­­Alex Chan

RapNet: Machine Learning for Hip-Hop Artist Classification

**Introduction**

Musical artists distinguish themselves in many ways, defining their own styles and genres to appeal to an audience. One of the ways artists accomplish this is through lyrics. Lyrics act as a medium for the artist to convey a message – whether it be a story, call to action, or anything in between. The goal of this paper is to develop a classifier to distinguish between lyrics of different artists, specifically in the hip-hop genre.

I found this problem particularly interesting because I was curious how well a machine learning classifier could do on the lyrics of authors within the same genre. Generally, these algorithms perform very well in genre classification, and I wanted to see if the same could be said about distinguishing between artists.

**Methodology**

Data Collection

The data analyzed in this project is song lyrics. Lyrics were retrieved using the Genius API to get the URLs to the lyrics of songs by an artist. Then, using the BeautifulSoup library, the data was parsed from the HTML of the page containing the lyrics. All song lyrics from the entire discography of each artist were collected and saved for use in our model.

Preprocessing

The lyric data was cleaned by normalizing the words. This was done by converting all of them to lowercase and removing non-alphanumeric characters from the text. Once the data was cleaned, it needed to be transformed into input for the model. First, the data was converted into a document-term matrix, where each row represented a sample (a song lyric) and each column represented a feature, with the sample’s corresponding count of that feature. The features consisted of all of the unigrams and bigrams (one and two word representations) from the lyric data. After that, the document term matrix was converted to a TFIDF matrix. TFIDF, which stands for Term Frequency-Inverse Document Frequency, is a way of assigning a measure of how important a word is in classifying a sample to a class. Once in this representation, the data was ready to be input into the model.



Model Selection

The data was evaluated by comparing the results from 3 models –Random Forest, SVM, and Neural Networks.

Table 1: Random Forest Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Actual | | | |
| Predicted |  | Kendrick Lamar | Logic | Tyler the Creator |
| Kendrick Lamar | 4 | 0 | 0 |
| Logic | 2 | 6 | 0 |
| Tyler the Creator | 2 | 2 | 8 |

Table 2: SVM Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Actual | | | |
| Predicted |  | Kendrick Lamar | Logic | Tyler the Creator |
| Kendrick Lamar | 6 | 0 | 0 |
| Logic | 1 | 8 | 3 |
| Tyler the Creator | 1 | 0 | 5 |

Table 3: Neural Network Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Actual | | | |
| Predicted |  | Kendrick Lamar | Logic | Tyler the Creator |
| Kendrick Lamar | 5 | 0 | 0 |
| Logic | 1 | 8 | 0 |
| Tyler the Creator | 2 | 0 | 8 |

The Random Forest, SVM, and Neural Network models perform with 73%, 80%, and 88% accuracies, respectively. The neural network outperforms the other two models at this task. Neural networks are known to perform exceptionally well for various classification tasks. Neural networks work by doing several computations through various “layers” of the model, computing dot products with labels in the network. The output is a set of probabilities that indicate the likelihood of a sample belonging to a particular class. The model improves its classification accuracy through the backpropagation algorithm, and algorithm that uses gradient descent to adjust the weights of the model to make better predictions. After tuning the hyperparmeters of our model, which include learning rate, hidden layer size, number of hidden layers, and activation functions, the model that performed best was a 4 layer neural network (with hidden layer sizes 256, 128, 128, and 64, respectively) that used a hyperbolic tangent activation function and a log loss function. Using our TFIDF representation of the data as the input, the model performs the series of computations in the network, and outputs its guess at which artist the song lyrics belong to.



**Results**

The training and test data consists of song lyrics by the artists Kendrick Lamar, Logic, and Tyler the Creator. The model was trained on 80% of the data, and then tested on the remaining 20% of the data. Grid search was used to tune the hyperparameters of the model (layer size, learning rate, batch size). Once the optimal hyperparameters were found, 5 fold cross validation was used to determine the success of our model. Our model achieves an 88.5% accuracy on the test set and a cross validation accuracy of 83.7%.

**Analysis**

An accuracy of 88.5% shows that our model does pretty well at predicting the artists for each song lyric into the three classes. The table below shows the most frequent words sung by each artist with stop words, common words that don’t contribute to the success of the model, removed.

|  |  |  |
| --- | --- | --- |
| **Kendrick Lamar** | **Logic** | **Tyler the Creator** |
| Love | Mind | Man/boy |
| Life | God | Time |
| Time | Everybody | Want/Wanna |
| Man | People | Girl |
| World | Feel | Love |
| High | Real | Watch |
| Black | Black | Feel |
| Money | Everything | Suck |
| God | Time | Gang |
| Mind | World | Couple |

As seen by the table of word frequencies and the general themes the artist, Tyler the Creator generally talks about more emotional topics like love and depression. Logic is more involved with topics like life and racial equality. Kendrick Lamar talks about a culmination of these topics, as well as “money” and “high”.

Interestingly, the model seemed to struggle particularly with Kendrick Lamar songs, which overlapped significantly with the themes of the other two artists. While the model was able to predict all of Tyler the Creator and Logic songs in the test set correctly, it misclassified several Kendrick Lamar songs. The song “Love” shares several of the central themes in Tyler the Creator’s songs, mainly including ideas about romantic relationships.

*“So give me a run for my money Sippin' bubbly, feelin' lovely, livin' lovely Just love me I wanna be with you, ayy”*

Another line from “Kush and Corinthians” talks about living a fulfilling life, which shows many similarities to many of the songs in Logic’s 2017 album, “Everybody”.

*“Die to it, live your life, live it right, be different, do different things”*

While the model has some shortcomings at differentiating songs with similar themes, overall it does very well at classifying song lyrics to their corresponding artists.

**Conclusion**

These results can hopefully provide a further step towards solving harder classification problems. While genre classification has had generally great success, classification of artists with semantically similar themes proves to be a harder problem to solve. The evaluated model wirjs well on artists with different themes; it was successfully able to distinguish between Logic and Tyler the Creator songs. However, the model fell short at predicting Kendrick Lamar songs, whose themes range across variety of different topics. Further studies include classifying more artists, and determining methods to make models even better at distinguishing between artists with similar themes, such as rhyme scheme and repeated phrases.