|  |
| --- |
| **Explicit Spotify Song Classification** |
|  |
| **David Chan** |
| **University of Maryland**  **Dchan107@umd.edu** |
|  |
|  |
|  |

Abstract

This project aims to determine whether a song is explicit by using various features that Spotify provides to its users when listing each song. Although there is a high accuracy when classifying whether a song is explicit, the model itself is not very precise and will not always provide the correct result. This project shows that although it is possible to create a model to classify whether a song is explicit, the created model is not very precise and makes classification mistakes for explicit songs.

Introduction

This project aims to use the selected features to determine whether a song contains explicit language. For example, the data set contains two instances of DJ Khaled’s intro song to the album “We The Best” named “Intro (We The Best)” with one being explicit and the other not. However, there are recorded differences in these two songs outside of the lyrics where the explicit version has a higher tempo as well as different values for things such as Spotify’s valence. By inputting these attributes, we hope to create a model which will allow us to determine if a song is explicit. In this dataset, about 7% of songs are explicit while the rest are not. The project will use information on approximately 1200000 songs on Spotify and use information from these songs such as the song name, the song’s album, the song’s artist, tempo of the song (the pace of a song), valence of a song (positiveness value), energy, song duration, time signature, and the year the song was released. Although some song names have the word “Explicit” in them for people to understand that they are explicit, not all songs do. In addition to this, not all songs that are explicit are marked as explicit.

Some of the attributes included in the data set are not used. For instance, the disc number of the song is unused because more than 1.1 million songs have the disc number 1. Some of the ids such as album id and artist id were also not used because they are repeats of the albums and artists and would not affect the classification. In this data set, every time the artist “Rage Against The Machine” showed up, they always had the same artist id which is expected. Since the same artist id is used for the artist, the columns are effectively the same and there would most likely be little to no effect on the classification by including the id columns in addition to the regular artist and album columns of the data set.

The experiment will be done using PySpark with multiple train-test splits such as 80-20, 75-25, and 70-30. Since the percentage of explicit songs in this set is so low, it is possible for a random test-train split to result in all the explicit songs being in the training set. Because of this, the experiment will be repeated at least once with a different seed to lower the chance that all the explicit songs are placed into either the training set or testing set. This issue will be a larger problem if many of the explicit songs are placed into the testing set because there won’t be enough training data for the model to determine which songs are explicit. We will use logistic regression to create each model before testing with the test set.

There was a recent paper about toxicity in online discussions. In that experiment, they focused on identifying areas of a post which would be responsible for marking the post as toxic. However, the same words said differently can mean very different things which makes it difficult for that research project to show similar results with this one. For example, emphasizing “dog” in “I didn’t say the *dog* ran away.” means something very different from emphasizing “away” in the same sentence. Because of this, this project uses the features of the song instead of the lyrics to test whether a song is explicit or not.

This project could be useful to parents of young children who feel that they do not want their children to hear explicit language. Another potential audience is people who prefer not to hear explicit language when listening to music. Although explicit language may not bother everyone, there are still cases of people being extremely offended by it. By creating this model to classify whether a song is explicit or not, people may be better able to determine whether a song is explicit or not even when the songs do not include the word in the song title.

Methodology

The purpose of this project was to determine whether it was possible or not to predict whether a song was explicit or not based off several features. When people see a song recommended for them, they see the song title, artist(s), and album. However, they do not see whether the song is explicit or not. Since not all songs will be marked as explicit when submitted to Spotify, we aim to find a way to determine for ourselves if a song is explicit. The features we used for this were the song’s name, the song’s album, the song’s artist(s), Spotify’s energy level for the song, the tempo of the song, Spotify’s valence for the song, the duration of the song, time signature, and the song’s release year. The data used for this was collected using Spotify’s API to collect information on approximately 1.2 million songs from about 165 thousand artists or groups of artists. For any data points that were missing data, the data point was skipped entirely as the missing data may have been important to determining whether the song was explicit or not. The earliest recorded songs used for this are from the album “Living Chicago Blues Vol. 3” from the year 1908 which includes songs from the artists A.C. Reed, Scotty and the Rib Tips, Lovie Lee, Lacy Gibson, and Sons of the Blues. Although not all this information is immediately in Spotify’s application, the Spotify API enables users to collect this information.

For the experiment, we created three training and testing sets that split the original data set in an 80-20 split, a 75-25 split, and lastly, a 70-30 split. To classify the data, a logistic regression was done to determine based off those features to determine if a song was explicit. A script was created using PySpark to first upload the data. In this script, we cleaned the data to create a data frame with only the features we planned on using. With the initially obtained data set, PySpark had parsed them all as strings and many features needed to be casted to the correct value. For example, name, album, and artists were all correct. However, energy and valence needed to be casted to be float types. After casting all the feature types, another issue came up when attempting to create the model. This issue required changing the values in the “explicit” column from being either true or false to 1 or 0 respectively to resolve. After that, we used estimators and transformers to create new data sets that could be used with logistic regression. The data was then randomly split using the same seed each time (1234) to ensure the reproducibility of the experiment. Next, we created a model for each train-test split and then reran them using another seed (12345). When performing logistic regression, we chose to set the value of “maxIter” in the “LogisticRegression” function to 3 because higher values such as 10 and 100 resulted in memory errors involving the Java heap stack as well as the GC Overhead limit being exceeded. After creating the model and using it to predict whether the songs in the test set were explicit, we created a confusion matrix for each train-test split and calculated the accuracies as well as the area under the ROC for the training set and testing sets.

Results and Discussion

We found that the experiments done with the largest training set had the highest accuracy at 96.77% with a test area under ROC of 98.31%. The second model had an accuracy slightly lower at 96.75% with a test area under ROC of 98.28%. The third model which had the smallest training set, had an accuracy of 96.74% and an area under ROC of 98.21%. Overall, the differences between these three models are extremely small and they will most likely correctly classify the same number of songs. To put it into perspective, after 10,000 songs, one might expect the second model to correctly classify one song more than the third model which had the smallest number of data points in its training set. Based off this accuracy, the best classification and model are very good. In addition to that, there is a high area under ROC for both the training set and testing set with the training set area under ROC being 99.93% and the testing set area under ROC being 98.31%

Chart

Description automatically generated

The problem with this classification is that the precision of the first model or the rate at which it determines whether a song is explicit only reached 60.10%. This means that although the most accurate model correctly classified 60.1% of the songs as explicit, it failed to correctly classify the other 39.9% of explicit songs as explicit and is only slightly better than flipping a coin.

Chart, waterfall chart

Description automatically generated

The results from this experiment show that it is possible to classify whether a song is explicit or not based off these features. If the information used for this experiment is provided for a song added to Spotify in the future, it will be possible for Spotify users to determine whether the song is explicit by using this model. Although this model was created by using approximately 1.2 million songs, Spotify has over 70 million tracks available. If more songs are used in the creation of a similar model, a more accurate model can be created and used to predict whether a song is explicit or not. As more songs are added, even if they are not initially marked as explicit, we will be able to determine whether they are explicit.  
  
From this experiment, Spotify users will be able to determine whether a song is explicit or not. As mentioned in the introduction, users who are parents of young children who do not want their children to hear explicit language will be able to stop their children from hearing it. Since not all songs are marked as explicit when they are initially uploaded, this experiment would allow for people to put a song and these values into the predictor to predict for them whether the song is explicit.

In future experiments, researchers may be able to use similar data with more or fewer features to maintain similar accuracy or even increase the accuracy when compared to the current three models created. In this current paper, some of the features used such as energy and valence are subjective to Spotify’s views on them. As a result, this model may not be completely accurate. Energy refers to a perceptual measure of intensity or activity with death metal for example having high energy while something such as a Bach piano prelude would have a low energy. The valence is the musical positiveness conveyed by the track with lower valences corresponding to more negative tracks. Since it is unknown how exactly these are determined unlike the tempo of a song or the song’s release year, a “better” model might involve not having these features included when being created. Instead, future researchers may only use features of the song itself as recorded on the sheet music.

1. Conclusion

Although this research project was a success and a classification for whether a song is explicit is possible based on the features mentioned, it may not be the most accurate classification possible. The most accurate model created is very accurate at 96.77%. However, it only has a precision of 60.1%. In addition to this, it uses some of Spotify’s own features for each song. For instance, the energy level of a song is used as one of the features to determine whether a song is explicit. However, since this number appears to be determined by Spotify, a better model would be one that does not need this. In the future, a better model might be able to be created based off only the features of a song that are determined the moment the song is written.