis including:
  1. Random Forest
  2. Multinomial Naïve Bayes
  3. Logistic Regression
  4. Neural Network
- Model performance will be evaluated on accuracy, as we are not particularly concerned about false positives or false negatives

DATA OVERVIEW

- Dataset was compiled using a combination of the Spotify API and the Genius API
- Process:
  1. Use Spotify API to download track names and artists from a number of genre based curated playlists for each of the following genres:
     - Hip-Hop
     - Punk
     - Country
  2. Genre tag each track based on its source playlist
  3. Use the Genius API and the list of track names and artists to search Genius and download each track's lyrics
- Final dataset consisted of 2,880 rows and four columns (track name, artist name, lyrics, and genre)
  - Predictive Variable: Lyrics
  - Target: Genre

CORPUS STATISTICS

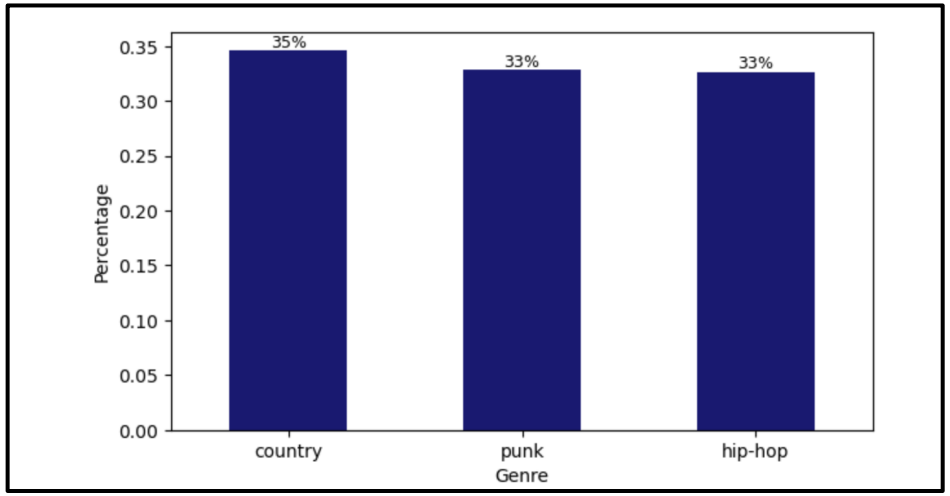


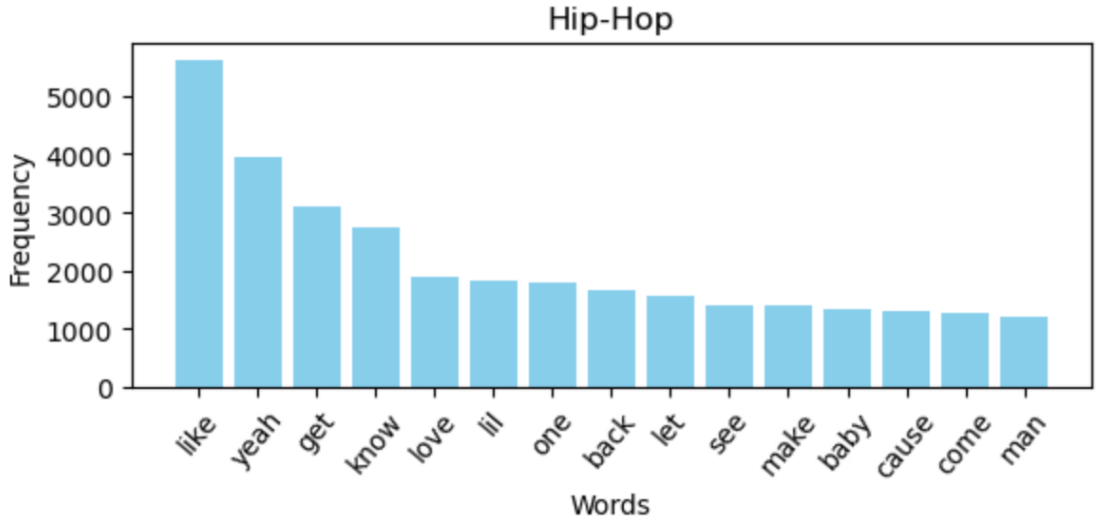
- Post tokenizing and removing stopwords:
  - Word Count: 760,567
  - Unique Words: 56,051

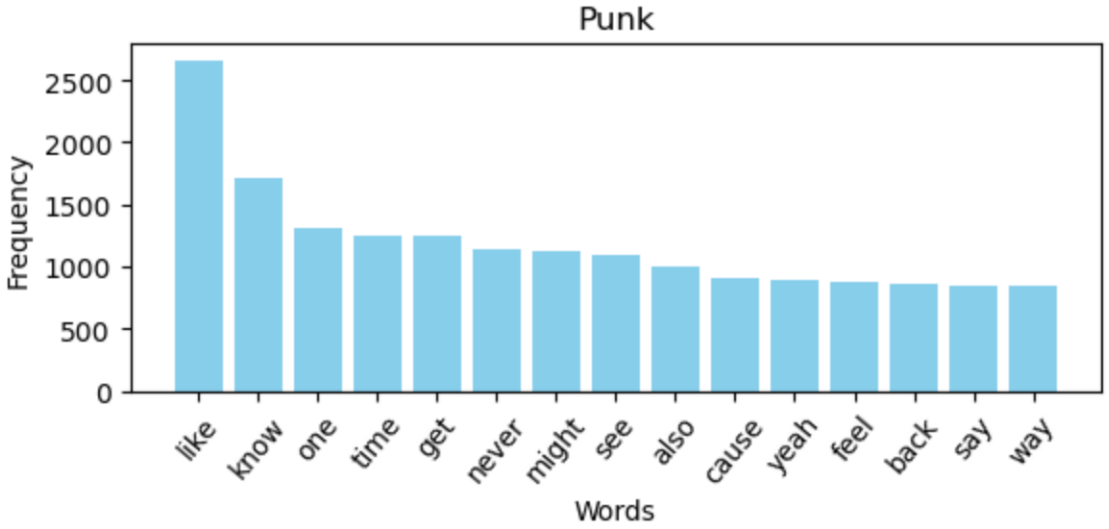
1) Post preprocessing: tokenized via RegexpTokenizer including only three+ letter words, stopwords removed, select inappropriate words removed.

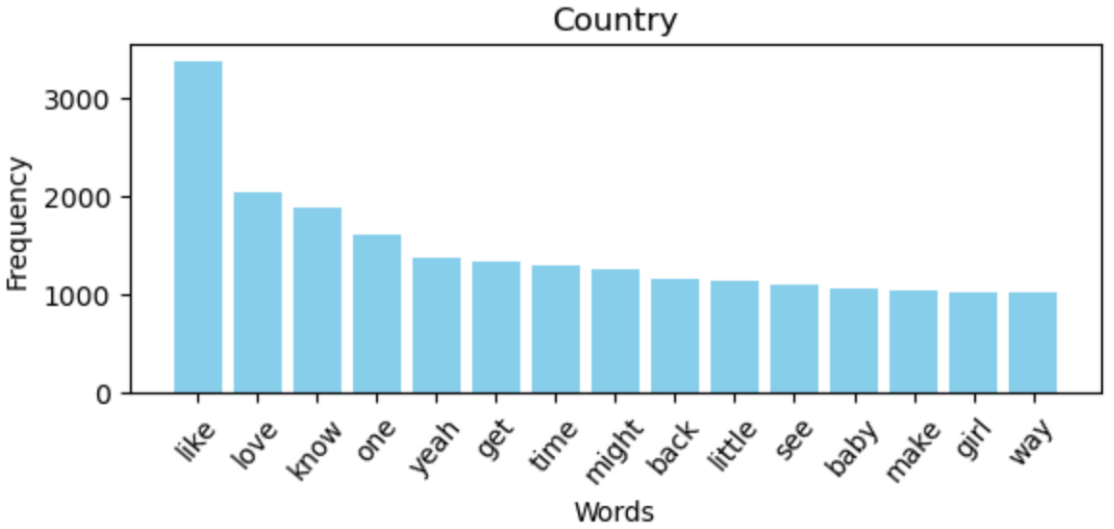
CORPUS STATISTICS BY GENRE

Genre Split:

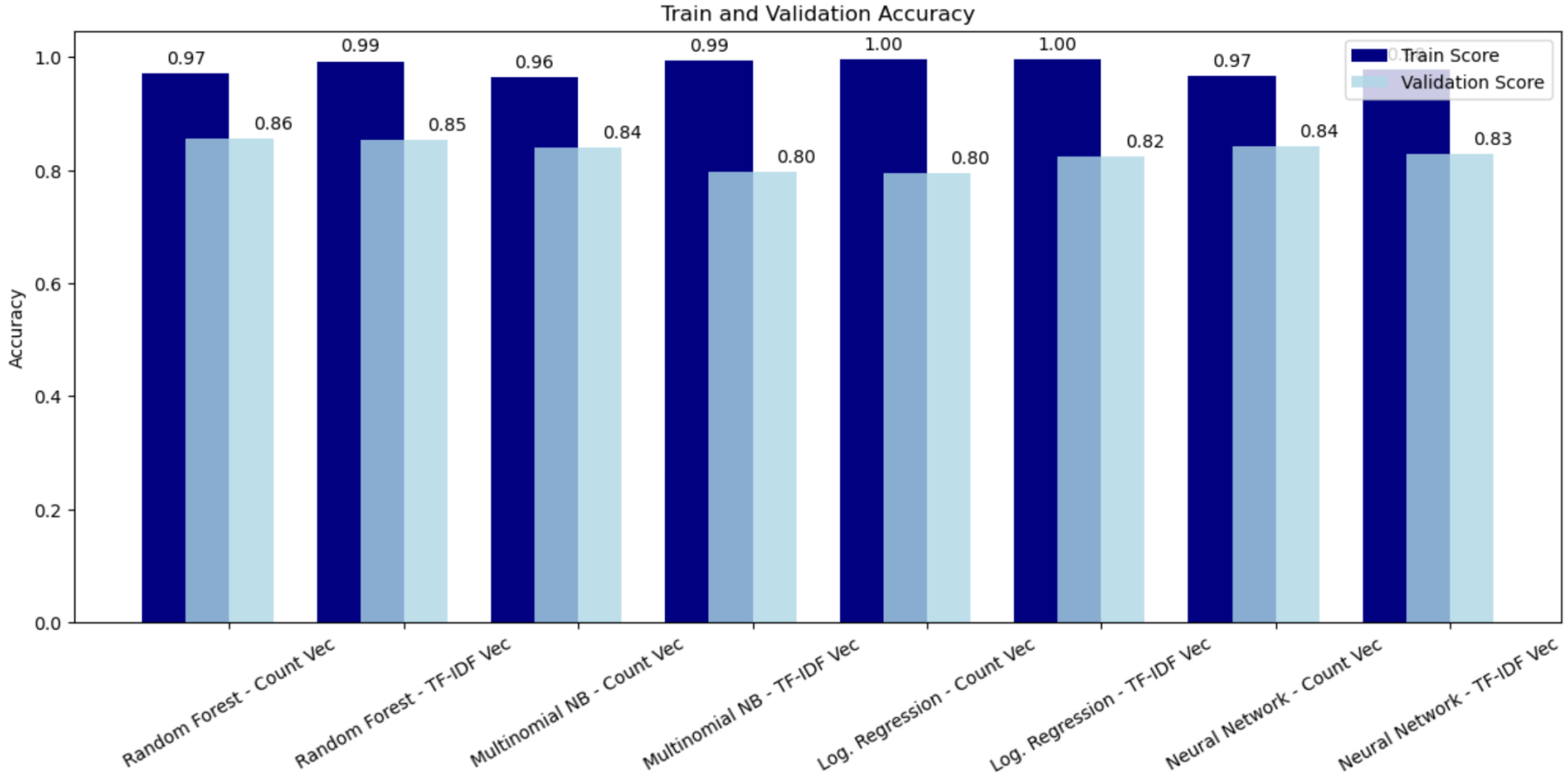






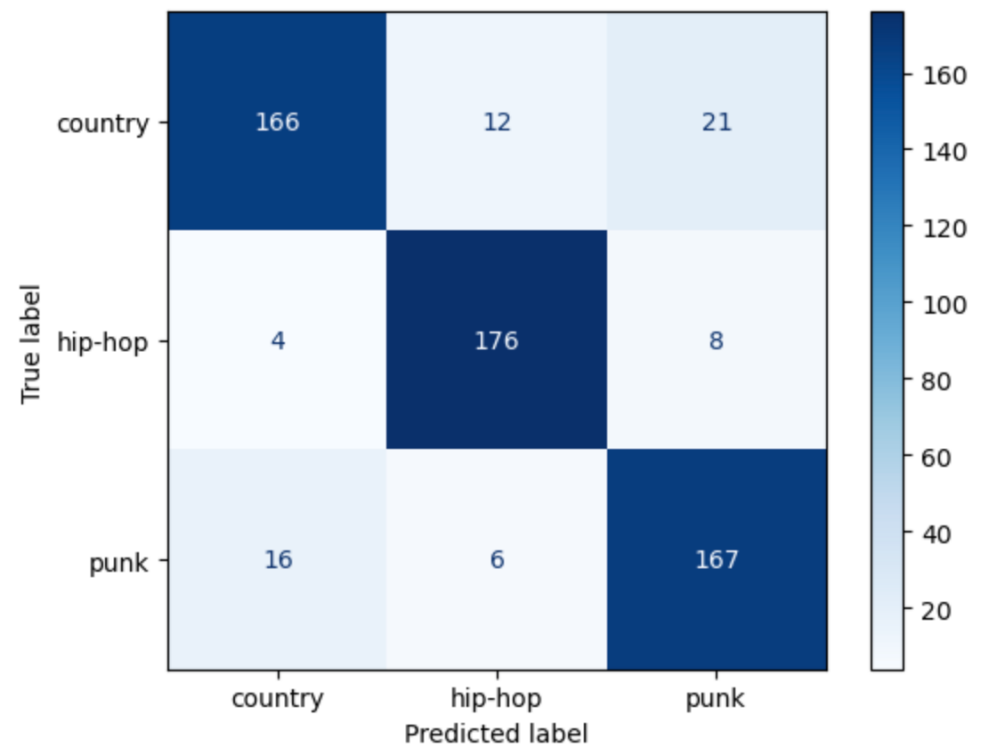


MODEL COMPARISON



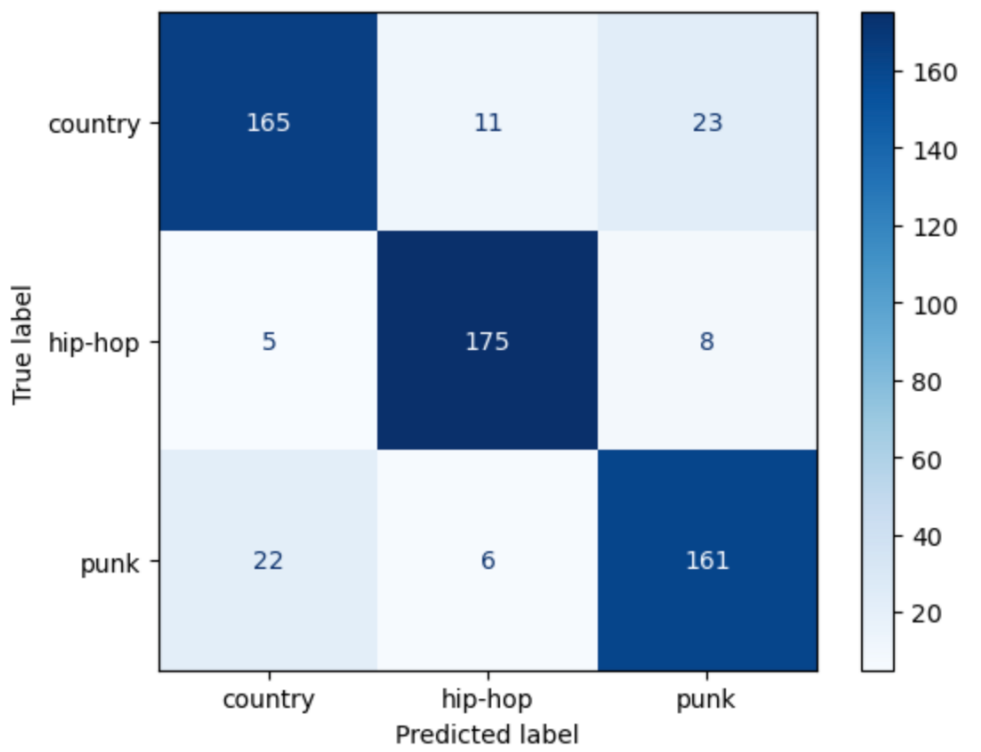
EVALUATION ON TEST DATA

Random Forest – Count Vectorized



- Accuracy: 88.4%
- Recall: 88.4%
- Precision: 88.4%
- F1: 88.3%

Random Forest – TF-IDF Vectorized



- Accuracy: 87.0%
- Recall: 87.0%
- Precision: 87.0%
- F1: 87.0%

CONCLUSION

- **Recommendations:**
  - This model can be used to create genre based playlists based on a set of songs or to evaluate genre of a song for genre tagging as songs are added to a given platform

Additional Considerations/Next Steps:

- Models across the board are still overfitting, one of the following is needed to improve the model for future use:
  1. This problem cannot be solved with lyrics alone and musical attributes should be added to better the model
  2. Train model on far larger dataset to improve model performance on unseen data. With this, add additional genres to the analysis to build a more comprehensive, useful model

BIOGRAPHY

Emily Krueger

Email: ekrueger1217@gmail.com

M: 732-403-4566

Github: https://github.com/ekrueger1217

LinkedIn: https://www.linkedin.com/in/emily-krueger-058513103/