Classification of Music Lyrics by Genre

Shiraz Chakraverty
schakraverty@berkeley.edu

Dan Price
dan.price@berkeley.edu

https://github.com/danonprice/LyricsGenreClassification

Abstract

In the field of Natural Language Processing (NLP), the machine classification of text into human taxonomies is a formidable task. Machine understanding of language in general is complex. We cannot yet fully understand how the human brain does it but strides have been made in NLP.

In the arts, the classification of music into genres inherits some of its nuance from the human cultural context the music originates. Yet, even casual listeners can often identify the assigned genre of a piece of music from written lyrics alone for many popular genres. Engendering machines with a similar ability presents with its challenges. With that understanding, newer and emerging algorithms and word embeddings are able to offer classifications useful for organizations doing language tasks at a large scale.

We explore the advantages and disadvantages of various methods in text classification papers we have selected and report our observations, learnings, and a recommendation. With the focus of deep learning, it is clear that neural networks tend to work better than previously used models. We would like to explore this path and provide comparative analysis. We would also like to offer commentary on the larger issue of bias in the setting of music genre classification.

1 Credits

We are grateful to Montero Lamar "Lil Nas X" Hill, Kiowa "YoungKio" Roukema, Atticus Ross, Trent Reznor, and Billy Ray Cyrus for stretching the bounds of social consensus and challenging bias.

2 Introduction

Rolling Stone magazine tells the story of a viral Internet hit song "Old Town Road" by a heretofore-unknown artist Lil Nas X. "Old Town Road" emerged on social video sharing app TikTok. It also sparked success on music sharing service SoundCloud. Partly motivated by promotion on social network Instagram from pop sensation Justin Bieber, the song simultaneously debuted on Billboard's Hot 100 Chart, Hot Country Songs chart, and the Hot R&B/Hip-Hop Songs chart. However, after some time and presumably deep reflection on the cultural implications, Billboard elected to remove the song from its Hot Country Songs chart claiming the song "does not merit inclusion on Billboard's country charts" because "it does not embrace enough elements of today's country music to chart in its current version" (Leight, 2019). Controversy over genre labeling ensued. Nevertheless, the novelty and notoriety of the hit single ultimately led to famed country artist Billy Ray Cyrus appearing on an "Old Town Road" remix thereby cementing the song's relationship with the country music genre - ostensibly in spite of the Billboard decision.

In its editorial, Rolling Stone signaled that antiquated definitions of music genre highly correlated to race are to blame for Billboard's classification controversy (Leight, 2019). In an age where codification of bias in artificial intelligence systems is under inspection, the conversation surrounding "Old Town Road" can be considered an important litmus test for the role of genre classification in music. As artificial intelligence weaves its way into modern life, it is not unreasonable to consider the contribution machines play in the classification of styles of music. Machines already play a significant role in the recommendation of music. Genre is an important feature in music recommendation systems.

Text classification is a broadly applied and heavily researched task in the field of natural language processing. Filtering of spam email and sentiment analysis are canonical, highly utilized examples of text classification in application. We seek to evaluate some of NLP's gains on music lyrics. In this paper, we aim to determine the applicability of machine learning and deep learning algorithms on the task of classifying music lyrics by genre.

Song lyrics exhibit structure distinct from other text documents. Lyrics are often organized in discrete sections such as intros, choruses, verses, bridges, and outros. Lyrics often align with a defined musical tempo with regularly occurring patterns. Various academic and industry teams have tried approaches to this space. Efforts to understand music, both sonically and semantically through sound, lyrics, and metadata have coalesced in a subfield known as Music Information Retrieval (MIR).

No single effort has been very successful in finding a stable method that performs significantly well to tackle the lyrical genre classification problem. Algorithms such as Support Vector Machines, k-Nearest Neighbors, and Naive Bayes have all been used in lyrical classification but they all have very low accuracy in comparison to other NLP tasks on other datasets. We explore the application of new and emerging algorithms and models to the lyrics genre classification task. We can then, perhaps, get a machine's take on how "Old Town Road" should be classified.

3 Dataset

Data for the lyrical classification problem is hard to come by due to copyright and other original content protection requirements. Artists and music labels do not usually publish lyrics with audio. With the rise of digital music and streaming, websites have emerged that build the infrastructure to publish crowd-sourced lyrics for ad revenue. Some services even distribute lyrics through application programming interfaces for profit to create data feeds for other consuming services. Despite the public sourcing of the lyrics, public access to the entire lyric datasets remains limited. Fortunately, a Kaggle user published a dataset of over 300,000 lyrics from a crawl of lyric website Metrolyrics.com (Mishra, 2016).

The initial 98-megabyte dataset is downloadable as a comma delimited file. The columns in the file include an index, song title, release year, artist, genre, and lyrics. We are interested in the text in the lyrics column for features and the genre column for labels. Ensemble methods that include analysis of the other columns and additional song metadata are a reasonable path to pursue but we consider it outside the scope of our current work. The lyrics are a string with carriage returns denoting an end of line.

Statistics specific to the overall structure of the lyrics, such as number of words, add value to the classification task. The dataset includes lyrics of varying word counts. The

word counts translate to the length of the sequences that we feed our models. We opted to only sample lyrics with at least 100 words.

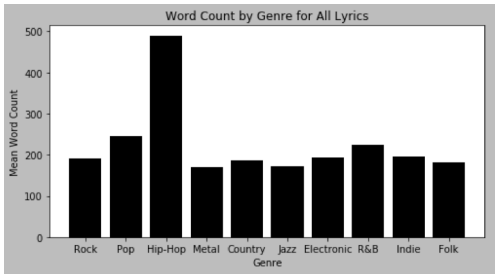

Figure 1: Word Count by Genre

The average word count for the Hip-Hop genre in comparison to the other genres may deserve some attention but we consider it outside the scope of this work.

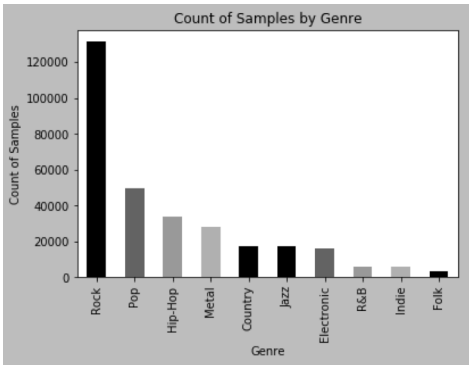

Figure 2: Number of Songs in Each Genre

The original dataset includes disproportionate representation of genres. The Rock genre comprises nearly one-third of the dataset with over 100,000 records. The least populated genre is folk with just over 2000 records. Due to the skewed distribution of classes, we sampled 1000 records from seven genres. The genres include Jazz, Metal, Electronic, R&B, Indie, Rock, Pop, Hip-Hop, Folk, and Country. We utilized scikit-learn's "train_test_split" method to create an 80/20 split of the train and test data.

4 Experiments & Results

Much is known and researched on the classification task in NLP. However, not much work has been published specifically on classification of a large corpus of music lyrics. In an effort to judge a minimum desired performance for more complex neural models, we applied our dataset to Bernoulli Naïve Bayes, Decision Tree, and Random Forest classifiers. We also constructed a simple Multi-layer Perceptron using skikit-learn's neural network library. We use term frequency - inverse document frequency (TF-IDF) as feature vectors and the genre classes as labels. We have also normalized the vector after applying the TF-IDF weighing scheme. None of these models achieved accuracy greater than 50%.

While linear and kernel models rely on good hand selected features, deep learning architectures attempt to prevent this by letting the model learn important features on its own. As was the case with non-neural methods, not much research has looked into the performance of these deep learning methods with respect to the genre classification task on lyrics. Here, we attempt to understand this by extending deep learning ideas on text classification to the particular case of lyrics.

An important idea in NLP is the use of dense vectors to represent words. To learn these word vectors a variety of methods have been proposed. A successful methodology proposes that similar words have similar context and thus that these vectors should be learned through context gained through training on large corpora. Examples of word embeddings explored include Word2vec, GloVe, and BERT.

Word2vec is a group of related models that are used to produce word embeddings from local context windows. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space.

Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space. Word2Vec's semantic style vectors preserve relevant information in text while having relatively low dimensionality (Mikolov, 2013).

The GloVe (Global Vectors for Word Representation) method combines the local context window used by Word2vec with global matrix factorization methods. Focus on local context alone omits the statistics of the corpus. GloVe includes corpus statistics by using a weighted least squares model trained on global word-word co-occurrence counts in its word vector space representation (Pennington, 2014).

Due to the expressive and poetic nature of music lyrics, we also applied BERT (Bidirectional Encoder Representations from Transformers) to the task of lyrics genre classification (Devlin, 2018). BERT's innovation to NLP is centered around the idea that its vector representations are conditioned on the left and the right context of the corpus. Prior representations represent only a unidirectional context. BERT also adds a masked language model whereby tokens are randomly masked with an object to predict masked words (Devlin, 2018).

We have applied the embeddings in the neural architecture known as the Convolutional Neural Network (CNN). CNNs have revolutionized image classification and computer vision by being able to extract features from images and using them in neural networks. The properties that made them useful in image processing makes them also handy for sequence processing. Convolving filters are applied to the local features of text. Our CNNs demonstrated improvement over our baseline models.

Algorithm	Embedding	Training Time (s)	Accuracy
Naïve Bayes	Word2vec	4	0.44
Decision Tree	Word2vec	72	0.40
Multi Layer Perceptron	Word2vec	442	0.45
Random Forest	Word2vec	97	0.48
Simple CNN	Word2vec	360	0.55
BERT Classifier	BERT	1800	0.47
Keras 10 Layer NN	Keras	1800	0.66
Keras 10 Layer NN	Pretrained GloVe	1500	0.72
CNN 10 Layer NN	Pretrained GloVe	1200	0.71

Table 1: Models applied to classification

5 Discussion

We have gone from a bag-of-words model with logistic regression to increasingly more advanced methods leading to convolutional neural networks. We found the use of pre-trained word embeddings improves accuracy on the classification task. We have also learned how to work with neural networks and how to use hyperparameter optimization to squeeze more performance out of the model. We implemented Grid Search but ran out of compute time attempting to train the model.

The accuracy improvement of deep learning models over baseline machine learning models is a testament to the power of backpropagation in its determination of important features of a dataset. However, given that notable NLP tasks are rewarded for accuracies greater than 90%, we cannot help but think that genre classification is determined by factors beyond the lyric text. An analysis of the audio and metadata in combination with lyric text may shed more

light on the classification problem. Methods for quantifying and analyzing the social context present at the time of music genre and style emergence may also improve accuracy on the classification task.

One big topic that we have not covered here left for another time is the use of recurrent neural networks, more specifically Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures. Those are other powerful and popular tools to work with sequential data like text or time series. Other interesting developments are currently in neural networks that employ Attention which are under active research and seem to be a promising next step since LSTM tend to be heavy on the computation.

Acknowledgments

We salute the labor of Kaggle user Gyanendra Mishra in assembling the dataset.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805

Elias Leight. 2019. *Lil Nas X's 'Old Town Road' Was a Country Hit. Then Country Changed Its Mind*. Rolling Stone. https://www.rollingstone.com/music/music-features/lil-nas-x-old-town-road-810844/

Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. 2013. *Efficient Estimation of Word Representations in Vector Space*. https://arxiv.org/pdf/1301.3781.pdf

Gyanendra Mishra. 2016. *380,000+ lyrics from MetroLyrics*. https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. *GloVe: Global Vectors for Word Representation.* https://nlp.stanford.edu/pubs/glove.pdf