Genre Classification Based on Song Lyrics

TDDE16 Report

Nicholas Sepp Löfgren, nicse725

Abstract

In this paper text classification of song lyrics has been studied with the goal of classifying lyrics into music genre classes. The genre classes were defined by reviews from the online music publication Pitchfork. The studied classifiers were classifiers based on count and tf-idf vectorizers and Multinomial Naive Bayes, Multilayer Perceptron Classifier and Support Vector Machine Classifier predictors. All of the studied classifier pipelines outperformed baseline results, which were two baseline classifiers using the 'stratified' and 'most frequent' sampling strategies. The Support Vector Machine Classifier was analysed closer and an exhaustive grid search cross-validation was conducted on this classifier to tune its hyperparameters, which increased its performance. From the results of the tuned classifier it was observed that the F1 score for the rap genre was significantly higher than the second highest F1. Further analysis revealed that the rap genre had the largest amount of total words and unique words per lyric on average among all studied genres. Furthermore, investigating the most significant features of a linear Support Vector Machine Classifier revealed that the most significant features when classifying rap were very unique, genre specific features. In conclusion rap is arguably the most lyrical among the studied genres.

Contents

1 Introduction

Pitchfork Media is a music publication website launched in 1995 initially as a blog which later expanded into a fully realised online publication. It is currently owned by Condé Nast (Pitchfork, 2015). Pitchfork is considered to be one of the most successful music publications of the digital era. Initially the publication earned its reputation for its extensive coverage of indie music, but has with time expanded into covering all kinds of genres and not only independent artists; big, established names in the music industry are often covered (Singer, 2014). Pitchfork has a reputation of being internet music era tastemakers, especially during the 90's and 00's, making or breaking an artist's career with a single review. However, such tastemaking influence of music journalism at large has diminished in recent years in the internet era (Barshad, 2018).

In this study a data set of over 18,000 reviews scraped from Pitchfork retrieved from Kaggle has been considered (Conway, 2017a). The primary question investigated was whether genre could be predicted from lyrics alone. Using the resulting classifiers a further analysis was conducted on the reasons behind the results. The genre classes for the classifier were defined by the genres which are reviewed by Pitchfork. An additional research question was to investigate whether or not Pitchfork as a publication is biased towards giving some genres higher or lower scores than what the data would suggest. All code used in this study can be found in a repository on GitHub.

2 Theory

2.1 Lyric Classification

In text classification the task is to categorise text documents into predefined classes and the standard pipeline is illustrated in figure 1. The documents for the lyric classifier is of course documents of lyrics, how these were collected is described in section 4.1 Lyrics Scraping. The classes were the 9 genres reviewed by Pitchfork. After retrieval the documents were vectorized, in this study a count vectorizer to produce bag-of-words vectors and a term frequency-inverse document frequency (td-idf) vectorizer to produce tf-idf vectors were used. The document vectors are then fed into a predictor which predicts the documents' class labels. The predictors studied were Multinomial Naive Bayes, Multilayer Perceptron Classifier and Support Vector Machine Classifier. These pipelines are trained using a subset of the total data as training data and the rest of the data is used as test data for the classification report to study the results of each classifier pipeline. Also in this study a grid search cross-validation has been used to tune the hyperparameters of a classifier, which exhaustively considers all parameter combinations when doing the cross-validation.

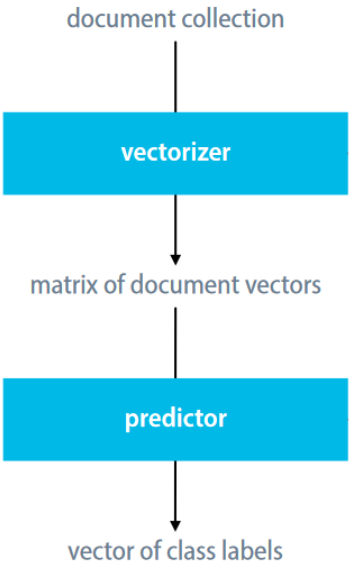


Figure 1: A standard text classification pipeline (Kuhlmann, 2020).

2.2 Genre Bias

The genre bias study was mostly a statistical consideration, so there is not much relevant theory to present. One aspect of the genre bias study was the smoothing of average score data points for each count of reviews by using the Savitzky-Golay filter. This filter works by fitting successive sub-sets of adjacent data points with a low-degree polynomial by the method of linear least squares (Schafer, 2011).

3 Data

3.1 Pitchfork Review Data

The primary data set used in this study was a set of 18,393 reviews scraped from Pitchfork from January 5th, 1999 to January 8th, 2017, which was retrieved from Kaggle (Conway, 2017a). The data set is provided as a sqlite database with the tables 'artists', 'content', 'genres', 'labels', 'reviews' and 'years'. All of these tables are more or less self-explanatory except for perhaps 'reviews'; this table contains the title of the project reviewed (most commonly an LP or EP), the artist's name, an URL to the review on the Pitchfork website, the score given, a boolean value whether the project was considered "Best New Music" when reviewed, the author's name, the relation the author has to the publication (eg. "contributor" or "senior staff writer"), date of publication, and which weekday the review was published on.

3.2 Lyrics Data

Using the Pitchfork data set a mapping of artist, album and genre was made for each genre included in the data set which were electronic, experimental, folk/country, global, jazz, metal, pop/RnB, rap and rock. For each genre a list of album/artist pairs were constructed, since versatile artists could produce albums which were reviewed in different genres the artist's name alone was not enough to determine genre belonging. These pairs were then used to scrape lyrics for each genre using the Genius API (Genius, 2021). Genius is another online music publication with a comprehensive database of lyrics that can be accessed through their API. The resulting data from this scrape was a data set of around 1,200 lyrics with corresponding genre labels. An excerpt of this data set is shown in figure 2 More details and limitations of how this data was generated are given in section 4.1 Lyrics Scraping.

	lyric	genre
686	confessor of the tragedies in man lurking in t...	metal
1061	get a load of me, get a load of you walking do...	rock
840	new, new attitude but it's a shame, better tha...	metal
292	like a freshly cut diamond like a freshly cut ...	rap
437	you know this place you know this gloom? we've...	electronic

Figure 2: An excerpt of the scraped lyrics data set.

4 Method

4.1 Lyrics Scraping

Using the aforementioned list of album/artist and genre pair, lyrics were scraped using the Genius API (Pai, 2019). From the Pitchfork data set 100 pairs were generated for each genre. When scraping lyrics, each artist's 5 most popular songs were scraped, where the popularity metric was determined by Genius. There were three important limitations in this approach, the first one was that scraping only considered the artist and not the album which meant that lyrics from versatile artists could be wrongly labelled, for example among one artist's 5 most popular songs there could be songs from an album that Pitchfork considered to be electronic and from another album which could have been considered rock. Taking album into consideration was not successfully implemented. The second limitation, partly caused by the first, was the lists of pairs contained duplicates of artists. Only one instance of these duplicates were processed to avoid copies of lyrics in the resulting lyrics data set. The third limitation was that the API occasionally was not able to access a song's lyrics and returned nothing in those cases. A simple preprocessing step was then applied; all lyrics were converted to lower case and empty lyrics for instrumental songs were not included in the final data set. The scraping limitations and the preprocessing step resulted in an uneven data set containing 1,224 lyrics. The uneven genre distribution is shown in figure 3.

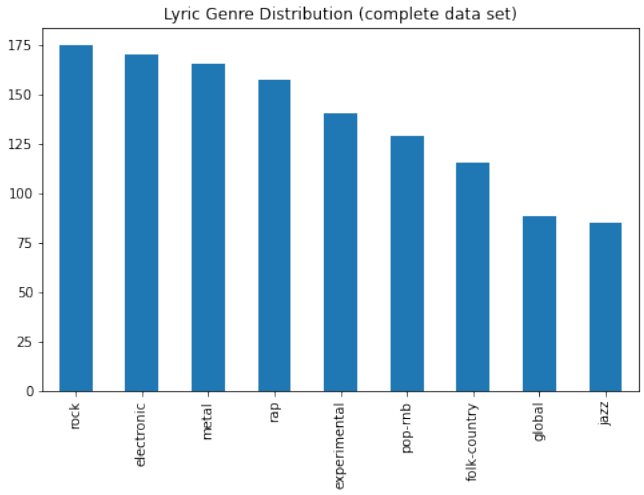


Figure 3: Genre distribution of lyrics of the complete lyrics data set.

4.2 Lyrics Classifiers

As a first step in training the classifiers the complete lyrics data set was randomly shuffled and divided into training and test data, in a 74/26 proportion. The training data set was undersampled to create a more balanced data set, which is shown in figure 4. After undersampling the training data and test data sets contained 567 and 323 lyrics respectively.

With these data sets, different classifiers were studied; pipelines of different combinations of vectorizers and predictors from the scikit-learn Python library were constructed and tested. The vectorizers studied were CountVectorizer and TfidfVectorizer, and the predictors tested were Multinomial Naive Bayes, Multilayer Perceptron Classifier and Support Vector Machine Classifier. At this point the default values were used for all parameters for both the vectorizer and predictor in all constructed pipelines. Two baseline classifiers were also constructed using the DummyClassifier function which were compared to the studied classifiers. These baselines used the sampling strategy 'stratified' and 'most frequent'. The performance metric used was the classification report in scikit-learn. All of the classifier pipelines performed better with the TfidfVectorizer except for the pipeline with the Multinomial Naive Bayes predictor. However, since that difference was insignificant, the results for the studied classifiers reported in section 5.1 Lyrics Classifiers all use the TfidfVectorizer, for the sake of consistency.

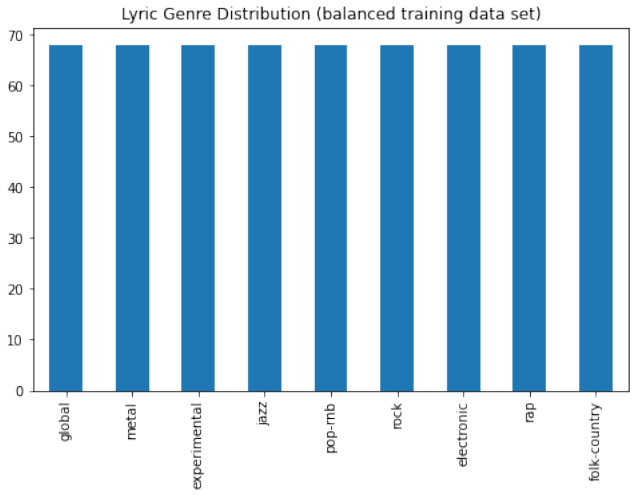


Figure 4: Genre distribution of lyrics of the lyrics training data set.

4.3 Genre Bias

The method used for investigating genre bias was based on a review score exploration done by the creator of the Pitchfork data set (Conway, 2017b). This method calculates and plots the average score depending on the number of reviews and also places each genre in relation to this average line. This result is shown in section 5.2 Genre Bias.

5 Results

5.1 Lyrics Classifiers

The first results for the lyrics classifiers are presented in table 1. The F1 score for each genre is reported as well as the total accuracy for the classifier in the last row. The two baselines in 1a were used as comparison for the classifiers in table 1b. The stratified baseline classifier generates predictions by respecting the training set's class distribution and the most frequent baseline classifier always predicts the most frequent class in the training set. Since the training set is downsampled so that there is an equal amount of data points for each class label, the most frequent baseline classifier will always predict the first class alphabetically, which is 'electronic'. The input for all classifiers presented in table 1b are tf-idf vectorized lyrics and the classifiers studied were the Multinomial Naive Bayes (Mult. NB), Multilayer Perceptron (MLP) and Support Vector Classifier (SVC) classifiers.

Table 1: Classifier results.

(a) Baseline classifiers.

	Stratified	**Most freq.**
Electronic	0.10	0.23
Experimental	0.22	0.00
Folk/Country	0.10	0.00
Global	0.03	0.00
Jazz	0.07	0.00
Metal	0.06	0.00
Pop/RnB	0.08	0.00
Rap	0.15	0.00
Rock	0.06	0.00
Accuracy	0.10	0.13

(b) Classifiers, tf-idf vectors.

	Mult. NB	**MLP**	**SVC**
Electronic	0.04	0.33	0.14
Experimental	0.08	0.15	0.15
Folk/Country	0.05	0.28	0.17
Global	0.15	0.35	0.25
Jazz	0.00	0.19	0.24
Metal	0.19	0.17	0.34
Pop/RnB	0.14	0.09	0.31
Rap	0.29	0.73	0.62
Rock	0.09	0.20	0.08
Accuracy	0.18	0.28	0.27

The results in table 1b suggested that SCV is an appropriate predictor for this task, it produced fairly good results and was very fast compared to MLP. The SVC classifier was therefore studied closer, and its hyperparameters were tuned with the help of a grid search with 5-fold cross-validation. The results given by the tuned classifier is given in table 2a and its corresponding hyperparameters are given in table 2b. These results are for the hyperparameter grid defined in table 3.

Table 2: Results for grid search with 5-fold cross-validation for SVC classifier.

(a) Tuned classifier results.

	Tuned SVC
Electronic	0.34
Experimental	0.18
Folk/Country	0.39
Global	0.25
Jazz	0.05
Metal	0.34
Pop/RnB	0.21
Rap	0.68
Rock	0.27
Accuracy	0.33

(b) Tuned hyperparameters.

SVC	**Parameter value**
C	1.5
kernel	'sigmoid'
Vectorizer	
binary	True
ngram_range	(1,1) (i.e. only unigrams)
stop_words	'english'

Table 3: The parameter grid from where the optimal hyperparameters were chosen.

SVC	**Parameter values**
C	[1, 1.1, 1.2, 1.3, 1.4, 1.5]
kernel	['linear', 'rbf', 'poly', 'sigmoid']
Vectorizer	
binary	[False, True]
ngram_range	[(1, 1), (1, 2), (2, 2)]
stop_words	['english', None]

5.2 Genre Bias

The produced genre bias plot is shown in figure 5. It is observed that the rock genre is the most reviewed genre and by a quite large margin; the rock genre has about 5,000 more reviews than the second most reviewed genre, electronic. All other genres were reviewed less than 2,000 times.

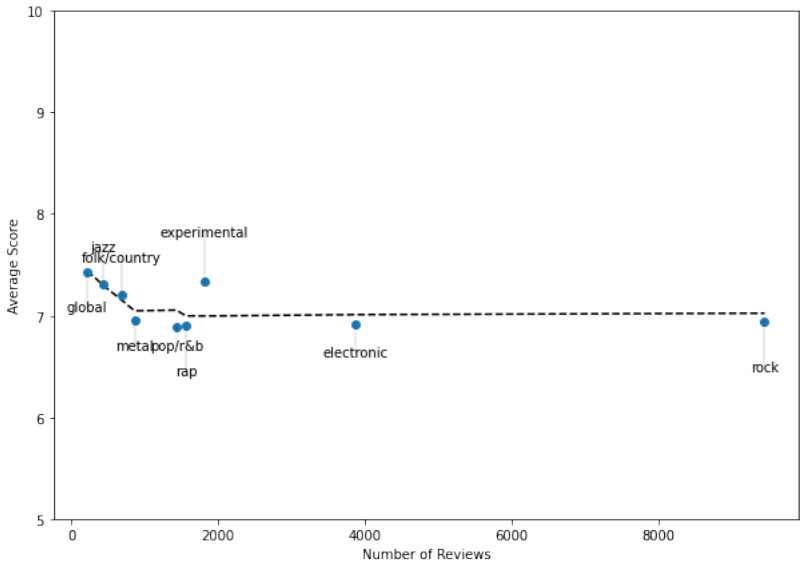


Figure 5: Plot of score ratings per genre (Conway, 2017b)

6 Discussion

6.1 Lyrics Classifiers

As an initial comment on the results it is observed from table 1 that all studied classifiers perform better than both baselines and that the cross-validated classifier

in table 2a outperforms the other classifiers. Furthermore from the results of the tuned SVC classifier presented in table 2a it seems like the rap genre's F1 score of 0.68 is an outlier, it is larger than the second highest F1 score of folk/country's 0.39 with a factor of almost 1.8. This observation is also highlighted by the confusion matrix for the tuned SVC classifier, which is shown in figure 6.

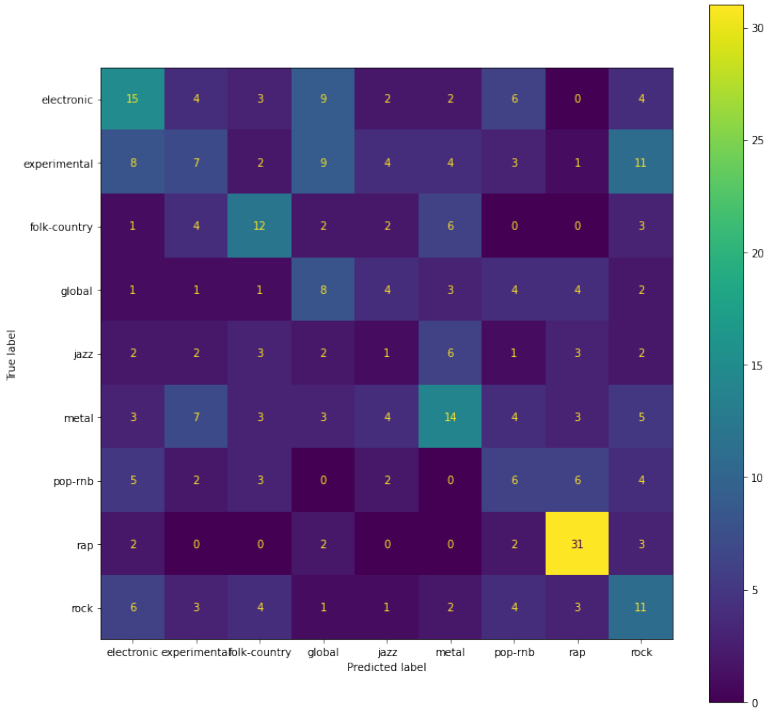


Figure 6: Confusion matrix for the tuned SVC classifier.

Further analysis of this outlying result is given by plots of the average amount of total and unique words per lyric for each genre, which are shown in figure 7 and figure 8. It is observed that rap has the largest amount of total and unique words per lyric among all studied genres on average, which could be indicative that the rap genre is the most easily correctly classified genre.

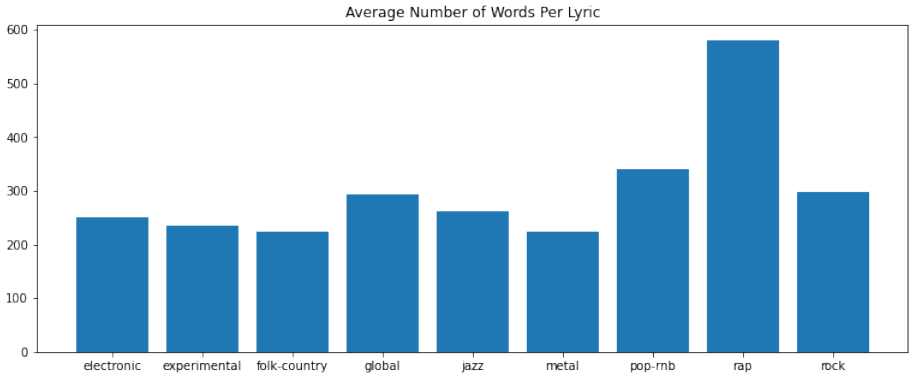


Figure 7: Plot of average amount of words per lyric for each genre.

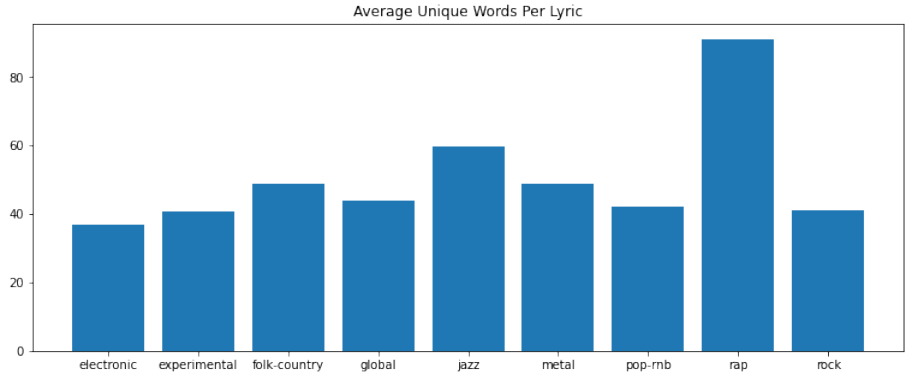
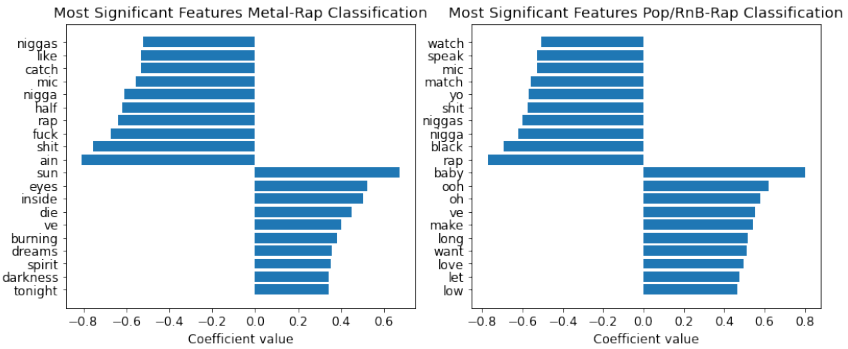


Figure 8: Plot of average amount of unique words per lyric for each genre.

One interesting aspect to study would be the most significant features of a linear classifier and to see if among these features were words which are heavily related to the rap genre. A SVC classifier was trained using the hyperparameters in table 2b except with a linear kernel. The coefficients in the trained classifier are returned in a matrix with the shape `(n_classes * (n_classes - 1) / 2, n_features)`, where in our case `n_classes = 9`, since we have 9 genre class labels (scikit-learn, 2020). Each row corresponds to a binary classifier. Using this it is possible to study the most significant features when the classifier does "one-vs-one" classifications between two genres. The most significant features when deciding between Metal-Rap and Pop/RnB-Rap are shown in figure 9. The metal and pop/RnB genres were chosen for comparison because they had the second and third highest F1 scores respectively as reported by the linear SVC classifier. Another observation worth noting is that the cultural roots of pop/RnB, especially RnB, are closely related to those of rap while metal and rap in general are fairly disconnected genres historically, but with significant exceptions of course.



(a) Most significant features in Metal-Rap classification.

(b) Most significant features in Pop/RnB-Rap classification.

Figure 9: Most significant features for the trained linear SVC classifier.

As is observed from subfigure 9a the words with positive coefficients are heavily associated with the metal genre, words such as 'darkness', 'die' and 'burning' seem very indicative. As is observed from subfigure 9b the words with positive coefficients are associated with the pop/RnB genre; words such as 'baby', 'love' and various vocalization sounds as 'ooh' are prevalent themes in this genre. In both subfigures the words associated with the negative coefficients are words heavily associated with rap music; these words mainly reflect the rebellious nature and African-American heritage of the rap genre. Furthermore, most of these words are rarely used in other genres and are more or less unique for the rap genre.

6.2 Genre Bias

As is observed from the plot of genre bias, which is shown in figure 5, there does not seem to be any particular genre bias when it comes to scores, except for perhaps the experimental genre. It seems like the experimental genre is scored above average at that count of reviews with somewhat of a margin. A closer look into whether or not this difference is statistically significant could be conducted, to determine if Pitchfork has a bias towards the experimental genre. If this difference would be shown to be statistically significant, it could perhaps be explained by the fact that the experimental genre is not, in general, as commercially viable as the other genres, both for music labels and reviewing publications, so Pitchfork might only review stand-out examples of this genre. The bias would then perhaps be a result of a compromise of maintaining credibility and integrity by reviewing experimental music while at the same time remaining commercially viable for advertisers and investors. However, more analysis on the actual data should be conducted as mentioned before any real conclusions can be made.

7 Conclusion

7.1 Lyrics Classifiers

As is observed from the results, lyric classification can be done to an extent that exceeds simple baseline results, which can be observed when comparing results for the baseline classifiers in table 1a with the results for the studied classifiers in table 1b. Furthermore, it is observed that the rap genre is particularly suitable for being classified, which, for example, can be observed from the confusion matrix in figure 6. Further analysis of this observation reveals that the rap genre has the largest amount of words per lyric and also the largest amount of unique words per lyric of all studied genres, as is shown in figures 7 and 8. When further analysing a linear SVC classifier and obtaining the most significant features used in classification, it is observed that the signifying features for the rap genre are very unique and recognisable, as is shown in figure 9. A conclusion can be made that rap is arguably the most lyrical of the studied genres, in the sense that lyrics play the largest role in the genre's identity.

7.2 Genre Bias

As mentioned in section 6.2 Genre Bias, before any conclusions can be made the difference in scores between the average and the experimental genre needs to be studied closer and determined whether or not it is statistically significant. This analysis was not done in the scope of this study.

8 Acknowledgements

I want to thank Marco Kuhlmann and lab assistants, in particular Riley Capshaw, for the course Text Mining, it has been a challenging, instructive and engaging course. Finally, one of my personal favourite rappers and lyricists MF DOOM was announced dead on New Year's Eve, December 31, 2020, and I would like to end this study by saying rest in peace MF DOOM, the illest villain.

References

Barshad, A. (01-05-2018). When Critics Could Kill. *Slate*. [Online; accessed 2021-01-12].

Conway, N. (2017a). 18,393 Pitchfork Reviews. https://www.kaggle.com/nolanbconaway/pitchfork-data. [Online; accessed 2021-01-14].

Conway, N. (2017b). Exploring the review scores data. https://github.com/nolanbconaway/pitchfork-data/blob/master/notebooks/review-score-exploration.ipynb. [Online; accessed 2021-01-15].

Genius (2021). Getting Started. https://docs.genius.com/. [Online; accessed 2021-01-14].

Kuhlmann, M. (2020). Lecture series in TDDE16. https://www.ida.liu.se/~TDDE16/labs.en.shtml. [Online; accessed 2021-01-12].

Pai, N. (11-12-2019). How to Scrape Song Lyrics: A Gentle Tutorial. https://medium.com/analytics-vidhya/how-to-scrape-song-lyrics-a-gentle-python-tutorial-5b1d4ab351d2. [Online; accessed 2021-01-14].

Pitchfork (13-10-2015). Pitchfork Acquired by Condé Nast. https://pitchfork.com/news/61621-pitchfork-acquired-by-conde-nast/. [Online; accessed 2021-01-12].

Schafer, R. W. (2011). What Is a Savitsky-Golay Filter? *IEEE Signal Processing Magazine*. [Online; accessed 2021-01-17].

scikit-learn (2020). Support Vector Machines. https://scikit-learn.org/stable/modules/svm.html#svm-multi-class. [Online; accessed 2021-01-17].

Singer, D. (13-11-2014). Music Critics See Their Role and Influence Waning in The Era of Digital Music. *American Journalism Review*. [Online; accessed 2021-01-12].