ISMIR 2022

Verse Versus Chorus: Structure-aware Feature Extraction for Lyrics-based Genre Recognition

Maximilian Mayerl, Stefan Brandl, Günther Specht,
Markus Schedl, Eva Zangerle

Core Idea and Findings

We investigate what impact lyrics structure has on the predictive power of genre recognition models. To that end, we train models on features extracted only from the verses and choruses of songs, and compare how those models perform overall, as well as for different music genres.

We found that, for rap and pop, verse features perform better, while for rock, chorus features do.

Data

We combine data from the LFM-2b dataset with lyrics and genre data from genius.com, leveraging crowd-sourced annotations provided by genius.com to segment song lyrics into structural parts. This gives us a dataset containing 295,416 songs with English lyrics, covering five musical genres (pop, rock, country, rap, and r&b).

Property	Value
Number of Songs	295,416
Number of Artists	39,357
Number of Tokens in Choruses	193,696,032
Number of Tokens in Verses	195,567,571
Average Number of Choruses per Song	3.46
Average Number of Verses per Song	2.37

Features and Models

We extract lyrics features separately for verses and choruses, covering a variety of features sets used in lyrics-based genre recognition:

- Bag-of-words features (separately for unigrams, bigrams, trigrams and (1,3)-grams)
- Rhyme features
- Readability features
- Lexical features
- Lexical diversity features
- Part-of-speech features
- Morphological features

Further, we used three different machine learning models, also chosen because of their common use in lyrics-based genre recognition:

- Random Forest
- Support Vector Machine
- Feed-forward Neural Network

Experiment Design

For our experiments, we separately trained and evaluated models for each combination of (1) model type, (2) feature set, and (3) song part (verse or chorus). This gives us a total of 60 concrete models that we trained and evaluated.

We then investigated the differences between models trained on verses against models trained on choruses. Furthermore, we computed those same performance differences for individual genres. For all our experiments, we performed significance tests to make sure the differences we observed were statistically significant.

To evaluate the performance of the models, we used to macro-averaged F1 score, since our dataset is imbalanced in the distribution of songs over genres.

Main Results

Feature Set	Random Forest Chorus	Random Forest Verse	Random Forest Diff.	Support Vector Machine Chorus	Support Vector Machine Verse	Support Vector Machine Diff.	Neural Network Chorus	Neural Network Verse	Neural Network Diff.
				F_1 macro					
unigrams	.377 (±.002)	.428 (±.002)	.0506†	.414 (±.002)	.502 (±.002)	.0880†	.419 (±.003)	.505 (±.003)	.0858†
bigrams	.364 (±.000)	.416 (±.003)	.0522†	.396 (±.002)	.485 (±.001)	.0888†	.396 (±.003)	.479 (±.005)	.0824†
trigrams	.328 (±.002)	.385 (±.001)	.0562†	.333 (±.002)	.422 (±.001)	.0894†	.342 (±.003)	.419 (±.003)	.0772†
(1,3)-grams	.379 (±.002)	.430 (±.002)	.0502†	.432 (±.002)	.511 (±.002)	.0792†	.424 (±.002)	.508 (±.003)	.0846†
rhyme	.254 (±.002)	.318 (±.002)	.0638†	.200 (±.001)	.280 (±.001)	.0796†	.230 (±.004)	.307 (±.009)	.0776†
readability	.263 (±.001)	.371 (±.002)	.1078†	.224 (±.001)	.355 (±.001)	**.1308**†	.264 (±.006)	.360 (±.002)	.0964†
lexical	.359 (±.002)	.400 (±.002)	.0412†	.291 (±.002)	.363 (±.001)	.0720†	.360 (±.004)	.403 (±.001)	.0430†
lexical diversity	.253 (±.002)	.351 (±.001)	.0980†	.195 (±.000)	.319 (±.000)	.1242†	.245 (±.002)	.333 (±.001)	.0878†
part-of-speech	.223 (±.001)	.297 (±.001)	.0738†	.180 (±.000)	.187 (±.001)	.0068†	.207 (±.005)	.289 (±.002)	.0816†
morphological	.201 (±.002)	.315 (±.001)	**.1146**†	.164 (±.002)	.169 (±.000)	.0056†	.181 (±.003)	.291 (±.005)	**.1100**†

The overall results show that there is a clear difference in performance between models trained on verses or choruses, respectively. In general, models trained and evaluated on features extracted from verses perform better than their chorus counterparts. In every case, this difference is statistically significant (shown by the † symbol behind the numbers).

Genre	readability RF	readability SVM	readability NN	lexical RF	lexical SVM	lexical NN
country	0.0054	0.0000	0.0011	-0.0448†	0.0001	-0.0076
pop	0.0298†	0.0367†	0.0360†	0.0110†	0.0271†	0.0193
r&b	0.0048	0.0000	-0.0021†	0.0271†	-0.0113†	-0.0583†
rap	0.4995†	0.6587†	0.4878†	0.3009†	0.3820†	0.2950†
rock	-0.0001	-0.0402†	-0.0412†	-0.0154†	-0.0383†	-0.0383

For investigating the per-genre differences, we looked at the feature sets for which we observed the biggest (readability) and smallest (lexical) overall differences. This table shows the performance difference between verse and chorus in terms of F1 score.

We observe that rap and pop songs show a consistently better performance for verse models. The opposite is true for rock songs, were models trained and evaluated on choruses consistently performed better.

Maximilian Mayerl (maximilian.mayerl@uibk.ac.at)