Learning from Lyrics: A musical and cultural perspective

Yutao Chen

Abstract

As a textual carrier for art, lyrics have always been an informative resource for text mining. The information contained in the lyrics not only relates to music themselves but also reflects the voice of society to a certain extent. In this research, I analyzed the information contained in lyrics from the perspective of music and society. For music, the lyrics are used to extract feature vectors for classifiers to identify the differences and similarities between different music genres. Features including td-idf vectors, sentiment polarity scores, word repeatability, and document embeddings are compared, and the temporal changes of these features are also explored. From the perspective of society, taking sexual liberation as an example, I explored the changing patterns of love-related and sex-related words in lyrics of Billboard Hot 100 songs, revealing the connection between lyrics and this social movement.

Keywords: music genre classification, text-mining, document-embedding, sex liberation.

1. Introduction

Lyrics, defined as the words of songs, are always an honest reflection of the emotions and attitudes of the singers and writers, and also an informative component of the music itself, providing hints on aspects like the genre of the music. Currently, many experiments have been conducted on identifying music styles or genres using methods like CNN and LSTM (Riyoichi Sawada Ueno & Furtado Silva, 2019). However, most of them constructed features based on audio elements or sound spectrums, and only a few considered the use of lyrics. Compared to other musical formats like audio or music notes (Pelchat & Gelowitz, 2019), lyrics are the most accessible and storable resources where classical text mining methods can be flexibly applied to. For different types of music, evidence (Li & Ogihara, 2004) shows that lyrics could become an alternative dimension to reveal their divergence other than acoustic features. For example, (heavy) metal music or hip-hop songs are found to have more negative, sex-related lyrics (Cheung & Feng, 2019) while the lyrics of folk or country music are generally more conventional. Other language characters, like the use of rhythms, are also found to have a different distribution among different types of music (Riyoichi Sawada Ueno & Furtado Silva, 2019), where folk music and hip-hop tends to employ a more frequent use of rhythms. Even though lyrics can provide some useful information about music genres, previous research also shows a limitation of using purely lyrics-based methods in music genre classification, considering metrics like accuracy and f1-score (Duggirala & Moh, 2020), so it is also important for us to understand the lyrics of which genre are more identifiable and the lyrics of which genres might be quite similar and result in confusion in classification so that we can find some directions to improve the classification results.

Moreover, when we process lyrics data in a time serial, they can also provide us with information about how different music genres update their textual content over years or decades. For instance, Herd (2008) discovered an increasing prevalence of drug-related words in rap songs during 1979-1997; Yasaman Madanikia and Kim Bartholomew, (2014) suggested that there was a significant increase in the proportion of songs with a theme focusing on lust but in the absence of love among the top 40 of Billboard hot 100 songs from 1971 to 2011. These inner changes of lyrics are also viewed as the signals of the variations in a social context and people's ideology in different periods by some scholars. They made direct comparisons between some experimental or socio-economic data and the statistics extracted from lyrics data and tried to build connections between the two parts. For example, Eastman & Pettijohn (2019) found that lyrics of R&B/Hip Hop music on Billboard are less likely to have leisure or fun words when the socioeconomic condition became unstable or declining; Qiu, Chan, Ito, & Sam (2020) claimed that unemployment rate plays a role in determining the degree of anger in pop song lyrics. Other good examples are the works of some famous artists like Bob Dylan (Czechowski, Miranda, & Sylvestre, 2016), U2 (Koller, 2018), which can be directly analyzed as an indicator of social issues and movements. Therefore, it is also possible for us to gain some insights into the connection between social movements and music through lyrics.

In this research, the analysis is divided into two parts: learning music from my lyrics' corpora and learning culture from my lyrics' corpora. For the first part, I intend to find out the difference or similarity between the lyrics of different music genres and how lyrics of different genres change over time. The research starts with a classification trial based on term frequency-inverse document frequencies (tf-idf) of words to see what proportion and kinds of the difference can be identified using the occurrence of words (Section 3). To understand the classification results, the similarity between tf-idf features of different genres is examined and the words that are the most important are checked and analyzed as well. Based on the classification result, I extracted other features like sentiment polarity scores and the repeatability of words to see how these two features can help to reveal the difference between genres (Section 4). As a comparison to word frequency features, the classification task is replicated with document embedding vector features in Section 5, and the difference between embeddings of different music genres is also examined. Additionally, these embedding vectors are applied to a projection analysis in three dimensions: happiness, gender, and romance, to further capture the difference of genres. After checking these cross-genre differences, we performed a time-serial analysis to see how these cross-genre differences evolve using different kinds of features mentioned before (i.e., tf-idf, word-embeddings, sentiment polarity scores, word repeatability, and three projected dimensions) in Section 6.

For the second part of this research (Section 7), a corpus containing billboard hot 100 songs from 1965 to 2015 is utilized to explore the impact of one big social movement, sex liberation. Especially, innovated by Yasaman Madanikia and Kim Bartholomew, (2014), the trends of love related and sex-related words are examined via word frequency calculation and dynamic topic modeling, and a comparison between the word trends, temporal marriage rates in the US is drawn to reveal more connections between music and our culture.

2. Data Description and Preprocessing

In this research, two different corpora are being used in different parts. For the first part of the research (Section 3-6), where we focused on analyzing the cross-sectional and temporal differences between music genres, we used a song-lyrics collection from Kaggle. This data set originally contains song items over 360000, ranging from 1968 to 2016. Each data sample is labelled with the genre of the song (genre), the year of the song (year), and the name of the artist (artist). To clean the data, I dropped those items with missing genres[1] or missing lyrics[2] and also cleaned out duplicated items. Still, some lyrics are detected to be in other languages than English via *langdetect* package. To simplify our analysis, these samples are also excluded from the dataset. After preprocessing, the remained dataset contains samples of 210642 songs with 10 music genres: Pop, Hip-Hop, Rock, Metal, Country, Jazz, Electronic, Folk, R&B and Indie.

In some raw lyrics, words annotating the section that the lyrics belong to are also included. These words include "intro", "verse", "pre", "post" "lift" "chorus", "bridge", "outro" and "instrumental"[3]. Though they might contain useful information about the positions of words in a song, we view them as special stop words and filtered them out from lyrics because only a part of the lyric samples has such annotations, which are even always incomplete.

For the second part of the research (Section 7), I turn to the billboard lyrics dataset to explore the connection between lyrics and sex liberation. Compared to other lyrics corpora, the lyrics of billboard top songs are assumed to be more culturally impactful and reflective by many studies (McAuslan & Waung, 2018), which makes it more suitable for cultural pattern identification. The billboard song collection (also from Kaggle) I used here contained Billboard hot 100 songs with their lyrics from 1965 to 2015. After filtering the samples, the cleaned data set contains 4827 songs altogether, and the number of samples included in each year ranges from 86 to 100, which is considered as sufficient to calculate the word trend. The song structure annotation words are also filtered from this Billboard lyric corpora. Besides, the time-serial data (1950-2019) of American married population[4] are also gathered to make comparison with the lyrics trend.

3. Music Genre Identification Based on Word Frequency

In this section, I would illustrate the music genre classification task based on the features extracted from word frequency. The goal for this process is not only to examine

[1] Songs labelled with genre "other" is also excluded because of genre ambiguity.
[2] The lyrics of a song is considered missing if the number of words is less than 10.
[3] Song structure source: https://en.wikipedia.org/wiki/Song_structure
[4] U.S. Marriage Data Source: https://www.census.gov/data/tables/time-series/demo/families/marital.html

the effectiveness (i.e., accuracy, f1-score) of using word frequency as features, but also to reveal the similarity or divergence of the word distribution among different music genres, and to understand how these similarities or differences influence the performance of the selected classifiers.

Feature Extraction: The features used here are derived from tf-idf vectors using scikit-learn. The calculation of tf-idf is defined as:

$$\frac{f_{w^*,d^*}}{\sum_{w \in d^*} f_{w,d^*}} \ (tf) * log \frac{N}{|\{d \in D: w^* \in d\}|} \ (idf)$$

Where

- w^*: the target word that we calculated the tf-idf for
- d^*: the document that w^* is in
- w: any word in d^*
- d: any document in the document collection D
- D: the collection of all the documents
- N: the number of all the documents

Compared to basic word frequencies, tf-idf can help to assign high scores to words that are the most distinctive for certain documents, decreasing the scores of words that commonly appear in all documents, which is thought to increase the difference between feature vectors of different categories. In my task, I set the vectorizer to include words with a document frequency between 0.001 and 0.8, to be the basis of the word frequency vectors (common stop words are filtered), and 5876 words are eventually kept in the dimension of the word frequency vectors. However, this number is still too large for a feature space and is likely to face the problem of overfitting, so I employed principal component analysis to reduce the dimension for the vectors. Different numbers of principal components are tested in the classification process, ranging from 10 to 200.

Classifiers: The classifiers tested for classification includes three types[5]. Their hyper-parameters are all tuned via grid-search and each model is validated by 10-folds cross-validation.

- Naïve Bayes: Using the kernel of Gaussian distribution, and fine-tuning the variance smoothing parameter between 10^{-20} to 10^{-10}.
- Logistic Regression: Using L2 regularization and fine-tuning the regularization strength parameter between 10^{-2} to 10^{2}
- Random Forest: Using Gini impurity as the criterion for spitting, and fun-tuning the max-depth parameter between 1 to 10.

For each of the above classifiers, the dataset is split by a 70/30 ratio. We also use cross-validation to decide the optimal class weights during the training process because of the imbalance among the numbers of samples in different classes (Country: 9897; Electronic: 4550; Folk: 1178; Hip-Hop : 15447; Indie: 2071; Jazz: 5029; Metal: 14636; Pop: 23691; R&B: 2287; Rock: 68663).

Metrics: For imbalanced dataset, using accuracy as the metrics is not suitable because accuracy is easily affected by the dominant class in the samples, so here I used average

[5] Supported Vector Machine (SVM) is also tested but proved to be too slow in classification with large-sample size, so we finally excluded it.

f1 score as our metric to evaluate the models and records the accuracy, the average f1 score and the average ROC-AUC score of each best model to make comparisons.

Results and Analysis:

Classifier	PC number	F1-score	Accuracy	ROC-AUC
NaiveBayes	10	0.361352	0.368031	0.656144
LogisticRegression	10	0.458969	0.548368	0.707829
RandomForest	10	0.471478	0.497476	0.728308
NaiveBayes	20	0.361189	0.361559	0.658721
LogisticRegression	20	0.472012	0.554824	0.727046
RandomForest	20	0.478206	0.51118	0.741242
NaiveBayes	50	0.326441	0.319323	0.643842
LogisticRegression	50	0.486616	0.556992	0.745664
RandomForest	50	0.481883	0.532733	0.749318
NaiveBayes	100	0.280805	0.281613	0.644051
LogisticRegression	100	0.501241	0.562926	0.759474
RandomForest	100	0.476259	0.544412	0.752827
NaiveBayes	200	0.253215	0.260282	0.649122
LogisticRegression	200	0.517843	0.574083	0.772277
RandomForest	200	0.463251	0.549017	0.751059

Table 1. Word Frequency Based Classification Result

From Table 1, I found that these three classifiers have different reactions to the number of principal components. For the Naïve Bayes classifier, the increasing number of principle components make it overfit the data and results in a continuous decrease in f1-score and accuracy. Random Forest and Logistic Regression classifiers always perform better than Naïve Bayes, but Logistic Regression performs better as the number of principal components increases. Generally, we are more interested in the best models, so I visualize the confusion matrix of the best estimators of the logistic regression and the random forest classifiers.

From Figure 1-2, we can find that three genres are more recognizable to the classifiers than others: Hip-Hop, Rock and Metal. Especially for the genre Rock, it is not only the type of genre that has the highest accuracy (.86) but also the type of genre that makes the classifiers most confused, i.e., get the largest number of false-positive cases. Other genres like Indie, R&B, and folk are largely misclassified as Rock. On the one hand, this might be related to the imbalance of the dataset. Even if we applied different class weights to samples of different genres, the bias is hard to be offset completely because of the original disproportion on information volumes from different genres. On the other hand, it is also worthwhile to check the similarity between tf-idf vectors of different music genres to see if the features of rock music are indeed close to those of other music genres. I calculated the tf-idf vector for each genre by averaging the sample vectors in each genre and visualized the heatmap of the cosine distances between the genre vectors in Figure 3. We can verify that the word frequency vectors of different genres, except Metal and Hip-hop, have great similarities (small cosine distance values).

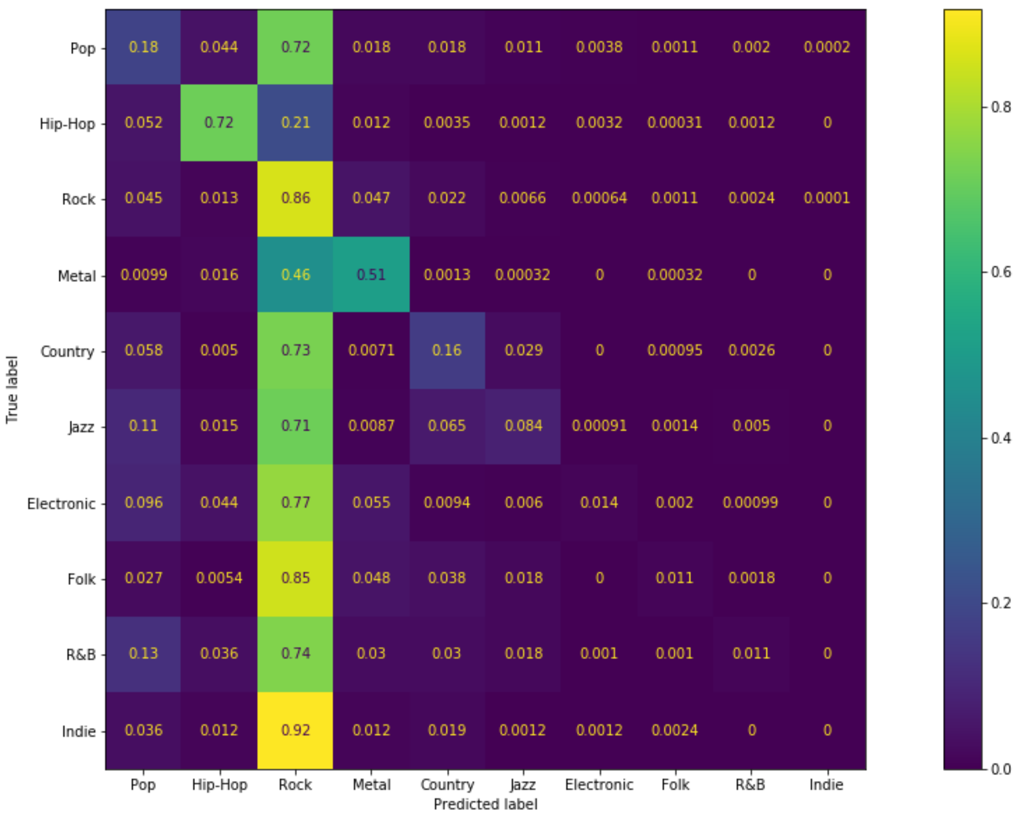


Figure 1. Word Frequency Based Classification Confusion Matrix-Logistic Regression (N = 200)

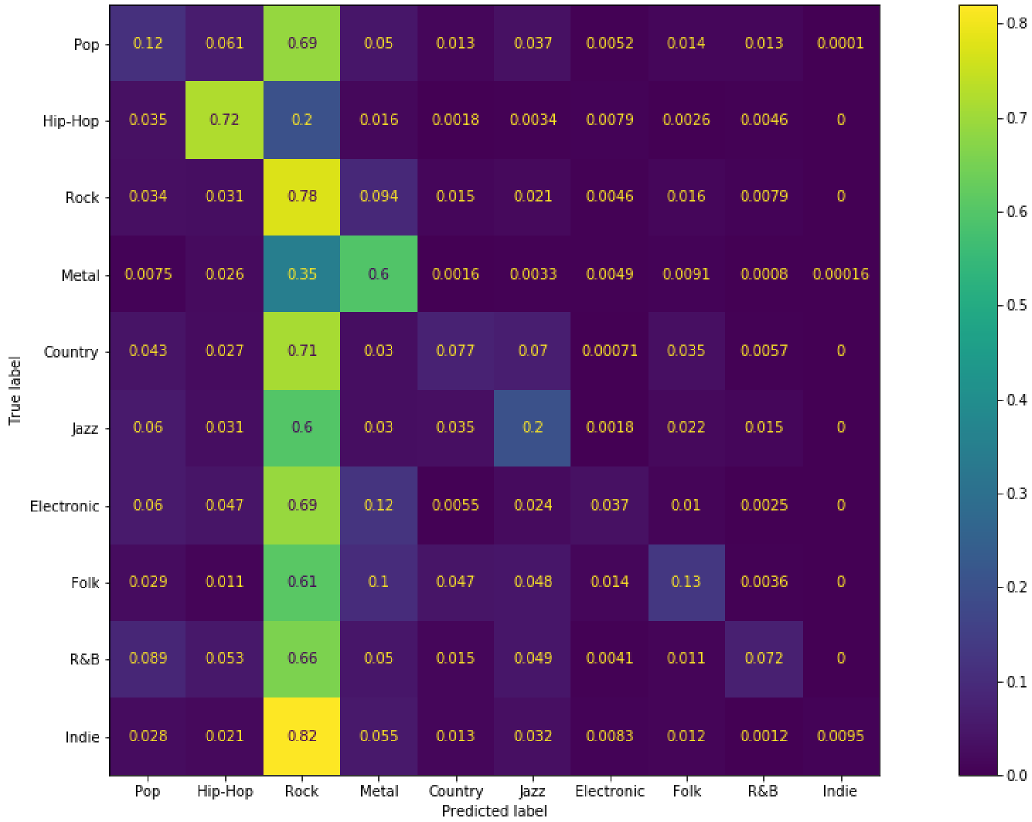


Figure 2. Word Frequency Based Classification Confusion Matrix-Random Forest (N = 50)

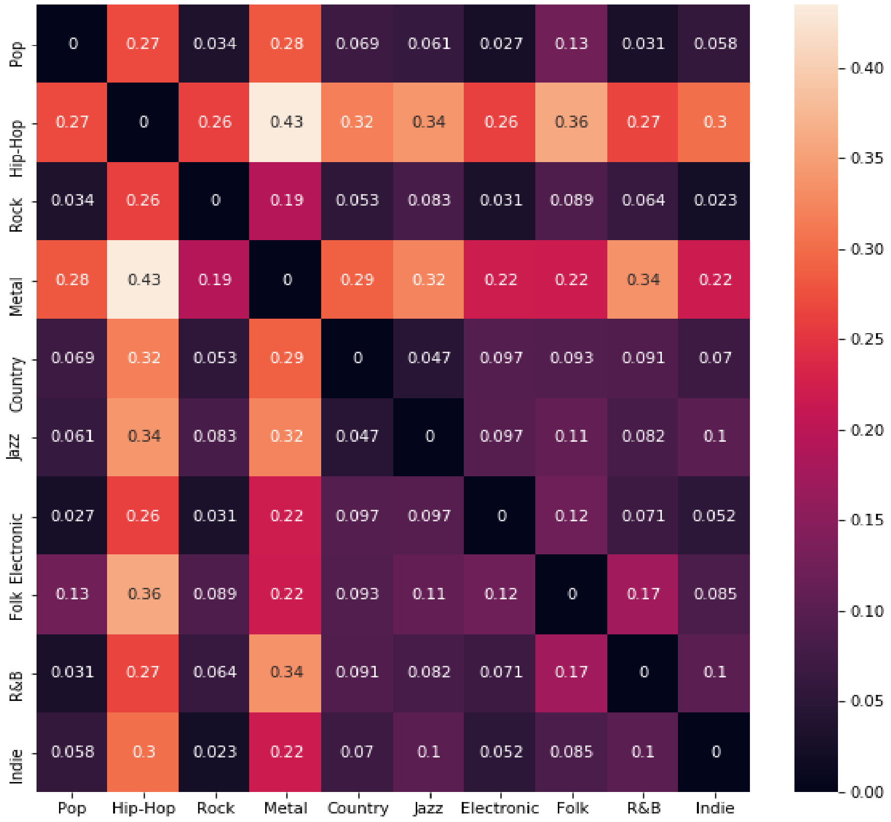


Figure 3. Word Frequency Based Features' Cosine Distance Heatmap

To exclude the impact of data imbalance, I re-test the cosine distances between a subset of samples without songs in Rock, Hip-hop, and Metal, the results also suggest that the tf-idf vectors of the remained genres are quite close.

To further understand how these similarities or differences affect the classifiers, I extracted the most important 10 principle components via scikit-learn permutation importance function and traced back to their tf-idf compositions to find out the exact words that are most responsive to the classification process. Using the best logistic regression classifier as an example, the 10 most important principle components with their Importance values are presented in Table 2.

Principal Component Index	Importance Value
3	0.0563
1	0.0456
0	0.0397
14	0.0161
8	0.0093
4	0.0091
6	0.0077
9	0.0049
12	0.0048
5	0.0038

Table 2. Top 10 principle components for best logistic regression classifier

To calculate the word importance, I average the absolute values of word coefficients in these top 10 components weighted by their importance values and extracted the top 10 words with the highest importance scores. They are "love", "don", "know", "baby", "oh", "ll", "like", "got", "want", and "nigga". The average tf-idf of these words provided firm evidence of the uniqueness of the word distribution in hip-hop and metal. In most cases, when other genres have a higher average tf-idf on certain words, Metal always has a lower average tf-idf value for these words (e.g. "love", "oh" see Figure 4). As for hip-hop, some top words are found to have exclusively high tf-idf values on it while having low values on other genres (e.g. "nigga", "like" see Figure 4). The most important words for the other two classifiers are also examined and proved to have little difference from the result of the logistic regression classifier.

From the analysis in this section, we can see that hip-hop and metal music have very different word frequency distribution from other genres, where hip-pop lyrics tend to have higher specificity on certain words like "nigga" and metal lyrics tend to exclude some common words appeared in other genres such as "love" and "oh". Apart from metal and hip-hop, the word frequency distribution in other genres tends to be similar, which can hardly be distinguished via features of td-idf. Additionally, it is worthy of attention that classification can not only be used to predict music genres but can also provide guidance to deeply understand the difference or similarities between the lyrics of different music genres.

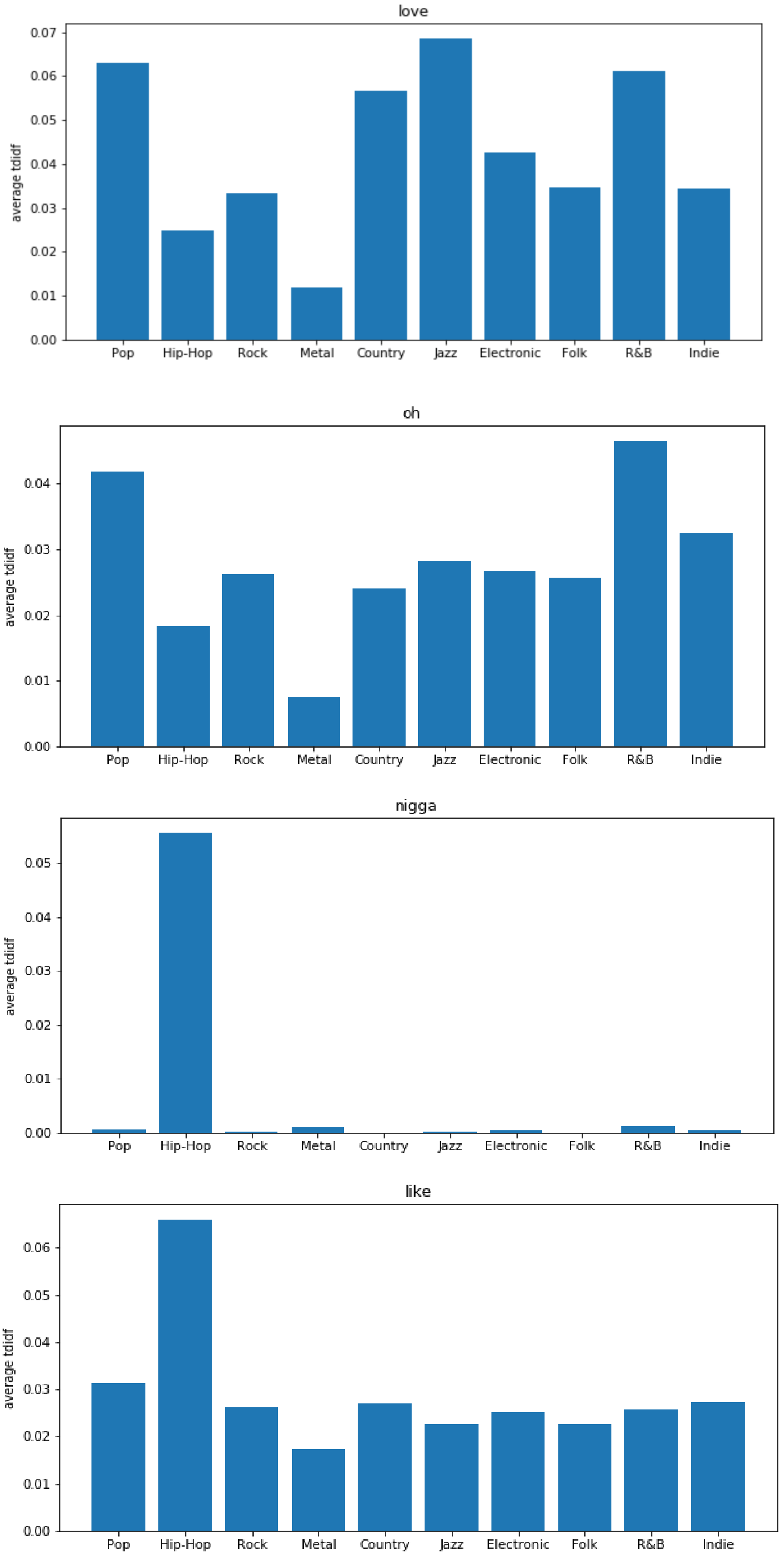


Figure 4. Average tf-idf for “love”, “oh”, “nigga” and “like” in different genres

4. Music Genre Identification with Sentiment and Word Repeatability

Based on the classification result in Section 3, we notice that using solely word frequency-based features could not give a good identification for music genres other than metal and hip-hop. Therefore, in this section, I tried to add other features to the feature spaces to see if they can help to improve the classification results. The new features I chose are sentiment polarity scores and word repeatability of each lyric documents. There are two reasons for choosing these two features: firstly, both sentiment polarity scores and word repeatability are attainable via simple analysis of the lyrics text; secondly, the information contained in these two features do not have strong correlations among different genres ($p \geq 0.2$) and are not largely overlapped by word frequency features.

Sentiment polarity scores are calculated with the Vader-Lexicon on NLTK, where the scores range from -1 to +1. I first check the average sentiment polarity distribution of each genre (see Figure 5), finding that hip-hop and metal music generally have negative sentiment values and other genres have positive sentiment values. Word repeatability is measured by $\frac{N(total\ words)}{N(unique\ words)} - 1$ for each lyrics document. I checked the average word repeatability of each genre (see Figure 6) and found out that electronic songs have an exceptionally higher word repeatability than other genres.

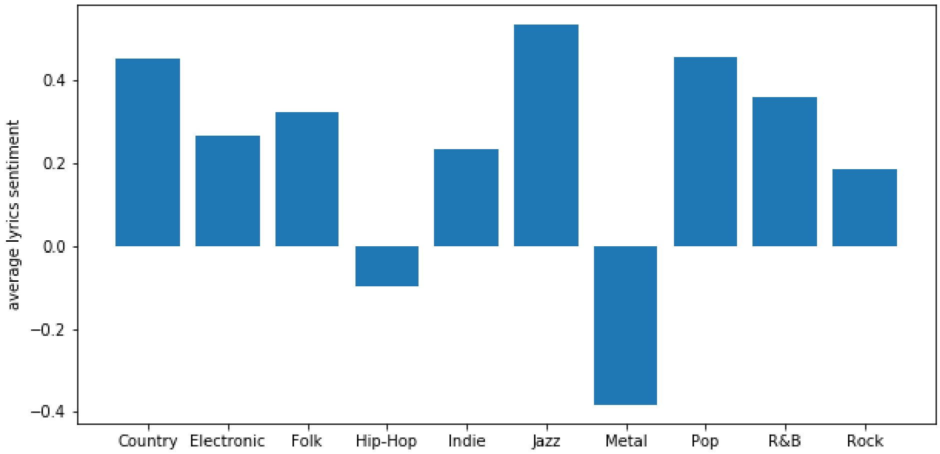


Figure 5. Average Sentiment Polarity for Different Genres

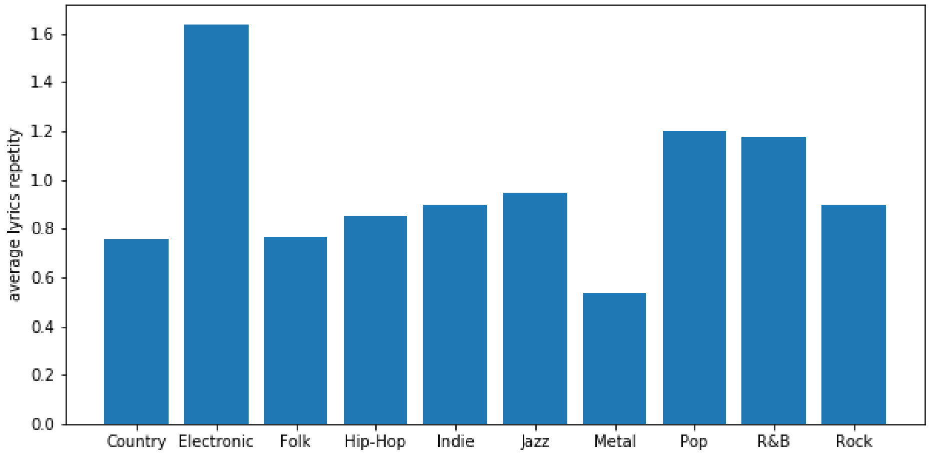


Figure 6. Average Word Repeatability for Different Genres

Using these two new attributes and the word frequency vectors extracted from Section 3, I replicated the classification work on 7 groups where each group has a different feature composition, the result of which is reflected in table 3. The hyper-parameters and class-weights are tuned the same as in Section 3.

Features	Classifier	F1-score	Accuracy	ROC-AUC
20 tf-idf PCs[6]	NaiveBayes	0.361729	0.362952	0.658969
	LogisticRegression	0.472110	0.554745	0.727411
	RandomForest	0.479135	0.511987	0.741335
20 tf-idf PCs sentiment	NaiveBayes	0.367145	0.366196	0.664919
	LogisticRegression	0.474507	0.556074	0.729458
	RandomForest	0.483879	0.517557	0.745405
20 tf-idf PCs repeatability	NaiveBayes	0.364168	0.360309	0.667379
	LogisticRegression	0.480435	0.558131	0.733844
	RandomForest	0.490368	0.517003	0.746504
20 tf-idf PCs sentiment repeatability	NaiveBayes	0.368348	0.362477	0.673194
	LogisticRegression	0.481030	0.559698	0.735844
	RandomForest	0.494489	0.520517	0.750792
sentiment	NaiveBayes	0.300277	0.469751	0.582951
	LogisticRegression	0.300277	0.469751	0.583318
	RandomForest	0.378074	0.429699	0.603503
repeatability	NaiveBayes	0.308899	0.464244	0.557555
	LogisticRegression	0.330075	0.418575	0.572891
	RandomForest	0.352927	0.398414	0.594483
sentiment repeatability	NaiveBayes	0.314658	0.465178	0.606326
	LogisticRegression	0.341590	0.407656	0.61256
	RandomForest	0.395977	0.420300	0.64091

Table 3. Mixed Feature Classification Results

From Table 3, we can see that simply using sentiment polarity scores or repeatability values does not yield a better result than using tf-idf features. However, when we combine these features, an increase was found on f1-scores and ROC-AUC scores of all three classifiers. At the same time, the accuracy of logistic regression classifiers, as well as random forest classifiers, are strictly increased when sentiment scores and repeatability are included in the feature space. Moreover, the improvement made by adding word repeatability features is greater than the improvement made by adding sentiment scores for these two better-performed classifiers (see Figure 7).

To further understand the improvement made by adding these two features, it is useful to examine the confusion matrix of the classifiers. Using the confusion matrix of the logistic regression model as an example (see Figure 8), I found that when adding the feature of sentiment polarity scores, the accuracy for Metal (0.47 to 0.48) and Country

[6] Other numbers of PCs (10,50,100,200) are also tested, suggesting the effect of using sentiment polarity and word repeatability as additional features is robust

(0.018 to 0.021) songs received an increase at a little cost of the accuracy of Jazz, Folk and R&B (a decrease at level 0.001); when we add the repeatability feature, Pop (0.11 to 0.13), Metal (0.47 to 0.49), Country (0.018 to 0.025), Electronic (0.0099 to 0.016) genres all received a boost on accuracy. When both features are added into the feature space, there is no wonder that most of the genres, except hip-hop and rock, all get an enhancement on accuracy. Besides, the permutation feature importance for the combined features is also calculated (Table 4), suggesting that both sentiment polarities scores and word repeatability values are ranked within the top 5 most important features.

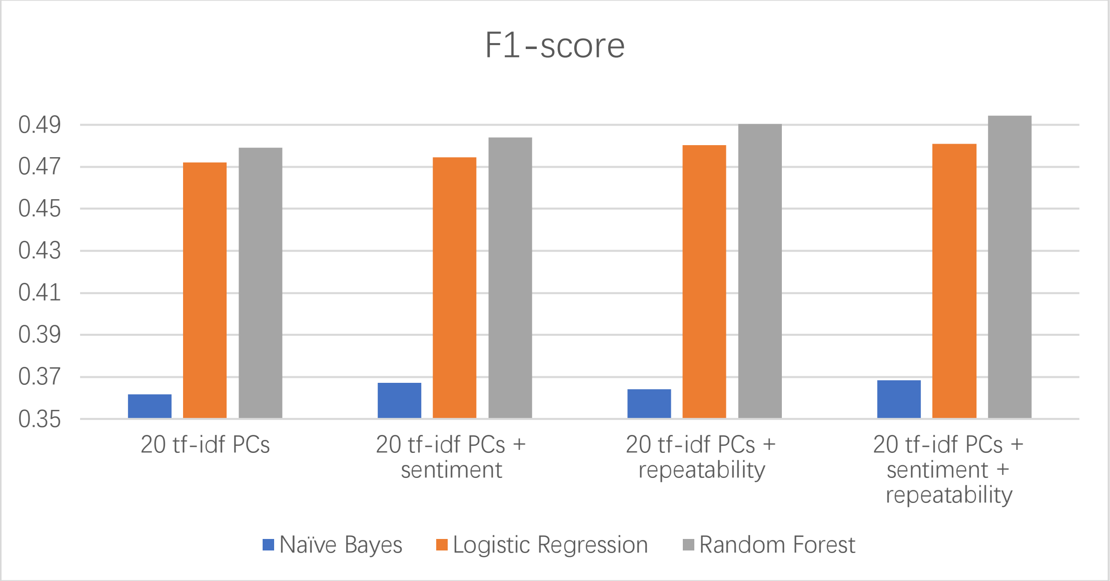
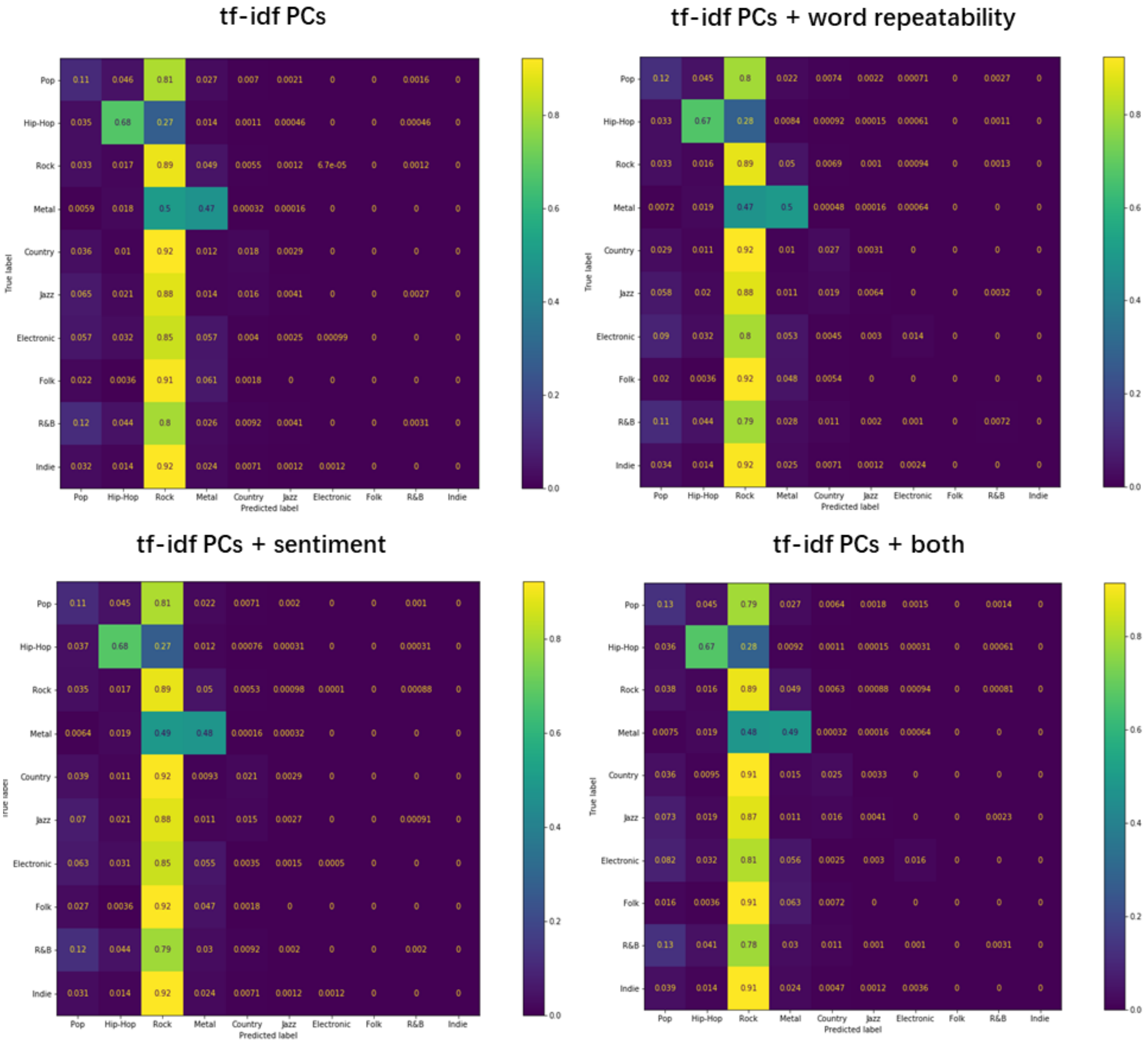


Figure 7. f1-score of classification using combined features

Feature Index	Importance Value
1	0.0491
3	0.0339
0	0.0211
20 (word repeatability)	0.0165
21 (sentiment polarity scores)	0.0115

Table 4. Top 5 principle components for logistic regression classifier

From these statistics, it is clear that sentiment polarity scores and word repeatability can give us additional information on the difference in music genres. Especially, this information is useful in distinguishing the genres that are under-estimated by-word frequency features. On the other hand, the importance of word frequency as features is also verified since purely using sentiment scores or word repeatability would yield a worse performance than using the tf-idf based feature vectors. Besides, the inspiration of this section also entails that when using lyrics as the data source to identify music genres, more attention needed to be paid on feature engineering. Other than barely using word frequency-based features, more information like sentiment and repeatability of the words are likely to help to make the differences between music genres more identifiable. Admittedly, the power of sentiment or word repeatability only improves the result in a limited degree, but as more innovative features being added, I think it is possible to make the feature space more informative regarding the difference between music genres.



tf-idf PCs

tf-idf PCs + word repeatability

tf-idf PCs + sentiment

tf-idf PCs + both

Figure 7. Classification Confusion Matrix Based on Combined Features-Logistic Regression

5. Music Genre Identification with Document Embeddings

Document embedding is another approach to build feature vectors for lyric samples. Unlike word frequencies from lyrics, the vectors produced by the document embedding method can provide us with some information on the relationships between the words in the same lyric files rather than merely the words themselves. In this section, we tested the effectiveness of using document embeddings as feature vectors to see that if document embeddings can reveal more differences about different genres. Besides, three-word embedding dimensions (happiness, gender, and romance) are built to analyze the projected difference of the document embeddings of different genres.

Feature Extraction: To obtain the document embeddings for each lyrics sample, the doc2vec model in genism is used for the lyric corpora. The model is tested for different vector size including 10, 20,50 and 100, and each model is trained for 10 epochs. The performance of the genism doc2vec model could be improved by increasing the epoch times of training. Large numbers of epoch times (i.e., 20,50,100) are also tested, but inconsistent with the 10-folds cross-validation used in section 3&4, I only discussed the case of 10 epochs.

Classifiers: The same classifiers used in Section 3 with the same hyper-parameters and class-weights fine-tuned via 10 folds cross-validation.

Metrics: Use the average f1-score as the metric for the training process and also record accuracy and ROC-AUC scores.

Results and Analysis:

Classifier	Vector Size	F1-score	Accuracy	ROC-AUC
NaiveBayes	10	0.348004	0.417372	0.614382
LogisticRegression	10	0.300277	0.469751	0.554056
RandomForest	10	0.404236	0.504581	0.641791
NaiveBayes	20	0.331414	0.384014	0.597425
LogisticRegression	20	0.300277	0.469751	0.581007
RandomForest	20	0.374744	0.491637	0.620523
NaiveBayes	50	0.274886	0.302644	0.580114
LogisticRegression	50	0.300277	0.469751	0.617683
RandomForest	50	0.362768	0.502935	0.595819
NaiveBayes	100	0.211050	0.232352	0.574385
LogisticRegression	100	0.300277	0.469751	0.612784
RandomForest	100	0.362824	0.500720	0.585646

Table 5. Document Embedding Based Classification Result

From Table 5, we see that the classification results of using document embedding features are less satisfactory than using word frequency-based features. Moreover, as the vector size increases, we even notice that the f1-score and ROC-AUC dropped sequentially. To understand this unexpected result (from other articles, document embeddings are more likely to perform better), the cosine difference between the document embeddings of different genres are visualized (see Figure 8). Compared to the cosine distance calculated by word frequency features (see Figure 3), the uniqueness of hip-hop lyrics is amplified, making the Metal lyrics less distinguishable, and the uniqueness of electronic songs is also mildly increased. Combining these changes, the overall difference between different genres is decreased, resulting in a less preferable classification result. Though the information revealed by document embedding does not improve the classification result, it is still useful for seeing the difference of music genres from another aspect. For hip-hop lyrics, the frequently occurred unique words like “nigga” give it an uncommon word combination than other features, and for electronic lyrics, though they have very similar word compositions as other genres, the combination of these words should be less alike.

What’s more, we can also gain other interesting insights about music genres via document embeddings like the representative words for each music genre. These representative words are calculated by finding the words whose embedding vectors are of the least cosine distance to the vector of music genres (see Table 9). For music genres like Pop, Jazz, and R&B, the representative words are obviously romance related. On the contrary, Rock and Metal have more emotionally negative representative words like “heartache” and “torment”. Particularly, Indie and Rock lyrics have the same representative word, which is consistent with the fact they are similar in both word frequency and document embeddings.

Music Genres	**Closet Word**	**Cosine Distance**
Pop	love	0.908828
Hip-hop	gansta	0.970134
Rock	heartache	0.951404
Metal	torment	0.945753
Country	lonely	0.973864
Jazz	sweetheart	0.964428
Electronic	destro	0.943048
Folk	weary	0.939134
R&B	darling	0.914521
Indie	heartache	0.946639

Table 9. Representative Words for Music Genres

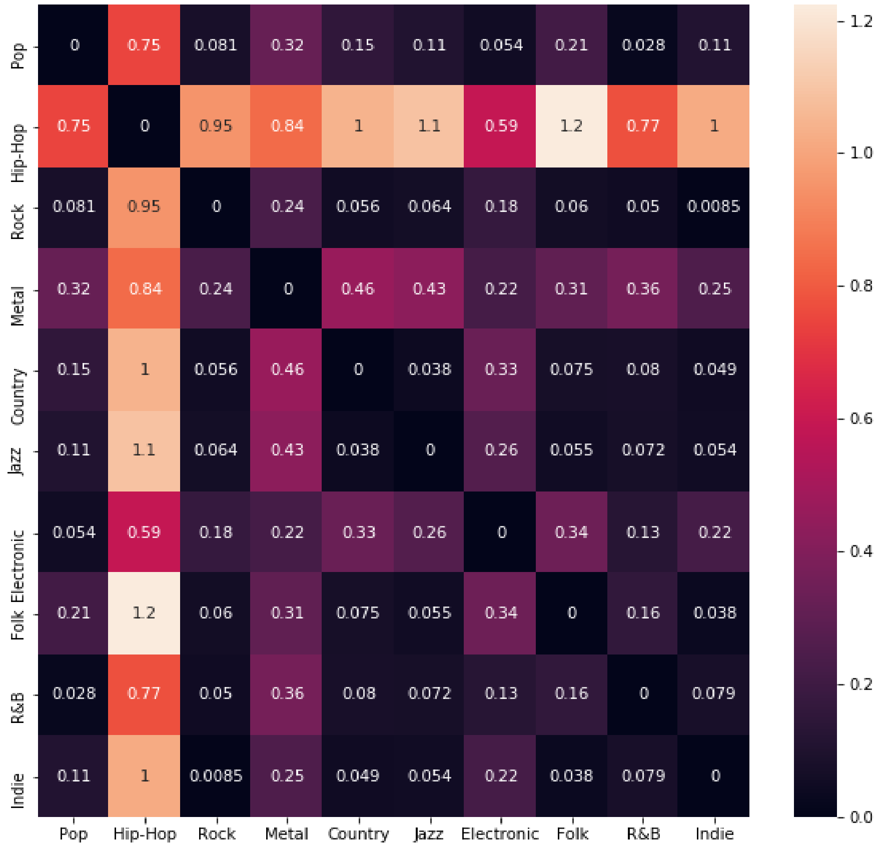


Figure 8. Document Embedding Based Features’ Cosine Distance Heatmap

Projection Analysis: From the examination of the representative words, we notice that different music genres have different emotions or content tendencies based on their document embeddings, so they are likely to reveal more differences if we project them on these sub-dimensions. Three dimensions are examined in this research, including *happiness, gender,* and *romance.* Each dimension is calculated by subtracting from the average positive polarity vector with the negative polarity vector, the composition of the polarity vectors is defined in Table 10:

Dimension	Positive Words	Negative Words
Happiness	happy, fun, joy, enjoy, glad, cheerful, smile, delight	sad, sorrow, cry, tear, unhappy, mournful, miserable, gloomy
Gender	man, boy, male, him, he	woman, girl, female, her, she
Romance	love, baby, kiss, heartbreak, romance	time, life, people, fate, world

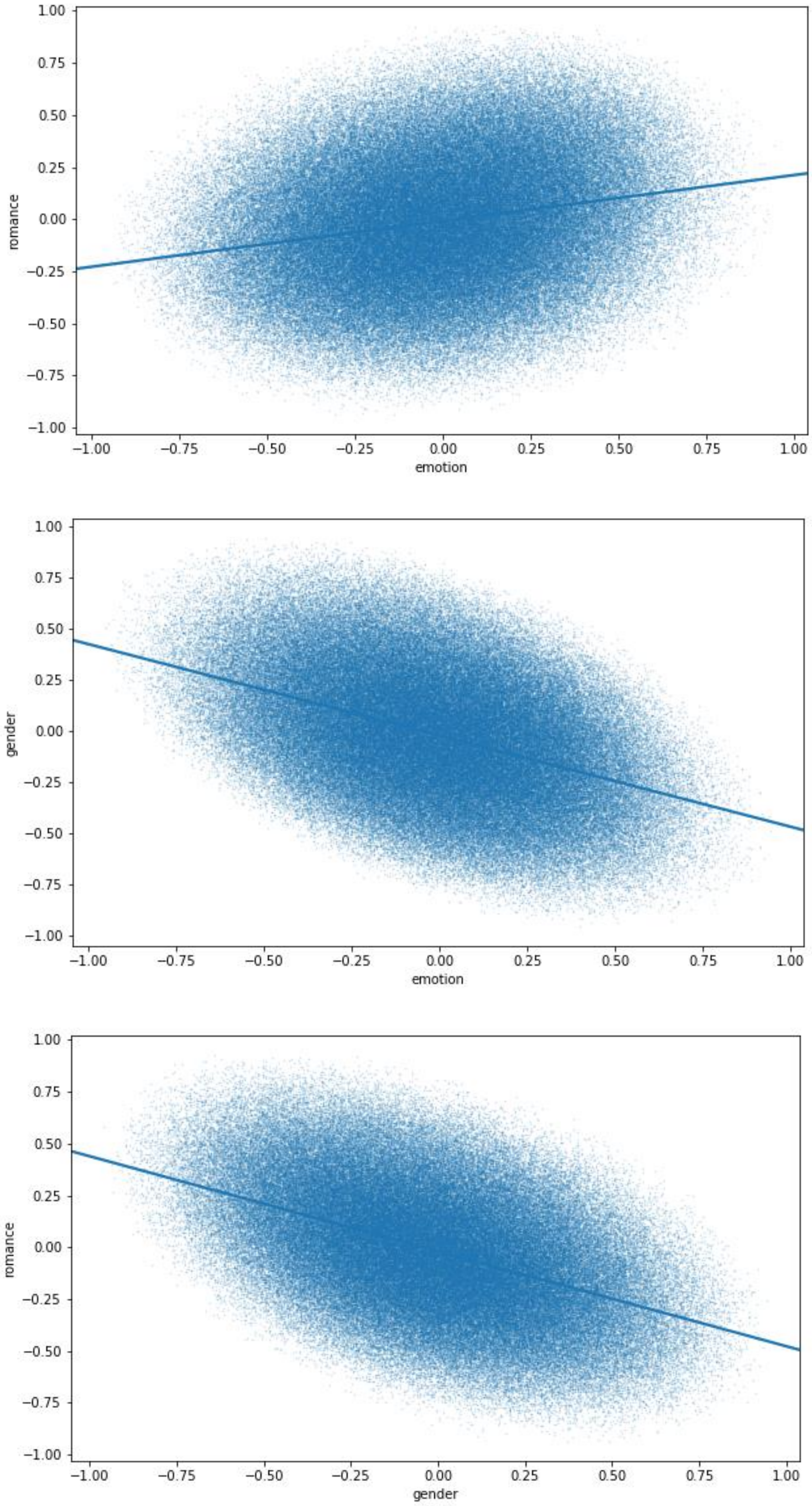
Table 10. Word Compositions for Dimensions

Considering the happiness dimension, hip-hop songs have the highest score on this dimension, indicating that lyrics in hip-hop songs are more likely to be happy-related. Noticing that in sentiment analysis, Hip-hop lyrics are found to have negative values on average, this contradiction might be related to the fact that hip-hop songs prefer to use dirty words as their lyrics. Typically, dirty words are associated with negative meanings in sentiment analysis but for hip-hop songs, the use of dirty words is more like a culture of ego-boosting. So, using the document embeddings is more accurate to reflect the emotional bias for hip-hop songs. Metal lyrics are found to have a low happiness score, which is coherent with previous studies. In the dimension of gender, Folk music tends to be centered with male-related descriptions followed by metal and country. These music genres are always regarded as more archaic compared with genres like electronic or indie, which tends to use more female-related words in their lyrics. This transitional difference might be related to social movements like sex liberation, which we would mention in Section 7. As for romance, Jazz becomes the top genre with romantic content while genres like hip-hop and metal, which are usually viewed as more rebellious, have the least romantic content (see Table 11).

	HAPPINESS	GENDER	ROMANCE
POP	0.071002	-0.177638	0.169118
HIP-HOP	0.183291	0.055826	-0.374592
ROCK	-0.178751	0.103498	0.102201
METAL	-0.459461	0.194164	-0.378693
COUNTRY	-0.101381	0.153019	0.224477
JAZZ	0.003789	0.023328	0.331921
ELECTRONIC	0.056582	-0.229384	0.055147
FOLK	-0.221769	0.220826	0.179044
R&B	-0.009551	-0.013505	0.151713
INDIE	-0.250462	0.166861	0.098442

Table 11. Projection Result Table

Besides exploring the difference in dimensions for different genres, the relationships between the three dimensions were also found to have significant linear correlations (see Figure 9). Lyrics that are more romantic-related tend to have a higher happiness (emotion) score ($r = .22, p = .001$). Male-related lyrics are associated with fewer happiness expressions ($r = -.44, p = .001$), and female-related lyrics are more likely to have romantic contents ($r = -.45, p = .001$).



Figu*re* 9. Correlation Plots between Projected Dimensions

6. Music Genres' Trend Analysis

Lyrics can not only reflect the differences between music genres but can also give us hints about how music genres change over time if the lyrics are tagged with time-related information. In my corpora, most lyric documents are tagged with year labels, but not all music genres have samples in each recorded year. To balance this disproportion, I picked up the years where all music genres have samples in and the remained sub-dataset contains 198621 lyrics documents from 2001 to 2016 (the samples in 2003 are dropped because of a vacancy in some genres). The objective features targeted in this time-serial analysis include all the features I have extracted above: word frequency vectors, sentiment polarity, word repeatability, document embeddings, and their projected dimensions.

Word Frequency: I used the same methods as in Section 3 to produce tf-idf vectors for each lyric document, and reduced their dimensions within 20 principle components. The documents are grouped by their genres and years, so the index for each subset would be a combination of (*genre, year*). To produce the vector representing each subset, I averaged the vectors in the subset by their elements. Figure 10 shows the cosine distance heatmaps of each genre over the selected years. Several interesting patterns are found among different music genres. For Pop and Rock lyrics, there is a turning point between 2007-2008 where the word frequency distribution before and after this point has a bigger change, resulting in that the word distribution before and after this point forms two "separate" groups. For Folk, Jazz and Indie, several special years (Folk: 2001, 2004; Jazz: 2002, 2005; Indie: 2001, 2002, 2004) have a more different word distribution, while the word distributions in other years are quite similar. For country and electronic lyrics, they have the traits of both types of patterns. The average yearly divergence of each genre is also measured suggesting that Indie is the most unstable genre (d = .58) and hip-hop is the most stable one (d = .02).

Besides checking the yearly change of each genre, it is also valuable to check the variation of the relationships between different music genres, i.e., the distance between different genres, to see if the distances between certain genres become closer or farther over time. These statistics could be viewed as a signal for potential fusions or separations of music genres. The pairwise cosine distances between the genres each year are measured and a Pearson correlation test is performed between the year index and the distances. Based on the significance of 99%, I found five pairs of genres have tendencies of being closer or farther during the years (see Table 12). We can see that (pop, electronic), (pop, rock), and (rock, indie) lyrics are becoming more similar over the years, while (hip-hop, indie) and (metal, electronic) grow to be more distant.

Genre Pairs	Coefficient	P Value
Pop, Electronic	-0.8032	0.0003
Pop, Rock	-0.7943	0.0004
Hip-hop, Indie	0.7797	0.0006
Rock, Indie	-0.7084	0.0031

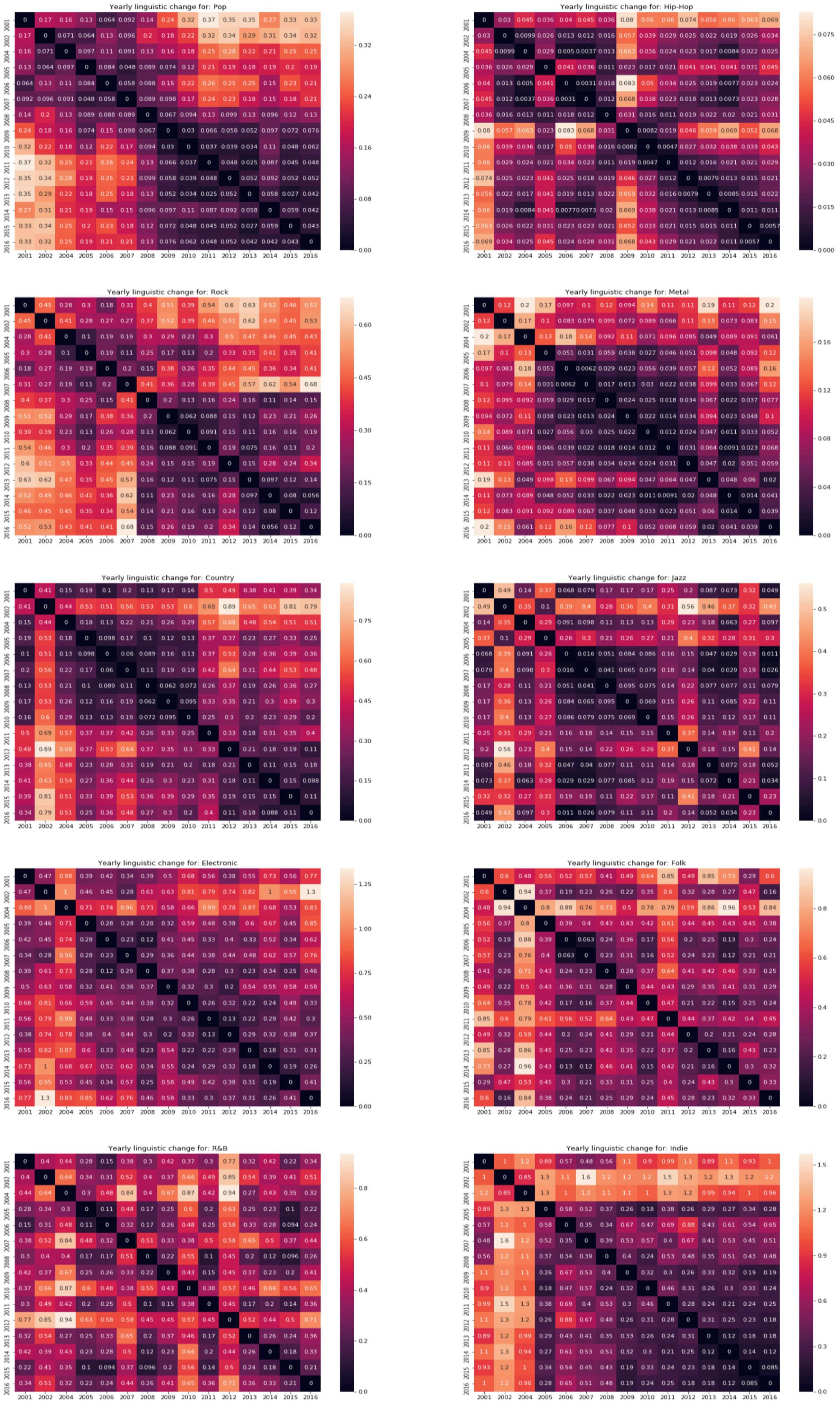
Metal, Electronic	0.6679	0.0065

Table 12. Genre Relationship Change based on Word Frequency

Sentiment and Word Repeatability: For the trend of sentiment polarity scores, I found that the sentiment polarity of all lyrics does not have a significant tendency of increasing or decreasing ($p > .50$). But for individual genres, Pop lyrics have a tendency of becoming more negative ($r = -.81, p = .001$). For word repeatability, the overall lyrics suggest that we are more likely to use repetitive words in lyrics as time goes by ($r = .68, p = .005$). Moreover, when examining the individual lyrics, it is found that Country ($r = .89$), Hip-hop ($r = .88$), Pop ($r = .77$) and even Metal ($r = .75$) have the tendency to increase the use of repetitive words on a significance level of 0.001. The increase of repetitive word use in lyrics suggests that more importance might be attached to other elements like melody or tempos, and the story-telling function of the lyrics is gradually declining.

Document Embeddings and Projections: To analyze the trend for document embeddings (10 size doc2vec model is used) of lyrics in different genres, the same analytical methods are used as in the analysis of word frequency vectors. Firstly, I checked the yearly change of the document embeddings for each music genre (see Figure 11), finding that the variations of document embeddings are much irregular than that of word frequency. One relatively obvious pattern we found is that for Pop, Hip-hop, Rock, and Metal, the embeddings become less divergent after 2009. When checking the yearly cosine distances of document embeddings from an average level, I found that all the genres are found to have a higher value compared to the distances of word frequency-based vectors. At the same time, Indie is still the genre that has the largest variation of all (d = .98) while Metal becomes one of the least variable embeddings over these years (d = .82). From this analysis, we can also know better about why using document embeddings are less effective than word frequency-based features in the classification of music genre. Classifiers are likely to be confused by the big difference for features in different years of the same genres, which rendering a poor performance.

The yearly changes of similarities between embeddings of different genres are also examined, and as expected, the significances of the relationships are less significant. But still, several pairs of genres have relatively more significant ($p < .05$) tendencies becoming closer or far away (see Table 12). The most surprising one is the pair of Folk and Indie, the negativity of the coefficient suggesting that they are approaching each other in document embeddings over these years. However, generally, we might think these are two quite opposite genres in music, where one is conventional (folk) and the other is relatively young (indie). On the one hand, we might expect to see some fusions of these two genres in the future. On the other hand, more data or evidence is needed to verify this tendency. Another pair needed to be emphasized is the pair of Hip-hop and indie, which appears as becoming distant for both word frequency vectors and document embeddings, suggesting the increasing separation of these two genres.



Yearly linguistic change for: Pop

Yearly linguistic change for: Hip-Hop

Yearly linguistic change for: Rock

Yearly linguistic change for: Metal

Yearly linguistic change for: Country

Yearly linguistic change for: Jazz

Yearly linguistic change for: Electronic

Yearly linguistic change for: Folk

Yearly linguistic change for: R&B

Yearly linguistic change for: Indie

Figure 10. Yearly Word Frequency Change for Music Genres

The yearly changes of similarities between embeddings of different genres are also examined, and as expected, the significances of the relationships are less significant. But still, there are several pairs of genres have relatively more significant (p < .05) tendencies becoming closer or far away (see Table 12). The most surprising one is the pair of Folk and Indie, the negativity of the coefficient suggesting that they are approaching each other in document embeddings over these years. However, generally we might think these are two quite opposite genres in music, where one is conventional (folk) and the other is relatively young (indie). On the one hand, we might expect to see some fusions of these two genres in the future. On the other hand, more data or evidence is needed to verify this tendency. Another pair needed to be emphasized is the pair of Hip-hop and indie, which appears as becoming distant for both word frequency vectors and document embeddings, suggesting the increasing separation of these two genres.

Genre Pairs	**Coefficient**	**P Value**
Hip-hop, Country	0.5790	0.0237
Hip-hop, Indie	0.5335	0.0405
Folk, Indie	-0.5264	0.0438
Jazz, Indie	0.5140	0.0500

Table 12. Genre Relationship Change based on document embeddings

For the three projected dimensions, several trends are found among the music genres. In the dimension of happiness, Rock lyrics are found to become happier ($r = .58, p = .02$), while R&B ($r = -.58, p = .02$) and Pop ($r = -.56, p = .03$) lyrics are found to become less happy. In the dimension of gender, Indie lyrics tend to become more female related ($r = -.53, p = .04$), despite it is originally a male-centered genre. On the contrary, R&B ($r = .49, p = .06$) and Pop ($r = .48, p = .07$) which are originally more female-related, becomes more manly in lyrics. Such anti-traditional trends also happened to the dimension of romance. Jazz ($r = -.50, p = .06$) and R&B ($r = -.54, p = .04$), which are overall have more romantic elements in lyrics than others, are found to have a decreasing trend on keeping these elements as before.

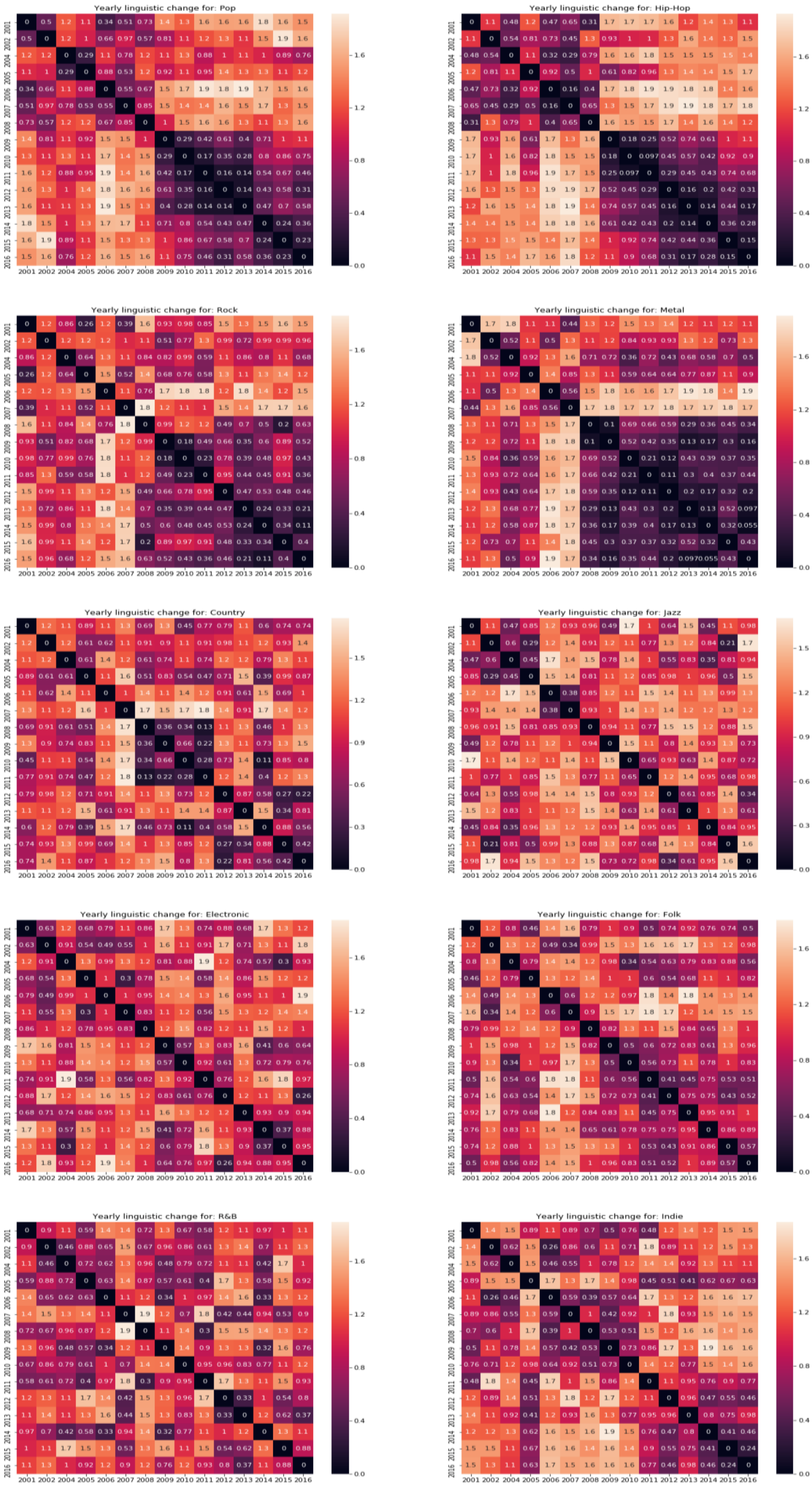


Figure 10. Yearly Document Embedding Change for Music Genres

7. Social Movements in Lyrics: An Example of Sex Liberation

In this section, the billboard hot 100 dataset is used to explore the connection between lyrics and the social movement of sex liberation. Sex liberation was understood as an "increased acceptance of sex outside of traditional heterosexual, monogamous relationships (primarily marriage)" (Escoffier, Jong, & McDarrah, 2003), which lasted from the 1960s to 1980s but pose a profound effect up to now. In this process, people have become less conservative to talk about sex-related topics and become more open to unconventional sexual activities. Meanwhile, the traditional marriage system has been weakened because for many people marriage is no longer a premise for sex, but even an obstacle to their enjoyment of free sex. To explore how this trend is reflected in lyrics, two groups of words are selected, where one is relevant to traditional love and the other is clinging to sex liberation.

- Traditional love word: love, darling, sweetheart, honey, marry
- Sex liberation word: sex, body, sexy, pussy, dick

Firstly, the average tf-idf frequencies for these two groups in billboard lyrics from 1965 to 2015 are examined (see Figure 11). Despite the fluctuations, we can still see a tendency for traditional love words to decline overall during these years ($r = -.71, p = .001$), while sex-related words become more frequent in billboard lyrics ($r = .83, p = .001$). For individual words, most sex-related words start to appear in the billboard lyrics since the 1980s, while the tf-idf frequency of love related-words started to drop at the same time or earlier.

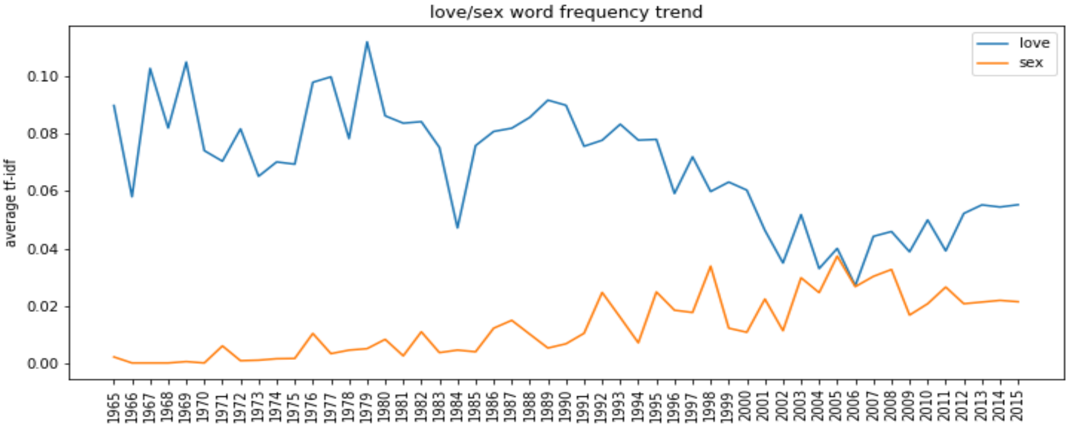


Figure 11. Love and Sex Related Word Trend in Billboard Hot 100 Lyrics

Besides, a further comparison was made between the US marriage rates and these two groups of words. The variations of the US marriage rate can be viewed as direct evidence of sex liberation, and using it as a benchmark could reveal the connection between the lyrics and our society more precisely. The result (see Figure 12 and Table 13) verified that lyrics can reflect this kind of trend in a certain degree, where the decreasing marriage rates for both male and female indicates a decline of people's willingness for traditional love formats.

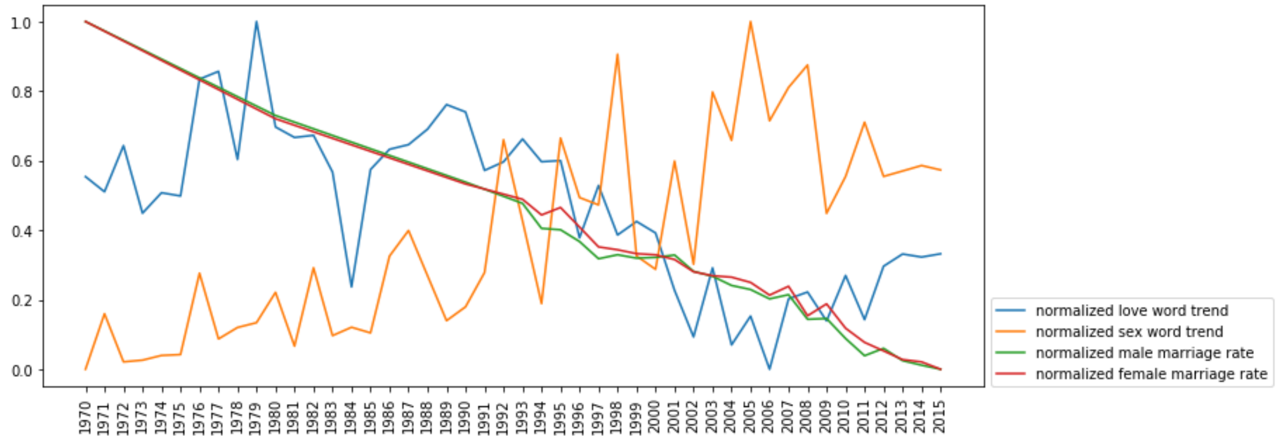


Figure 12. Love and Sex Word Trend with Marriage Rate in US

Correlation	Male Marriage Rate	Female Marriage Rate
Traditional love words	$(r = .64, p = .0004)$	$(r = .64, p = .0004)$
Sex liberation words	$(r = -.61, p = .001)$	$(r = -.59, p = .001)$

Table 13. Pearson Correlation between Marriage Rates and Words

Another evidence is from the topics I extracted through dynamic topic modeling using genism models. The word "love" and the word "sexy" are found in to have a sequential appearance in the topic loadings. "love" drops from 0.06 loading level (the 1960s) to 0.04 (the 2010s), while "sexy" firstly appears in the loadings of 1970s as the 20th loading with a level of 0.005, but increase to the 7th loading in 2010s with a level of 0.011. Based on these results, though not perfectly matching, we can still view music lyrics as a strong signal reflecting social movement. After all, lyrics are produced by humans and reflect their ideologies, which are easily impacted by social movements. In particular, the songs on billboard hot 100 are always closer to the preference of the mainstream, so it is expectable that their lyrics are representative of the social culture to a certain degree.

8. Conclusions

In this research, I examined the function of lyrics in telling the difference or similarities between music genres and reflecting the culture of our society using the song corpora from Kaggle. For the identification of music genres, I found that using word-frequency based features could identify genres of relatively unique word distributions like Hip-hop and Metal, while other genres, which are found to have great similarity in word frequency are hard to be correctly identified via this method. Moreover, adding features like sentiment polarity scores and word repeatability can help to increase the accuracy of identifying different music genres, especially those that are hard to be identified with word frequency-based features. This experience gives us suggestions that different types of features from lyrics could reveal non-overlapping information, which helps to classify music genres. Document embedding features are also checked for identifying music genres but are found to be less effective compared with word frequency-based features. Still, it reveals that genres with similar word frequency-based features could have more differences in document embeddings (Electronic), suggesting a different way of the word combination. Besides, the embeddings are projected to three dimensions happiness, gender, and romance, to find more specific differences between music genres.

In the following part, the time serial changes in lyrics are analyzed using all types of extracted features. Word frequency-based features are found to be more stable over the examined years than document embedding features, which can help to explain why it performs better in classification. Several temporal trends are found and discussed for different music genres. Also, music genre pairs like Pop and Electronic are found to have a decreasing distance in a word frequency distribution, indicating a possibility for the fusion of different music genres. While genre pairs like Indie and Hip-hop are believed to become more separate because of the increasing distance. Additionally, it is found that lyrics as a whole tend to contain more repetitive words than before.

Finally, the connection between social movements (i.e., sex liberation) and lyrics are checked by tracing the evolution of frequency and topic-loadings about love-related and sex-related words, and further verified through a comparison with the US marriage data series.

9. Future Work

For the identification of music genres, I am trying to apply these features to more powerful classifiers like LSTM and Bret. Currently, the best accuracy suggested by LSTM is 61% using coded ids as inputting features, which is comparable to the results of previous studies. Since these super models usually do not use the same input features as plain classifiers, I do not make a comparison of them here, but wish to see how they can make a difference in the future. Besides, other features including some phonological attributes of the words (like rhythms) could be tried on to see if they can improve the classification results.

For examining the connection between lyrics and social movements, other specific topics like feminism could be tested. As other research suggests, the effect of economic conditions can also be checked in further analysis. Plus, the connection here is revealed as a correlation, but if we can build a causal inference between the statistics from lyrics and socioeconomic index, I think it would be more valuable for understanding the link between art and society.

Reference

[1]. Cheung, J., & Feng, D. (. (2019). Attitudinal meaning and social struggle in heavy metal song lyrics: a corpus-based analysis. *Social Semiotics*. doi:10.1080/10350330.2019.1601337

[2]. CzechowskiKonrad, MirandaDave, & SylvestreJohn. (2016). Like a rolling stone: A mixed-methods approach to linguistic analysis of Bob Dylan's lyrics. Psychology of Aesthetics, Creativity, and the Arts, 10, 99–113. doi:10.1037/aca0000045

[3]. DuggiralaSharan, & MohTeng-Sheng. (2020). A Novel Approach to Music Genre Classification using Natural Language Processing and Spark. 2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM), Ubiquitous Information Management and Communication (IMCOM), 2020 14th International Conference on, 1–8. doi:10.1109/IMCOM48794.2020.9001675

[4]. EastmanT.Jason, & PettijohnF. I. I.Terry. (2019). Good times and endless love: Billboard R&B/Hip Hop songs of the year across social and economic conditions. Psychology of Popular Media Culture, 8, 243–250. doi:10.1037/ppm0000176

[5]. EscoffierJ., JongE., & McDarrahW.F. (2003). Sexual Revolution. Running Press.

[6]. HerdDenise. (2008). Changes in drug use prevalence in rap music songs, 1979-1997. Addiction Research & Theory, 16, 167–180. doi:10.1080/16066350801993987

[7]. KollerChristina. (2018). Rewriting the past, shaping the present - U2's rock lyrics as social and political activism.

[8]. KrauseE.Amanda, & NorthC.Adrian. (2019). Pop music lyrics are related to the proportion of female recording artists: Analysis of the United Kingdom weekly top five song lyrics, 1960–2015. Psychology of Popular Media Culture, 8, 233–242. doi:10.1037/ppm0000174

[9]. LiT., & OgiharaM. (2004). Music Artist Style Identification by Semi-supervised Learning from both Lyrics and Content. (364). ACM Press.

[10]. McAuslanPam, & WaungMarie. (2018). Billboard Hot 100 songs: Self-promoting over the past 20 years. Psychology of Popular Media Culture, 7, 171–184. doi:10.1037/ppm0000118

[11]. PelchatNikki, & GelowitzM.Craig. (2019). Neural Network Music Genre Classification. 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE), Electrical and Computer Engineering (CCECE), 2019 IEEE Canadian Conference of, 1–4. doi:10.1109/CCECE.2019.8861555

[12]. QiuLin, ChanHian MaySarah, ItoKenichi, & SamYan TingJoyce. (2020). Unemployment rate predicts anger in popular music lyrics: Evidence from top 10 songs in the United States and Germany from 1980 to 2017. Psychology of Popular Media. doi:10.1037/ppm0000282

[13]. Riyoichi Sawada UenoLuiggyCaio, & Furtado SilvaDiego. (2019). On Combining Diverse Models for Lyrics-Based Music Genre Classification. 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), Intelligent Systems (BRACIS), 2019 8th Brazilian Conference on, BRACIS, 138–143. doi:10.1109/BRACIS.2019.00033

[14]. SprankleL.Eric, EndM.Christian, & BretzN.Miranda. (2012). Sexually Degrading Music Videos and Lyrics: Their Effect on Males' Aggression and Endorsement of Rape Myths and Sexual Stereotypes. doi:10.1027/1864-1105/a000060

[15]. Yasaman Madanikia and Kim Bartholomew. (2014). Themes of Lust and Love in Popular Music Lyrics From 1971 to 2011. doi:10.1177/2158244014547179