CSYS 300 Final Project: Exploratory Analysis of Rap/Hip-Hop Genre

Justin Clark
Department of Complex Systems
The University of Vermont
(Dated: 12/09/2020)

The music industry as a whole and more recently, the rap/hip-hop genre, has consistently displayed a close tie with contemporary pop culture. Trends and relationships between this genre of music and song popularity have influence on many social and economic factors within the United States. Spotify and Genius API was used to extract audio/lyrical features of 2000+ popular songs in the years 1993 to 2019 as we look examine popularity related to rap/hip-hop music through the use of both exploratory analysis and machine learning techniques. Three algorithms were applied towards popularity classification and results indicated a general importance of track release and lyrical repetition in predicting song popularity. General trends of audience attention towards the recency of music was discovered and specific factors of this genre were examined in relationship to features within the data set.

I. Introduction

The rap and hip-hop music industry, once associated with low-income urban communities, has recently become an integral part of contemporary pop culture within the United States. A music industry that began as a framework for critiquing America's failure to uphold constitutional ideals of equality and disseminate the long history of Black suffering within the United States has transformed into a billion dollar industry [7]. The historically recent emergence of this genre in combination with its expression of cultural messages and sentiment define a clear impact on this genre of music and economic and social factors within the United States. The importance of music within contemporary culture can not be overstated.

Music has been shown to have numerous social psychological impacts and meanings for people at different periods of development[2]. Adolescents commonly use music as inspiration to develop social identity where popular musicians becomes role-models and influence adolescents through observation and imitation of specific behavior [3, 4].

A fundamental comprehension of indicators of song popularity has implications to numerous business's/fields that operate through the use of popular music. These fields include both physical and digital music marketplaces such as streaming services, radio stations, record labels, and business related to music revenue.

Machine learning, or statistical learning, refers to a vast set of tools for understanding hidden relationships within data. The application of machine learning methods within a supervised task attempts to build statistical models for predicting or estimating a labeled output variable based on a a set of input or feature variables. The use of these supervised learning techniques finds applications within a variety of fields including economics, health-care, government policy, and education. We look to apply machine learning techniques within the music sector, specifically to the popularity of both historical and contemporary rap/hip-hop tracks.

Within this paper, we look to perform an exploratory analysis of both historical and contemporary rap/hip-hop music to determine trends in lexical/sentiment features conveyed through both lyrical and audio features of a track. Moreover, we look to determine features of interest in predicting song popularity and compare feature trends over numerous decades of popular rap/hip-hop music through the use of three unique statistical learning algorithms.

II. Methods

A. Data Acquisition and Text Pre-processing

The initial step in obtaining appropriate data for this rhetoric analysis of contemporary rap/hip-hop music was determining popular songs released for each year within the research space. Billboard Top 100 charts were found for the years 1993 to 2019 and equivalent Spotify playlists were found through novel search. Spotipy [1], a lightweight Python library for the **Spotify Web API** [2], was used to obtain full access to all of the music data provided by the Spotify platform based on the given Spotify playlist. This included artist and track data such as album title, track features, song duration and other audio features. Next, **Genius API**[3] was used to extract song lyrics from Genius.com for each song within the data set using artist name and track title. Basic data preprocessing steps were applied to lyric data to obtain proper accounts of word frequency and lyric sentiment for each track. These basic steps included conversion of text to lower-case, removal of punctuation,apostrophes and single characters, removal of stop words, conversion of numbers to text, and stemming of text into root words. Some preprocessing methods were repeated to ensure clean and tidy data for later use. Specific track identifiers such as [Chorus] and [Intro] were also removed. The above described song

[1] https://github.com/plamere/spotipy/blob/master/LICENSE.md
[2] https://developer.spotify.com/documentation/web-api/
[3] https://docs.genius.com/

resource contains over 2000 songs produced by over 500 unique artists.

B. Text Analysis

The language generated by artists within this genre display many distinct features that are not native to normal text or lyrics of songs categorized within a different genre of music. This is partly due to the use of 'new' vocabulary that is sufficiently coded/complex such that meaning varies according to context.These features of rap language may adversely affect analysis involving the structural fragmentation of texts such as Sentiment Analysis and Lemmitization. Some common examples of these features include neologisms (slang), expletives, abbreviations and colloquial words.

C. Sentiment Analysis

Sentiment Analysis methods were applied to track data in attempt to measure the attitude, sentiments, and emotions of an artist/artists song based on the computational treatment of subjectivity within the track lyrics. More specifically, VADER (Valence Aware Dictionary for Sentiment Reasoning) from the NLTK package was implemented due to its application on unlabeled text data that is sensitivity to both the polarity and intensity of emotion relayed by the text.

D. Feature Pre-processing

Audio features related to each track included track duration, key, and year of release. More abstract features such as energy, acousticness, and danceability were also generated by Spotify API. A complete list of features can be found in Table II. Lexical features related to word frequency and sentiment were also calculated such as the average sentiment of the track, number of words and proportion of unique words. These features were all numeric variables and therefore did not require any transformation through the use of One-Hot Encoding. Feature scaling techniques were applied to the data set to create uniformity in feature ranges. Variable feature ranges will make Gradient Descent algorithms take longer to converge [1] thus the normalization of feature ranges will improve the learning rate of desired algorithms. The data was standardized using sklearn preprocessing library to ensure features of uniform range [-1,1]. This type of transformation was applied to instances within both the training and prediction sets. The target variable, popularity, was initially numeric with values in the range (1,99). This variable was transformed for binary classification such that the upper 1/3 of popular songs were labeled as popular. This split occur at popularity values close to 50. Additionally, an 80/20 split was used for training/testing instances in model implementation.

E. ML Techniques

Three unique machine learning techniques were applied to popularity classification.

The first was Logistic Regression with l_1 regularization. l_1 regularization serves to minimize over-fitting of the model and increase generalizability to unseen data through the use of a shrinking. Shrinking regularizes feature coefficient estimates by shrinking coefficients towards zero. In short, l_1 regularization eliminates non-essential features by minimizing the magnitude of the features coefficient. Moreover, logistic regression selects the parameter θ that maximizes the likelihood function:

$$L(\theta) = \prod_{i=1}^{n} P(Y = y^{(i)} \mid X = x^{(i)}) \quad (1)$$

$$L(\theta) = \prod_{i=1}^{n} h(\theta^T x^{(i)})^{y^{(i)}} (1 - h(\theta^T x^{(i)}))^{(1-y^{(i)})} \quad (2)$$

$$h(\theta^T x^{(i)}) = \frac{1}{1 + e^{-\theta^T x^{(i)}}} \quad (3)$$

The second statistical learning method implemented was support vector machines (SVM) based on their minimization of the upper bound of generalization error. Support Vector Machines look to maximize the margin around a separating hyper plane through that use of support vectors. This constrained optimization bypasses restrictions to linearly-separable data through the use of numerous kernel functions

The third and final statistical learning method implement was Decision Trees/Random Forest Classification. The Decision Tree, hereby abbreviated DT, is a tree-based hierarchical representation of the data distinguishing values and rules based upon notably simple conditional statements. The algorithm used for constructing decision trees worked top down, by choosing a variable at each step that best splits the training instances based on the largest expected information gain [5]. The Gini Impurity was used to quantify the best variable for each split such that

$$Gini = 1 - \sum_{i=1}^{k} p_k^2 \quad (4)$$

where p_k denotes the proportion of instances belonging to class k (*POCS Note:* Tsallis Entropy with deformation coefficient $q = 2$). Based on their definition, DTs inherently suffer from high variance and we look to reduce this variance through the use of numerous, decorrelated decision trees. Therefore, a Random Forest (RF) is an ensemble method averaging the results of many individual Decision Trees. Decision Trees within a RF are built using bootstrapped training samples such that each time a split in a tree is considered, a random sample of m predictors is chosen as possible candidates from the full set of p predictors. We define the size of $m = \sqrt{p}$ such that the number of predictors considered at each split within the DTs is approximately equal to the square root of the total number of predictors.

F. Performance Metrics

The above discussed classification methods were evaluated through the use of numerous performance metrics based on the distribution of the target variable, popularity. More specifically, given the unequal distribution of song popularity, it was determined that accuracy would

not be strong indicator of model performance. In order to capture this idea, classification performance measures such as precision, recall, F1 score, and AUC were considered. Precision is calculated as the proportion of testing instances classified as popular that are truly popular. On the other hand, recall, also known as sensitivity, is calculated as the proportion of testing instances that actually are popular out of all testing instances classified as popular by the discussed algorithm. Next,we looked to combine the analysis of precision and recall into one unbiased estimator, the F1-score. The F1-score is defined as the harmonic mean of precision and recall such that

$$F_1 = \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{5}$$

This value is constrained between [0,1] where a value of 1 indicates a perfect balance between precision and recall. Finally, the last evaluated performance metric was AUC, or Area Under the ROC curve. This value represents the probability the classifier ranks a rank positive example higher than a random negative example. AUC was seen as a desirable performance metric is it is both scale-invariant and classification-threshold invariant.

G. Hyper-parameter Tuning

Multiple of the discussed machine learning models require the optimization of hyper-parameters to maximize above discussed performance metrics. One parameter, C, was tuned for the Logistic Regression model while two parameters, C and γ, were optimized with respect to the utilized kernel for our SVM model. Hyper-parameter C looks to determine the trade-off between classification of target instances with the general simplicity of hyper plane separation. A higher value of this regularization parameter implies less violations within the classifier margin. Hyper-parameter γ looks to weight individual training instances. Optimization of this parameter leverages the bias-variance trade-off within model prediction. Parameters were optimized by cross-validated grid-search over the parameter grid.

III. Results/Discussion

A. Exploratory Analysis

The initial portion of research looked to examine general trends of word frequency and sentiment over the discussed time period (1993-2019) within this genre of music. More specifically, relationships between audio/lyrical features were examined based on groupings by artists and year. Word-frequency of the entire data set was examined with a large-majority of the most frequent words being stop words or expletives.

A general skew in popularity was recognized with greater values in inherently correlated with the year of track release, with a correlation coefficient of 0.27. Moreover, a small, negative correlation coefficient quantified the relationship between average track sentiment and year of song release indicating a decrease in song sentiment over the examined time frame. The results of Fig. 1 visualizes a consistent decrease in both the proportion of lyrics within a track and the proportion of unique words within a track from the years 1993 to 2019. The results of this figure indicates a decrease emphasis on lyrical value amongst popular songs in contemporary rap/hip-hop music when compared to historical rap/hip-hop music. These results signify a potential shift in the characteristics of the audience and preferences related to popular music with less attention to artist creativity in recent years.

B. Popularity Classification

Initial diagnostics of popularity classification looked to quantify feature importance within the discussed models. Feature importance was calculated using a 250 randomized decision trees (i.e. extra-trees) on numerous sub-samples of the data set and averaging of these trees looked to improve predictive accuracy and control for over-fitting. Feature importance was calculated at each split through the use of the discussed decision criterion, Gini Impurity, and features that maximized this value were labeled as more important features to popularity classification.

The results of this feature selection are displayed in Fig. 2. The total Gini Impurity for each feature are displayed as the size of the bars with the standard deviation each feature represented through the gray error bar. Examining the figure, there appears to be a large discrepancy between the importance of specific features and classification of song popularity.Fig. 2 indicates the year in which the song was released and the proportion of unique words are most indicative of song popularity. It is important to note these features also show the largest amount of variation amongst all model features. Examination of this plot also raises some concerns about the target variable and relationship to audio/lyrical features of a song. The most important feature,year, is only semi-related to song attributes indicating a possible skew or underlying correlation between song popularity and year of song release. This factor points to a potential drop-off in continuous appreciation of older rap/hop-hop tracks amongst listeners and a greater focus on recent artists,

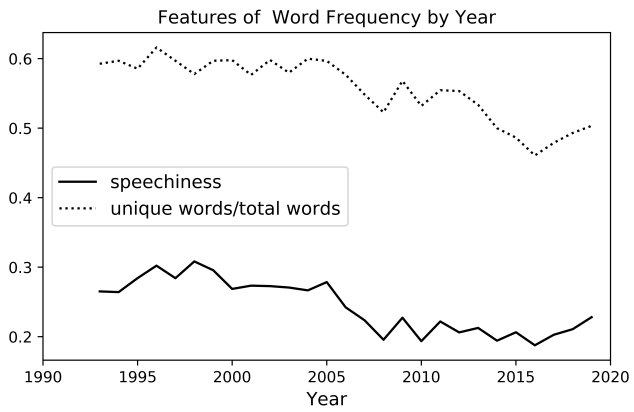

FIG. 1. Track speechiness and proportion of unique words averaged by year over the time frame.

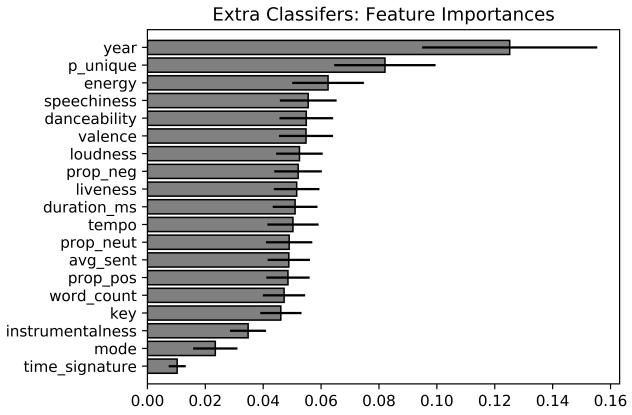

FIG. 2. Extra trees classifier feature importance calculated through the use of 250 de-correlated decision trees. Feature importance was calculated as overall Gini Impurity and aggregation of these values are displayed.

music, and song releases.

The machine learning algorithm with the highest recorded F1-score was tied between the Random Forest and Random Forest using Feature Selection (See Table I. The RF with feature selection minimizes model over fitting through the use of smaller set of features. Larger decision trees lose generalization capability due to being more specific to the training set. Thus, we determine the optimal model in terms of leveraging precision,recall, and over-fitting to be the Random Forest using feature selection. Moreover, amongst compared kernel, SVMs using an 'rbf' kernel achieved the highest precision, recall, and F1-score implying a non-linear relationship between the audio/lyrical features of a track and track popularity. General comparison of model performance, displayed in Table I, indicates no significant differences in popularity classification dependent on model choice. This is demonstrated by similar ranges of performance metrics. For example, all recall and AUC scores fell in the range of 0.51 to 0.68. The results of model testing also inherently displays the nuances of specific classification algorithms. For instance, poor performance of the SVM classifier is primarily caused by the increase importance of data points closer to the separation hyper plane. On the other hand, logistic regression and DTs/RFs place equal weighting on all training instances. Additionally, increased model performance can be attributed to the popularity threshold discussed above. More specifically, definition of the top 33% of songs as popular for classification creates training instances with more defined feature characteristics in combination with enough instances of both class to properly training models for popularity classification.

IV. Conclusion

Through the use of three unique classification algorithms, we are able to delineate the most influential model features, specifically the year of song release, the proportion of unique words, and energy/speechiness of the track.

TABLE I. Classification Results for Different Models with performance metrics, prediction, recall, F1-score and AUC

Model	Precision	Recall	F1	AUC
Logistic Regression (C=10)	0.68	0.65	0.65	0.65
SVM (RBF,C=1,$\gamma = 0.1$)	0.79	0.51	0.38	0.51
DT	0.62	0.62	0.62	0.55
DT w/ Feature Selection	0.66	0.66	0.66	0.65
RF (n=1000)	0.73	0.67	0.67	0.68
RF w/ Feature Selection	0.71	0.67	0.67	0.66

Moreover, model results visualize the increased importance of lyrical/metadata features to audio features of a track. This conclusion demonstrates a relationship between feature importance and the examined music genre, rap/hip-hop. Comparatively, audio features of rap/hip exhibit a large amounts of variation with respect to examined track, a feature inherent to the evolving framework of popular rap music. A wide variety of acoustical features are displayed amongst popular artists making it difficult to extract audio features of rap/hip-hop songs that represent popularity of the genre.

A. Limitations

Specific limitations of data analysis arise when examining methodology and conclusions of discussed research. First and foremost, possible bias arise through the method of data aggregation. Data was acquired through the use of user Spotify playlists which allows for a certain threshold of user error to occur. These playlists may not exactly resemble Top 100 Billboard charts of assigned year causing possible outliers in the popularity of specific years or individual tracks. Additionally, the target variable popularity appears to be be heavily reliant on the recency of song release. Creation of a new, normalized metric of song popularity such as total Spotify streams or normalized track revenue (within year of release) may properly account for the popularity of less recent tracks by accounting for popularity closer to track release. Finally, computational limitations restricted the acquisition of large amounts of track data possibly constraining model results and conclusions.

B. Further Work

Future work within the realm of music analysis and more specifically historical trends of the rap/hip-hop genre displays numerous possibilities for extensions of discussed research. For example, a possible extension of this research may utilize audio/lyrical features of tracks and topic modeling to allow for artist/track recommendations and creation of user specific Spotify playlists for later recreational enjoyment. Additionally, further exploration into the field of machine learning and music prediction through the use of different machine learning techniques and feature creation may indicate new, important features not discussed within this piece. The Millions Song Dataset [6] highlights an opportunity for strong data collection as a freely-available collection of audio features and metadata for a million contemporary popular music

tracks.

A. Audio Features

TABLE II. Audio Features from Spotify for Developers

Key	Type	Description
duration_ms	int	The duration of the track in milliseconds
key	int	Estimated overall key of the track (integers map to pitches using Pitch Class notation) -1 if no key detected
mode	int	Modaility(major or minor) of track (major = 1, minor = 0)
time_signature	int	An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).
acousticness	float	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
danceability	float	Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
energy	float	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
instrumentalness	float	Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
loudness	float	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.
liveness	float	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
speechiness	float	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
valence	float	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
tempo	float	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.

[1] Ioffe, S., Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. ArXiv:1502.03167 [Cs]. http://arxiv.org/abs/1502.03167

[2] North, A., Hargreaves, D. (2008). The Social and Applied Psychology of Music. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198567424.001.0001

[3] North, A. C., Hargreaves, D. J., O'Neill, S. A. (2000). The importance of music to adolescents. British Journal of Educational Psychology, 70(2), 255–272. https://doi.org/10.1348/000709900158083

[4] Raviv, A., Bar-Tal, D., Raviv, A., Ben-Horin, A. (1996). Adolescent idolization of pop singers: Causes, expressions, and reliance. Journal of Youth and Adolescence, 25(5), 631–650. https://doi.org/10.1007/BF01537358

[5] Rokach, L., Maimon, O. (2005). Top-Down Induction of Decision Trees Classifiers—A Survey. IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), 35(4), 476–487. https://doi.org/10.1109/TSMCC.2004.843247

[6] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.

[7] Tibbs, D. F. (2012). From Black Power to Hip Hop: Discussing Race, Policing, and the Fourth Amendment Through the "War on" Paradigm. The Journal of Gender, Race, and Justice, 15(1), 47-79. https://search-proquest-com.ezproxy.uvm.edu/scholarly-journals/black-power-hip-hop-discussing-race-policing/docview/1508067282/se-2?accountid=14679