



Master Thesis
Data Science: Business and Governance

Music genre classification using audio features and lyrics sentiment features with multinomial logistic regression

Name:	Anneloes Bal
EMPLID:	1256860
Email:	
Supervisor:	dr. M.M. van Zaanen
Second reader:	dr. A. Alishahi
Project period:	February 2018 – July 2018
Project theme:	Human Learning and Decision Making
Institute:	Tilburg School of Humanities and Digital Sciences
Word count:	11.232

Abstract

This study focused on classifying songs from the Million Song dataset to their musical genre by performing multinomial logistic regression on audio features, lyrics sentiment features and a combination of both. The audio features were extracted from the Million Song dataset, while the words in the lyrics from the MusiXmatch dataset were matched to their emotional scores on three emotional dimensions, being: valence, arousal and dominance, which were extracted from the extended ANEW dataset (Warriner et al., 2013). This study aimed to investigate whether adding lyrics sentiment features to the classification model would lead to increased classification performance. As a result, the model including both audio features and lyrics sentiment features produced the highest overall accuracy of 41.5%. By including lyrics sentiment features into the model, music genre classification performance was increased by roughly two percent. On their own, lyrics sentiment features also proved to be an important music genre classifier with an overall accuracy of 39.6%. This contradicted findings from related work by Liang et al. (2011) which states that including lyrics sentiment features does not influence classification performance at all.

Keywords: music genre classification, audio features, lyrics sentiment features, Million Song data, ANEW, arousal, valence, dominance, multinomial logistic regression, multiclass classification.

Table of contents

Abstract	1
1. Introduction	4
1.1 Context	4
1.2 Relevance	5
1.3 Problem statement	6
1.4 Methodology	6
2. Related work.....	7
2.1 Music genre classification	7
2.2 Music genre classification using audio features	8
2.3 Music genre classification using lyrics features	9
2.4 Lyrics sentiment features.....	11
2.5 Current state of the art	11
3. Method.....	13
3.1 Datasets	13
3.1.1 Million Song dataset.....	13
3.1.2 Tagtraum genre annotations	14
3.1.3 MusiXmatch dataset	15
3.1.4 ANEW dataset.....	15
3.2 Combining and preprocessing the datasets.....	16
3.3 Descriptive statistics.....	17
3.3.1 Million Song data	17
3.3.2 Genre data.....	18
3.3.3 ANEW data	19
3.3.4 Lyrics data	21
3.3.5 Combined dataset	22
3.4 Cross-validation.....	22
3.5 Imbalanced classes	23
3.6 Multinomial logistic regression.....	23
3.7 Evaluation metrics.....	25
3.7.1 Macro averaging vs. micro averaging	25
4. Results	26
4.1 Summary of results.....	26
4.1.1 Baseline	26
4.1.2 10-fold cross-validation results	27
4.1.3 Full multinomial logistic regression model.	27

4.1.4 Undersampling, oversampling and SMOTE	28
4.1.5 Only audio features and only lyrics sentiment features.....	28
4.2 Classification performance across classes	29
5. Discussion	32
5.1 Results of similar studies.....	33
5.2 Limitations.....	34
5.3 Future research	34
5.4 Scientific and practical implications	35
6. Conclusion.....	36
References	37
Appendix I.....	41
Appendix II	43

1. Introduction

1.1 Context

Recently, the music industry has shifted more and more towards digital distribution through online music stores and streaming services such as iTunes and Spotify. The quantity of available music is growing rapidly. For example, online music streaming service Spotify has over 30 million songs in their catalog in 2018. As a result, efficient and accurate music information processing of these large music libraries is becoming increasingly important (Oramas et al., 2017).

A way of organizing music information is by applying automatic music genre classification. Automatic music genre classification is a fundamental component of music information retrieval systems (Li, Ogihara & Li, 2003). Automatic music genre classification is the classification of a piece of music into its corresponding genre by using a computational algorithm (Ahrendt, 2006). However, automatic music genre classification is not completely indefectible. Precisely defining a music genre proves difficult, because many musical sounds sit on boundaries between genres. Music is an art that evolves, where the artists are often influenced by music in other genres (Li et al., 2003). Nevertheless, it has been shown that audio features of music belonging to the same genre share certain characteristics, because they are composed of similar types of instruments, having similar rhythmic patterns, and similar pitch distributions (Tzanetakis & Cook, 2002). This means that genre can be used to establish music similarity.

In addition to audio features, text-based features extracted from the song lyrics can also be used for music classification (Neumayer & Rauber, 2007). Song lyrics cover semantic information about a song's contents on a level that could never be covered by audio features only (Neumayer & Rauber, 2007). For example, Christmas songs can only be detected by including textual features, since they occur across many different musical genres (Neumayer & Rauber, 2007). Furthermore, lyrical features also contain relevant semantic information. Logan, Kositsky & Moreno (2004) discovered natural genre clusters when performing Latent Semantic Analysis on song lyrics. In addition to clustering songs based on their similarity in their lyrics, previous research also focused on the affective meaning of the words in the lyrics.

Dodds & Danforth (2010) measured the sentiment of songs by matching song lyrics to the Affective Norm of English Words (ANEW) dataset by Bradley & Lang (1999), concluding that that valence can help distinguish between genres when a large number of songs are considered. Other research found that certain emotional labels apply to music in particular genres, such as "angry" for punk music, "depressed" for slow blues, and "happy" for children's music (Li & Ogihara, 2006). These findings lead to believe that semantic information extracted from song lyrics and matching lyrics to an affective lexicon to conduct sentiment analysis on the lyrics can both be used in music genre classification.

Liang, Gu & O'Connor (2011) employed multinomial logistic regression with L2 regularization on the Million Song dataset with audio features, lyrics bag-of-words features and lyrics sentiment scores derived from the ANEW dataset by Bradley & Lang (1999). However, they indicate that lyrics sentiment features provide zero value when added to the model. It is assumed that this is due to the use of a relatively small ANEW dataset which only contains emotional scores for 1,034 words. In the present study, the expanded ANEW dataset by Warriner et al. (2013) is used which includes about 13 times more emotional scores than the ANEW dataset by Bradley & Lang (1999).

1.2 Relevance

This study provides important practical relevance. Since the available quantity of music keeps growing rapidly and a need still exists for automatic music analysis techniques, this research provides practical relevance with regard to organizing these large music repositories. Genre similarity algorithms make an important contribution to providing a scalable way to index and recommend music (Logan et al., 2004). For example, automatic music recommendation allows listeners to discover new music that matches their taste (Van den Oord, Dieleman & Schrauwen, 2013). With automatic music genre classification, listeners can receive recommendations based on their genre preferences. Content-based recommendation systems try to recommend items similar to those a given user has liked in the past (Lops, De Gemmis & Semeraro, 2011). For example, in a music recommendation application, the system may recommend songs to a user that are of the same genre as the songs already listened to by this user (Lops et al., 2011). Therefore, the results of this research can contribute to the improvement of recommending songs and generating playlists based on genre preferences.

In addition, this research also makes a scientific contribution. In this study, music genre classification using audio features extracted from the Million Song dataset and sentiment lyric features will be performed. The lyrics sentiment features are extracted by matching the corresponding lyrics to an expanded version of the ANEW dataset. Warriner, Kuperman & Brysbaert (2013) extended the ANEW dataset to include 13,915 English lemmas, including gender, age and educational differences in emotion norms. This extended ANEW dataset has not (yet) been used in any scientific publications concerning music genre classification. In the study by Liang et al. (2011), the smaller original ANEW dataset was used to generate lyrics sentiment features. This is probably why they report that including lyrics sentiment features in the classification model adds zero value. This research will hopefully result in different findings by using a larger ANEW dataset to generate lyrics sentiment features. In result, this study is assumed to show that lyrics sentiment features do contribute to music genre classification performance. With that, this study contributes to scientific research on the added value of including lyrics sentiment features in music genre classification.

1.3 Problem statement

The aim of this research is to improve music genre classification of songs contained in the Million Song dataset, by using different audio features such as timbre, pitch and tempo features, and lyrics sentiment features extracted from the song lyrics and the ANEW dataset. The focus of this study lies in investigating whether lyrics sentiment features provide added value when included in the classification model. Consequently, the following problem statement is developed: does including lyrics sentiment features improve music genre classification performance?

The audio features included in the classification model are: Duration, Key Signature, Mode, Segments Pitches, Segments Timbre, Tempo and Loudness. In this study, music genre classification using both audio features and lyrics sentiment features will be performed. In addition, a classification model with only audio features will be employed as well as a classification model using only lyrics sentiment features. It is assumed that the combination of both audio and lyrics sentiment features will yield the highest performance on this music genre classification task.

1.4 Methodology

First, a thorough literature review will be conducted to map out which audio features are best suited to classify songs into genre. Following this literature study, a number of audio features assumed to affect music genre classification will be chosen to include in the classification model. Before data analysis can be conducted, data first needs to be gathered. The songs used in this study are part of the freely-available Million Song dataset provided by the Echo Nest. The genre labels will be collected from the Tagtraum genre annotations dataset which is also publicly available online. The MusiXmatch dataset containing lyrics in bag-of-words format will be used to extract the words in the song lyrics. Lastly, the extended ANEW dataset by Warriner, Kuperman & Brysbaert (2013) will be used to match the words in the lyrics to the words in this affective lexicon. These words will produce a score on three dimensions: valence, arousal and dominance which will be incorporated into the classification model. These various datasets need to be downloaded, converted into the same format and joined together into one dataset. Then, data analysis will be conducted by employing multinomial logistic regression models aiming to solve this multiclass music genre classification problem. Multinomial logistic regression is applied because this method is also used in the study by Liang et al. (2011) which this study aims to partly replicate. The performance of the classification models will be evaluated by using various performance metrics, such as accuracy, macro averaged precision, recall and F1 score.

2. Related work

In this chapter, related work on music genre classification will be discussed. First, music genre classification will be defined and explained further in section 2.1. Then, music genre classification using audio features and music genre classification using lyrics features will be described in sections 2.2 and 2.3 respectively and their combined explanatory value highlighted. In addition, section 2.4 talks about textual sentiment features and its application to song lyrics. Lastly, an overview of the current state of the music genre classification research field will be given.

2.1 Music genre classification

Music classification is important in dealing with large collections of music when retrieving and classifying information (Oramas et al., 2017). There is an increased need for tools which help in automatically organizing music for easy retrieval (Nanni et al., 2016). Music streaming services such as Spotify offer catalogs of millions of tracks, for which music classification is important since it makes music retrieval and music recommendation easier (Oramas et al., 2017).

Music genre is perhaps the most common classification scheme used to distinguish music, since populations worldwide reach agreement on broad genre definitions (Ajoodha et al., 2015). Genres are often treated as a proxy for music similarity (Pálmason et al., 2017). The assumption is that songs very similar to each other belong to the same genre (Pampalk, Flexer & Widmer, 2005). However, many artists have a very individual mix of several styles which is often difficult to classify into one genre which can make it difficult to classify a song into just one genre (Pampalk et al., 2005).

Music genres are a widely used concept to manage large music collections (Pampalk et al., 2005). Music genre is determined by two complementary approaches being (a) a group of experts or musicologists and (b) a collaborative effort of a community by social tagging (Sordo et al., 2008). One such expert-based music genre taxonomy by MP3.com (2005) identified 744 genres with a maximum depth of six levels. This depth of different levels refers to sub genres within a music genre hierarchy. To illustrate this point, figure 1 demonstrates a small part of the depth of a music genre classification hierarchy, where the sub genres of classical and jazz music are shown. Musical genres have no strict definitions and boundaries as they are formed through a complex interaction between the public, marketing, historical, and cultural factors (Tzanetakis & Cook, 2002). This makes multiclass music genre classification a difficult process. In this study, each song only has one genre label. In reality, assigning just one genre to a song can be more challenging when there is some degree of possible overlap between certain genres.

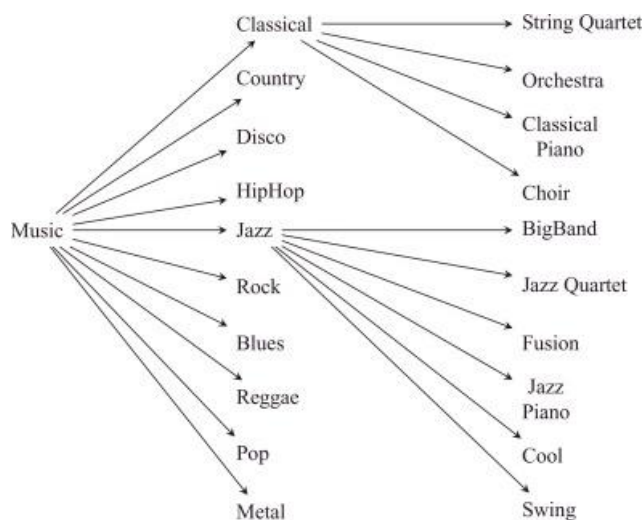


Figure 1.

Example of a small part of a genre classification hierarchy (Tzanetakis & Cook, 2002)

2.2 Music genre classification using audio features

Songs which share a particular musical genre also share certain characteristics typically related to the instrumentation, timbral texture, rhythmic structure, and pitch content of the music (Tzanetakis & Cook, 2002). Timbre is considered as the “color” or the “quality” of a sound (Wessel, 1979), with timbral features being a set of components that are used to differentiate mixture of sounds that have similar pitch and loudness contents (Scaringella, Zoia & Mlynek, 2008). Rhythm may be defined as the way in which one or more unaccented beats are grouped in relation to an accented one (Cooper & Meyer, 1963). Pitch is the quality that makes it possible to judge sounds as “higher” and “lower” in the sense associated with musical melodies (Plack & Oxenham, 2005). Pitch is a major auditory attribute of musical tones, along with duration, loudness, and timbre (Patterson, Gaudrain & Walters, 2010). Loudness is the subjective perception of sound pressure. More formally, it is defined as “that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud” (American National Standards Institute, 1973). Since pitch, duration, loudness and timbre are important attributes in distinguishing between musical tones, they need to be included in the classification model in the present study.

In addition, the key the song is in and the mode (minor or major) are assumed to also have an effect on music genre classification and are therefore also included in the classification model. Generally, songs in major sound like happy songs while songs in minor do not. In the western musical tradition major music is played at times of celebration (Mendelssohn’s ‘Wedding March’ or ‘Happy Birthday’), whereas minor music is more often used to mark mourning (Chopin’s ‘Funeral March’), heartache (‘Back To Black’ by Amy Winehouse) and despair (‘Hurt’ by Johnny Cash).¹ When people

¹ <http://www.nme.com/blogs/nme-blogs/the-science-of-music-why-do-songs-in-a-minor-key-sound-sad-760215>

talk about the ‘key signature’ or ‘key’ of a song, they are referring to the tonal center of the song. The key signature will also determine what notes and chords make a song sound consonant/pleasant.²

Approaches that use audio feature extraction combined with machine learning techniques achieved promising results a decade ago (Lu, Liu & Zhang, 2006; Mandel, Poliner & Ellis, 2006; Laurier et al., 2007; Skowronek, McKinney & van de Par, 2007; de Leon & Martinez, 2012). However, these audio-based techniques can only achieve satisfying results to a certain extent (Laurier, Grivolla & Herrera, 2008). Complementing audio features with lyrics information in the context of genre classification yields increased classification accuracy over using only audio features (Neumayer & Rauber, 2007; Hu & Downie, 2010).

2.3 Music genre classification using lyrics features

Many musical genres are rather defined by the topics they deal with than a typical sound (Mayer, Neumayer & Rauber, 2008). Lyrics describe songs in terms of content words. Song lyrics differ from traditional text documents in the sense that song lyrics exhibit a certain structure, as they are organized in choruses and verses (Mayer et al., 2008). Lyrics are often composed in rhyming verses and have different frequencies for certain parts-of-speech (Mayer et al., 2008). Different part-of-speech categories are for example nouns, verbs, articles or adjectives (Mayer et al., 2008). An example of the difference between certain part-of-speech features across genres is the strong fluctuation of relational pronouns between genres (Mayer et al., 2008). However, nouns, verbs and adjectives appear to be similar across genres (Mayer et al., 2008).

A rhyme is a linguistic style, based on consonance or similar sound of two or more syllables or whole words at the end of one line (Mayer et al., 2008). These rhyming patterns which are reflected in a song’s lyrics and easier to detect from text than audio (Mayer et al., 2008). For example, when thinking of very catchy songs, often simple rhyme structures are the common denominator (Mayer et al., 2008). Furthermore, lyrics may use “slang” language in Hip-Hop or Rap music. In addition, other statistical information such as (average) line lengths or words per minute can be derived from lyrics. Mayer et al. (2008) showed that rhyme, part-of-speech and simple text statistic features can be used for genre classification.

Another often used approach to music genre classification using lyrics features is the bag-of-words method. This method takes each unique word occurring in any of the lyrics as a feature. The frequency of the occurrences of the terms is used to populate the feature vectors (Mayer et al., 2008). In this bag-of-words approach, it is only considered whether a term is present or not. If the importance of certain terms is relevant, a term weighting scheme can be used, such as the popular $tf \cdot idf$ (term frequency * inverse document frequency) weighting scheme (Salton & Buckley, 1988). Features like

² <https://insights.spotify.com/it/2015/05/06/most-popular-keys-on-spotify/>

tf*idf have been used in lyrics classification, but they are known to not work well with texts of short length and large vocabulary, and it makes no use of semantic similarities between words (Jamdar, Abraham, Khanna & Dubey, 2015).

Jamdar et al. (2015) also state that n -grams are not suited for music genre classification because they often fail to capture the context and poetic expressions of lyrics (Jamdar et al., 2015). An n -gram is a contiguous sequence of n words from a given sample of text or speech. An n -gram of size 1 is referred to as a "unigram"; size 2 is a "bigram" and size 3 is a "trigram". To illustrate this, figure 2 shows a sequence of one word for unigrams ($n = 1$), a sequence of two words for bigram ($n = 2$) and a sequence of three words for a trigram ($n = 3$). Studies on music mood classification used bag-of-words features in various unigram, bigram and trigram representations. Combinations of unigram, bigram and trigram tokens performed better than individual n -grams, indicating that higher-order bag-of-words features captured more of the semantics useful for mood classification (Hu & Downie, 2010).

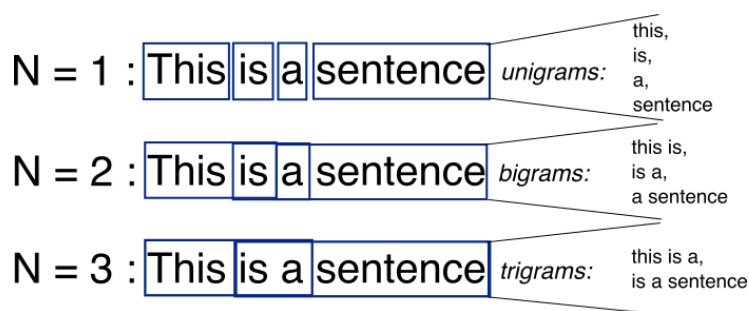


Figure 2.

N-grams. Source: <http://recognize-speech.com/language-model/n-gram-model/comparison>

Textual features of song lyrics such as textual analysis of parts-of-speech can contribute to discovering music genre similarity. Statistics such as the average word or line length are calculated (Mayer et al., 2008). Related to the audio beats-per-minute (BPM) descriptor, a words-per-minute (WPM) measure is constructed to identify music similarity. The rationale behind WPM is that it can capture the "density" of a song and its rhythmic sound in terms of similarity in audio and lyrics characteristics (Mayer et al., 2008). Text stylistic features often include interjection words such as "ooh" or "ah", special punctuations and text statistics such as the number of unique words or length of the words among other things (Hu & Downie, 2010). The combination of textual features such as bag-of-words, part-of-speech, rhyme features and text stylistic features demonstrated a greater performance on music genre classification than using only audio features as well as combining audio and only bag-of-words lyric features (Mayer et al., 2008).

2.4 Lyrics sentiment features

Moving beyond just using bag-of-words lyric features, Hu, Chen & Yang (2009) extracted features based on an affective lexicon translated from the Affective Norm of English Words (ANEW) (Bradley & Lang, 1999). This lexicon contains 1,034 unique English words with scores on valence (ranging from unpleasant to pleasant), arousal (a scale from calm to excited), and dominance (ranging from submissive to dominant). All dimensions are scored on a scale of 1 to 9 (Bradley & Lang, 1999).

Dodds & Danforth (2010) also used features extracted from lyrics and the ANEW dataset to measure the sentiment of songs among other things. They calculate the average valence for each instance as a measure of happiness, resulting that valence can help distinguish between genres when a large number of songs are considered (Dodds & Danforth, 2010). The genres scoring highest on valence are gospel/soul, pop, reggae, and rock. Followed by rap/hip-hop, punk, and lastly metal/industrial with the lowest score on valence (Dodds & Danforth, 2010).

Warriner, Kuperman & Brysbaert (2013) extended the ANEW dataset to include 13,915 English lemmas, including gender, age and educational differences in emotion norms. Nearly all previous research on measuring sentiment of song lyrics has been based on the ANEW norms collected by Bradley and Lang (1999) for 1,034 words for the purpose of music mood classification (Hu et al., 2009; Hu & Downie, 2010; Oh, Hahn & Kim, 2013). However, some related research has been conducted using the extended ANEW dataset by Warriner et al. (2013) for mood classification in music (Wohlfahrt-Laymanna & Heimbürgerb, 2017). Currently, there are no studies (yet) which use the extended ANEW dataset for the purpose of music genre classification. Since this ANEW dataset by Warriner et al. (2013) is roughly 13 times larger than the original ANEW dataset, it can be assumed that the use of this extended ANEW dataset will yield better classification performance.

2.5 Current state of the art

The International Society of Music Information Retrieval (ISMIR) is a non-profit organization seeking to advance the access, organization, and understanding of music information³. Currently, ISMIR is directing its efforts towards deep learning techniques for the purpose of automated music genre classification. Of particular interest has been the investigation of music descriptors that are based less on classical notions of musical features (such as timbre, rhythm and pitch) and more on the information extracted from some state-of-the-art descriptors that have already proven powerful in other machine learning classification problems (Nanni et al., 2018). These descriptors automatically discover and extract information from samples in a dataset, for example the Convolutional Neural Network (CNN). CNNs have been successfully applied on both discriminative and generative modeling for music-related tasks, including musical genre classification tasks (Sigtia & Dixon, 2014; Chiliguano & Fazekas, 2016;

³ <http://www.ismir.net/society.php>

Geng, Ren & Ogihara, 2017). However, these approaches use CNNs that are given an input of short segments of the music audio signal.

Music genre classification using only lyrics can be treated as a natural language processing problem (Tsaptsinos, 2017). Deep learning methods such as Recurrent Neural Networks (RNNs) or CNNs can be applied to genre classification tasks on lyrics. In fact, neural-based methods can improve on the genre classification accuracy (Tsaptsinos, 2017). This recent research stresses the importance of combining audio and lyrical data to build a stronger classifier. Employment of neural-based methods in the task of mood classification via sentiment analysis is also named as an important area of future research (Tsaptsinos, 2017). This leads to believe that combining audio features and lyrical features with sentiment analysis will lead to promising results.

However, developing neural networks for this purpose is beyond the scope of this study because the present study focuses on the contribution of lyrics sentiment features on music genre classification performance. Since Liang et al. (2011) employed multinomial logistic regression on audio features, lyrics bag-of-words features and lyrics sentiment features and concluded that lyrics sentiment features added zero value, the present study aims to replicate parts of the study by Liang et al. (2011) to see if the use of the extended ANEW dataset will lead to different findings.

3. Method

In this chapter, the dataset and experimental procedure are described in detail. First, a description of all datasets used in this study is given in section 3.1. In section 3.2, details are provided about how these datasets are joined together and the data preprocessing steps being taken. Next, descriptive statistics are given in section 3.3. Section 3.4 talks about the application of cross-validation in this study, followed by section 3.5 on imbalanced outcome classes. Lastly, the multinomial logistic regression procedure is explained in section 3.6.

3.1 Datasets

First, the contents of the Million Song dataset are described. The process of retrieving the Million Song data and the extraction of the features is explained. Then, a description of the Tagtraum genre annotations is given. Next, the MusiXmatch lyrics dataset is described, followed by the extended ANEW dataset by Warriner et al. (2013). Lastly, the process of joining all the datasets together is explained.

3.1.1 Million Song dataset

The songs used for genre classification in this study are part of the Million Song dataset⁴, a large collection containing music meta-data and audio features provided by The Echo Nest (Bertin-Mahieux, Ellis, Whitman & Lamere, 2011). It is important to stress that the dataset does not contain raw audio signals, only audio features. A detailed description of all 54 features that are part of the complete Million Song dataset can be found in Appendix I.

Meta-data consists of author, album, title, year, and duration. There are two major sets of audio features that are described as ‘Mel Frequency Cepstral Coefficients (MFCC) like’ and ‘Chroma like’ and a number of additional descriptors including tempo, loudness, key and some high level features (e.g. danceability, hottness) (Schindler & Rauber, 2012). Chroma features refer to the twelve different pitch classes. MFCCs are a way of representing the spectral information in a sound. Each coefficient has a value for each frame of the sound (Loughran et al., 2008). Obtaining MFCCs involves analyzing and processing sound according to the steps in figure 3. MFCCs have been dubbed as ‘timbral similarity’ by Aucouturier & Pachet (2002).

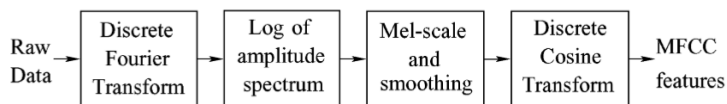


Figure 3.

Mel frequency cepstral coefficients feature extraction process (Meng et al., 2007)

⁴ <https://labrosa.ee.columbia.edu/millionsong/>

The variables ‘Pitch’ and ‘Timbre’ are extracted as 12-dimensional vectors producing a score on pitch and timbre for each segment of a song. Segments_timbre is described as ‘Mel-frequency cepstral coefficients (MFCC)-like features’ on the Million Song data website. These 12-dimensional vectors contain unbounded values centered around 0. Segments_pitches is described as ‘Chroma-like features’. This is a normalized 12-dimensional vector ranging from 0 to 1 corresponding to the 12 pitch classes C, C#, to B (Nürnberg et al., 2014).

The average score of all segments for ‘Timbre’ and ‘Pitch’ was calculated for each track ID, because this approach is also commonly used among other related scientific work and produces decent accuracy scores (Liang, Gu & O’Connor, 2011; Nürnberg et al., 2014).

For this study, not all features from the Million Song dataset were necessary to take into account. The eight features that were extracted from the complete dataset are: Track ID, Duration, Key Signature, Mode, Segments Pitches, Segments Timbre, Tempo and Loudness. Track ID was needed to match the genre and lyrics dataset to the Million Song dataset and the rest of the features were assumed to contribute to classifying the songs to genre.

The original dataset contains 54 features for over a million songs and has a size of 280 GB. The tracks are available as raw HDF5 files. To analyze the data, these HDF5 files needed to be converted into the same format as the other datasets in order to be joined together. A pre-existing Python script⁵ was used to convert the HDF5 data. This Python script made use of the Python wrapper "HDF5_getters.py", written by Thierry Bertin-Mahieux at Columbia University. This file makes use of the Python libraries numpy (Numerical Python) and tables (PyTables/Python Tables) which are capable of dealing with HDF5 formats.

After converting the HDF5 files into one CSV dataset and removing redundant features, the dataset contained exactly one million songs and 30 features: Duration, Key Signature, Tempo, Loudness, Mode, Track ID, timbre0 through timbre11, and pitch0 through pitch11.

3.1.2 Tagtraum genre annotations

The genre labels used in this study are extracted from the Tagtraum genre annotations dataset⁶, which consists of three separate datasets that are based on the Last.fm dataset, the Top-MAGD dataset and the beaTunes Genre Dataset (BGD). All three of these datasets are downloaded and converted into the same format as the extracted Million Song data. Finally, the three datasets were merged together. Rows occurring more than once were removed so that only unique track IDs remained. If track IDs had more than one corresponding genre annotation, this second genre annotation was removed from the dataset in result. This was done to maintain comparability among track IDs. This resulted in a dataset of 381,285

⁵ <https://github.com/amgreenstreet/Million-Song-Dataset-HDF5-to-CSV>

⁶ http://www.tagtraum.com/msd_genre_datasets.html

track IDs with their respective genre annotation. Within this dataset, there are 18 unique genres: Blues, Country, Electronic, Folk, International, Jazz, Latin, Metal, Pop, Pop_Rock, Punk, Rap, Reggae, RnB and Rock.

3.1.3 MusiXmatch dataset

The lyrics data used in this study comes from the MusiXmatch dataset⁷ which is directly linked to the Million Song dataset. This dataset contains song lyrics in bag-of-words format containing lyrics for 237,662 tracks out of the one million tracks present in the Million Song dataset. The remaining tracks were omitted due to either copyright restrictions, a given track not containing lyrics, or duplication. All words within the lyrics were already stemmed (e.g. kneel, kneeled, and kneeling, were mapped to the same stem “kneel”) (Gossi & Gunes, 2016).

Each track is described as the word-counts for a dictionary of the top 5,000 frequent words present in the dataset, which account for approximately 92% of the complete set of unique words. This limited dictionary is chosen because many of the words outside this dictionary are noisy and unusable, or used too infrequently in the dataset to be of much statistical value (Gossi & Gunes, 2016).

The data is provided in an SQLite database. This mxm_dataset.db database consist of two tables: lyrics and words. The table ‘words’ contains one column: ‘word’ that contains the unique top 5000 words. The table ‘lyrics’ contains five columns: the track ID from the million song dataset, ‘mxm_tid’ is the MusiXmatch track ID, ‘word’ is one of the words in table ‘words’, ‘count’ is the word count for word in that track (non null), ‘is_test’ tells you if a track is in the test set (value 1) or not (value 0). This is because this SQLite version of the data was created from the two text files of the MusiXmatch dataset (train and test). However, in this study the train and test set were merged together so that one full dataset with the top 5,000 words remained. The table ‘lyrics’ was converted into the same format as the extracted Million Song dataset. The columns ‘mxm_tid’, and ‘is_test’ were removed because they had no added value for this study. This resulted in a CSV file with the size of 19,045,332 rows and three features: Track ID, Word and Count.

3.1.4 ANEW dataset

In this study, the expanded ANEW dataset by Warriner et al. (2013) was used. This dataset is based on the Affective Norms for English Words (ANEW) norms dataset collected by Bradley and Lang (1999) for 1,034 words. This dataset distinguishes between three components of emotions: valence (the pleasantness of the stimulus), arousal (the intensity of emotion provoked by the stimulus), and dominance (the degree of control exerted by the stimulus) (Warriner et al., 2013). Participants had to rate words on one single dimension only. The scale of the ratings ranged from 1 (happy [excited; controlled]) to 9 (unhappy [calm; in control]), depending on which dimension the participant had to rate.

⁷ <https://labrosa.ee.columbia.edu/millionsong/musixmatch>

Warriner et al. (2013) extended the database to 13,915 English lemmas, including information on gender, age and educational differences in emotion norms⁸. Since emotional values of words generalize to modulated forms (e.g. sings, sang, sung, singing) only the base form of the words (lemmas) are part of this dataset (Warriner et al., 2013). Since all words in the MusiXmatch lyrics dataset are already stemmed, these datasets can be well linked together.

This dataset can be downloaded in CSV format, consisting of 13,915 rows and 65 columns. For each word, the overall mean (Mean.Sum), standard deviation (SD.Sum) and number of contributing ratings (Rat.Sum) is reported for each of the three dimensions (valence, arousal, dominance). In addition, these values are also reported for each group difference (male, female, older, younger, high education, low education). These group differences are not to be considered relevant in the present study, and thus they were removed from the dataset. This resulted in a dataset of 13,915 words and 9 features, being the Mean.Sum, SD.Sum and Rat.Sum for valence, arousal, and dominance.

3.2 Combining and preprocessing the datasets

In this study, there were four separate datasets that needed to be joined together, namely the Million Song dataset, the Tagtraum genre annotations dataset, the MusiXmatch lyrics dataset and the ANEW dataset. To combine the datasets, the dplyr operation ‘inner_join’ was used, which only keeps rows with matching keys in all datasets.

Out of the million songs present in the Million Song dataset, only the songs for which genre labels and lyrics were available were used in the present study. The key used for joining the datasets was the variable ‘track_id’, which is present in the Million Song dataset, the Tagtraum genre annotations dataset and the MusiXmatch lyrics dataset. The key used for joining these datasets and the ANEW dataset was the feature ‘word’. Each word within the lyrics that did not have a score on the three emotional dimensions in the ANEW dataset was deleted. Out of the 5,000 unique words in the MusiXmatch lyrics dataset, only 1,979 unique words remained after words which did not have an emotional score on valence, arousal and dominance were deleted.

The individual words from the lyrics that remain in the dataset all had a score on arousal, valence and dominance. Their combined average scores on these emotional dimensions indicated the mean score on arousal, valence and dominance per song. So for each word within one song, the score on arousal, valence and dominance was summed up and divided by the number of total matched words within the song. This resulted in an average score on the three emotional dimensions for the summed up words in the lyrics for each of the songs in the dataset, which were used as input in the logistic regression classifier.

⁸ <http://crr.ugent.be/archives/1003>

This resulted in the final dataset of 94,482 rows (track IDs) and 34 features: Duration, Key Signature, Tempo, Loudness, Mode, Track ID, Genre, Mean Valence, Mean Arousal, Mean Dominance, timbre0 through timbre11, and pitch0 through pitch11.

3.3 Descriptive statistics

In this section, descriptive statistics of features from the combined dataset are given in order to provide a better understanding of the data.

3.3.1 Million Song data

Table 1 shows the mean and standard deviation of the audio features ‘Duration’, ‘KeySignature’, ‘Tempo’ and ‘Loudness’ within the combined dataset. ‘Duration’ is given in seconds. ‘KeySignature’ is defined as the key the song is in, which ranges from 0 to 11 in the data. The 12 keys are: A, A#/Bb, B, C, C#/Db, D, D#/Eb, E, F, F#/Gb, G, and G#/Ab. ‘Tempo’ is the estimated tempo in beats per minute, and ‘Loudness’ is the overall loudness measured in decibel. In addition, the frequency of unique values for the feature ‘Mode’ is listed. ‘Mode’ is either minor (Mode = 0) or major (Mode = 1).

Table 1

Descriptive statistics of audio features in combined dataset

	Mean	Standard deviation
Duration	240.83	88.71
Key Signature	5.30	3.57
Tempo	127.08	34.14
Loudness	0.56	0.16
Frequency	Mode = 0	Mode = 1
Mode	30,882	63,600

‘Pitch’ and ‘Timbre’ each consist of 12 dimensions. In table 2, the mean and standard deviations of these 12 dimensions is given. ‘Timbre’ consists of 12-dimensional vectors containing unbounded values centered around 0. ‘Pitch’ is a normalized 12-dimensional vector ranging from 0 to 1 corresponding to the 12 pitch classes C, C#, to B (Nünberger et al., 2014).

Table 2

Mean and standard deviation of timbre and pitch per dimension

	Mean	SD		Mean	SD
Timbre0	45.34	5.16	Pitch0	0.45	0.15
Timbre1	10.27	43.04	Pitch1	0.45	0.15
Timbre2	10.47	27.61	Pitch2	0.37	0.12
Timbre3	-2.48	11.87	Pitch3	0.31	0.11
Timbre4	-9.96	20.82	Pitch4	0.37	0.12
Timbre5	-11.89	10.86	Pitch5	0.33	0.11
Timbre6	-2.06	12.07	Pitch6	0.34	0.11
Timbre7	-2.50	6.63	Pitch7	0.36	0.12
Timbre8	4.47	8.92	Pitch8	0.33	0.11
Timbre9	1.65	6.09	Pitch9	0.36	0.12
Timbre10	-0.41	3.54	Pitch10	0.30	0.11
Timbre11	1.93	7.33	Pitch11	0.35	0.12

3.2.2 Genre data

The genres are ranked by size in table 3. This table shows how many times these genres occur within the combined dataset. From this table it is visible that the distribution of songs per genre is imbalanced.

Table 3

Summary of unique genres and their frequency in the combined dataset

Genre	Frequency
Pop_Rock	37,120
Rock	21,831
Pop	5,180
Metal	4,661
Country	4,201
Rap	4,174
RnB	3,803
Electronic	3,601
Latin	2,091
Folk	1,920
Reggae	1,537
Jazz	1,296
Punk	1,261
Blues	862
World	433
International	249
New Age	194
Vocal	68
Total number of songs in the dataset	94,482

3.2.3 ANEW data

Table 4 contains the mean, standard deviation, minimum and maximum of the average score on valence, arousal and dominance per song within the combined dataset.

Table 4

Mean, standard deviation, minimum and maximum of valence, arousal and dominance in combined dataset

	Mean	SD	Min	Max
Valence	5.66	0.44	1.67	8.00
Arousal	4.09	0.25	2.29	7.74
Dominance	5.51	0.26	2.84	7.62

In order to better understand exactly which words from the lyrics result in a higher average score on valence, arousal or dominance per song, table 5 displays the top 10 highest scoring words within the combined ANEW/lyrics dataset after removing words that did not have any score on the three emotional dimensions. These words from the lyrics score the highest on valence, arousal and dominance. A higher score on valence corresponds to “happy”, a higher score on arousal to “excited”, and a higher score on dominance to “in control” (Warriner et al., 2013).

Table 5

Top 10 highest scoring words in the ANEW/lyrics dataset

Valence		Arousal		Dominance	
Fun	8.37	Gun	7.74	Win	7.86
Free	8.25	Sex	7.60	Self	7.74
Hug	8.23	Lover	7.45	Smile	7.72
Delight	8.21	Nuclear	7.30	Project	7.63
Joy	8.21	Rape	7.24	Kind	7.62
Honest	8.16	Snake	7.24	Shower	7.47
Laughter	8.05	Thrill	7.19	Style	7.47
Lover	8.05	Fuck	7.14	Calm	7.44
Love	8.00	Scare	7.10	Strength	7.42
Prize	8.00	Attack	7.05	Winner	7.42

In return, table 6 displays the top 10 lowest scoring words from the lyrics on each of the three emotional dimensions. These words score the lowest on valence, arousal and dominance. A lower score on valence corresponds to “unhappy”, a lower score on arousal to “calm”, and a lower score on dominance to “controlled” (Warriner et al., 2013).

Table 6

Top 10 lowest scoring words in the ANEW/lyrics dataset

Valence		Arousal		Dominance	
Murder	1.48	Grain	1.60	Doom	2.42
Rape	1.54	Calm	1.67	Rape	2.47
Wreck	1.62	Dull	1.67	Bastard	2.59
Die	1.67	Quiet	1.95	Slave	2.61
Virus	1.71	Scene	1.95	Drown	2.65
Killer	1.74	Asleep	2.00	Panic	2.65
Stress	1.79	Tea	2.05	Dead	2.66
Kill	1.81	Level	2.15	Afraid	2.71
Death	1.89	North	2.15	Frighten	2.71
Worthless	1.89	Low	2.21	Worthless	2.71

3.3.4 Lyrics data

To calculate which words occur most across all songs, the lyrics data was grouped by ‘word’ after all words which did not have a score on the three emotional dimensions of the ANEW dataset were deleted. Table 7 shows the top 10 most frequent words across all songs in the combined dataset and their score on valence, arousal and dominance.

Table 7

Top 10 most frequent words in the lyrics/ANEW data

Word	Frequency	Valence	Arousal	Dominance
Do	582,114	5.41	3.67	6.35
Will	442,645	5.32	2.90	6.61
Be	406,074	6.18	3.43	5.50
Have	366,618	5.86	3.52	5.72
Love	361,020	8.00	5.36	5.92
Know	338,244	6.82	3.24	5.78
Like	277,630	7.44	4.40	6.28
Get	235,006	6.09	3.67	5.65
Can	232,350	6.41	3.14	6.44
Go	227,437	6.32	4.86	5.33

3.3.5 Combined dataset

To see which genres score the highest or lowest on each of the three emotional dimensions from the ANEW data based on the content of their lyrics, table 8 was constructed. This table provides information on the three highest and lowest ranking genres on arousal, valence and dominance.

Table 8

The lowest and highest ranking genres on arousal, valence and dominance

	Arousal		Valence		Dominance	
Highest score	Metal	4.20	RnB	5.91	RnB	5.64
	Rap	4.16	Jazz	5.86	Jazz	5.61
	Punk	4.15	Country	5.85	Country	5.59
Lowest score	World	3.98	Latin	5.04	Latin	5.24
	International	4.00	Metal	5.44	Metal	5.38
	Folk	4.02	Rap	5.56	Rap	5.47

3.3 Cross-validation

To evaluate the performance of the classification algorithm used in this study, k-fold cross-validation was applied. K-fold cross-validation aims to check how the classification method will perform on future unseen data. Generally, when the amount of data is large, k-fold cross-validation should be employed to estimate the accuracy of the model, because the accuracy resulting from the training data of the model is usually too optimistic (Wong, 2015). For example, when 80% of the data is used for training and 20% of the data as test set, it is possible that the test set happens to contain data points that are particularly easy (or hard) to predict. By using k-fold cross-validation, each data point ends up in the test fold exactly once.

Based on related scientific work concerning music genre classification algorithms, 10-fold cross-validation was applied in this study (Trohidis et al., 2008; Hu, Downie & Ehmann, 2009). The combined dataset was randomly partitioned into ten equal sized subsamples. Out of these subsamples, one single subsample was then set aside as the validation data for testing the model and the remaining nine subsamples were used as the training data. This process was repeated $k = 10$ times with each of the ten subsamples used once as the validation data. These ten results were then averaged to produce a single accuracy estimation. This way, all observations were used for both training and validation.

Since the purpose of k-fold cross-validation is model evaluation, the resampled models were discarded after performing the 10-fold cross-validation and extracting the averaged accuracy estimations. For the purpose of prediction, the final classification model was re-fitted on the whole dataset.

3.4 Imbalanced classes

When looking at the distribution of the classes in table 3, there is a clear variation between majority classes and minority classes. Such an unequal distribution of the dependent variable can influence the performance of the classification algorithm. It causes classifiers to get biased towards the majority class. In addition, classification algorithms are often accuracy driven and therefore they aim to minimize the overall error, to which the minority classes contribute very little. Multinomial logistic regression does not aim to maximize classification accuracy, but instead finds coefficients that minimize the cross-entropy cost, which approximates classification accuracy.

To deal with the unequal class distribution of the dependent ‘genre’ variable in this study, either oversampling or undersampling can be applied to make the training data more balanced (Weiss, McCarthy & Zabar, 2007). Oversampling is used to duplicate observations of the minority classes to obtain a more balanced dataset, while undersampling is used to drop observations for the same purpose. Oversampling and undersampling can be used to alter the class distribution of the training data and both methods have been used to deal with class imbalance (Weiss et al., 2007). This is not without risk. Random undersampling can potentially remove certain important examples, and random oversampling can lead to overfitting (Chawla, Japkowicz & Kotcz, 2004).

In this study however, 10-fold cross-validation was applied. The distribution of classes was therefore assumed to be roughly equal in each training and test split. This means that the distribution of classes was likely to not cause any problems in training the model. To be sure, both undersampling and oversampling were introduced to the classification model. In addition, SMOTE was also applied to sampling. This is a hybrid approach of both oversampling and undersampling. SMOTE applies oversampling in the minority classes and undersampling in the majority classes for better classifier performance (Chawla et al., 2002). Finally, the accuracy estimates of the models were compared to the model without any sampling corrections.

3.6 Multinomial logistic regression

Multinomial logistic regression is used to predict categorical placement in or the probability of category membership on a dependent variable based on multiple independent variables (Starkweather & Moske, 2011). Multinomial logistic regression uses maximum likelihood estimation to evaluate the probability of categorical membership (Starkweather & Moske, 2011). In principle, multinomial logistic regression works like a series of logistic regressions, each one comparing two classes of the dependent variable.

In this study, the reference category chosen was the genre “Pop_Rock”, because this is the largest class. This was done because the reference category is often the most frequent or the most common category. In multinomial logistic regression, the probability of being in any of the other categories is compared to the probability of being in the reference category. These relative probabilities are the predicted log odds (Sperandei, 2014). Separate odds ratios are calculated for all independent

variables for each category of the dependent variable with the exception of the reference category, which is omitted from the analysis. The exponential beta coefficient represents the change in the odds of the dependent variable being in a particular category versus the reference category, associated with a one unit change of the corresponding independent variable (Sperandei, 2014).

Multinomial logistic regression was applied as the classification algorithm in this study, because it is used in the study by Liang et al. (2011). The current study aims to replicate parts of their analysis by using the same classification method. In addition, multinomial logistic regression is suited to solve multiclass problems with more than two possible discrete outcomes. It performs well when categories are unordered, which the genre categories are in this case. One of the assumptions of multinomial logistic regression is that the odds ratio of any two categories are independent of all other response categories. The multinomial logistic model also assumes that the dependent variable cannot be perfectly predicted from the independent variables for any case. These reasons led to the choice of using multinomial logistic regression for this particular classification problem.

The process of conducting multinomial logistic regression was done with the use of the package “nnet” in R. When performing 10-fold cross validation, the multinomial logistic regression model was trained ten times on 90% of the complete dataset as training data, and evaluated on 10% test data. For each of the ten folds, their accuracy estimation was averaged into one single accuracy estimation. Then, a confusion matrix was constructed. With the help of this confusion matrix, a number of other performance metrics were calculated to evaluate the performance of the multinomial logistic regression classifier, such as precision, recall and F1 score. Precision is defined as the fraction of correct predictions for a certain class. Recall is the fraction of instances of a class that were correctly predicted. The F1 score is defined as the harmonic mean (weighted average) of precision and recall.

First, a multinomial logistic regression model was employed on both audio features and lyrics sentiment features. The audio features ‘Duration’, ‘Key Signature’, ‘Tempo’, ‘Loudness’, ‘Mode’ and the twelve dimensions on ‘Timbre’ and ‘Pitch’ were incorporated into the model, as well as the lyric sentiment features ‘Mean Arousal’, ‘Mean Valence’ and ‘Mean Dominance’. In this model 10-fold cross-validation was applied and no other sampling methods were used.

For the second model, the same multinomial logistic regression model was employed, but with undersampling in the training data. For the third model, oversampling was applied instead of undersampling. In the fourth model, the hybrid approach SMOTE was used as the sampling method. The accuracy scores of these three models were compared to the accuracy score of the first model.

In addition, separate multinomial logistic regression models were employed where one model only incorporated the audio features and the other model only used the lyric sentiment features as input. The accuracy scores of these two models were compared to the first model again. It was assumed that the complete model with both audio features and lyric sentiment features combined performed better than the models only using audio features or only using lyric sentiment features.

3.7 Evaluation metrics

For each multinomial logistic regression model, the accuracy, macro-averaged precision, recall and F1 score, micro-averaged F1 score and the Kappa statistic were calculated.

Accuracy is the ratio of number of correct predictions to the total number of input samples. Precision is the number of correct positive results divided by the number of positive results predicted by the classifier. Recall is defined as the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). The F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

Kappa is a metric ranging from 0 to 1 that compares the observed accuracy (OA) with the expected accuracy (EA) obtained if the prediction model follows random chance. It is calculated as $K = (OA - EA) / (1 - EA)$. The Kappa statistic measures how closely the instances classified by the model match the data labeled as ground truth, controlling for the accuracy of a random classifier. Thus, Kappa provides an accuracy score that is normalized by random chance. The random chance accuracy is 0 and scores above zero show an improvement over predicting by random chance.

3.7.1 Macro averaging vs. micro averaging

Precision, recall and F1 score can be averaged over all the classes, being the macro-averaged precision, recall and F1 score. Macro averaging gives equal weight to each class, whereas micro averaging will aggregate the contributions of all classes to compute the average metric. Because the F1 measure ignores true negatives and its magnitude is mostly determined by the number of true positives, large classes dominate small classes in micro averaging. Micro averaged results are therefore really a measure of effectiveness on the large classes in a test collection. To get a sense of effectiveness on small classes, macro averaged results are more appropriate (Manning et al., 2008).

In multiclass settings, micro-averaging produces the same precision, recall and F1 score. The micro-averaged F1 measure is also equal to the accuracy of the model. In this study, the macro-averaged precision, recall and F1 score were used, as well as the micro-averaged F1 score.

4. Results

4.1 Summary of results

In this section, the classification performance of various models on the complete dataset after performing 10-fold cross-validation is described. The first classification model is the multinomial logistic regression on both audio features and lyrics sentiment features. Then, the results of the same model but with undersampling, oversampling and SMOTE sampling applied are listed. Next, the results for the classification model with only audio features are provided. Lastly, the results for the classification model with only lyrics sentiment features is given. The results are displayed in table 9.

Table 9

Summary of results on multinomial logistic regression models

Model	Accuracy	Macro Precision	Macro Recall	Macro F1 score	Micro F1 score	Kappa
MLR Full model	0.415	0.242	0.142	0.187	0.415	0.155
MLR Full model + undersampling	0.190	0.166	0.299	0.161	0.190	0.129
MLR Full model + oversampling	0.193	0.173	0.326	0.166	0.193	0.145
MLR Full model + SMOTE	0.296	0.155	0.148	0.125	0.296	0.085
MLR + only audio features	0.400	0.224	0.147	0.187	0.400	0.148
MLR + only lyrics sentiment features	0.397	0.162	0.062	0.160	0.397	0.021

4.1.1 Baseline

In order to compare the results of the models listed in table 9, a certain baseline had to be established. In this study, the baseline classifier used is the Zero rule. This rule simply predicts the majority class in the dataset. If the classifier would only predict the majority class ‘Pop_Rock’, the predictions would be accurate 39.3% of the time. The accuracy score of the majority class ‘Pop_Rock’ is 0.393. The majority class metrics can be found in table 10.

Table 10

Majority class metrics

Accuracy	Recall	Precision	F1
0.39	1	0.393	0.564

4.1.2 10-fold cross-validation results

10-fold cross-validation was performed for each of the classification models. After performing 10-fold cross-validation, ten accuracy scores were produced for each of the ten folds. These accuracy estimates, the averaged accuracy score over the ten folds and the standard deviation for each of the six classification models can be found in table 11.

Table 11

10-fold cross-validation accuracy estimates

Fold	Full model	Full model + undersampling	Full model + oversampling	Full model + SMOTE	Audio features only	Lyrics sentiment only
1	0.413	0.184	0.195	0.273	0.406	0.397
2	0.406	0.198	0.204	0.256	0.397	0.396
3	0.409	0.198	0.185	0.298	0.400	0.397
4	0.418	0.184	0.215	0.269	0.408	0.397
5	0.413	0.186	0.210	0.261	0.404	0.396
6	0.422	0.179	0.196	0.282	0.408	0.395
7	0.412	0.203	0.224	0.310	0.403	0.401
8	0.411	0.191	0.224	0.326	0.406	0.393
9	0.409	0.201	0.190	0.257	0.407	0.397
10	0.422	0.175	0.183	0.238	0.411	0.395
Mean	0.414	0.190	0.203	0.277	0.405	0.396
SD	0.005	0.010	0.015	0.027	0.004	0.002

4.1.3 Full multinomial logistic regression model.

When looking at table 9, the accuracy, or the fraction of classes that are classified correctly, is 0.415 for the full model containing both audio features and lyrics sentiment features. This accuracy of 41.5% is higher than the 39.3% majority class accuracy estimate. This means that this classification model is only a slight improvement compared to the model only predicting the majority class.

The macro-averaged precision of 0.242 indicates that on average, 24.2% of all positive predictions across all classes were correctly classified as positive. The macro-averaged recall of 0.142 indicates that on average, 14.2% of all positive items (including false negatives) across all classes were actually classified as positive. And the macro-averaged F1 of 0.187 indicates that the weighted mean of the macro-averaged precision and macro-averaged recall is 18.7%. Since an F1 score reaches its best value at 1 (meaning perfect precision and recall) and its worst value at 0, this value of 0.187 can be considered somewhat low. In addition, the micro-averaged F1 score is included in table 9. The micro-averaged F1 score is equal to the accuracy estimate of the model. In this model, Kappa = 0.155, which means that the model provides a 15.5% improvement over predicting by random chance.

4.1.4 Undersampling, oversampling and SMOTE

Undersampling, oversampling and the SMOTE sampling methods resulted in an overall accuracy of 0.190, 0.193 and 0.296 respectively, as can be seen in table 9. Concluding from these results, the full model (accuracy = 0.415) performed worse when sampling corrections are applied. In addition to the lower accuracy estimates, the Kappa statistic of the undersampling, oversampling and SMOTE model was also lower than the Kappa statistic of the full model. Based on the accuracy scores and the kappa statistics, the full model performed better than the models where sampling correction was applied. Therefore, these sampling corrections were disregarded when performing the classification model on only audio features and the classification model on only lyrics sentiment features.

4.1.5 Only audio features and only lyrics sentiment features

The model where only audio features were incorporated produced an accuracy score of 0.400, while the full model incorporating both audio features and lyrics sentiment features had an overall accuracy score of 0.415. The macro-averaged F1 score for the audio features only model was the same as the F1 score for the full model. However, the Kappa statistic was slightly lower than the Kappa statistic of the full model. Based on the slightly higher values of the accuracy estimate and the Kappa statistic of the full model, it can be concluded that the model including both audio features and lyrics sentiment features performed better than the model with only audio features.

The model which only included the lyrics sentiment features produced an overall accuracy score of 0.397, which was slightly lower than the accuracy score of the full model. The macro-averaged F1 score of 0.160 was lower than the F1 score of the full model (0.187). In addition, the Kappa statistic for the lyrics sentiment features model was remarkably lower (0.021) than that of the full model (0.155). Based on the higher accuracy score and the higher Kappa statistic, the full model outperformed the lyrics sentiment features model.

Finally, likelihood ratio tests were performed to compare the goodness-of-fit between the audio features only model and the full model, and the lyrics sentiment features only model and the full model. Resulting from these tests, the full model fits the data best at a significance level of $p < 0.05$.

4.2 Classification performance across classes

To check if classification performance differs across classes, precision, recall and F1 scores per class were computed. The accuracy metric could be misleading because it could predict the dominant class most of the time and still achieve a relatively high overall accuracy but low precision or recall for other classes. This way, predictions made per class can be evaluated.

In addition, random weighted guess scores of the class proportions were added as a baseline to compare the class scores from the full model with their random weighted guess scores (table 12). In this study, the dependent classes were imbalanced which means that the probability of random guessing a class was not the same for all classes. Therefore, the weighted random guess precision, recall and F1 scores were equal to the class proportions.

If the per-class F1 scores were lower than the per-class random weighted guess F1 score, this would mean that the full multinomial logistic regression model was no better in predicting this class, than when predicting by random guessing would have been applied. By looking at table 12, per-class predictive performance was compared to random guessing F1 scores to examine which classes were not being classified properly by the classification model.

In addition, a confusion matrix visualizing the percentage of per-class predictions is provided in figure 4. The complete confusion matrix listing the number of predictions of the model can be found in Appendix II. In figure 4, the colors represent the percentage of instances classified, with steel blue being the highest percentage of instances classified and white being the lowest percentage of instances.

The confusion matrix in figure 4 shows that the majority of the classes are often misclassified as 'Pop_Rock', which is probably due to it being the majority class. When looking at figure 4, it is clear that 'Vocal' and 'World' were not classified at all, because the squares indicate 0.0 percentage of instances classified to these classes. The classes which have a low number of instances being classified correctly are 'Blues', 'International', 'New Age', 'Pop', 'Punk' and 'Reggae', which can be derived from the light color of the squares and the low numbers of percentages in the diagonal. If classes were correctly classified, the squares in the diagonal should be steel blue and the percentage should be higher.

Table 12

Summary of results per class on the full multinomial logistic regression model

Class	Precision	Recall	F1	Random weighted guess precision, recall and F1
Pop_Rock	0.448	0.790	0.572*	0.393
Blues	0.333	0.001	0.002	0.009
Country	0.216	0.068	0.103*	0.044
Electronic	0.465	0.186	0.266*	0.038
Folk	0.189	0.067	0.099*	0.020
International	0.000	0.000	-	0.003
Jazz	0.223	0.099	0.138*	0.014
Latin	0.192	0.197	0.194*	0.022
Metal	0.276	0.200	0.232*	0.049
New Age	0.093	0.021	0.034*	0.002
Pop	0.120	0.002	0.003	0.055
Punk	0.000	0.000	-	0.013
Rap	0.433	0.574	0.494*	0.044
Reggae	0.248	0.018	0.033*	0.016
RnB	0.271	0.142	0.187*	0.040
Rock	0.374	0.201	0.261*	0.231
Vocal	0.000	0.000	-	0.001
World	0.000	0.000	-	0.005

* F1 > random weighted guess F1

Based on table 12, there are four classes which were not predicted at all, being 'International', 'Punk', 'Vocal' and 'World'. These classes received a score of zero on precision and recall. The classes 'Blues' and 'Pop' received a lower F1 score than when their class would have been predicted by random guessing. The performance of the full classification model on the classes 'Pop_Rock', 'Country', 'Electronic', 'Folk', 'Jazz', 'Latin', 'Metal', 'New Age', 'Rap', 'Reggae', 'RnB' and 'Rock' was higher than the random weighted guess scores, which means that the full model better classified songs into these classes than when random guessing would have been applied.

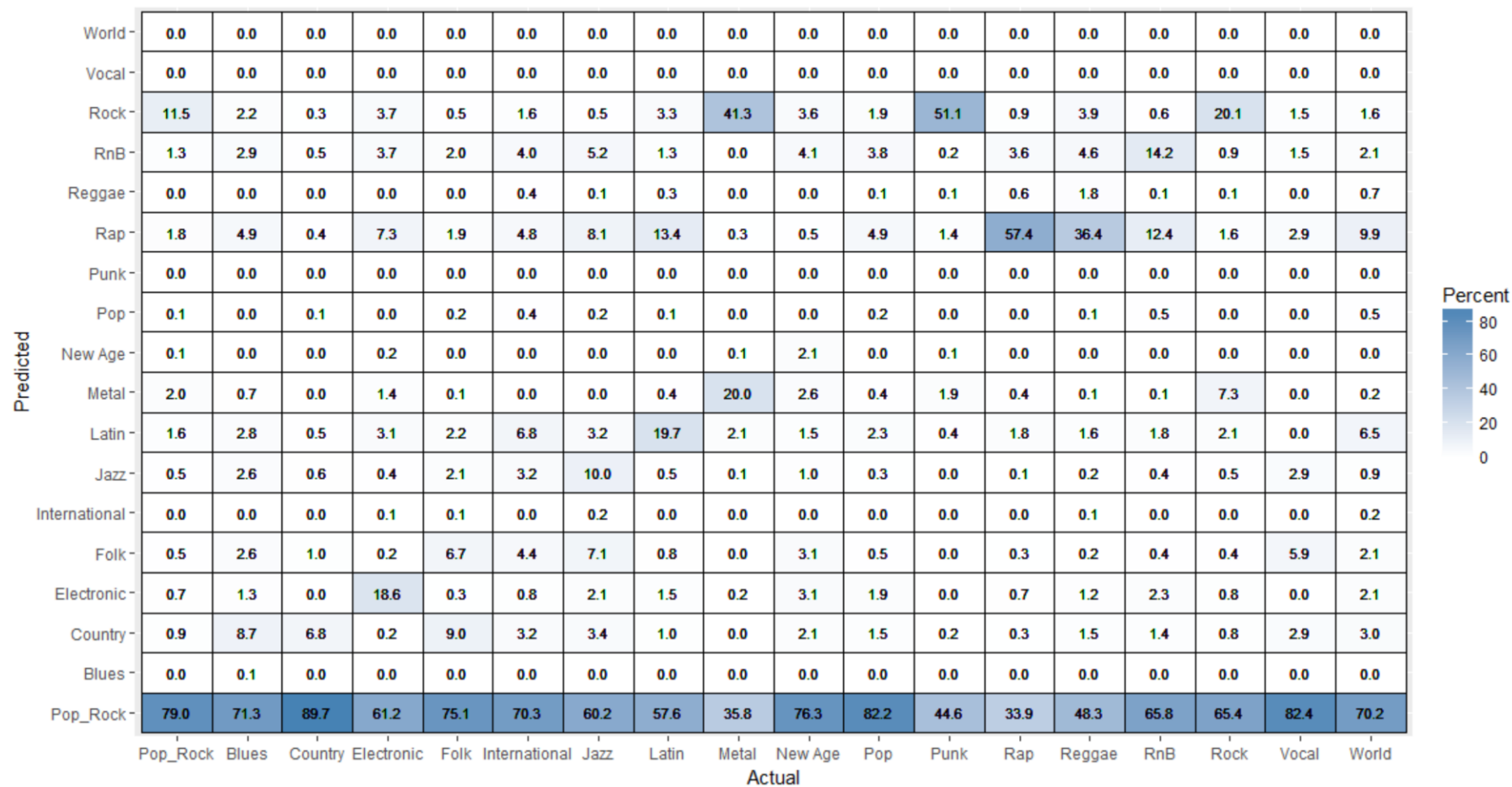


Figure 4.
Confusion matrix classification performance on whole dataset per class indicated in percentages

5. Discussion

The goal of this study was to examine if lyrics sentiment features contribute to higher music genre classification performance. In answer of this research question: a multinomial logistic regression model on a combination of both audio features and lyrics sentiment features yields the highest classification performance with an overall accuracy of 41.5%. In addition, the multinomial logistic regression model containing only lyrics sentiment features performed well on its own with an accuracy score of 39.6%. This leads to believe that including lyrics sentiment features does contribute to higher music genre classification performance. However, the best performing model includes both lyrics sentiment features and audio features.

Out of the various multinomial logistic regressions performed, the full model containing both audio features and lyrics sentiment features yields the highest performance. Within the full model, twelve out of the eighteen outcome classes were predicted more accurately than when random guessing was applied. The classes 'Blues' and 'Pop' received a lower F1 score than when their class would be predicted by random guessing, thus these classes were not predicted more accurately by the model. Furthermore, the classes 'Vocal' and 'World' were not predicted at all. 'Vocal', and 'World' each had a low representation in the dataset (< 500), so perhaps this had effected the model's ability to classify songs to these genres.

Some interesting findings emerged when looking at the confusion matrix in figure 4. The confusion matrix shows that all classes are often misclassified as 'Pop_Rock', which was probably due to it being the majority class. The class which was classified correctly most often is 'Rap'. The classification model classified 'Rap' songs correctly 57.4% of the time.

Another interesting finding for instance, is that 'Latin' songs were mostly classified into 'Pop_Rock' and 'Latin', but proportionally also to 'Rap'. This leads to wonder which characteristics 'Latin' songs have in common with 'Rap' songs. Unfortunately, there is no way to tell which features exactly caused 'Latin' to be sometimes classified as 'Rap'. When studying the summarized genre scores on each of the 34 features used in this study, 'Rap' and 'Latin' only scored similar on four features, being 'Duration', 'Valence', 'Dominance' and 'timbre0'. Both 'Rap' and 'Latin' scored around the mean on 'Duration', they both were in the bottom three lowest scores on 'Valence' and 'Dominance', and they scored very similar on the first dimension of 'Timbre'. After studying the coefficients in the model and transforming them to z-scores for interpretation, it became clear that 'Rap' and 'Latin' can be considered similar by the model based on their scores on the 'Valence' dimension. 'Rap' ($z = -10.0$) and 'Latin' ($z = -17.4$) produced the largest negative z-score coefficients on 'Valence' compared to the other genres. Their z-score coefficients on 'Dominance' did not deviate as much from the z-scores of the other genres. Therefore, it can be assumed that the model sometimes misclassified 'Latin' as 'Rap', because these genres both scored extreme on 'Valence'.

‘Metal’ got misclassified as ‘Rock’ a considerable number of times. However, this makes intuitively more sense since rock and metal songs can sound similar in a certain way. They scored almost the same on ‘Key Signature’, ‘Tempo’, ‘Loudness’ and multiple ‘Timbre’ and ‘Pitch’ dimensions. ‘Punk’ also got misclassified as ‘Rock’ over 50% of the time. ‘Punk’ and ‘Rock’ score almost the same on ‘Loudness’, ‘Valence’, ‘Arousal’, and ‘Dominance’. Moreover, ‘Reggae’ got sometimes misclassified as ‘Rap’. The only feature on which these genres scored similar is ‘Key Signature’. Otherwise, there is no way to tell what exactly caused the model to classify ‘Reggae’ songs as ‘Rap’ songs around 36% of the time.

5.1 Results of similar studies

Liang et al. (2011) employed multinomial logistic regression with L2 regularization on the Million Song dataset with audio features, lyrics bag-of-words features and lyrics sentiment scores based on the smaller version of the ANEW dataset by Bradley & Lang (1999). Their highest performing model produced an accuracy score of 38.6% which is roughly comparable with the present study. However, they indicated that lyrics sentiment features provide zero value when added to the model. This is interesting, because in the present study the model with only lyrics sentiment features yields an accuracy of 0.396. This is probably explained by the use of the relatively larger ANEW dataset by Warriner et al. (2013) which includes about 13 times more emotional scores than the ANEW dataset by Bradley & Lang (1999). Furthermore, the model including both audio features and lyrics sentiment features performs the best with an accuracy of 0.415. Therefore, it can be concluded that lyrics sentiment features do provide value when added to the model, contradicting the findings by Liang et al. (2011).

Hu & Downie (2010) performed mood classification among songs by combining audio features and lyrics features, among which emotional scores on the ANEW dataset, which they expanded by using WordNet. In this study, the combination of ANEW scores and text stylistic features on the lyrics alone achieved the second best performance among all feature types and combinations. They call this a surprising finding, because this model did not include any audio features. However, their best performing model included audio features as well as lyrics features. This is more in line with the findings in the present study. Including both lyrics sentiment features and audio features in the model increases music genre classification performance.

In addition, including lyrics sentiment features in the model produced some interesting results. As mentioned in the part above, due to incorporating lyrics sentiment features in the classification model, genres can be compared based on their emotional information extracted from the lyrics. Based on only audio features, ‘Latin’ songs and ‘Rap’ songs would not have much in common. However, concluding from the results it becomes clear that ‘Latin’ and ‘Rap’ were similar in the sense that the emotional information from their lyrics showed that they scored low on ‘Valence’. Interestingly, low

valence indicates unpleasantness. This would mean that ‘Latin’ and ‘Rap’ use the most unpleasant words in their lyrics out of all other genres included in this study.

5.2 Limitations

This study experienced some limitations, where the main limitation was probably the unequal distribution of the genre classes. The small number of songs belonging to the genres ‘World’ and ‘Vocal’ was probably the cause of their inability of being predicted by the model. In contrast, the majority class ‘Pop_Rock’ had such a large number of instances in this class that this caused most songs to be misclassified as ‘Pop_Rock’. In an attempt to counter these effects of imbalanced classes, undersampling, oversampling and SMOTE sampling was performed on the multinomial logistic regression models. However, this resulted in lower model performance than when no sampling methods were applied.

Another limitation could be the loss of valuable information when merging the datasets used in this study and deleting songs that did not have the necessary genre labels and lyrics information. Starting with one million songs, this was reduced to a little over 380,000 songs because songs that did not have a genre label were deleted from the dataset. Then, songs which did not have any lyrics were deleted as well. This resulted in a dataset of roughly 94,000 songs, which is about one tenth of the size of the original Million Song dataset. However, removing songs which did not have any information on genre and lyrics was necessary for comparability to reach unambiguous results after performing classification.

In addition, this study only performed multinomial logistic regression on the classification task at hand. No other classifiers were constructed to determine classification accuracy. It could be that other classification models would outperform the multinomial logistic regression models used in the present study. However, the reason why multinomial logistic regression was used in the present study was to replicate the method used in the study by Liang et al. (2011).

Lastly, the audio information present in the Million Song dataset is not exhaustive. It is possible that other features could help to improve music genre classification performance. In addition, this study has not included all available features of the Million Song dataset in the analysis. This was done because of theoretical support saying that the rest of the available audio features would not lead to improved classification performance.

5.3 Future research

Future research can benefit from including more songs if genre labels are present. Therefore, the genre label data needs to be extended. In addition, more lyrics sentiment information could have been included as well. In the present study, only 1,979 unique words from the lyrics with a score on the emotional dimensions by Warriner et al. (2013) were used, due to other words in the lyrics not having a score on

these emotional dimensions. Perhaps, future research could be focused on using the existing words in the ANEW dataset to derive emotional scores on words that are not present in the ANEW dataset but are similar to words that are present in this dataset. Bestgen and Vincze (2012) showed that emotional ratings of words can be used to automatically estimate the emotional values of new words by comparing them to validated words which are semantically related. Following this assumption, the emotional value of words in the lyrics used in this study which do not occur in the ANEW dataset could be automatically estimated based on other semantically related words which are present in the ANEW dataset. This way, word similarity by calculating some form of distance between words, can be used to generate emotional scores for more words within the song lyrics.

Future research should also direct attention to other audio features which are not currently present in the Million Song dataset. Other features could possibly contribute to increased genre classification performance. Currently, studies on music information retrieval are already directing their efforts towards song analysis on the audio segments level. This could lead to new findings which could be linked to the Million Song dataset. Then, there would be a greater number of audio features to perform classification on, hopefully resulting in better classification performance.

5.4 Scientific and practical implications

This study makes an important contribution to scientific research on music genre classification in the sense that lyrics sentiment features were successfully employed in the classification model. In contrast to other scientific work, adding lyrics sentiment features in the present study did improve classification performance. Though, including these lyrics sentiment features only increased the accuracy of the best performing model by 1.9% and can therefore not be considered groundbreaking. However, likelihood ratio tests showed that the goodness-of-fit of the model including audio and lyrics sentiment features improved significantly when lyrics sentiment features were added to the model ($p < 0.05$).

Genre similarity algorithms are important in recommending music. Automatic music recommendation lets people discover new songs that match their taste. By creating the best possible music genre classifier, genres can be more accurately predicted. Music recommendation systems rely on these genre similarity algorithms and therefore, a well performing classification model will contribute to a more enjoyable listening experience for people who use music recommender systems based on genre similarity. Since the overall accuracy produced by the best performing model in this study is only 41.5%, it can be assumed that online streaming services like Spotify employ better performing genre classification algorithms, and thus there is not much practical relevance tied to this study. However, it could also be used by music hobbyists who would like to quickly sort their music collection on genre. For such a scenario, an accuracy score of 41.5% is better than guessing the genre for each song by yourself.

6. Conclusion

The goal of this study was to examine if including lyrics sentiment features would lead to higher music genre classification performance. By including lyrics sentiment features based on emotional valence, arousal and dominance scores on the words in the lyrics, more valuable information can be used as input to classify songs into genre.

In an attempt to answer this research question, multinomial logistic regression models were employed on audio features, lyrics sentiment features, and a combination of both audio features and lyrics sentiment features. Since the distribution of the dependent classes in this sample was imbalanced, sampling methods such as undersampling, oversampling, and SMOTE sampling were used to correct for the imbalanced dependent classes. However, the classification models where sampling corrections were applied performed worse than the model including both audio features and lyrics sentiment features.

The model including both audio features and lyrics sentiment features resulted in the highest classification performance with an overall accuracy of 41.5%. This contradicts other findings by Liang et al. (2011) that lyrics sentiment features have no added value in classifying musical genres. Therefore, this study makes an important contribution to the research field of music information retrieval and music genre classification. However, further research needs to be conducted into the use of lyrics sentiment features. Future research should direct its efforts towards expanding the ANEW dataset in order to study if this will lead to improved music genre classification performance.

References

- Ahrendt, P. (2006). Music genre classification systems-a computational approach (Doctoral dissertation).
- Ajoodha, R., Klein, R., & Rosman, B. (2015). Single-labelled music genre classification using content-based features. In *Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, 66-71. doi: 10.1109/RoboMech.2015.7359500
- American National Standards Institute. Committee on Bioacoustics, S3, & Sonn, M. (1973). *American national standard psychoacoustical terminology*. American National Standards Institute.
- Aucouturier, J. J., & Pachet, F. (2002). Music similarity measures: What's the use?. *ISMIR*. 13-17.
- Bertin-Mahieux, T., Ellis, D. P., Whitman, B., & Lamere, P. (2011). The Million Song Dataset. In *Ismir*. 2(9), 10.
- Bestgen, Y. & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. Advance online publication. *Behavior Research Methods*, 44(4), 998–1006. doi: 10.3758/s13428-012-0195-z
- Bradley, M. M. & Lang, P. J. (1999). Affective Norms for English Words (ANEW): Stimuli, Instruction Manual and Affective Ratings. *Technical report C-1*. University of Florida.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357. doi: 10.1613/jair.953
- Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1), 1-6. doi: 10.1145/1007730.1007733
- Chiliguano, P., & Fazekas, G. (2016). Hybrid music recommender using content-based and social information. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference*, 2618-2622. doi: 10.1109/ICASSP.2016.7472151
- Cooper, G., & Meyer, L. B. (1963). *The rhythmic structure of music*. University of Chicago Press.
- de Leon, F., & Martinez, K. (2012). A music genre classifier combining timbre, rhythm and tempo models. In *TENCON 2012-2012 IEEE Region 10 Conference*. 1-5. doi: 10.1109/TENCON.2012.6412211
- Dodds, P. S., & Danforth, C. M. (2010). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of happiness studies*, 11(4), 441-456. doi: 10.1007/s10902-009-9150-9
- Geng, S., Ren, G., & Ogihara, M. (2017). Transforming Musical Signals through a Genre Classifying Convolutional Neural Network. *arXiv preprint arXiv:1706.09553*.
- Gossi, D., & Gunes, M. H. (2016). Lyric-Based Music Recommendation. *Complex Networks VII*. 301-310. Springer, Cham.

- Hu, Y., Chen, X. & Yang, D. (2009). Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR'09)*. 123-128.
- Hu, X., & Downie, J. S. (2010). Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th annual joint conference on Digital libraries*, 159-168. doi: 10.1145/1816123.1816146
- Hu, X., Downie, J. S., & Ehmann, A. F. (2009). Lyric text mining in music mood classification. *American music*, 183(5), 2-209.
- Jamdar, A., Abraham, J., Khanna, K., & Dubey, R. (2015). Emotion analysis of songs based on lyrical and audio features. *arXiv preprint arXiv:1506.05012*. doi: 10.5121/ijaia.2015.6304
- Laurier, C., Herrera, P., Mandel, M., & Ellis, D. (2007). Audio music mood classification using support vector machine. *MIREX task on Audio Mood Classification*, 2-4.
- Laurier, C., Grivolla, J., & Herrera, P. (2008). Multimodal music mood classification using audio and lyrics. In *Machine Learning and Applications, ICMLA'08. Seventh International Conference*, 688-693. doi: 10.1109/ICMLA.2008.96
- Li, T., Ogihara, M., & Li, Q. (2003). A comparative study on content-based music genre classification. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 282-289. ACM.
- Li, T., & Ogihara, M. (2006). Toward intelligent music information retrieval. *IEEE Transactions on Multimedia*, 8(3), 564-574.
- Liang, D., Gu, H., & O'Connor, B. (2011). Music genre classification with the million song dataset. *Machine Learning Department, CMU*.
- Logan, B., Kositsky, A., & Moreno, P. (2004). Semantic analysis of song lyrics. *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference*, 2, 827-830.
- Lops, P., De Gemmis, M., & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*, 73-105. Springer, Boston, MA.
- Loughran, R., Walker, J., O'Neill, M., & O'Farrell, M. (2008). The Use of Mel-frequency Cepstral Coefficients in Musical Instrument Identification. In *ICMC*.
- Lu, L., Liu, D., & Zhang, H. J. (2006). Automatic mood detection and tracking of music audio signals. *IEEE Transactions on audio, speech, and language processing*, 14(1), 5-18. doi: 10.1109/TSA.2005.860344
- Mandel, M. I., Poliner, G. E., & Ellis, D. P. (2006). Support vector machine active learning for music retrieval. *Multimedia systems*, 12(1), 3-13. doi: 10.1007/s00530-006-0032-2
- Manning, C. D., Raghavan, P., & Schu'tze, H. (2008). Introduction to Information Retrieval. Cambridge, UK: Cambridge University Press.
- Mayer, R., Neumayer, R., & Rauber, A. (2008). Combination of audio and lyrics features for

- genre classification in digital audio collections. In *Proceedings of the 16th ACM international conference on Multimedia*, 159-168.
- Mayer, R., Neumayer, R., & Rauber, A. (2008). Rhyme and Style Features for Musical Genre Classification by Song Lyrics. In *Ismir*, 337-342.
- Meng, A., Ahrendt, P., Larsen, J., & Hansen, L. K. (2007). Temporal feature integration for music genre classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5), 1654-1664. doi: 10.1109/TASL.2007.899293
- Nanni, L., Costa, Y. M., Lumini, A., Kim, M. Y., & Baek, S. R. (2016). Combining visual and acoustic features for music genre classification. *Expert Systems with Applications*, 45, 108-117. doi: 10.1109/ICMLA.2011.48
- Nanni, L., Costa, Y. M., Aguiar, R. L., Silla Jr, C. N., & Brahnam, S. (2018). Ensemble of deep learning, visual and acoustic features for music genre classification. *Journal of New Music Research*, 1-15. doi: 10.1080/09298215.2018.1438476
- Neumayer, R., & Rauber, A. (2007). Integration of text and audio features for genre classification in music information retrieval. In *European Conference on Information Retrieval*, 724-727. Springer, Berlin, Heidelberg.
- Nürnberger, A., Stober, S., Larsen, B., & Detyniecki, M. (2014). Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation: *10th International Workshop, AMR 2012, Copenhagen, Denmark, October 24-25, Revised Selected Papers*, 8382. Springer.
- Oh, S., Hahn, M., & Kim, J. (2013). Music mood classification using intro and refrain parts of lyrics. *Information Science and Applications (ICISA), 2013 International Conference*, 1-3. doi: 10.1109/ICISA.2013.6579495
- Oramas, S., Nieto, O., Barbieri, F., & Serra, X. (2017). Multi-label Music Genre Classification from Audio, Text, and Images Using Deep Features. *arXiv preprint arXiv:1707.04916*.
- Pálmason, H., Jónsson, B. Þ., Schedl, M., & Knees, P. (2017). Music genre classification revisited: An in-depth examination guided by music experts. *Proceedings of CMMR*.
- Pampalk, E., Flexer, A., & Widmer, G. (2005). Improvements of Audio-Based Music Similarity and Genre Classification. In *ISMIR*, 5, 634-637.
- Patterson, R. D., Gaudrain, E., & Walters, T. C. (2010). The perception of family and register in musical tones. *Music perception*, 36, 13-50. Springer, New York, NY. doi: 10.1007/978-1-4419-6114-3_2
- Plack, C. J., & Oxenham, A. J. (2005). Pitch perception: Neural Coding and Perception. In C. J. Plack, A. J. Oxenham, A. N. Popper, & R. R. Fay (Eds.), *Pitch: Neural Coding and Perception*, New York, NY: Springer.
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5), 513-523.
- Scaringella, N., Zoia, G., & Mlynek, D. (2006). Automatic genre classification of music content: a

- survey. *IEEE Signal Processing Magazine*, 23(2), 133-141. doi: 10.1109/MSP.2006.1598089
- Schindler, A., & Rauber, A. (2012). Capturing the temporal domain in echonest features for improved classification effectiveness. *International Workshop on Adaptive Multimedia Retrieval*, 8382, 214-227. Springer, Cham. doi: 10.1007/978-3-319-12093-5_13
- Sigtia, S., & Dixon, S. (2014). Improved music feature learning with deep neural networks. *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference*, 6959-6963. doi: 10.1109/ICASSP.2014.6854949
- Skowronek, J., McKinney, M. F., & Van De Par, S. (2007). A Demonstrator for Automatic Music Mood Estimation. *ISMIR*, 345-346.
- Sordo, M., Celma, O., Blech, M., & Gaus, E. (2008). The quest for musical genres: Do the experts and the wisdom of crowds agree?. *ISMIR*, 255-260.
- Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia medica: Biochemia medica*, 24(1), 12-18. doi: 10.11613/BM.2014.003
- Starkweather, J., & Moske, A. K., (2011). Multinomial logistic regression. Retrieved from <http://www.readbag.com/unt-rss-class-jon-benchmarks-mlr-jds-aug2011>
- Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. P. (2008). Multi-Label Classification of Music into Emotions. *ISMIR*, 8, 325-330.
- Tsaptinos, A. (2017). Music Genre Classification by Lyrics using a Hierarchical Attention Network.
- Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5), 293-302. doi: 10.1109/TSA.2002.800560
- Van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. *Advances in neural information processing systems*, 2643-2651.
- Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45(4), 1191-1207. doi: 10.3758/s13428-012-0314-x
- Weiss, G. M., McCarthy, K., & Zabar, B. (2007). Cost-sensitive learning vs. sampling: Which is best for handling imbalanced classes with unequal error costs?. *DMIN*, 7, 35-41.
- Wessel, D.L. (1979). Timbre space as a musical control structure. *Computer Music Journal*, 45-52.
- Wohlfahrt-Laymanna, J., & Heimbürgerb, A. (2017). Content Aware Music Analysis with Multi-Dimensional Similarity Measure. *Information Modelling and Knowledge Bases XXVIII*, 292-303.
- Wong, T. T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9), 2839-2846. doi: 10.1016/j.patcog.2015.03.009

Appendix I

Million Song dataset description list of all 54 features

Table 1

Million Song Data features and a brief description

Field name	Type	Description
analysis sample rate	float	sample rate of the audio used
artist 7digitalid	int	ID from 7digital.com or -1
artist familiarity	float	algorithmic estimation
artist hotttnesss	float	algorithmic estimation
artist id	string	Echo Nest ID
artist latitude	float	latitude
artist location	string	location name
artist longitude	float	longitude
artist mbid	string	ID from musicbrainz.org
artist mbtags	array string	tags from musicbrainz.org
artist mbtags count	array int	tag counts for musicbrainz tags
artist name	string	artist name
artist playmeid	int	ID from playme.com, or -1
artist terms	array string	Echo Nest tags
artist terms freq	array float	Echo Nest tags freqs
artist terms weight	array float	Echo Nest tags weight
audio md5	string	audio hash code
bars confidence	array float	confidence measure
bars start	array float	beginning of bars, usually on a beat
beats confidence	array float	confidence measure
beats start	array float	result of beat tracking
danceability	float	algorithmic estimation
duration	float	in seconds
end of fade in	float	seconds at the beginning of the song

energy	float	energy from listener point of view
key	int	key the song is in
key confidence	float	confidence measure
loudness	float	overall loudness in dB
mode	int	major or minor
mode confidence	float	confidence measure
release	string	album name
release 7digitalid	int	ID from 7digital.com or -1
sections confidence	array float	confidence measure
sections start	array float	largest grouping in a song, e.g. verse
segments confidence	array float	confidence measure
segments loudness max	array float	max dB value
segments loudness max time	array float	time of max dB value, i.e. end of attack
segments loudness max start	array float	dB value at onset
segments pitches	2D array float	chroma feature, one value per note
segments start	array float	musical events, ~ note onsets
segments timbre	2D array float	texture features (MFCC+PCA-like)
similar artists	array string	Echo Nest artist IDs (sim. algo. unpublished)
song hottnesss	float	algorithmic estimation
song id	string	Echo Nest song ID
start of fade out	float	time in sec
tatums confidence	array float	confidence measure
tatums start	array float	smallest rythmic element
tempo	float	estimated tempo in BPM
time signature	int	estimate of number of beats per bar, e.g. 4
time signature confidence	float	confidence measure
title	string	song title
track id	string	Echo Nest track ID
track 7digitalid	int	ID from 7digital.com or -1
year	int	song release year from MusicBrainz or 0

		Actual								
Predicted		New Age	Pop	Punk	Rap	Reggae	RnB	Rock	Vocal	World
	Pop_Rock	148	4260	563	1413	743	2504	14268	56	304
	Blues	0	0	0	0	0	0	0	0	0
	Country	4	79	2	11	23	52	179	2	13
	Electronic	6	96	0	30	19	86	170	0	9
	Folk	6	26	0	11	3	16	91	4	9
	International	0	0	0	1	2	0	6	0	1
	Jazz	2	14	0	4	3	16	108	2	4
	Latin	3	120	5	75	25	70	463	0	28
	Metal	5	19	24	17	1	2	1583	0	1
	New Age	4	0	1	0	0	0	9	0	0
	Pop	0	9	0	2	1	20	5	0	2
	Punk	0	0	0	0	0	0	0	0	0
	Rap	1	252	18	2394	560	471	350	2	43
	Reggae	0	7	1	27	27	3	13	0	3
	RnB	8	198	2	151	70	541	206	1	9
	Rock	7	100	645	38	60	22	4380	1	7
	Vocal	0	0	0	0	0	0	0	0	0
	World	0	0	0	0	0	0	0	0	0