# Language Representation Models for Music Genre Classification Using Lyrics


Hasan Akalp*
Hacettepe University, Ankara, Turkey
hasan.akalp@hacettepe.edu.tr

Enes Furkan Çiğdem*
Hacettepe University, Ankara, Turkey
enescigdem@hacettepe.edu.tr

Şeyma Yılmaz
Hacettepe University, Ankara, Turkey
seymayilmaz@hacettepe.edu.tr

Necva Bölücü
Hacettepe University, Ankara, Turkey
necva@cs.hacettepe.edu.tr

Burcu Can
Hacettepe University, Ankara, Turkey
burcucan@gmail.com



## ABSTRACT

There are various genres of music available in every period and field of human life. Every music genre represents a set of shared conventions. Today people have the opportunity to listen to any genre of music they want using various music platforms. However, with the increasing number of music genres, the management of these platforms becomes difficult. Language representation models such as BERT, DistilBERT have been proven to be useful in learning universal language representations. Such language representation models have achieved amazing results in many language understanding tasks. In this study, we apply language representation models for music genre classification using song lyrics. We examine whether language representation models are better than traditional deep learning models for music genre classification by comparing results and computation times. Experimental results show that BERT outperforms other models on one-label and multi-label classification with accuracy of 77.63% and 71.29% respectively. On the other hand, considering the time taken for one epoch, BERT runs 4 times faster than DistilBERT.



## CCS CONCEPTS

• **Computing methodologies**; • **Information extraction.**;

## KEYWORDS

Language Representation, Music Genre, Lyrics, BERT, DistilBERT, Classification

**ACM Reference Format:**
Hasan Akalp, Enes Furkan Çiğdem, Şeyma Yılmaz, Necva Bölücü, and Burcu Can. 2021. Language Representation Models for Music Genre Classification Using Lyrics. In *2021 International Symposium on Electrical, Electronics and Information Engineering (ISEEIE 2021), February 19–21, 2021, Seoul, Republic of Korea.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3459104.3459171


*Both authors contributed equally




*ISEEIE 2021, February 19–21, 2021, Seoul, Republic of Korea*





ACM ISBN 978-1-4503-8983-9/21/02…$15.00
https://doi.org/10.1145/3459104.3459171


## 1 INTRODUCTION

There are millions of songs on the music platforms which require organization usually based on their genres. Therefore, it becomes difficult for admins to organize songs on such music platforms. Machine learning techniques have been used for music genre classification using various types of data extracted from songs. While some studies used audio signals [7, 30], others used song lyrics for music genre classification considering the task as a text classification task. Since lyrics contain great information about the genre of the music, we try to tackle the music genre classification problem with lyrics in this study.

Text classification is a well-known problem in natural language processing (NLP). The aim of the task is to assign predefined categories to a given text. Previous studies used neural models for text categorization such as Convolution Neural Network (CNN) [10, 27], Recurrent Neural Networks (RNN) [15, 34] and attention mechanism [13, 32]. Recently, pre-trained language models have been used in learning language representations using a large amount of unlabeled data. There are two recent language representation models: BERT [5] and DistilBERT [24]. BERT is based on a multilayer bidirectional Transformer [31] and DistilBERT is a "distilled" version of BERT, which is comparably smaller and faster than BERT.

In this study, we investigate the usage of BERT and DistilBERT in music genre classification by comparing the results with a traditional deep neural network based on a BILSTM (bidirectional long short-term memory network) [26]. Additionally, we compare two types of models in terms of their complexities by analyzing their computation times.

Therefore, we analyze results in two folds: 1- accuracy, and 2- computation times. If we aim to employ the models in a real-time application, response and training times would be crucial. Considering both aspects (accuracy and computation time) in such models, the results show that BERT would be a reasonable solution in a real-time and real-world application.

The remainder of this paper is organized as follows: Section 2 overviews the related work on text categorization using language representation models, as well as the related work on music genre classification; section 3 describes the proposed BERT and DistilBERT models for music genre classification; section 4 presents the experimental results and finally, section 5 concludes the paper with the future goals.

## 2 RELATED WORK

In this study, we propose a model for music genre classification using lyrics. In the literature, various text classification studies are using deep recurrent neural networks. In this section, we will cover previous studies on music genre classification using machine learning algorithms and language representation models used for text classification.

As machine learning algorithms, Howard et al. [6] utilize the Naive Bayes algorithm for the music genre using a multilingual dataset. Ying et al. [33] use various classification algorithms such as Naïve Bayes, k-NN (k nearest neighbor), SVM (support vector machines) for the categorization of music genres and moods in a song. Oramas et al. [21, 22] propose a convolutional neural network (CNN) model employing features extracted from audio, images, and text of each song. Lima et al. [2] propose a BILSTM model to classify a set of Brazilian song lyrics. Analogously, Tsaptsinos use a hierarchical attention network (HAN) [29] for also lyric-based music genre classification. Other studies investigate the usage of audio and lyrics [16, 18, 20]; or rhymes and styles in lyrics [17] for music genre classification.

Extreme multi-label text classification is a challenge for labeling a text based on an extremely large label set, which is generally more than thousands. Chang et al. [3] utilize BERT for that purpose. Munikar et al. [19] and Li et al. [12] make use of BERT for sentiment classification. Sun et al. [28] conduct experiments by fine tuning BERT using different methods and perform experiments on the uncased BERT-base model for English text classification and the Chinese BERT-base model for Chinese text classification. Distillation is also applied to text classification. Chia et al. [4] apply distillation by using Open AI GPT [23] as a teacher and a BILSTM network, a shallow CNN network, a novel CNN structure are used as students. Adhikari [1] proposes a distillation of BERT-large to small LSTMs, thereby using 30x less number of parameters.

## 3 MUSIC GENRE CLASSIFICATION USING LANGUAGE REPRESENTATION MODELS

In this study, we aim to analyze different models on the classification of music genres using lyrics. For this purpose, we employ three different models: BILSTM [25], BERT, and DistilBERT.

### 3.1 BILSTM Model

BILSTM is a recurrent neural network used in text classification [9]. The model involves 4 layers: word embedding layer, BILSTM layer and two additional dense layers. The dimension of the input layer is determined by the number of unique tokens in the dataset. In the dense layers, *relu* and *softmax* are used as activation functions. In the output layer, the dimension of the output is determined by the number of music genre categories (i.e. *country*, *hip-hop*, *metal*, *pop*, *rock*, and *other*).

A softmax classifier and sigmoid function are employed on the top dense layer to predict the probability of the music genre $c$ or genre list $list_c$ for one-label classification and multi-label classification respectively as follows:

$$p(c|h_2) = soft\max(W_{lstm}h_2) \tag{1}$$

$$p(list_c|h_2) = sigmoid(W_{lstm}h_2) \tag{2}$$

where $W_{lstm}$ is the weight matrix and $h_2$ is the output of the second dense layer. Here, $c$ is a music genre (i.e. *pop*, *alternative*, *country*, *hip-hop*, *rock*, *R & B*) and $list_c$ is the list of music genres.

### 3.2 BERT Model

BERT, designed by Google AI Language, uses transformer to learn the contextual relationships between words in a text. A transformer contains two mechanisms; an encoder to read a text and a decoder to generate the predictions. Since the purpose of the model, only the encoder mechanism is required. The architecture of the proposed BERT model for music genre classification is given in Figure 1. BERT-base model contains an encoder with 12 transformer block, 13 self-attention heads, and a hidden size of 768. BERT receives an input of a sequence of no more than 512 tokens and outputs the representation of the sequence. The sequence has one or two segments where the first token of the sequence is always [CLS] which contains the special classification embedding and another special token [SEP] is used for separating segments.

For the classification task, we incorporate two dense layers after the final hidden state $h$ of the first token [CLS] to be used as the representation of the whole sentence. A softmax classifier and sigmoid function are used on top of the second dense layer to predict the probability of the music genre $c$ or genre list $list_c$ for one-label classification and multi-label classification respectively as follows:

$$p(c|b_2) = soft\max(W_{bert}b_2) \tag{3}$$

$$p(list_c|h_2) = sigmoid(W_{bert}h_2) \tag{4}$$

where $W_{bert}$ is the weight matrix, $b_2$ is the output of the second dense layer, $c$ is a music genre and $list_c$ is the list of music genres.

### 3.3 DistilBERT Model

The DistilBERT model, designed by HuggingFace [8], utilizes DistilBERT embeddings. The model uses knowledge distillation which is a compression technique where a small model is trained to reproduce the behavior of a larger model. The name of the model is "DistilBERT-base-uncased" which is distilled from the BERT model "bert-base-uncased". The model involves 6 layers, 768 dimensions, and 12 heads with a total number of 66 million parameters. The architecture of DistilBERT for music genre classification is given in Figure 2

A softmax classifier and sigmoid function are used on top of the final hidden state $d$ to predict the probability of music genre $c$ or genre list $list_c$ for one-label classification and multi-label classification respectively as follows:

$$p(c|d) = soft\max(W_{db}d) \tag{5}$$

$$p(list_c|d) = sigmoid(W_{db}d) \tag{6}$$

where $W_{db}$ is the weight matrix, $d$ is the output of the final hidden state, $c$ is a music genre and $list_c$ is the list of music genres.

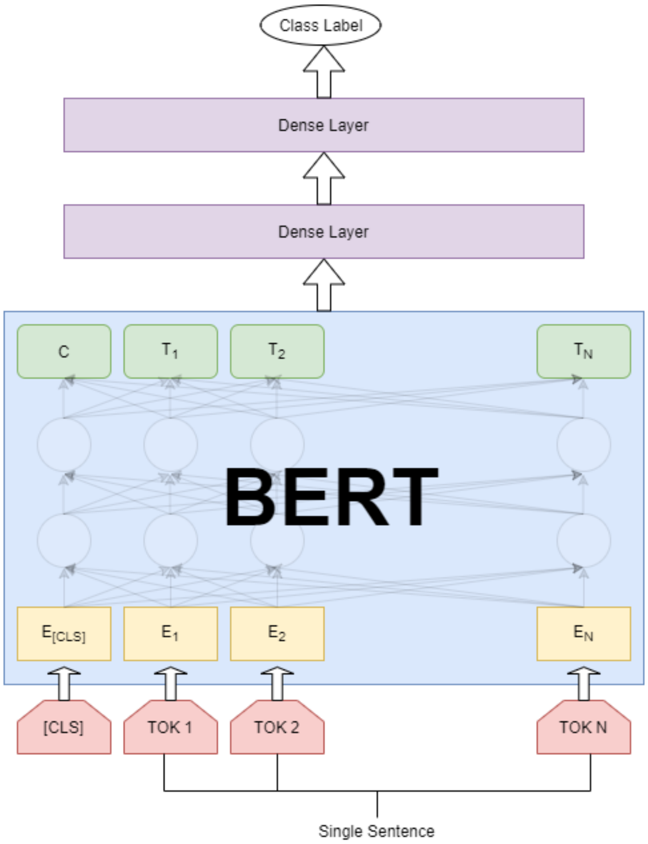



**Figure 1: The architecture of the BERT model.**

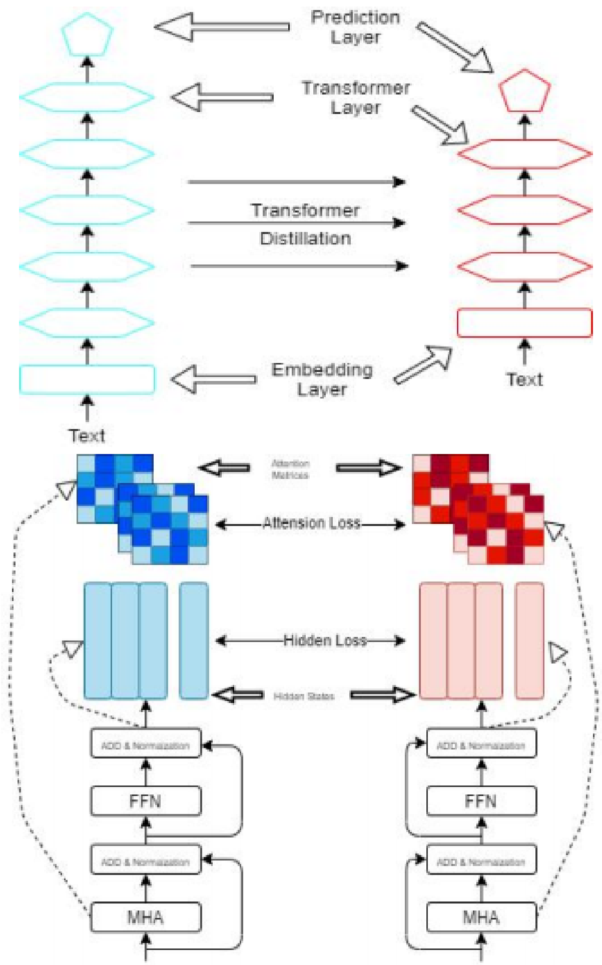



**Figure 2: The overview of the DistilBERT model.**

# 4 EXPERIMENTS & RESULTS

## 4.1 Dataset

In this study, we used the combination of two datasets along with multi-labels of music genres: *250,000+ lyrics over 2K singers*[1] *and 380,000+ lyrics from MetroLyrics*[2]

The distribution of the categories are given for both datasets and the combined dataset in table 1. Since the datasets are unbalanced, we merged some categories. After merging the datasets, the total number of classes was reduced to 13. The final categories are metal, pop, punk, alternative, blues, country, indie, jazz, hip-hop, Electronic, Folk, Rock, R&B, and Other. Moreover, the size of the data set is reduced further by deleting the same songs and the songs that do not have assigned genres. Finally, a total number of 6 genres are obtained, that are alternative, country, pop, R&B, rock, and hip-hop. The distribution of the final categories is given in table 2.

## 4.2 Hyper-parameters

Table 3 summarizes the hyperparameters used in each model. In addition to all these parameters, we tried various threshold values to convert the output probabilities into a binary format during the inference of the multi-label classification. The highest results are obtained with a threshold value of $t = 0.4525$ for the multi-label classification.

## 4.3 Experimental Results

We conducted the experiments for both one-label classification and multi-label classification with the thought that a song may have one or more genres. The accuracy results obtained from BILSTM, BERT and DistilBert models are given in Table 4. The results show that for each task BERT achieves the highest accuracy score among three approaches for music genre classification.

The complex structure of the BERT model enables the hidden features to be learned well. For this reason, higher accuracy was achieved using the BERT model compared to other two models. On the other hand, the DistilBERT model is a simplified version of the BERT model by sacrificing some of the BERT model's features. Therefore, DistilBert reduces the complexity of the BERT by also reducing the computation time. For this reason, the accuracy achieved by DistilBert is relatively lower than the BERT. Since the BERT model and the DistilBERT model are more advanced models

[1]https://www.kaggle.com/detkov/lyrics-datasetsongs_dataset.csv
[2]https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics.

**Table 1: # of instances in each genre.**

| Genre | 380k | 250k | Combined |
|---|---|---|---|
| Country | *17286* | *16133* | *33419* |
| Electronic | *16205* | *0* | *16205* |
| Folk | *3241* | *3054* | *6295* |
| Hip-Hop | *33965* | *64713* | *98678* |
| Indie | *5732* | *14699* | *20431* |
| Jazz | *17145* | *0* | *17145* |
| Metal | *28408* | *4185* | *32593* |
| Other | *23683* | *0* | *23683* |
| Pop | *49444* | *123079* | *172523* |
| Punk | *0* | *5123* | *5123* |
| R&B | *5935* | *17268* | *23203* |
| Rock | *131377* | *66214* | *197591* |
| Blues | *0* | *3540* | *3540* |
| Not Available | *29814* | *0* | *29814* |

**Table 2: # of instances in each genre in the final categories after preprocessing.**

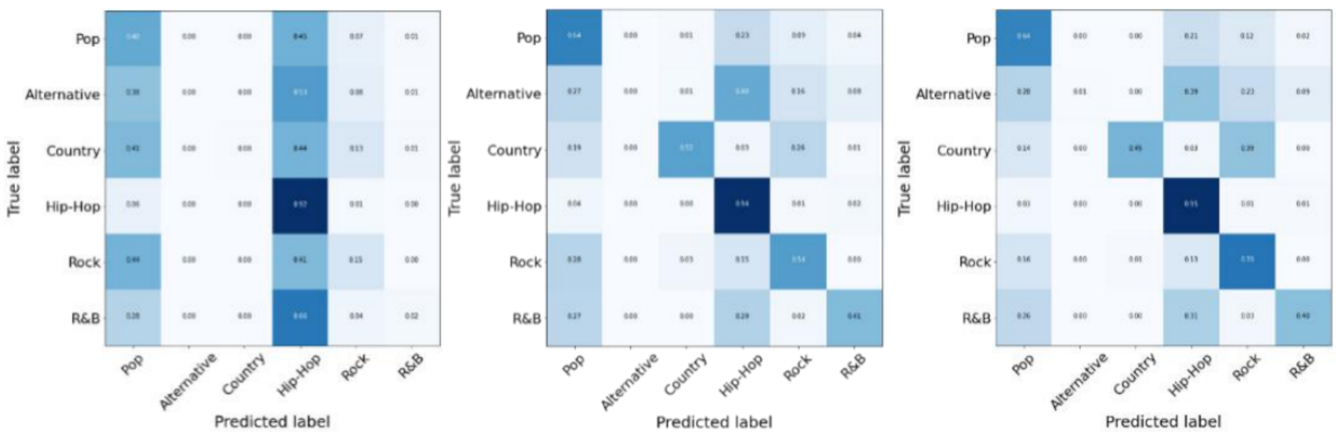
| Genre | One-Label | Multi-Label |
|---|---|---|
| Pop | *15698* | *17071* |
| Alternative | *779* | *1797* |
| Country | *975* | *1232* |
| Hip-Hop | *45982* | *50696* |
| Rock | *7678* | *7833* |
| R&B | *3728* | *6035* |

**Table 3: Hyperparameters used in the study.**

| **BiLSTM** | **BERT** | **DistilBERT** |
|---|---|---|
| tokenization of words: 400 | vocab size: 400 | vocab size: 400 |
| batch size: 128 | hidden size: 768 | max position embeddings: 512 |
| embedding dim: 64 | num hidden layers: 12 | n layers: 6 |
| validation split: 0.2 | num attention heads: 12 | n heads: 12 |
| epoch: 8 | intermediate size: 3072 | dim: 768 |
| Optimizer: Adam [11] | hidden dropout prob: 0.1 | hidden dim: 4 * 768 |
| | attention probs: 0.1 | dropout: 0.1 |
| | dropout prob: 0.1 | attention dropout: 0.1 |
| | max position embeddings: 512 | initializer range: 0.02 |
| | type vocab size: 16 | qa dropout: 0.1 |
| | initializer range: 0.02 | seq classif dropout: 0.2 |
| | epoch (one-label): 1 | epoch (one-label):9 |
| | epoch (multi-label): 1 | epoch (multi-label): 15 |
| | Optimizer: Adam [11] | Optimizer: AdamWarmup [14] |

with a better capability of handling sequences compared to the BiLSTM model, the accuracy achieved is higher than the accuracy of the BiLSTM.

For each genre, one-label and multi-label results are given in Table 6 and 7 respectively. As seen in Table 6, hip-hop songs are predicted accurately using all models. On the contrary, successful predictions have not been made for alternative songs with BILSTM and BERT models. This case has been slightly improved with the DistilBERT model. As can be seen in Table 7, no songs other than hip-hop genres were successfully predicted with the BILSTM model. This deficiency was improved with BERT and DistilBERT model. This situation proves that both BERT and DistilBERT models are better than the BILSTM model.




(a) BiLSTM (b) BERT (c) DistilBERT

**Figure 3: Confusion Matrices for the models (a) BiLSTM (b) BERT (c) DistilBERT.**

**Table 4: Experimental results of the models for one-label and multi-label classification.**

| Model | One-label Accuracy | Multi-label Accuracy |
|---|---|---|
| BiLSTM | 68.11 | 64.41 |
| BERT | 77.63 | 71.29 |
| DistilBERT | 74.38 | 70.48 |

**Table 5: Running times of the models for one-label and multi-label classification.**

| Model | One-label | Multi-label |
|---|---|---|
| BiLSTM | 7m | 9m |
| BERT | 94m | 147m |
| DistilBERT | 205m | 411m |

Confusion matrices of all models are given in Figure 3 for one-label classification. Because of the imbalance dataset, all models tend to tag the genres as *hip-hop*. However, BERT and DistilBERT models are comparably more robust for this problem. As seen in Figure 3, the BILSTM model cannot learn well and it is affected negatively by the imbalanced data problem. Hence, the LSTM model has a bias to predict lyrics either as pop or hip-hop. BERT and DistilBERT models give better results compared to the BILSTM model. However, the imbalanced data problem also affects these models. For instance, as seen in the Figure, 48% of *alternative* songs taken from confusion matrices in Figure 3 are estimated to be *hip-hop* in the BERT model. Moreover, 39% of *alternative* songs were estimated as *hip-hop* in the DistilBERT model. Although there is a high margin of error, the type of error has been slightly improved in the DistilBERT model compared to the BERT model.

Computation times are given in Table 5 for one-label and multi-label classification for three models. Although the LSTM model is more convenient in terms of time and calculation when performing multi-label classification, more accurate results are obtained with the transformers models. While there is not much difference between BERT and DistilBERT in terms of accuracy results, BERT is much faster than DistilBert in terms of running times.

### 4.4 Error Analysis

The imbalanced dataset is the major drawback in this study. Although the combined and preprocessed dataset was used to mitigate the imbalance problem, it could not be resolved completely. For this reason, the correct classification rates of some classes are very low. Besides, precision, recall and F1-score values for punk, alternative, blues, alternative, jazz, and electronic genres are 0 in one-label classification. In multi-label classification, precision, recall and F1-score values are also measured as 0 for blues, alternative, jazz, electronica, folk and other classes.

To resolve this problem, the classes with few examples are gathered in the *other* class. However, model accuracy has not improved significantly. For this reason, similar classes are combined and the *other* class is completely removed. Finally, a total number of 6 genres (alternative, country, pop, R&B, rock, and hip-hop) are used in all experiments. All results are obtained as a result of those preprocessing tasks.

## 5 CONCLUSION

In this work, we compare language representation models (BERT and DistilBERT) with traditional recurrent neural networks for music genre classification for one-label and multi-label classification problem. In addition to resolving the imbalanced dataset problem, we combined two different music datasets for the task. After merging the datasets, the imbalance problem is partially solved by combining similar classes with low density and by deleting some classes. Results show that despite being advantageous in terms of time and calculation, the desired accuracy is not achieved in the BILSTM model in contrast with BERT and DistilBERT models. All results are compared and it is seen that the best results are obtained

**Table 6: One-label classification results of the models.**

| | BiLSTM | | | BERT | | | DistilBERT | | |
|---|---|---|---|---|---|---|---|---|---|
| Genre | Precision | Recall | F1-Score | Precision | Recall | F1-Score | Precision | Recall | F1-Score |
| Pop | .48 | .48 | .48 | .63 | .61 | .62 | .73 | .64 | .68 |
| Alternative | .00 | .00 | .00 | .00 | .00 | .00 | 1.00 | .01 | .01 |
| Country | .12 | .00 | .01 | .48 | .47 | .47 | .73 | .45 | .56 |
| Hip-Hop | .76 | .92 | .83 | .86 | .93 | .89 | .88 | .95 | .92 |
| Rock | .36 | .15 | .21 | .61 | .51 | .56 | .64 | .70 | .67 |
| R&B | .24 | .02 | .03 | .49 | .39 | .44 | .62 | .40 | .49 |

**Table 7: Multi-label classification results of the models.**

| | BiLSTM | | | BERT | | | DistilBERT | | |
|---|---|---|---|---|---|---|---|---|---|
| Genre | Precision | Recall | F1-Score | Precision | Recall | F1-Score | Precision | Recall | F1-Score |
| Pop | .00 | .00 | .00 | .65 | .66 | .66 | .68 | .61 | .64 |
| Alternative | .00 | .00 | .00 | .62 | .04 | .08 | .58 | .03 | .06 |
| Country | .00 | .00 | .00 | .72 | .40 | .51 | .76 | .31 | .44 |
| Hip-Hop | .64 | 1.00 | .78 | .90 | .93 | .92 | .89 | .93 | .91 |
| Rock | .00 | .00 | .00 | .61 | .49 | .54 | .61 | .57 | .58 |
| R&B | .00 | .00 | .00 | .62 | .49 | .55 | .62 | .45 | .52 |

with the BERT model, with an accuracy of **77.63%** in one-label classification and **71.29%** accuracy in multi-label classification.

As future work, we aim to work with a new more balanced dataset. Moreover, we want to lean over to the parameter optimization problem of models.

## REFERENCES


[1] Adhikari, Ashutosh, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. "Docbert: Bert for Document Classification." *arXiv Preprint arXiv:1904.08398*.

[2] Araújo Lima, Raul de, Rômulo César Costa de Sousa, Simone Diniz Junqueira Barbosa, and Hélio Cortês Vieira Lopes. 2020. "Brazilian Lyrics-Based Music Genre Classification Using a Blstm Network." http://arxiv.org/abs/2003.05377.

[3] Chang, Wei-Cheng, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit Dhillon. 2019. "X-Bert: EXtreme Multi-Label Text Classification with Using Bidirectional Encoder Representations from Transformers." *arXiv Preprint arXiv:1905.02331*.

[4] Chia, Yew Ken, Sam Witteveen, and Martin Andrews. 2019. "Transformer to Cnn: Label-Scarce Distillation for Efficient Text Classification." http://arxiv.org/abs/1909.03508.

[5] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." http://arxiv.org/abs/1810.04805.

[6] Howard, Sam, Carlos N. Silla Jr, and Colin G. Johnson. 2011. "Automatic Lyrics-Based Music Genre Classification in a Multilingual Setting." In *Thirteenth Brazilian Symposium on Computer Music*. https://kar.kent.ac.uk/33266/.

[7] Huang, Derek A, Arianna A Serafini, and Eli J Pugh. n.d. "Music Genre Classification."

[8] "Hugging Face – on a Mission to Solve Nlp, One Commit at a Time." n.d. *Hugging Face – on a Mission to Solve NLP, One Commit at a Time.*https://huggingface.co/.

[9] Johnson, Rie, and Tong Zhang. 2016. "Supervised and Semi-Supervised Text Categorization Using Lstm for Region Embeddings." http://arxiv.org/abs/1602.02373.

[10] Johnson, Rie, and Tong Zhang. 2017. "Deep Pyramid Convolutional Neural Networks for Text Categorization." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 562–70.

[11] Diederik P Kingma and Jimmy Ba. 2014. Adam:Amethodforstochasticoptimization. arXiv preprint arXiv:1412.6980 (2014).

[12] Li, Wenting, Shangbing Gao, Hong Zhou, Zihe Huang, Kewen Zhang, and Wei Li. 2019. "The Automatic Text Classification Method Based on Bert and Feature Union." In*2019 Ieee 25th International Conference on Parallel and Distributed Systems (Icpads)*, 774–77. IEEE.

[13] Lin, Zhouhan, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. "A Structured Self-Attentive Sentence Embedding." *arXiv Preprint arXiv:1703.03130*.

[14] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019. On the variance of the adaptive learning rate and beyond. arXiv preprint arXiv:1908.03265 (2019).

[15] Liu, Pengfei, Xipeng Qiu, and Xuanjing Huang. 2016. "Recurrent Neural Network for Text Classification with Multi-Task Learning." *arXiv Preprint arXiv:1605.05101*.

[16] Mayer, Rudolf, Robert Neumayer, and Andreas Rauber. 2008a. "Combination of Audio and Lyrics Features for Genre Classification in Digital Audio Collections." In *Proceedings of the 16th Acm International Conference on Multimedia*, 159–68.

[17] Mayer, Rudolf, Robert Neumayer, and Andreas Rauber. 2008b. "Rhyme and Style Features for Musical Genre Classification by Song Lyrics." In *Ismir*, 337–42.

[18] Mayer, Rudolf, and Andreas Rauber. 2011. "Musical Genre Classification by Ensembles of Audio and Lyrics Features." In *Proceedings of International Conference on Music Information Retrieval*, 675–80.

[19] Munikar, Manish, Sushil Shakya, and Aakash Shrestha. 2019. "Fine-Grained Sentiment Classification Using Bert." In*2019 Artificial Intelligence for Transforming Business and Society (Aitb)*, 1:1–5. IEEE.

[20] Neumayer, Robert, and Andreas Rauber. 2007. "Integration of Text and Audio Features for Genre Classification in Music Information Retrieval." In *European Conference on Information Retrieval*, 724–27. Springer.

[21] Oramas, Sergio, Francesco Barbieri, Oriol Nieto, and Xavier Serra. 2018. "Multimodal Deep Learning for Music Genre Classification." *Transactions of the International Society for Music Information Retrieval. 2018; 1 (1): 4-21*.

[22] Oramas, Sergio, Oriol Nieto, Francesco Barbieri, and Xavier Serra. 2017. "Multi-Label Music Genre Classification from Audio, Text, and Images Using Deep Features." http://arxiv.org/abs/1707.04916.

[23] Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. "Improving Language Understanding with Unsupervised Learning." *Technical Report, OpenAI*.

[24] Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. "DistilBERT, a Distilled Version of Bert: Smaller, Faster, Cheaper and Lighter." http://arxiv.org/abs/1910.01108.

[25] Schuster, Mike, and Kuldip Paliwal. 1997a. "Bidirectional Recurrent Neural Networks." *Signal Processing, IEEE Transactions on* 45 (December): 2673–81. https://doi.org/10.1109/78.650093.

[26] Schuster, Mike, and Kuldip K Paliwal. 1997b. "Bidirectional Recurrent Neural Networks." *IEEE Transactions on Signal Processing* 45 (11): 2673–81.

[27] Shen, Dinghan, Yizhe Zhang, Ricardo Henao, Qinliang Su, and Lawrence Carin. 2018. "Deconvolutional Latent-Variable Model for Text Sequence Matching." In *Thirty-Second Aaai Conference on Artificial Intelligence*.

[28] Sun, Chi, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. "How to Fine-Tune Bert for Text Classification?" In *China National Conference on Chinese Computational Linguistics*, 194–206. Springer.

[29] Tsaptsinos, Alexandros. 2017. "Music Genre Classification by Lyrics Using a Hierarchical Attention Network." In. ICME.
[30] Tzanetakis, George, and Perry Cook. 2002. "Musical Genre Classification of Audio Signals." *IEEE Transactions on Speech and Audio Processing* 10 (5): 293–302.
[31] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*, 5998–6008.
[32] Yang, Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. "Hierarchical Attention Networks for Document Classification." In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–9.
[33] Ying, Teh Chao, Shyamala Doraisamy, and Lili Nurliyana Abdullah. 2012. "Genre and Mood Classification Using Lyric Features." In*2012 International Conference on Information Retrieval & Knowledge Management*, 260–63. IEEE.
[34] Yogatama, Dani, Chris Dyer, Wang Ling, and Phil Blunsom. 2017. "Generative and Discriminative Text Classification with Recurrent Neural Networks." *arXiv Preprint arXiv:1703.0189*