Can You Hear Me Now?

Predicting song genre from song lyrics using deep learning

by

Nick DeMasi

An honors thesis submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Science

Undergraduate College

Leonard N. Stern School of Business

New York University

May 2019

Professor Marti G. Subrahmanyam	Professor Panos Ipeirotis
Faculty Adviser	Thesis Adviser

Abstract

This paper seeks to evaluate the performance of traditional machine learning classification algorithms (Naïve Bayes, Random Forest, and multiclass logistic regression) against that of deep learning algorithms (a simple single layer ANN) with respect to the task of automatic multiclass genre classification. Feature sets were derived from lyric data sourced from the musiXmatch.com dataset (one of the ancillary datasets attached to the well-known Million Song Dataset) and consisted of various statistical representations of said data such as aggregated metrics, tf-idf calculations, and word vectors. The best performing model proved to be an ANN trained on tf-idf features which achieved accuracy of approximately 41.6%. We note however that this accuracy was indeed not as high as certain blended models (models which were train using both audio and lyric features) already found in the literature. As a result, we recommend that companies wishing to implement this technology (whether it be for the purpose of advertising as outlined here, or something else entirely) employ a blended model and that researchers interested the topic of automatic genre classification explore the possibility of applying deep learning to a blended feature set.

1. Introduction

A quick Google search for “music genre classification” brings up a near endless stream of results linking to various research papers, informative blog posts, and GitHub repos addressing the topic. Genre classification belongs to the inter-disciplinary field of Music Information Retrieval (MIR) which focuses on analyzing music data (such as a song’s lyrics, pitch, tempo, etc.) for the purpose of extracting useful and relevant information. Typical use cases for MIR research comes in the form of music recommendation systems (whereby an algorithm is used to predict what songs a listener might be interested in hearing next based on their listening history), audio recognition software (such as Shazam which is able to identify a song based on a brief audio profile), music generation (where a machine learns how to create its own music by examining and learning from pre-existing pieces), and many others.

Automatic genre classification is of interest to researchers because of the ambiguity surrounding these genre labels. Music genre is, after all, a human-made construct of categorization, which is to say that it is not an empirically observable fact that a song belongs to a certain genre. Rather these genres are typically ascribed to a song by human listeners who upon hearing it, recognize certain traits or features that are characteristic of a pre-existing body (aka genre) of music. However, because this is a subjective process, listeners can disagree on which songs they feel belong to which genres, throwing the whole purpose of the exercise into flux, for what good is a taxonomy if it changes depending on who you ask? Typically, this problem is solved by either looking to the creator/distributor of the music, who will typically have a view of what genre it belongs to, or polling a broad group of listeners about their opinions and moving forward with the position taken by the majority. This polling can happen informally (such as

when a rock radio station takes calls from listeners about what songs it should play) or formally (such as on sites like Last.fm where users are able to tag songs with genres). Regardless, there is clearly room for a more definitive and objective classification process for music genre, which is precisely why the field of automatic genre classification has such a large community interested in the problem.

Researchers interested in automatic genre classification have typically used either a song's audio features (i.e, pitch, tempo, timbre, etc.), lyrics, metadata, or a combination of the three to train their models and make their predictions. The focus of this paper is to use lyrics alone as the basis for genre predictions. The driving motivation behind this particular decision has been outlined previously by other papers and includes reasons such as:

- A song's content may be linked to its genre in a way that is unique from audio. Think of the phrase 'sex, drugs, and rock and roll", the first two features of the list are common song topics within the genre but don't creep up quite as often (or in the same manner) in pop songs. (Mayer & Rauber, 2010)
- Lyrical data is oftentimes easier to process than audio data. Furthermore, there still exists a fair amount of discourse regarding the best way to process audio data, whereas while there may be multiple ways of representing text data (some of which are explored in this paper) the data itself is not subject to fundamentally change.
- Lyrics can serve as a proxy for the melody and rhythm of a song (note this is only possible when lyrics are represented in a manner that retains structure). (Fell & Sporleder, 2014)

(not the research as a whole) is a larger interest in the field of natural language processing (NLP), the use of statistical-based methods to give machines the ability to interpret and understand human language, held by the author.

2. Literature Review

What follows is a comprehensive but by no means exhaustive, review of recent research which has been conducted in the field of automatic genre classification. Focus is given to papers which have used song lyrics to make these predictions, though sometimes these may have been combined with other features such as audio and release data. Research is discussed in chronological order for organizational purposes.

Early research involving lyric-based song classification often did not focus explicitly on music genre. Instead papers such as (Logan, Kositsky, & Moreno, 2004) and (Mahedero, MartÍnez, Cano, Koppenberger, & Gouyon, 2005) used text data to predict a song's themes and/or artist similarity. Neither paper achieved results significant enough to recommend the sole use of lyrics for music classification tasks, in fact the authors of both end their discussions advocating for a blended approach that takes advantage of both audio and lyric data. However, it should be noted that these papers employed traditional machine learning algorithms (Naïve Bayes and Probabilistic Latent Sentiment Analysis) which have since been surpassed in effectiveness by modern deep learning approaches. Additionally, the number of songs used in each analysis (500 words and 15,000 words respectively) is much smaller than those found in other research, as well as in this paper.

In 2007, researchers from the Vienna University of Technology (Neumayer & Rauber, 2007) trained Support Vector Machines on the audio and lyric data of close to 10,000 songs. Model performance was shown to improve as the dimensionality of the lyric vectors did, and while blended models again outperformed those that just used lyrics, the difference in classification accuracy for the best performing variant of each (the 3000-dimension lyric only

model and the 2880-dimension blended model) was less than 0.015. This result indicates that better performance for lyric only models may be obtained through better representations of the text data.

Examples of innovative text representation can be found in (Mayer, Neumayer, & Rauber, 2008) and (Fell & Sporleder, 2014). For (Mayer et al., 2008), researchers trained a variety of traditional machine learning algorithms (k-Nearest Neighbors, Decision Trees, Naïve Bayes, and Support Vector Machines) on word, rhyme structure, parts-of-speech, and text statistic features derived from lyric data. Dimensionality of the various combinations ranged from 6 (for the rhyme only representation) to 3,336 (for the representation which combined all four features). The best performing model was the SVM trained on a combination of the rhyme, parts-of-speech, and text statistic features, but the representation which led to the most consistently well performing model across each algorithm was that which was comprised of just the text statistics. In (Fell & Sporleder, 2014), researchers developed 13 different feature representations tied to a song's lyrical content, some of which (such as parts-of-speech and rhyme schema) can be found in (Mayer et al., 2008). These 13 features are divided into 5 separate classes (vocabulary, style, semantics, orientation, and song structure) and used to train SVMs for three separate classification tasks (genre, best v. worst, and publication time). In each case, models trained using the 13 features performed better than the baseline which consisted of the top 100 n-grams (n less than or equal to 3) ranked according to term frequency-inverse document frequency (tf-idf).

(Mayer & Rauber, 2010) and (Liang, Gu, & O'Connor, 2011) are both papers which used ensemble models to predict music genre from a combination of lyric and audio data. The (Mayer & Rauber, 2010) paper shared the same authors as (Mayer et al., 2008) and represented their text

data in largely the same manner as was seen in that research. They also employed many of the same algorithms with the addition of a random forest model, some variants of nearest neighbors, and both linear and polynomial SVMs. Models are created for each combination of algorithm and feature (a cartesian product of the two subspaces) and weighted according to a variety of different schemas. Classification accuracies of as high as 74.1% were achieved. (Liang et al., 2011) is of particular interest because its lyric and audio data comes from the Million Song Dataset (the same one used in this paper). Here, the authors represent audio and lyric features in a variety of different ways including some which are learned from the data itself (eg. Hidden Markov models are used to generate genre probabilities from audio features, Canonical Correction Analysis is used to reduce the dimensionality of the audio and lyric features to one shared space). Their results are not as promising as those from (Mayer & Rauber, 2010), classification accuracy fails to surpass 40% for any combination of features.

Most recently, a researcher at Stanford University published a paper (Tsaptsinos, 2017) which set about applying some of the most cutting edge deep learning algorithms (Hierarchical Attention Networks or HANs, Long Short-Term Memory or LSTMs) and NLP representation techniques (word embeddings) to the task of lyric-based genre classification. Models were trained on a dataset consisting of nearly 500,000 lyrics across 117 genres. Two classification tests were conducted, one which only sought to predict the top 20 genres and the other the full set of 117. For each task, neural-based methods were able to achieve accuracy above 45%, a significant improvement over previous lyric-only models.

3. Motivation

The motivations behind this paper are not purely exploratory. As the Literature Review has already revealed, there is already a healthy body of research which is attached to the subject, another solely exploratory paper would be white noise (barring any major predictive breakthroughs). Rather, this paper is motivated by a marketing use case for sophisticated automatic genre classifiers. The hope is that by providing a clear argument for the application and potential benefits of this research within the marketing field, more will study it.

3.1 Recent Advancements in Personality Prediction

Before we may begin discussing how automatic genre classification can assist marketers, some background is needed on recent developments in the field of personality prediction, in particular those having to do with social media data.

One of the earliest studies which attempted to predict people's personalities based on their digital footprints is (Marcus, Machilek, & Schütz, 2006), which examined the content of personal websites as means of effective personality prediction. A self-other agreement study was conducted in which personal website owners were asked to submit a personality profile of themselves. Then, university students, unacquainted with the website owner, were asked to judge the their personalities as well, but only based on the content of their website. Results indicated that yes, effective personality prediction based on personal website content alone was possible for some personality types, but to turn this into a machine learning model the question of data collection and representation remained. In 2010, Tal Yarkoni conducted research which showed significant correlations between various characteristics of blogs (such as writing style, subject, and word choice) and the personality of the author (Yarkoni, 2010). This research provided

further evidence of the potential for deriving a person's personality from their digital data, as well as a source and method of collection of said data. However, the research lacked any discussion of how to model these relationships and thus make predictions.

(Golbeck, Robles, & Turner, 2011) moves away from blogs and instead uses Facebook data on 279 Facebook users to make personality predictions. User's friend networks, personal information, activities, preferences, and posts/writings were used to create 74 features which were then used in regressions. These regressions were found to correctly predict personality with only 11% error. Those same researchers would go on to conduct a similar study using Twitter data, near identical results were achieved (Golbeck, Robles, Edmondson, & Turner, 2011). A recent 2015 study, found that not only was social media data useful in predicting personality, but that these predictions were more accurate than those made by peers and sometimes even the person themselves. The study had over 70,000 participants and two separate samples of approximately 17,000 and 14,000 people were used to create prediction models based off of Facebook liked data (what objects, personalities, entertainment they liked on Facebook) (Youyou, Kosinski, & Stillwell, 2015).

3.2 The Potential Marketing Applications of Personality Predictions

A massive step towards demonstrating the application of the above research was taken in 2017 by faculty from the business schools at Columbia, Stanford, Penn, and Cambridge. In (Matz, Kosinski, Nave, & Stillwell, 2017), these authors examined whether the personalities predicted by the social media data in the aforementioned studies, could be used to effectively target people with advertisements on those very platforms.

3.2.1 Historical problems with personality-based targeting

Before the results and conclusions of (Matz et al., 2017) can be made, one must first note the rocky standing of personality based psychographic segmentation in the marketing research tradition. The concept of psychographic targeting and segmentation was first introduced to the marketing and advertising field in the 1970s. Unlike demographic segmentation, which divided consumers into groups based on easily observable traits such as age, sex and income, psychographic segmentation divided consumers based on psychological characteristics such as values, attitudes, beliefs, and, personalities. In fact, personalities were the primary focus of early psychographic segmentation work, motivated by the rise of self-congruency theory which stated that a consumer's perception of brand image (or brand personality) is directly related to and influenced by their own self-image (Vyncke, 2002). Intuitively the theory makes sense, consumers want to buy products which they can envision themselves using, but over the years attempts to prove its validity over the years have proven largely inconsistent. Of particular interest is a 2009 study which found that the Big Five personality traits (those same ones used in each piece of personality prediction research cited in this paper) were "not strong enough to be reliable predictors of brand preferences." (Mulyanegara, Tsarenko, & Anderson, 2009). Thus, there exists a healthy degree of skepticism within the field of marketing research when it comes to personality-based targeting.

However, there is evidence that digital data may serve as a remedy to the problems that have plagued this line of research over the years. One of the major knocks against personality-based marketing research has been that the methods used to measure participants' personalities are improper and lack validity. Oftentimes researchers employ personality tests which were validated exclusively in either academic or medical contexts, the assumption that these same tests would prove useful in the context of a more general consumer setting is tenuous at best.

Furthermore, the way these tests are typically administered (newspaper or direct mail survey) fail to create an environment in which honest or accurate responses are likely to occur (Kassarjian, 1971). These tests may yield results which are less indicative of what someone's actual personality is and more so what that person would like to believe (or like you the researcher to believe) their personality is. Deriving personality using digital data deals with both issues. Firstly, personality assessments made by the models trained on digital data are not privy to the same bias that's introduced when someone sits down at their coffee table to fill out a personality survey in the newspaper (the traditional method by which this data was collected). In the latter scenario, the quality of the data is dependent upon the quality of the answers given by the participant, who may or may not be fully engaged in the task of self-reflection which stands before them. The personality models discussed in the papers from section 3.1 of this paper use people's observed online behavior as a basis for their predictions, and while on some levels users may be aware that their actions are monitored, the fact is not as directly apparent as in the survey scenario, thus it is less likely to modify their behaviors.

3.2.2 Results of (Matz et al., 2017)

Returning to the study addressed at the beginning of this subsection (3.2), and confirming the point made in subsection 3.2.1, the authors did concluded that their targeting was effective. In the study, individuals were targeted with online advertisements based on their personalities as described by the Five Factor model (more specifically, the Openness and Extraversion traits of the Five Factor model, which were selected because previous studies have shown that they exhibit strong correlation with Facebook likes). Personality traits were derived using the *myPersonality.org* database to arrive at a set of Facebook likes (10 for each category) which indicate high levels of Extraversion, low levels of Extraversion, high levels of Openness, and

low levels of Openness. Additionally, four test ads were created, with each meant to target one of the four personality categories. Two studies were conducted (one for each trait) in which participants were served with ads that fell in to one of four scenarios: they exhibited high levels of the trait in question and were served congruent ads, they exhibited high levels of the trait in question and were served incongruent ads, they exhibited low levels of the trait and were served congruent ads, or they exhibited low levels of the trait and were served incongruent ads. While ad click-through-rates (how many people click on an ad per x number of people it's served to) may have remained fairly stable across the categories, conversion rates (how many people made a purchase per x number of people an ad is served to) were consistently higher for categories of congruency. These results suggest that serving consumers ads which speak to their personalities is an effective method of targeting.

3.3 Recent Event-Driven Issues

Applications of the above research have been carried out in the real-world, one of the most high-profile (and damning) being the recent partnership between the Trump 2016 election campaign and the now defunct British research firm Cambridge Analytica.

When Cambridge Analytica first burst on to the American political scene in 2015, its founder Alexander Nix marketed the company as a new-age data-driven political research organization which had developed detailed personality profiles on almost every US voter. According to Nix, each profile was built on nearly 5,000 different individual data points and provided insight into how persuadable a voter was to various campaign messages. On behalf of the Trump campaign, the company used these profiles to target potential voters and donors with online messages meant to sway them towards either voting for or donating to the candidate

(Lapowsky, 2017). Since the Cambridge Analytica-Facebook scandal broke in early 2018 and it was revealed that most of the data the company used to build its famed personality profiles had been harvested from Facebook, there has been some question as to whether or not the company truly provided any insight beyond that already available to users of the Facebook ad platform (Granville, 2018). Whatever the case, the triumph of this approach remains, in a surprise upset Trump bested his opponent Hilary in the 2016 election and many analysts credit his robust online ad targeting campaign as the reason why.

But, while the 2016 election may have served as a great triumph for big data and its potential for personality-based psychographic targeting, it may have also served as its death knell. Since the specifics of what occurred between the Trump campaign and Cambridge Analytica have come to light, sweeping data privacy reforms have been implemented (Europe's successful adoption of GDPR) and consumer's have become weary of many large online platforms, Edelman's 2018 Trust Barometer revealed that only 60% of consumers indicated that they no longer trusted social media platforms with personal data (*2018 Edelman Trust Barometer Special Report: Brands & Social Media*, 2018). Less consumer data means less accurate personality predictions which translates to less effective ad targeting. Now that consumers are no longer directly providing their detailed personal information to these digital platforms, the question becomes where to find supplemental data so that these platforms can maintain their ability to approximate a user's personality.

3.4 Opportunity for Online Content Providers

The link between personality and aesthetic preference has been studied since at least the 1960s when Irvin L. Child, a psychology professor at Yale, conducted a study which linked

personality to aesthetic preferences (Child, 1962). Since then research has linked personality differences to people's preferences for art (Furnham, Avison, & Differences, 1997), movies, television shows (Weaver III, 1991), and, of greatest interest to this paper, music (Rentfrow & Gosling, 2003). This research suggests that if the ability to use consumer's personal data to predict personality is becoming less and less feasible, their content consumption habits may provide a serviceable proxy. Content providers such as the movie streaming service Netflix, the video network YouTube, and the music streaming service Spotify, could create detailed profiles of their user's consumption habits and map these profiles to different personality traits in accordance with the correlations found by the studies from above. These derived personality profiles can be offered to ad buyers as an extra suite of features to target users on. The presence of an extra, valuable targeting feature can serve as a competitive advantage for these content providers when selling ad space (should that be part of their business model).

Considering the musical focus of this paper, let's examine how a process, such as the one previously describes, may look at a company like Spotify:

1. The company aggregates a user's listening history by genre
2. Genres which said user listens to the most (according to either streams, total minutes, or some combination of the two) are chosen as being representative of that user's musical preference
3. Based on the genres they prefer, their personality is then inferred based on the table of correlations provided by (Rentfrow & Gosling, 2003) (eg. users who listen to a lot "Upbeat & Conventional" music such as pop and country, would be labelled as Agreeable, Extraverted, and somewhat closed off).

4. Once labelled, Spotify's ad buyers could then target these consumers by specifying which pieces of copy should be played for which personality groups.

For the above system to work, Spotify (and other streaming music service platforms) need a consistent, accurate, and scalable method for classifying songs into genres, one that does not rely on the subjective opinion of the artist, or requires mass surveying of the crowd. Automatic genre classification models checks the box for each of these requirements, making them a crucial and foundational piece in this content-based personality targeting system.

4. Data & Methodology

4.1 The Million Song Dataset

A collaboration between researchers at Columbia University and the music intelligence firm Echo Nest (since acquired by Spotify), the Million Song Dataset (MSD) was formally introduced in 2011 and serves as the primary source of data for this research paper (Bertin-Mahieux, Ellis, Whitman, & Lamere, 2011). The MSD consists of audio and metadata on one million songs/files which were legally available to Echo Nest at the time of creation. The core dataset spans 44,745 unique artists and amounts to a total size of 280 GB. Primary features are numerical representations of a song's pitch, timbre, and loudness which are provided for each song segment. Additionally, the MSD has numerous ancillary datasets which provide data on lyrics (musiXmatch dataset), song-level tags (Last.fm dataset), genre annotations (tagtraum genre annotations), and more. Of interest to this paper are the musiXmatch and Last.fm datasets which provided the input and output data used to train the machine learning models.

4.1.1 musiXmatch Dataset

The musiXmatch dataset contains lyric data for 237,662 tracks within the core MSD. Song lyrics are represented in bag-of-words format using the top 5,000 words according to number of occurrences across the entire set of songs. Note that these 5,000 words were calculated post-stemming and account for roughly 92% of all words seen. A bag-of-words model, for those unfamiliar with the term, is a vectorized representation of text data whereby each element in the vector corresponds to a specific word in a dictionary or vocabulary. The value found at each element corresponds to the frequency with which that word occurs in the text.

The musiXmatch dataset also contained a feature called `is_test` which researchers suggested be used as a way of splitting the dataset into testing and training segments. This feature was not used in the paper's analysis of the data.

4.1.2 The Last.fm Dataset

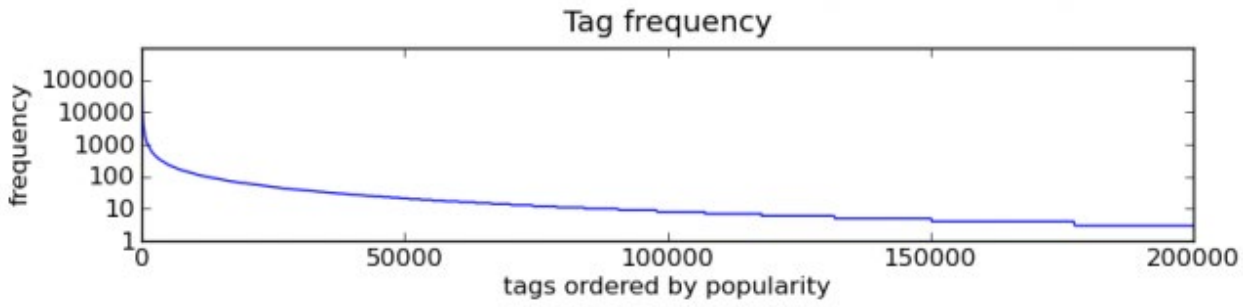
The Last.fm dataset contains two additional pieces of data for tracks present in the MSD, tags and similar tracks, this analysis only uses the former. Last.fm tags are user-generated keyword which users can ascribe to tracks using their own discretion. Furthermore, tags can be ascribed to songs multiple times, though Last.fm's API only reports on this number for up to 100 occurrences. There are 522,366 unique tags in the dataset which span across 505,216 tracks. As seen in Figure 1, tag data has an extremely long tail as a majority of tags are only associated with a small number of tracks.

4.2 Dataset Creation

The dataset used in this paper is a subset of the MSD, To arrive at this subset the following steps are taken:

1. All duplicate tracks are removed from the Last.fm dataset using the list of duplicates supplied by the MSD's originators.

Figure 1: Last.fm Tag frequency



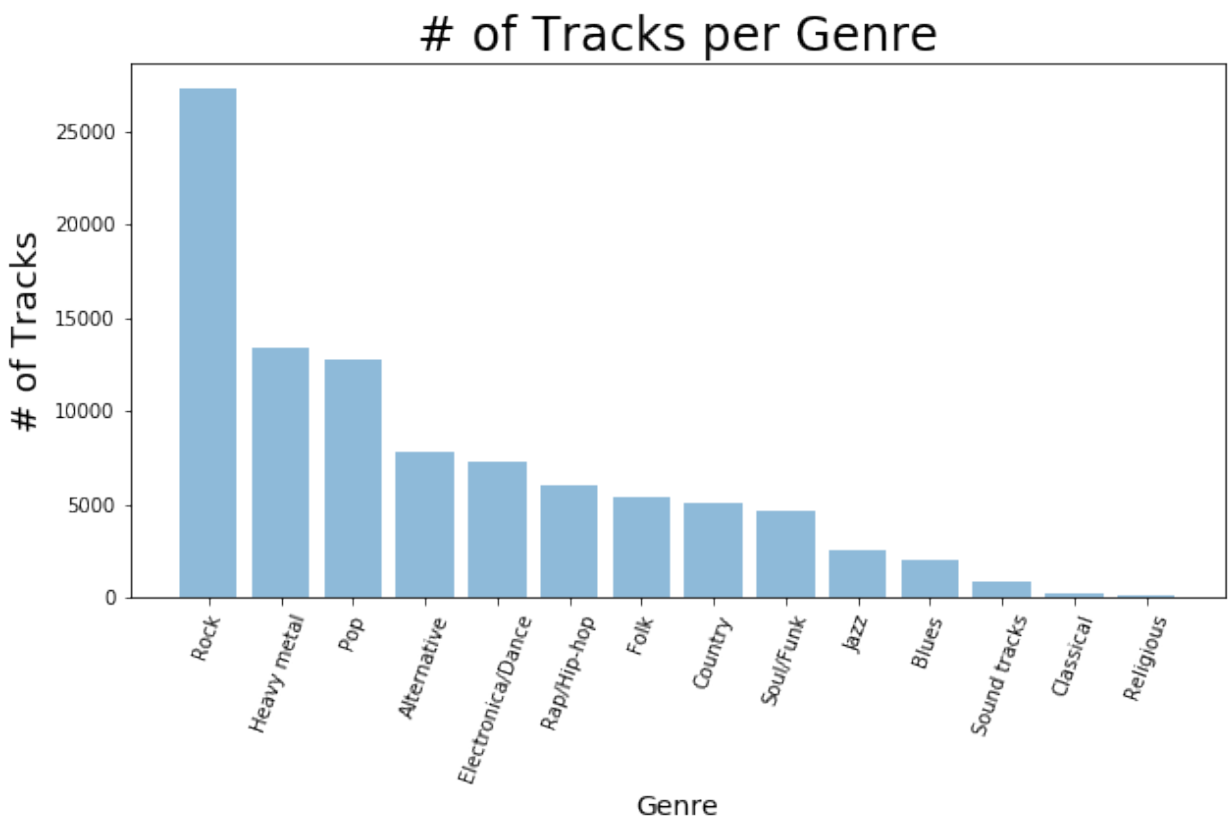
2. Not all tracks in the musiXmatch dataset are found in the Last.fm dataset and vice versa, therefore only tracks belonging to the intersection of these two datasets are used, approximately 163,820.
3. All tracks not tagged with one of the 11 genres found in the Rentfrow & Gosling (R&G) study linking music genre to personality are filtered out. Because tags are user-generated a dictionary is used to map each of the 11 genres to common misspellings and or alternate spellings (see Table 1).

The resultant dataset consists of either 96,661 or 106,213 songs which are classified as either one or multiple genres respectively. When songs are classified as one genre, the genre which was tagged with the highest frequency is used. In cases where songs have multiple genres tagged with the same frequency, the song is removed from the dataset (thus the lower number of tracks for single classification representation). The distribution of these tracks across the 14 genres can be seen in Figure 2.

Table 1: Mapping of R&G genres to Last.fm tags

Genre	Tags
Classical	Classical
Jazz	jazz
Blues	blues, bluesy
Folk	folk
Alternative	Alternativo, alternativ, alternativa, alternative
Rock	rock
Heavy metal	mental, metal, heavy metal
Country	contry, country
Pop	pop
Religious	religious
Sound tracks	Soundtrack, Soundtracks
Rap/Hip-hop	Hip-Hop, hip hop, hiphop, rap
Soul/Funk	soul, funk, r & b, r&b
Electronica/Dance	-electronic-, electronic, electronica, eletronic, dance, dancey

Figure 2: Distribution of tracks by genre in final dataset (multi-classification)



4.3 Feature Set

4.3.1 Lyric Features

Lyrics are represented in three ways:

Term Frequency – Inverse Document Frequency (`tf-idf`)	5,000 element vectors weighted according to term frequency-inverse document frequency metric below: $$f(t_n) = \frac{x_{nt}}{\sum_{t=0}^{4999} x_{nt}} \times \ln\left(\frac{N}{n_t}\right)$$
50 Dimension Word Vector (`wv50`)	50 dimension word vector trained on the Wikipedia 2014 + Gigaword 5 corpus using the Stanford GloVe unsupervised learning algorithm. The weights of each word vector are multiplied by the

	frequency of the corresponding word in a set of lyrics. The scalar multiplied word vectors are then summed across all the words in a song. See equation below: $$f_{lyric-vector}(i) = \sum_{t=0}^{4999} x_{it} * v_t$$
300 Dimension Word Vector (`wv300`)	Same as above but with 300 dimension word vectors.

4.3.2 Text Statistic Features

The following text statistics are calculated and used to create the `stat` feature set:

- total_words: the total number of words in a song's lyrics
- unique_words: the number of unique words in a song's lyrics
- unique_word_ratio: the number of unique words divided by the total number of words in a song's lyrics
- character_per_word: the average number of characters per word in a song's lyrics

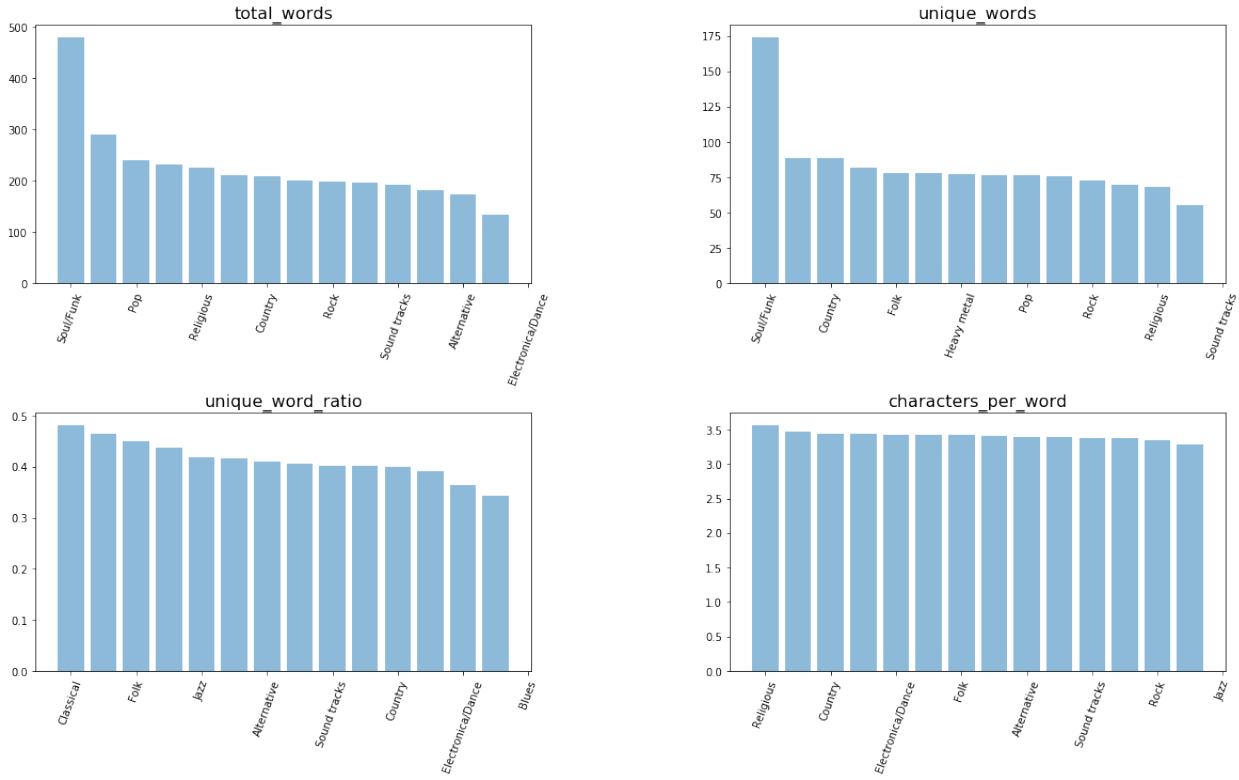
Average metric values for each genres are plotted in Figure 3.

4.4 Models

4.1 Multiclassification Algorithms

Four algorithms were used to make multiclass genre predictions across six different combinations of the feature sets described in section 4.3. These included: a Naïve Bayes classifier, a Random Forest with 80 decision tree estimators, a multi-class Logistic Regression model (structured as a neural network with no hidden layers) and an Deep Neural Network

Figure 3: Average metric per genre



(DNN) with one hidden layer consisting of 500 nodes. Both the Logistic Regression and the DNN used categorical cross-entropy as their loss functions.

4.2 Multilabel Algorithms

A Deep Neural Network was used to make multilabel genre predictions according to the same six feature set combinations used for multiclass classification. This DNN also employed the same number of hidden layers (1) and nodes (500) as that used for multiclass classification, but swapped out categorical cross-entropy for binary cross-entropy as its loss function.

5. Results

5.1 Multiclassification Results

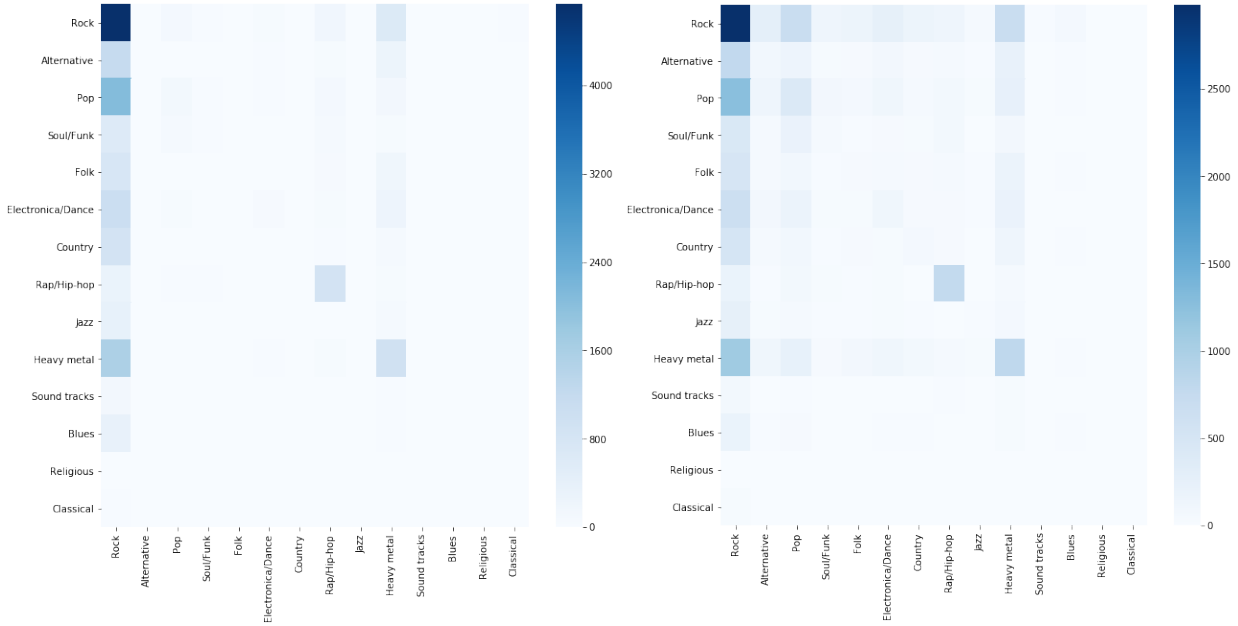
The mean accuracy and standard deviation for the various multiclass classification models can be found in Table 2. Each combination of algorithm and feature set was tested using K-Fold Cross Validation with 10 splits. The best performing model of the 24 is the DNN trained on the `tf-idf` feature set, which produced a mean accuracy of 0.416, besting the next closest model (the Logistic Regression trained on `stat+td-idf` feature set) by 0.03. However, the DNN did not produce the best performing model for any of the other feature sets.

We also note that that the Naïve Bayes and Random Forest models trained on the `stat` feature set produced significantly better mean accuracies than other models which were trained using the same algorithms. A closer inspection of the confusion matrices for each model (which can be seen in Figure 4) reveals that they were in fact classifying each song as the most popular genre ("rock"), except for those which it could more easily recognize ("heavy metal" and "rap/hip-hop"). Combined the three genres account for approximately a third of the songs in the dataset, thus resulting in the 0.346 mean accuracy of the two models.

Table 2: Multiclass Classification Model Accuracies

	Naïve Bayes	**Random Forest**	**Logistic Regression**	**Deep Neural Network (DNN)**
stat	0.3464 (*0.0032*)	0.3463 (*0.0029*)	0.1610 (*0.0179*)	0.13768
tf-idf	0.1840 (*0.0042*)	0.3039 (*0.0034*)	0.3623 (*0.0035*)	**0.4155**
wv50	0.1653 (*0.0066*)	0.3235 (*0.0037*)	0.2422 (*0.0185*)	0.2837
wv300	0.1329 (*0.0034*)	0.3203 (*0.0036*)	0.2757 (*0.0256*)	0.3181
stat + tf-idf	0.1448 (*0.0027*)	0.3100 (*0.0032*)	0.3862 (*0.0046*)	0.3787
stat + wv300	0.1355 (*0.0036*)	0.3244 (*0.0039*)	0.2883 (*0.0408*)	0.2837

Figure 4: Confusion matrices for the Naïve Bayes and Random Forest multiclass classification algorithms trained on the `stat` feature set



Naïve Bayes Random Forest

5.2 Multilabel Results

The results, which can be seen in table 3 below, of multilabel classification were exceedingly underwhelming. Only one of the models (that trained on the `tf-idf` feature set) produced a classifier which achieved better accuracy than that of the null classifier, and even then it was only a 1.2% difference.

Table 3: Results of multilabel classification

Null Accuracy		0.8711
	Accuracy	**Accuracy-Null Accuracy**
stat	0.8712	-0.0001
tf-idf	0.8834	0.0121
wv50	0.8718	0.0007
wv300	0.8669	-0.0045
stat + tf-idf	0.8762	0.0041
stat + wv300	0.8726	0.0000

6. Conclusion & Discussion

Based on the multiclass classification results observed in section 5, we can say that deep learning algorithms have the potential to outperform traditional machine learning algorithms at the task of automatic genre classification if provided with the right feature set. However, the accuracy of this model failed to best the accuracy achieved in (Mayer & Rauber, 2010) (74.1%) which used a combination of audio and lyric features. Furthermore, the relatively basic deep learning techniques employed here (a simple DNN) came close to achieving a similar level of accuracy (41.6% v. 49.5%) as models trained using far more advanced algorithms which took contextual information into account (Tsaptsinos, 2017). All of this suggests that either lyric data, or the methods by which we currently represent lyric data is not enough to train an effective automatic genre classifier model. As for why this is the case, the reasons could be many. Given the relative success of blended feature models it is likely that audio data contains a unique piece of information about the relationship between song and genre that is not captured by lyrics alone. Similarly, there may lack sufficient differentiation between the lyrics of certain genres for any algorithm (no matter how smart) to discern a difference. Of course, the case may simply be that we have yet to discover the proper techniques which can accomplish this task, there is always the prospect of a better, smarter algorithm coming along and rendering all this doubt for naught.

As for multilabel automatic genre classification, there is no evidence that ANNs can train an effective model for the task, regardless of what feature set is used, each model failed to perform better than a random null classifier by more than 1.2%. This poor performance may be due to poor data representation. Currently each multilabel genre vectors follow a one-hot encoding methodology whereby each genre tag is given a weighting of one if at least one Last.fm

user tagged the song as such, though obviously genres which are tagged to a song multiple times by multiple users are more likely to be that song's true genre, compared to those that are tagged just once. A frequency cutoff may be useful to implement, whereby songs are only classified as belonging to genres if said genre has been tagged a certain number of times by a certain number of users (essentially this is already being done but with a threshold of 1). This frequency cutoff would provide a more accurate idea of what genres a song belongs to and will reduce the amount of noise in the data.

With an eye towards the future, we recommend that researchers interested in this subject focus on blended models, or at the very least ensemble learning methods which contain both lyric and audio models. Also, considering the success of the ensemble learning methods, which were comprised entirely of models trained using traditional machine learning algorithms, one may find it advantageous to apply these same methods to a group of deep learning models, considering their displayed outperformance of traditional algorithms.

Music and media companies which are considering implementing an automatic genre classifier (whether that be for the purpose of advertising as posited in the Motivation section of this paper, or any other reason), should consider the following before making a final decision:

- To achieve the best possible accuracy both lyric and audio features must be used to train the model. This naturally requires more upfront engineering and computational costs than a lyric only model as the data pipeline and models will be more complex. Whether or not this cost is worth taking on should be evaluated by the company.
- Even the best automatic genre classifiers only achieve an accuracy of approximately 70%, thus the company must also decide if the technology is better than current

alternatives available to them (eg. having the artist self-report genre, outsourcing the task to upworkers, etc.)

- Automatic genre classification is currently only able to handle multiclass representations, thus if the problem or opportunity being addressed requires multilabel representation for maximum effectiveness, the company should look beyond computational methods for categorization.

Bibliography

2018 Edelman Trust Barometer Special Report: Brands & Social Media. (2018). Retrieved from https://www.edelman.com/sites/g/files/aatuss191/files/2018-10/2018_Trust_Barometer_Brands_Social_Media_Special_Full_Report.pdf

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., & Lamere, P. (2011). The million song dataset.

Child, I. L. (1962). Personal preferences as an expression of aesthetic sensitivity.

Fell, M., & Sporleder, C. (2014). *Lyrics-based analysis and classification of music.* Paper presented at the Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers.

Furnham, A., Avison, M., & Differences, I. (1997). Personality and preference for surreal paintings. *23*(6), 923-935.

Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011). *Predicting personality from twitter.* Paper presented at the 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing.

Golbeck, J., Robles, C., & Turner, K. (2011). *Predicting personality with social media.* Paper presented at the CHI'11 extended abstracts on human factors in computing systems.

Granville, K. (2018, 3/19/2018). Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens. *The New York Times*. Retrieved from https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html

Kassarjian, H. H. (1971). Personality and consumer behavior: A review. *Journal of Marketing Research, 8*(4), 409-418.

Lapowsky, I. (2017, 10/26/2017). What Did Cambridge Analytica Really Do For Trump's Campaign? *Wired Magazine*.

Liang, D., Gu, H., & O'Connor, B. (2011). Music genre classification with the million song dataset.

Logan, B., Kositsky, A., & Moreno, P. (2004). *Semantic analysis of song lyrics.* Paper presented at the 2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763).

Mahedero, J. P., MartÍnez, Á., Cano, P., Koppenberger, M., & Gouyon, F. (2005). *Natural language processing of lyrics.* Paper presented at the Proceedings of the 13th annual ACM international conference on Multimedia.

Marcus, B., Machilek, F., & Schütz, A. (2006). Personality in cyberspace: personal Web sites as media for personality expressions and impressions. *Journal of Personality, 90*(6), 1014.

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences, 114*(48), 12714-12719.

Mayer, R., Neumayer, R., & Rauber, A. (2008). *Rhyme and Style Features for Musical Genre Classification by Song Lyrics.* Paper presented at the Ismir.

Mayer, R., & Rauber, A. (2010). *Building ensembles of audio and lyrics features to improve musical genre classification.* Paper presented at the 2010 International Conference on Distributed Frameworks for Multimedia Applications.

Mulyanegara, R. C., Tsarenko, Y., & Anderson, A. (2009). The Big Five and brand personality: Investigating the impact of consumer personality on preferences towards particular brand personality. *Journal of Brand Management, 16*(4), 234-247.

Neumayer, R., & Rauber, A. (2007). *Integration of text and audio features for genre classification in music information retrieval.* Paper presented at the European Conference on Information Retrieval.

Rentfrow, P. J., & Gosling, S. D. (2003). The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of Personality, 84*(6), 1236.

Tsaptsinos, A. (2017). Lyrics-based music genre classification using a hierarchical attention network.

Vyncke, P. (2002). Lifestyle segmentation: From attitudes, interests and opinions, to values, aesthetic styles, life visions and media preferences. *European Journal of Communication, 17*(4), 445-463.

Weaver III, J. B. (1991). Exploring the links between personality and media preferences. *Journal of Personality, 12*(12), 1293-1299.

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality, 44*(3), 363-373.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences, 112*(4), 1036-1040.