
Classifying Music Producers and Composers

— NLP Models Using Reddit Data —

Introduction

Background and Data Science Problem

- Music producers use Digital Audio Workstations (DAWs) such as Ableton
- Composers use scorewriter software such as Sibelius
- DAW market size in 2021: **\$2.49 Billion**¹, with CAGR estimated at **9%**
- Problem: Using message data from persons interested in making music, determine whether they're a **"Producer"** or a **"Composer"** so that we can properly market software

1. <https://www.fortunebusinessinsights.com/industry-reports/digital-audio-workstation-market-100150>

Data and Methods

- We train classification models to guess whether a Reddit post came from **r/composer** or **r/musicproduction**
- Using **PRAW**, we scrape over 1000 posts from each subreddit
- We use posts' **title texts and body texts** to build “bag of words” models
 - **URL preprocessing** using Regular Expressions
 - Count vectorization
- Scoring based on cross-val **accuracy**; validation with test data accuracy

Example Post: Composer or Producer?

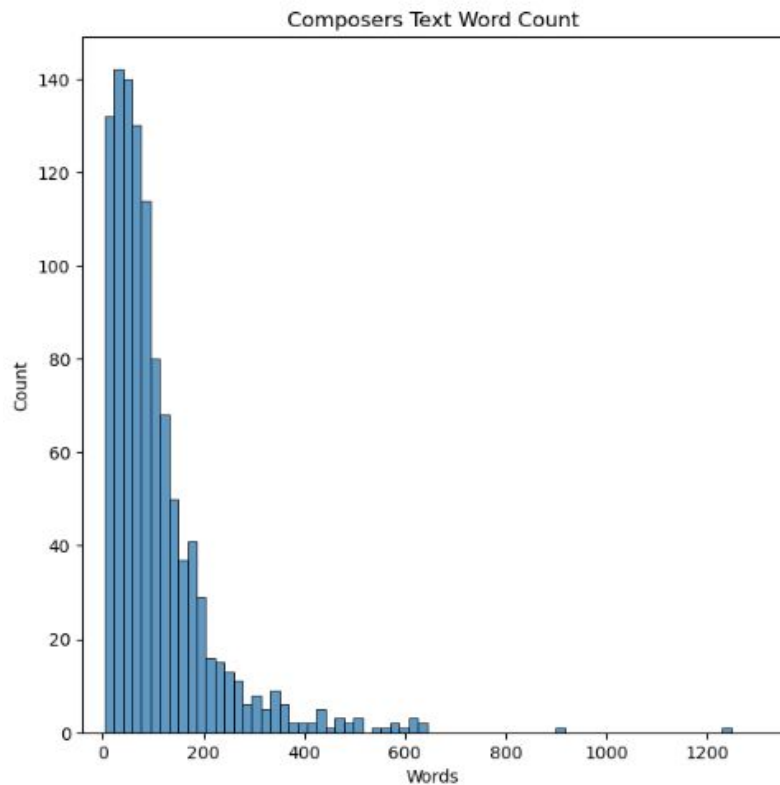
help with evolving

=====

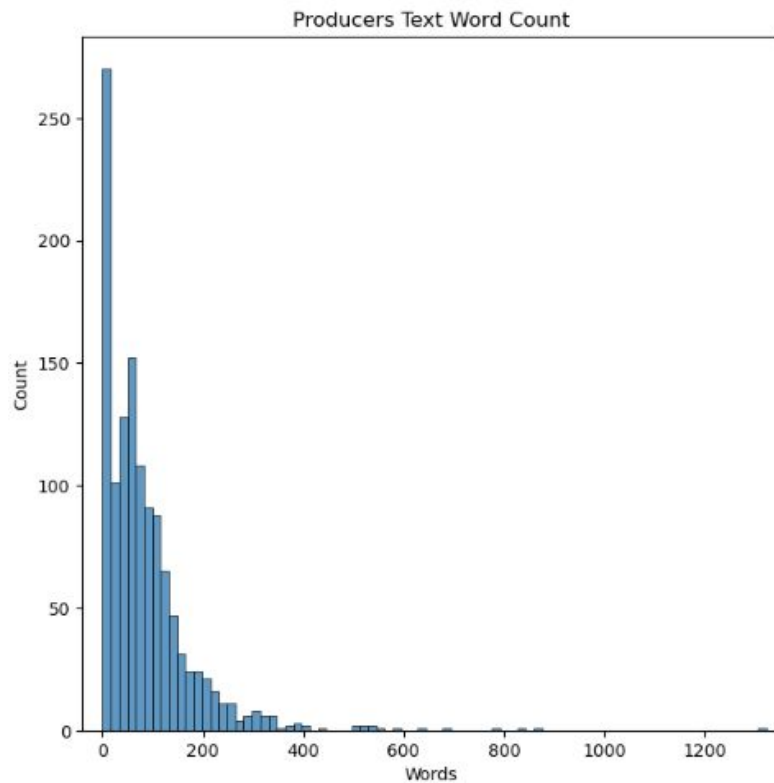
i have a problem. i can make a melody and so. but when im done with the melody, i often find it hard to continue on it. what do you guys do when you have something you wanna work on but dont know how to continue or what to do next?

Exploratory Data Analysis

EDA: Word Count

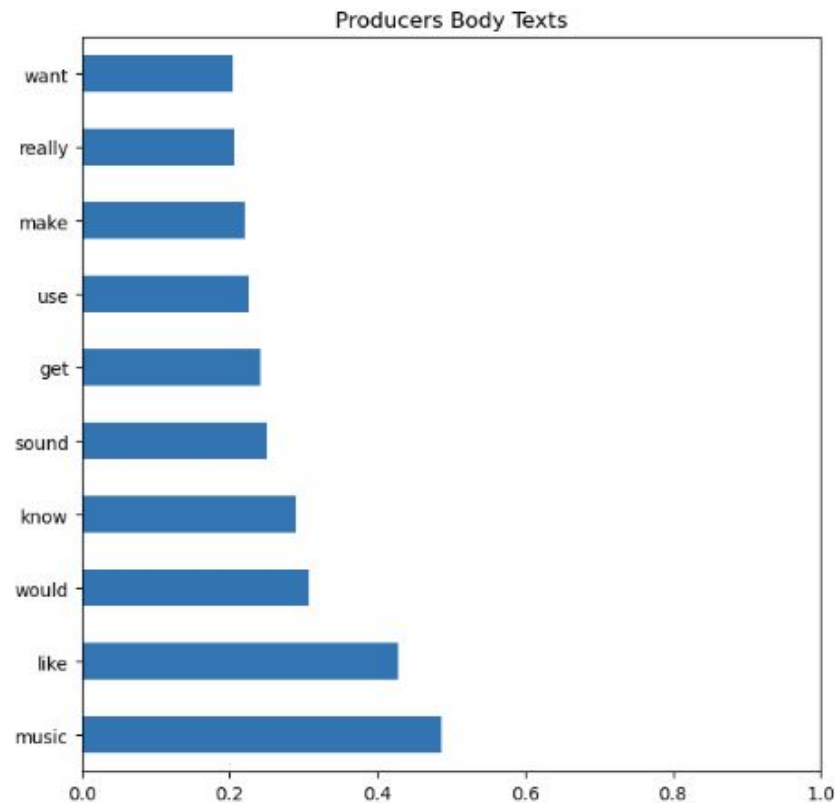
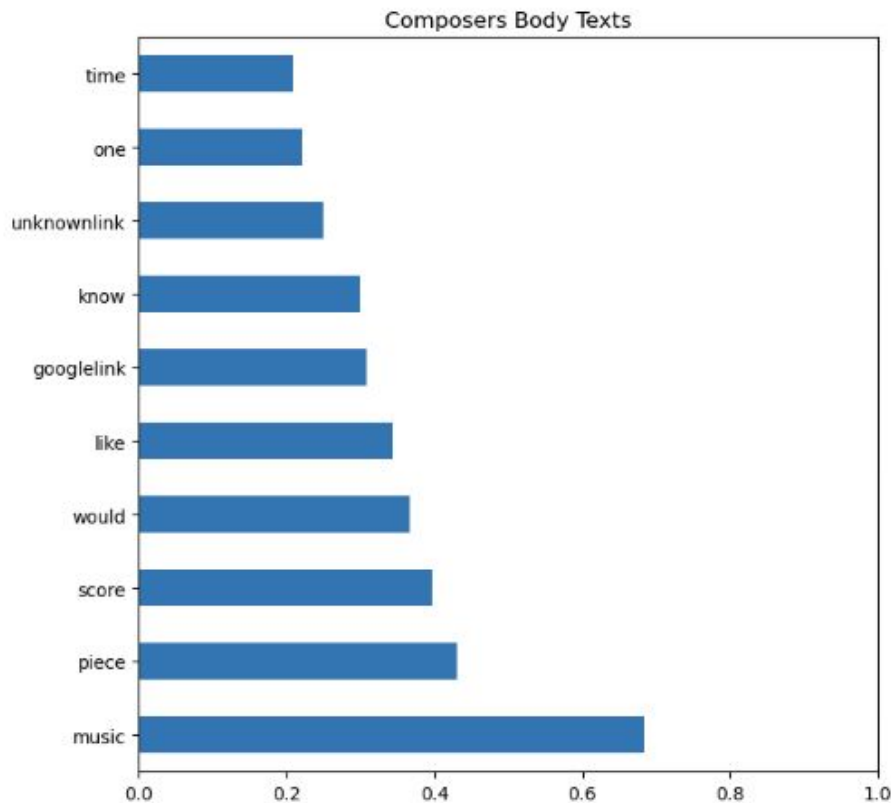


Mean: 106, SD: 108



Mean: 86, SD: 100

EDA: Most Common Words (after only URL processing)



x-Axis = Average Frequency Per Post

Modeling

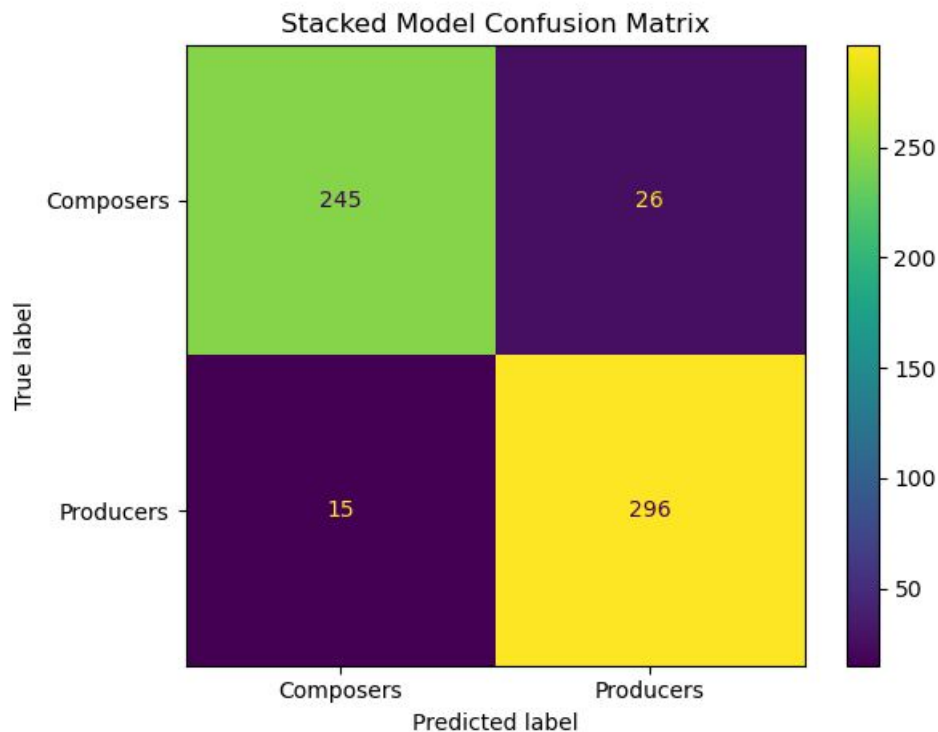
Modeling Techniques

- Gridsearch to tune hyperparameters
- Some methods of preprocessing/modeling helped a lot:
 - **Preprocessing URLs:** misclassification ↓ 15-25%
 - **Stacking best models:** misclassification ↓ 15%
 - Removing stop words
- Others did not help:
 - Stemming / Lemmatizing
 - TF-IDF word encoding
 - Adding bigrams
 - Including word counts as features
 - Considering only titles (no body text)

Best Models

	Crossval Score	Test Data Score
Stacked Model	0.940325	0.929553
Naive Bayes	0.936311	0.917526
Random Forest	0.922524	0.914089
Gradient Boost	0.913938	0.908935
Logistic Regression	0.911074	0.903780

Test Data Classifications



Classification Examples

Example Classifications (Stacked Model)

Looking for a virtual instrument like this. Can anyone tell me what it is (video link)

=====

Looking to buy a a virtual instrument to whatever that is being played in video? Anyone know which instrument and what's a good library that carries this?

[SIDEWAYS] (<https://youtube.com/clip/UgkxypJBTTWFtamIb60WBQEb8KZznanwNw8D>)



Prediction



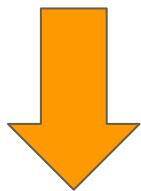
"Composers" (52.7%)



Example Classifications (Stacked Model)

help with evolving

i have a problem. i can make a melody and so. but when im done with the melody, i often find it hard to continue on it. what do you guys do when you have something you wanna work on but dont know how to continue or what to do next?



Prediction



"Producers" (94.4%)



Room for Improvement?

Failed Classification (Stacked Model)

Question about Metropolis Ark 1 choirs

If anyone here has used Metropolis Ark 1, you'll notice the excellent choir samples, both Male and Female. But the thing I wanted to point attention to is the Marcato and Staccato, where with each note is punctuated with a different syllable, It doesn't seem to be "Do re me la...", and I just wanted to know if there's a specific name for this style of singing? Okay, obviously when you choose the note it chooses a random syllable, but it sounds impressive and I want to learn a bit more about it, and if there's a way to notate such singing (or would you consider it chanting?)



Prediction



"Producers" (97.2%)



What went wrong?

Question about Metropolis Ark 1 choirs

=====

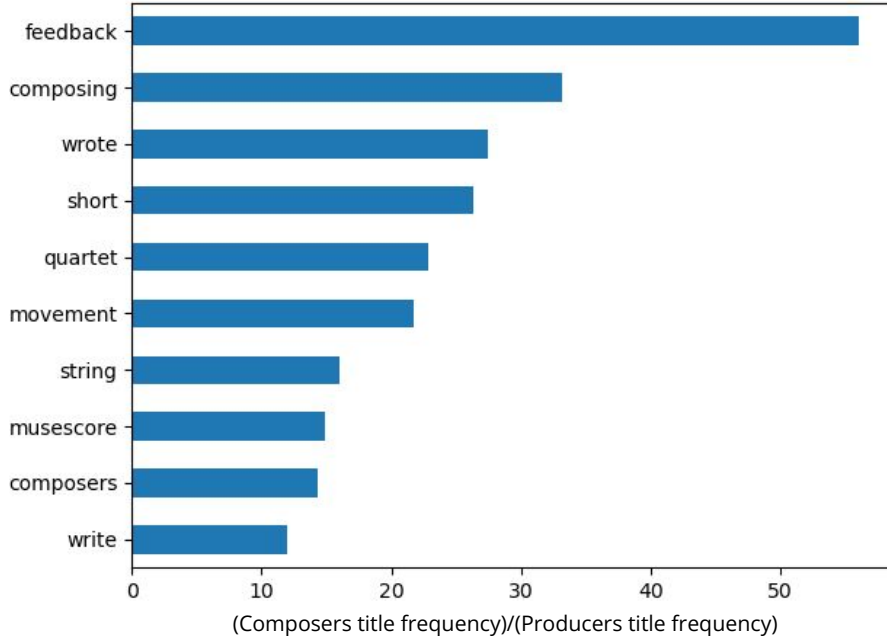
If anyone here has used Metropolis Ark 1, you'll notice the excellent choir **samples**, both Male and Female. But the thing I wanted to point attention to is the **Marcato** and **Staccato**, where with each note is punctuated with a different syllable, It doesn't seem to be "Do re me la...", and I just wanted to know if there's a specific name for this style of singing? Okay, obviously when you choose the note it chooses a random syllable, but it sounds impressive and I want to learn a bit more about it, and if there's a way to notate such singing (or would you consider it chanting?)

Rare in corpus (not selected as a feature)
but ***relatively more common among Composers***

**Common among
Producers**

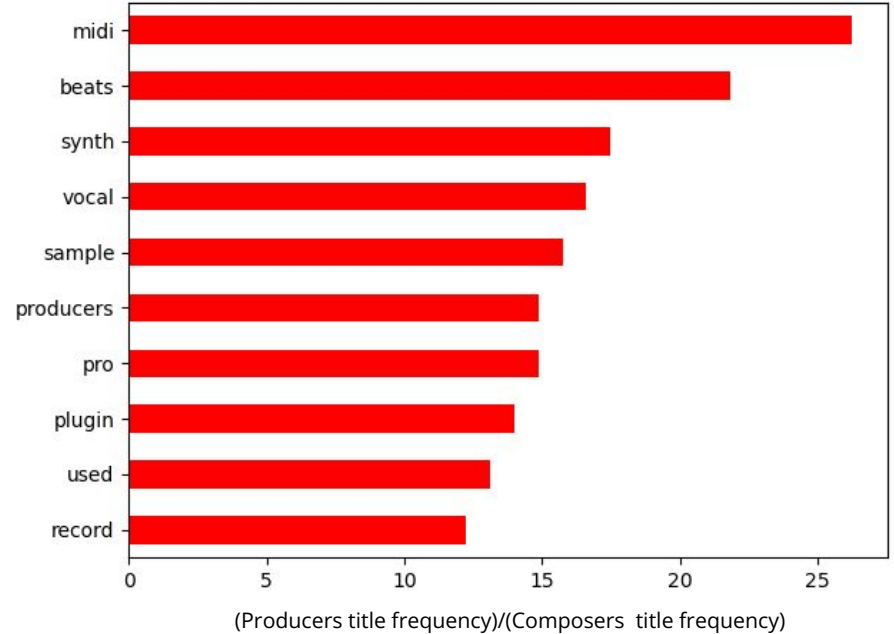
Words with Top Frequency Ratios

Frequency Ratios of Top Composers Words
(among words that appear in at least 1% of titles)



Infinite Ratio Words: fugue, sonata, love, theme, piece, score, wind, prelude, composer, violin, symphony, flute, composition, orchestra, concerto, minor, quintet

Frequency Ratios of Top Producers Words
(among words that appear in at least 1% of titles)



Infinite Ratio Words: artists, plugins, mix, effect, producer, beat, ableton, interface, FL, laptop, samples, production, got, logic, vocals, headphones, mixing, DAW, mic

Conclusions

Conclusions

- “Bag of words” NLP classification models can perform well with < 1000 data points in each class
- Potential music makers can be sorted into “Producers” or “Composers” with over 93% accuracy based on their Reddit posts
- Tuning different models then stacking them improves performance
- Proper preprocessing is essential
 - URL processing for Reddit posts
 - Frequency ratios rather than overall frequencies?