(2021), 2, 1–6
doi:

Berkeley
UNIVERSITY OF CALIFORNIA

RESEARCH NOTE

Music Lyrical Classification and Analysis Using BERT for Music Recommendation Systems

Casey McGonigle and Elias E. Saravia

School of Information, University of California, Berkeley, CA, USA

Abstract

Contemporary music recommendation systems on music streaming platforms provide users with curated playlists and suggested songs. These systems, based on musical attributes, have positively impacted sales, user experience, and subscriptions, but do not leverage all lyrical attributes or utilize high-powered language models. Therefore, with lyrics from Billboard Hot 100 songs (1960 - 2020), we fine-tuned a pre-trained BERT model to classify the decade to which the song belongs to improve existing music recommendation systems. Our final BERT model achieved a validation accuracy of 0.4605 and testing accuracy of 0.40525, both significant improvements over any baseline model. In this paper we present, to the best of our knowledge, the first use of BERT to classify lyrics by their decades, an important first step toward utilizing lyrics in music recommendation systems.

Keywords: Natural language processing; data science; classification; BERT

1. Introduction

Contemporary music platforms (e.g. Spotify, Pandora, Apple Music) create music recommendation systems in order to curate playlists, recommend artists, and suggest songs to users based on their music listening patterns. These recommendation systems have been successful in creating a "positive impact on sales volume leading to increased firm revenue and web usage" and "have been shown to influence sales diversity." (Dokyun Lee et al., 2014). However, these recommendation systems are largely dependent on extracting features from the music (e.g. danceability, valence, timbre, BPM, etc.), and not the lyrics.

The innovation of modern NLP model architectures such as transformers presents an opportunity to apply high-powered language models to augment recommendation systems. Therefore, we suggest an application of Bidirectional Encoder Representations from Transformers (BERT) to song lyrics as a supplement to existing music recommendation systems. Thus, our proposed model will contain tokenized Billboard Hot 100 song lyrics from 1960 to 2021 as inputs. These tokenized lyrics will be inputted into a pre-trained BERT model to fine-tune and ultimately classify the decade to which the song belongs to or was created in.

The potential of the information from decade classification will allow music streaming platforms to complement and improve their existing music recommendation systems. For example, an avid 60s and 70s music streamer on Spotify can acquire new, contemporary artist or song recommendations that are similar in their lyrical context. Our BERT-based model is a first-step toward making that a reality.

2. Background

Bidirectional Encoder Representations from Transformers (BERT)(as presented in Devlin et al. (2018)). has been utilized in many different contexts to achieve state-of-the-art results for natural language processing tasks. Adhikari et. al (2019) have demonstrated its effectiveness in document classification (i.e. DocBERT). Kexin Huang et al (2020) have created a model to predict hospital readmission utilizing clinical notes and discharge summaries (i.e. MedicalBERT). However, to the best of our knowledge, there currently is no BERT representation for musical lyrics.

Research in lyrical recommendations using Natural Language Processing (NLP) is sparse. Lawrence Technological University Researchers, Napier and Shamir, used sentiment analysis to show that lyrics with feelings of anger, disgust, fear, sadness, and conscientiousness have increased significantly over the years while joy, confidence, and openness expressed in pop song lyrics have declined (Napier Shamir, 2018). Researchers Pettijohn and Sacco discovered trends in meaningful, comforting, and romantic lyrics during difficult social and economic times (Pettijohn et al., 2009). Furthermore, researchers at University of Illinois at Urbana-Champaign conducted multi-modal mood classification on 5,296 songs based on lyrics and found their model to significantly outperform audio spectral features across 7 mood categories (Xiao Hu et al., 2010). In summary, researchers conducting sentiment analysis, trend analysis, and mood classification have shown that lyrics in different eras are differentiable. But, to our knowledge, those lyrical attributes have not been leveraged for recommendations yet.

Therefore, we propose a new model that can be utilized for understanding information from lyrics. Our proposed model will contain tokenized Billboard Hot 100 song lyrics from 1960 to 2021 as inputs. These tokenized lyrics will be inputted into a pre-trained BERT model to fine-tune and ultimately classify the decade to which the song belongs to or was created in.

3. Methods

3.1 Introduction to the Data: Billboard Hot 100 Songs

We sought a comprehensive dataset covering popular western music from the past half century and, as in (Napier Shamir, 2018), found the Billboard Hot 100 Chart to be the best source of pop music tracks since 1958. Before the dominance of streaming (circa 2011), the songs on the Hot 100 were defined based on 3 core metrics: sales, jukebox plays, and radio plays. In the past decade, Billboard has added data from streaming services, video apps, and social media to make their chart more accurately reflect the top pop songs in a 21st century media music ecosystem. Based on its comprehensive approach and reputation as the best source on pop music, we chose to move ahead using the Billboard Hot 100 from each year since 1960 to define our dataset.

The yearly Hot 100s provided us roughly 6,200 pop songs from 1960 to 2021. However, their respective lyrics were still needed. Therefore, we utilized the Genius API, the world's biggest collection of song lyrics and musical knowledge. We conducted a series of data cleaning by filtering out instrumental songs with no lyrics, utilizing regular expressions to format the lyrics, and matched the lyrics to their respective songs.. Our final dataset contained 3,547 songs and was utilized to fine-tune our BERT-based models.

Our final dataset contained 3,547 songs with the following features: song title, artist, year, decade, and number of words. Before modeling, we further explored the data and found most songs contained around 300 words (Figure 1.1). Looking at the tails of this distribution, we can see the shortest song contained approximately 100 words and the longest song contained just over 1,000 words – still, this data has a long right tail and over 90 percent of our songs have fewer than 512 words (which is the cutoff for using BERT).

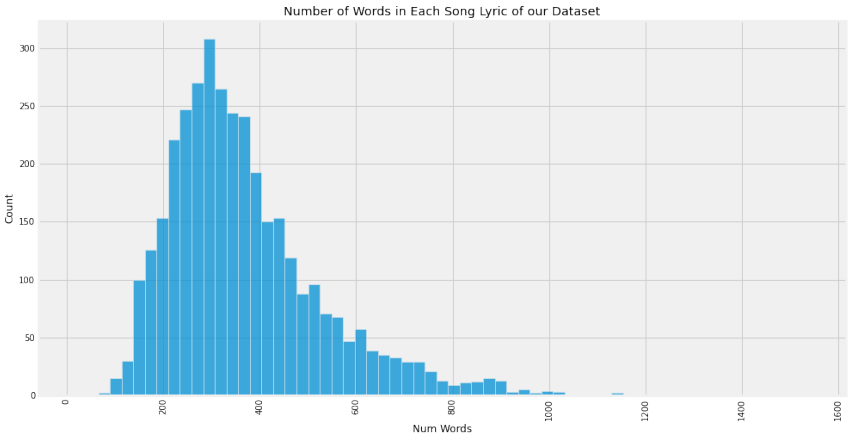
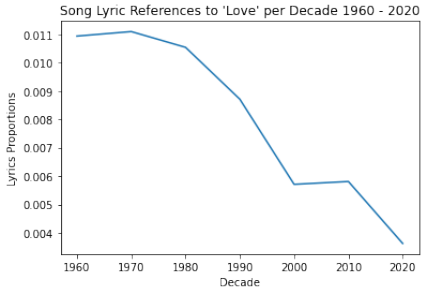
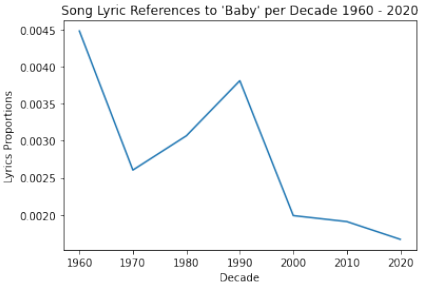


Figure 1. Figure 1.1. Distribution of Number of Words in Each Song Per Decade in Dataset (1960 - 2020).



(a) Frequency of the Word "Love"



(b) Frequency of the Word "Baby"

Figure 2. Figure 1.2. Frequency of the Popular Words Across Each Decade (1960 - 2020)

3.2 Baseline Models

Our baseline models had the following inputs: the decade was used as the y-values and a vectorized embedding of the tokenized lyrics served as our feature vector. We evaluated these baseline models (and all our models) based on accuracy as in Devlin et al. and Adhikari et al., which naturally fits multi-class classification (i.e. higher accuracy reflects predicting the correct class more often).

We pursued 2 different types of baseline models – a naive majority class predictor and non-transformer ML algorithms. We elected for this approach to find differences in information about our data across each type. The majority classifier served as a weak base model – if our models scored an accuracy at or below, we would know that our model is not learning from the data provided. Upon training our non-transformer ML algorithms, we found that they had improved performance over the majority classifier.

We chose to build this second class of baseline models for 3 main reasons. First, it would give us a more strict definition of success; quite simply, the 18 percent accuracy that our majority classifier provided is very low. We did not want to be satisfied with 25 percent accurate classifier that only works ¼ of the time. Second, Adhikari et. al (2019) suggest and use sklearn's Logistic Regression and Support Vector Machines as baselines for their document classification task which closely resembles our task (our classification task simply focuses on a specific type of document: song-lyrics). Third, it gave us an opportunity to build models with our lyrics data before diving into BERT.

3.3 BERT Model Configuration and Training

Before building our BERT model, we selected which lyrics to focus on. Base BERT (as presented in Devlin et al. (2018)) is limited to 512 tokens, which serves as a problem for documents. We considered following the work done by Adhikari et al. (2019) to allow BERT to intake full documents, but realized that with such a small percent of our songs having over 512 lyrics (Figure 1.1), this was adding unnecessary complexity. Moreover, Adhikari et al. notes that their top-performing model (KD LSTM reg) underperforms BERT large (and in most cases, Bert base). Luckily, song lyrics are relatively short documents that have a large number of repetitions. We chose to keep only the first 512 words of each song and expected very little non-repeated information to be lost.

We used 512 token BERT embeddings as inputs with categorical cross-entropy as our loss function, softmax as our activation function, and a learning rate of 0.00005, allowing BERT to find and learn a multiclass decision boundary. We fine-tuned BERT on roughly 2,000 of our pop songs from 1960 to 2021 and set our other 1,500 songs aside for validation testing.

We were very cautious about overfitting our BERT models, due to our large sparse input data and our results from our baseline models (which included some significant overfitting). We would have liked to use large batch sizes (eg. 128 as in Adhikari et al. and 32 as in Devlin et al.), but consistently ran up against computation limits in Google Colab Pro when using any batch size over 8. Thus, we settled on a batch size of 8 with 5 epochs and attempted to combat overfitting within the model architecture.

Our first approach (hereafter called "BERT Model1") included no hidden layer in order to discourage an overly-complex architecture. This model simply takes in the BERT-tokenized song lyrics as an input to BERT, then passes BERT's output into a Dense layer with 7 outputs (1 for each decade in our data) and then uses the softmax function to classify the most likely decade. We also attempted to reduce overfitting by introducing a higher dropout rate. Standard BERT (and our first model) has a 0.1 dropout rate. In this second model, (BERT Model 2), we used a dropout rate of 0.25. Otherwise, our model architecture matched our BERT Model 1.

4. Results and Discussion

The first 3 rows of Table 1.1 present the results from our baseline models. As expected, the Majority Classifier scores lower than our other models. Thus, our other models are able to learn the differences between lyrics from different decades. The Logistic Regression model is the best performing of the 3 baselines, with roughly 38 percent accuracy on test data. That's over 2 times better than the majority classifier, validating our desire to include more complex baselines. However, we noticed that our baseline models score lower than 40 percent accuracy on the validation and testing data and predicted the wrong class over 60 percent of the time. Therefore, in theory, a more powerful BERT-based language model will help drive better results on unseen validation testing data.

Table 1.1 Baseline and BERT Models Training, Validation, and Testing Accuracies

Model	Training Accuracy	Validation Accuracy	Testing Accuracy
Majority Classifier	0.190572	0.184210	0.187617
Logistic Regression	0.80458	0.3665	0.382739
SVM	0.471394	0.280075	0.288931
BERT Model 1	0.6853	0.4492	0.4353
BERT Model 2	0.8775	0.4605	0.40525

That theory is validated by our first BERT-based model (BERT Model 1). It has a higher validation testing accuracy than any of our baseline models while simultaneously maintaining a lower training accuracy than the Logistic Regression model (suggesting BERT Model 1 is much more generalizable).

On the other hand, the BERT Model 2's results had surprising results. Although it had the highest validation accuracy of any model, it considerably overfitted than BERT Model 1, even after expecting a higher dropout rate to encourage generalizability. The high validation accuracy that doesn't hold over to the test dataset, on which this model is nearly 6 percent less accurate.

These BERT-based models are definitely an improvement over the baseline non-language model approaches. However, they misclassify over 50 percent of the time. These results are a good reminder that this is a difficult task – for every song that seems to live specifically in 1 era lyrically, there's plenty others that could come from many decades. In addition, we've drawn arbitrary boundaries at the decade-level. These boundaries suggest that songs written in 1989 are more similar lyrically to songs from 1980 than those from 1990. This can potentially be untrue and may limit the performance of any classifier on this task.

Finally, we analyzed the outputs of our model to understand how it interacts with our data. Specifically, we were interested in answering: how does our model miss? When it incorrectly classifies a song to be from the 1990s, is it more likely to be from the 80s and 00s than from the 1960s? Figure 2.1 certainly suggests this. For example, of the 137 songs in our test set that BERT Model 2 classified as a 1990s song, only 54 were actually from the 90s. But 110 were from the 90s or the adjacent 80s and 00s. This pattern is encouraging – our model appears to notice differences in lyrics across time, but the issue of the arbitrary decade cutoff remains. Since our task is novel, we didn't find any other researchers with these patterns in their data.

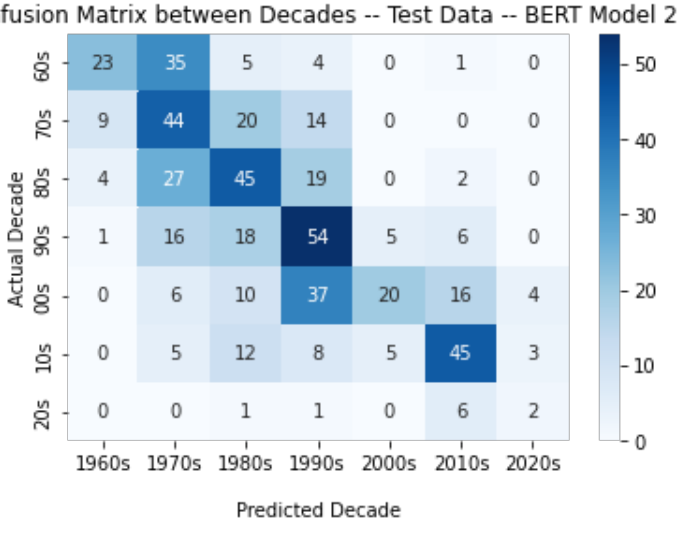


Figure 3. Figure 2.1. Confusion Matrix between Actual versus Predicted Decade for BERT 2 Model.

5. Conclusion

Both of our BERT-based models showed significant improvement over baseline naive models for classifying lyrics by decade. Our work suggests that lyrics from different eras are differentiable and are an overlooked aspect of music recommendations.

Still, there's plenty of room for further improvement and iteration. We'd begin by trying to come up with a cleaner and more complete dataset than the one we ended up using – whether simply including all 6,000+ Hot 100 pop songs since 1960 or expanding to include non-pop songs, more data could help make the model more generalizable. Likewise, we'd attempt to use larger GPUs in order to increase the batch size of our BERT models. Finally, though we chose to focus on classifying decades, which may have served as an arbitrary cutoff. We'd like to encourage further work better defining "eras" of songwriting and using that as the classes for the model.

Appendix 1. Bibliography

Adhikari, Ashutosh, et al. "DocBERT: BERT for Document Classification." ArXiv.org, 22 Aug. 2019, arxiv.org/abs/1904.08398.

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Huang, Kexin, et al. "Clinicalbert: Modeling Clinical Notes and Predicting Hospital Readmission." ArXiv.org, 29 Nov. 2020, https://arxiv.org/abs/1904.05342.

Hu, X., Stephen Downie, J. (2010). When lyrics outperform audio for music mood classification: A feature analysis. In Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010 (pp. 619-624). (Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010).

Lee, Dokyun Hosanagar, K.. (2014). Impact of recommender systems on sales volume and diversity.

Li, Zhuohan, et al. "Train Large, Then Compress: Rethinking Model Size for Efficient Training and Inference of Transformers." ArXiv.org, 23 June 2020, arxiv.org/abs/2002.11794.

Napier, Kathleen, and Lior Shamir. "Quantitative Sentiment Analysis of Lyrics in Popular Music." Journal of Popular Music Studies, vol. 30, no. 4, 2018, pp. 161–176., doi:10.1525/jpms.2018.300411.

Pettijohn, Terry F., and Donald F. Sacco. "The Language of Lyrics." Journal of Language and Social Psychology, vol. 28, no. 3, 2009, pp. 297–311., doi:10.1177/0261927x09335259.

Sun, Chi, et al. "How to Fine-Tune BERT for Text Classification?" Lecture Notes in Computer Science Chinese Computational Linguistics, 2019, pp. 194–206., doi:10.1007/978-3-030-32381-3_16.

Vaswani, Ashish, et al. "Attention Is All You Need." ArXiv.org, 6 Dec. 2017, arxiv.org/abs/1706.03762.

Yang, Zichao, et al. "Hierarchical Attention Networks for Document Classification." Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, doi:10.18653/v1/n16-1174.