Separating Music Consumers Through Lyrics

Sven van den Beukel
Faculty of Exact Sciences, VU University
De Boelelaan 1081
1081 HV Amsterdam, The Netherlands
sbl530@student.vu.nl

ABSTRACT

This study investigates whether differences in the lyrical content of songs in the listening history of users, can be used to predict their gender, nationality and age. The results show that extracting song topics and stylistic features (e.g. the number of unique - and function words used, or the sentiment polarity value) can be used by classifiers to separate user groups. The proposed features are most useful for predicting the gender of a user, but also help predict its age and nationality. The experiments described in this paper were designed to gather meaningful results that can be compared to the State-of-the-Art on user trait prediction using music listener events[22]. Increased knowledge of lyrical preferences of specific demographic user groups, could improve the accuracy of recommender systems, and help them serve new users and market new music pieces to their audience more effectively.

1. INTRODUCTION

Arguably the biggest challenges in the music industry are helping the customers search, sort, filter and share the music of their preference. A recommender system (RS) is a system designed to help users do these sort of things with any type of online content [30], effectively helping decrease the information overload experienced by consumers [14]. The most commonly used approach used in RSs is collaborative filtering, which recommends items to an active user, based on items that other users with similar tastes have liked in the past (or: user X likes A, B and C, so user Y who likes A and C probably also likes B) [34]. However, collaborative filtering approaches suffer from the "cold-start problem", which means that they cannot effectively recommend music to new users, as it is not possible to identify similar users without a listening history of significant size [34]. In a similar manner, it is just as impossible to effectively recommend newly released songs to users because they do not feature in anyone its listening history yet. To overcome these obstacles, hybrid systems are required that can recommend music based on knowledge about the songs and users. This study investigates whether differences exist in the music consumption behaviors of demographic user groups, focusing specifically on topics and writing styles extracted from song lyrics. Ultimately, if these differences are found to be significant among demographically separable user groups, this knowledge can be used in a hybrid RS to help reduce the cold-start problem. The actual realization of such a system lies outside of the scope of this paper.

Previous work on the automatic content-based clustering of music, has focused largely on similarities in audio-based features [28, 38, 43]. However, recent work has revealed that there is a "glass ceiling" in spectral-based Music Information Retrieval (MIR), because a lot of high-level (e.g. stylistic or semantic) music features cannot be represented using spectral-only techniques [32]. As a result, non-audio based approaches have recently regained attention, with researchers extracting meta data-based features such as artist mainstreaminess [2, 35] or text-based features such as TF-IDF scores [31]. Furthermore, Bag-Of-Words and POS-tag ratios have been successfully used for music mood classification [18] and might therefore also be useful for lyric-based MIR.

Some researchers reported that music subject classifiers trained on user-interpretations of songs outperform the classifiers that only use song lyrics [9], yet this approach is unlikely to work for songs that are still quite unknown to the big public, as there are little user interpretations available for these songs. This approach therefore might result in a bias towards songs that are already mainstream, effectively reducing its scalability. More scalable approaches that have gained traction over the past decade, require derivation of features from music lyrics and using these for example for categorizing music pieces based on some measured similarity. Studies have found that combining lyrical content features (such as tf-idf Bag of words) with audio features significantly improves the performance at tasks such as music mood classification [18, 24], music emotion classification [29] and music genre detection [31, 27]. Others have used the lyrics more in-depth by exploring the usefulness of lyrical stylistic features for music mood classification, concluding that they can significantly help improve systems that use only audio features, for several moods [17]. However, these features have not yet been analyzed for differences between demographic user groups, nor were they deployed on a dataset as large as the LFM-1B dataset.

The State-of-the-Art research that is most comparable to this study, has investigated whether the music listening history of users on social music platform last.fm can be used to predict their age, gender, and nationality [22]. The researchers used Term Frequency-Inverse Document Frequency (TF-IDF) on artist listening information as well as artist tags, and combined this with 42 additionally extracted features that are available in the used LFM-1B dataset (for example artist or genre mainstreaminess, average number of listening events per day of the week) [35]. This study extends on the research by Krismayer et al. [22] by using the same LFM-1B data subset and the 42 originally provided features as a starting point. In addition, lyrics are gathered for all songs that occur in the listening events, which are then used to extract features that capture the topics and writing style of songs. Knowing that user traits such as age and gender can be predicted by the contents and writing style of social media posts written by a user [36], the style of texts consumed by these users can be expected to differ as well. The classification results achieved with the proposed extended set of features is compared to the State-of-the-Art results [22].

Research questions

At the end of the final paper, the following research questions need to be answered.

1. What correlations exist between user demographics and song topics?

2. What correlations exist between user demographics and a song its writing style?

3. Can topics and text style features that are extracted from lyrics, be used to improve on the classification accuracy reported in the State-of-the-Art [22]?

Answering the first research question results in knowledge about whether, and to what extent, song topics can be used to classify the gender, age and nationality of users. The answer to the second research question tells us whether the extracted stylistic features are useful for user trait classification, and which ones are more useful than others. Thirdly, the classification results are evaluated and compared to the State-of-the-Art so that the usefulness of the proposed features can be evaluated. The answers to these questions help evaluate whether the used features help reduce the ”cold start”-problem in RSs based on Collaborative Filtering (CF). Finally, it further explores whether neural networks can improve results in this domain.

The next section explores related work in the field of Music Subject Classification (MSC) and usertrait prediction. Subsequently, an approach is proposed for extracting the most useful features to represent the topic and writing style for each song. This section includes rationales for the deployed topic modeling approach and for the extracted stylistic features, as well as a complete list of the features that are eventually used as input for user trait prediction. Next, the evaluation section discusses the setup of the performed experiments. Section 5 elaborates on the datasets that were used, how they were gathered and what data preprocessing was required. Subsequently, the results are presented and interpreted, after which the research questions are answered and the implications of this study are discussed in the Discussion.

2. RELATED WORK

In this section we discuss work related to music subject (topic) classification and the prediction of user traits.

2.1 Music Subject Classification

The field of Music Information Retrieval (MIR) is concerned with *”the extraction and inference of meaningful features from music (from the audio signal, symbolic representation or external sources such as web pages), indexing of music using these features, and the development of different search and retrieval schemes (for instance, content-based search, music recommendation systems, or user interfaces for browsing large music collections)”* [11]. Within this field, the first part of this study focuses on the task of classifying the subject of pieces of music using textual features, commonly referred to as Music Subject Classification (MSC). The work done on MSC is scarce, partially due to the absence of freely available lyrical datasets as a result of copyrights being spread among a large number of different copyright holders. Some attempts at improving MSC found that systems using user interpretations of songs performed better than systems using just lyrics [9, 7]. Earlier work explored the use of semi-supervised topic clustering by combining Term Document Matrices with non-negative matrix factorization [21]. Although the latter reports significant annotator agreement($p \geq 0.05$) on the most relevant tags for 31 out of 60 clusters, it remains unclear how good the actual clustering was. To our knowledge, no attempts have been made to analyze variety in music subject preferences in users of different ages, gender and nationalities.

2.2 User trait prediction

The idea of using text to predict user traits is not new. Previous attempts range from the prediction of an internet user its gender using text on teen blogs [19] to the prediction of its age using its facebookposts[20]. This idea has also found its way to the research field of MIR. Previous age, gender and nationality classification attempts using the listening events from music platform Last.fm, have used inputs varying from (i) music meta-data alone [45], (ii) a combination of timestamps, song and artist metadata and audio-signal features of songs [25] to (iii) TF-IDF for artist listening information and artist tags [22], combined with additionally extracted features that are included in the LFM-1B dataset [35] (e.g. on artist mainstreaminess or the user listening activity per hour of the day). Previous work on the use of personal user characteristics to predict music preference suggests that age has less impact on music selection than gender, which in turn has less impact than nationality [6]. This study aims to increase the understanding of the correlation between the writing style and topics found in song lyrics on one hand, and user demographics such as age, gender and nationality on the other hand. To our knowledge, no similar attempts have been done before.

3. METHODS

The main idea of the proposed approach is to understand what song lyrics are about, as well as how the style of the lyrics of the songs correlates with demographic user characteristics. The focus lies on what the content and style of the lyrics do with users of specific genders, nationalities and ages. This section elaborates on the features that are derived before running the experiments. First, an unsupervised approach is proposed for finding topics in songs. Next, we discuss the approach for extracting stylistic features that we expect to be preferred among certain groups of music consumers.

3.1 Topic Modeling: Latent Dirichlet Allocation

Topic Modelling was implemented in Python through Latent Dirichlet Allocation (LDA). LDA is a generative probabilistic model of a text corpus, in which documents are represented as random mixtures of latent topics and in which each topic is represented as a distribution over words [4]. It is an unsupervised method that can be deployed to find and observe the topics present in a corpus of text documents. In its simplest explanation, it does so by finding terms (words) that co-occur often among multiple documents, but that rarely occur in others. The assumption made about these words is, that they represent a certain topic the document is about, and that they are unlikely to occur in document written about topics other than this one. Each topic is thus represented by a set of words that repeatedly co-occurred in documents.

The parameters used in LDA are the Alpha and Beta hyperparameters and the number of iterations (I). Lets first discuss the Alpha and Beta hyperparameters. The first represents document-topic density, the second represents topic-word density. As the value of Alpha rises, documents get composed of more topics. Lowering the Alpha-value reduces the number of topics per document. Secondly, raising the value for Beta results in topics being composed of a higher number of words from the corpus. Lowering the value of beta, causes topics to be composed of less words.

The alpha parameter concerns the number of topics to be extracted from the corpus. One of the most common approaches for determining the optimal number of topics, is to use the Kullback Leibler (KL) divergence score [1]. In short, KL divergence calculates the difference between two distributions p and q [10]. The KL distance measures the inefficiency of assuming that the distribution is q, rather than the actual true distribution p. Although there are several effective ways of using KL divergence to optimize the parameters, recent studies have shown that online Variational Bayes (VB) outperform previously known methods in speed, without sacrificing on performance [16, 40]. Therefore, this method is also implemented in the experiments performed for this study. Python package Gensim[1] was used to implement the online VB algorithm as proposed by Hoffman et al. [16].

Next, there is the number of topic terms (beta), which is the number of terms present in each topic. The value chosen for this parameter depends on the requirement. If the main focus lies at extracting themes or concepts, it is recommended to pick a higher number. If on the other hand the problem concerns extracting features or terms, it is recommended to pick a lower number. As we are mainly interested in extracting features, this parameter is expected to require a lower number in this study.

The final parameter is the maximum number of iterations allowed to the LDA algorithm to converge. Setting this parameter too low results in a set of sub-optimal topics. The final parameters were determined through experiments and optimized to consist of sensible topics without overlapping terms, and the maximum number of iterations was chosen roughly at convergence point. The final setup can be found in Table 1. The topic model is provided in Appendix A.

Alpha	Beta	I
5	10	500

Table 1: LDA parameters

3.2 Text style features

The features used to quantify the text style of the lyrics, are split in the categories (i) text statistics, (ii) sentiment polarity and (iii) rhyme and alliteration. All features and approaches are the result of a thorough literature research of related fields (e.g. music genre and mood classification). All the used features (including topic modeling) are summarized and separated by category in Table 2. Eventually, the style-feature array for each user consists of the mean value and standard deviation of each feature over all English songs it listened to (and we got the lyrics for).

3.2.1 Text statistics

The first category of features are ones that concern statistics derived from simple counts of (parts of) the text. This category includes the easier obtainable features, such as the number of unique words divided by the number of words in a song lyric (Unique Word Ratio) [27], the number of function words (also referred to as stopwords) divided by the total number of words in the lyric (Function Word Ratio) [18] and the number of song title mentions in lyrics [26]. The function words used, are those listed by Python Natural Language Processing library NLTK[2] as stopwords.

3.2.2 Rhyme and Alliteration

Rhyme-based features were found to be usable for music genre classification [27]. Although the authors did not include alliteration features in their study, they did suggest that it could be useful future work. The rhyme/alliteration features that are extracted and used in this study are i) the unique rhyme and alliteration words used as a fraction of all rhyme and alliteration words ii) the percentage of lines that contain alliteration/rhyme and iii) the number of Rhymes-AA, Rhymes-AABB, Rhymes-ABBA and Rhymes-ABAB schemes in a song, of which the occurrence frequencies were found to be very different between genres. Therefore, there might also be differences in the demographics that appreciate these schemes.

A phoneme dictionary is required for recognition of rhymes and alliteration, since sometimes letters do not rhyme even though they are written the same. For example, there is no rhyme in ”harass” vs. ”brass”, even though both words end with the same four letters. On the other hand, there are also words that are written differently that do rhyme with each other, such as ”tie” and ”why”. The same holds for alliterations; the letters ”c” in the words ”chill” and ”club” do not alliterate, even though they start with the same letter. A phoneme dictionary resolves both these issues. For this task, the Python library CMUdict[3] is used, which contains

[1] https://radimrehurek.com/gensim/models/ldamodel.html

[2] www.nltk.org

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

phonemes for over 100.000 English words.

3.2.3 *Sentiment*

Sentiment polarity is the total sentiment score of a sentence. In this study, the Senticnet 4 package for Python[5] is used. The sentiment intensity scores ranging from very negative (-1) to very positive (1), are used to calculate the total sentiment polarity of a sentence. A song containing both positive and negative text with similar intensity, might thus result in a neutral score. Previous findings suggest that this average of the positive and negative sentiment score is the most informative and thus most useful sentiment score feature for mood classification of songs [23]. In order to capture the mere presence of sentiment in the text as well, another feature is introduced using only absolute values (negative values are added as if positive). This feature represents the absolute sentiment polarity found in a song.

It has been suggested that sentiment analysis for regular text documents does not necessarily perform well on song lyrics [46], but the main concern (negations not being considered) is resolved in this study through using Senticnet, which does involve surrounding words. Therefore, it is assumed that the current approach results in useful sentiment representations for the songs.

Table 2: Recap of the used features

Category	Feature(s)
Text statistics	1. Unique Word Ratio (UWR) 2. Function Word Ratio (FWR) 3. Song title mentions (STM)
Sentiment	4. Sentiment polarity value (SPV) 5. Absolute sentiment polarity (ASP)
Rhyme and Alliteration	6. Fraction of Unique Rhyme words (FUR) 7. Fraction of Unique Alliteration words (FUA) 8. Percentage of lines containing rhyme (PLR) 9. Percentage of lines containing alliteration (PLA) The number of: 10. Rhymes-AA schemes (RAA) 11. Rhymes-AABB schemes (RAABB) 12. Rhymes-ABBA schemes (RABBA) 13. Rhymes-ABAB schemes (RABAB) 14. Alliteration-AA schemes (AAA) 15. Alliteration-AABB schemes (AAABB) 16. Alliteration-ABBA schemes (AABBA) 17. Alliteration-ABAB schemes (AABAB)
Topic Modeling (LDA)	18. Topic 1 19. Topic 2 20. Topic 3 21. Topic 4 22. Topic 5

4. EVALUATION

Although it cannot be guaranteed that the extracted topics and stylistic features are optimal, all effort was put in place to implement (in Python 2.7) the best practices reported in relevant literature. The evaluation of the experiments take place two-fold. Firstly, we apply statistical (T-)tests to measure whether differences in topics and writing styles correlate with differences in user traits. Secondly, several classification and regression algorithms are deployed to find the optimal ones for the tasks of predicting the users' gender, age and nationality. A Java-based software package for machine learning called Weka[13] is used to test the predictive power of the LDA-generated topics, textstyle-features and LFM-1Bs additional features on their own respectively, as well as the combination of all three. Although it is expected that the complete featureset contains the most information and thus achieves the best prediction performance, this allows us to compare the predictive value of the newly proposed style-features and topics.

4.1 T-tests

The first step of assessing whether the newly proposed features hold informative value, is to run T-tests on the groups means for each style- and topicfeature. This should present some preliminary insight into the features that appear to be most different between two selected groups. On visual inspection, all extracted features seem to be normally distributed. Three of these T-tests are ran on three group-pairs, each representing one target variable (gender, age or nationality). Some noise comes from the 121 users for whom no song features could be extracted, either because they listen only to non-English music or because they listen to very unpopular songs. Non-english natives can be expected to have slightly lower feature values, as the odds are greater that they do not listen to any English songs, and thus contain more users with zero-vectors. Therefore, the first T-test is a comparison of Russian users versus non-Russian users. Using Russians rather than Americans, should reduce the chance of a bias that positively skews the results. We found that the group of Russian users is big enough to contain a lot users that have listened to at least one English song, and a small enough subset of non-English natives to have enough of them remain in the group of non-Russians. The other two groups that were T-tested are separated on gender (males versus females) and age ('Adolescents' aged 24 or younger, versus 'Adults' aged 25 and older). These age groups have previously been used in age classification using a different Last.FM user dataset [25], and was repeated on the LFM-1B dataset with competitive results as well [22].

4.2 Machine Learning

Secondly, the features are used by machine learning algorithms to evaluate their (combined) predictive power. 5-fold cross validation is deployed to train the classifiers on all datapoints and to use each datasample at least once for testing. The selection of machine learning algorithms deployed in this study to classify the users' gender, age and nationality, are based on the State-of-the-Art [22]. The three algorithms that performed best on the prediction of at least one of the classes, were the Support Vector Machine (SVM) classifiers (optimized through Sequential Minimal Optimization, or SMO [33, 47]) with RBF kernel (lowest error on age prediction) and polynomial kernel (most accurate on balanced gender prediction), as well as the simple logistic regression (highest accuracy on country prediction). The age of a user is predicted using regression, which means that a specific SMO algorithm optimized for regression is required (SMOreg [37, 39]). In addition to the ones used in the State-of-the-Art, the performance of Multilayer Perceptrons (MLP) are evaluated, as different types of neural networks have been reported to achieve high accuracy on related problems, such as music mood and genre prediction[8]. This MLP is optimized using the Iterative Classifier Optimizer in Weka. Neural networks have successfully outperformed more classical machine learning algorithms such as Naive Bayes and SVMs in related MIR tasks, such as music recommendation using audio features [44] or music genre classification [41]. To our knowledge, neural networks have not been tested before on the task of user trait prediction.

4.3 Baseline results

Originally, we purposefully selected the user subset provided in the State-of-the-Art paper [22] to perform our experiments on, so that our results could be compared against theirs. However, as the first two rows in Table 3 show, the Majority Vote (MV) baselines for our dataset are different from those of the State-of-the-Art. MV baselines are either A) in the case of classification, equal to the relative frequencies at which the most common label per class occur or B) in case of regression, the error when predicting the average age for all instances. Further inspection revealed that the user subset published in the State-of-the-Art is incompatible with the dataset characteristics described in the paper. For country, the most common label (US) occurs in 27.30% (not 19%) of the cases. The other frequencies are also not the same as the reported ones; the dataset contains 11.45% Russians (not 8.9%), 9.36% Germans (not 8.4%), 8.75% Englishmen (not 7.8%), 8.10% Poles (not 7.8%), 8.09% Brazilians (not 7.9%), and only 2.2% Dutchmen (not 2.8%). This means that especially Americans and Russians are overrepresented in this subset in comparison to the complete LFM-1B dataset. The same irregularities show up for gender (79.1% males instead of 72.50) and age (average of 24.9 rather than 25.6). It is unclear what user subset the authors have actually used, but it seems very unlikely that it is the one they published.

Because of this unlikelihood, the results of this study are best compared to the MV-baseline for our dataset. However, to put our results into perspective, we still report the results obtained by the State-of-the-Art. For gender classification, the authors achieved an accuracy of 77.06% (with a standard deviation of 0.24) on a balanced subset of the data and 81.36% on the unbalanced dataset. The baseline Mean Absolute Error for age is 4.13 and the baseline accuracy for nationality classification is 69.37%. All baselines can be found in Table 3.

Baseline	**Age**	**Gender (U)**	**Gender (B)**	**Country**
Majority Vote	3.85	79.11%	50.00%	27.30%
Majority Vote State-of-the-Art	6.23	72.50%	50.00%	19.03%
Krismayer et al.[22]	4.13	81.36%	77.06%	69.37%

Table 3: Baseline performance

5. DATA

5.1 Data Gathering & Selection

5.1.1 User data

In order to study different music preferences in user groups and to find possible correlations with their demographic traits (such as age, country and gender), a dataset is required that contains both user listener events and user characteristics information. The biggest such dataset available, is (to our best knowledge) the LFM-1b set[4], which contains over a billion listener events from approximately 120,000 users of streaming website last.fm [35]. Of this set, a user subset was selected that contains only those users of which age, gender and nationality are known and of which we have at least 500 listener events. More elaborate discussion of the data and how it was selected, can be found in the State-of-the-Art study [22]. The remainder subset contains 12,181 users and 6,736,824 listener events. For the task of gender classification, another subset needs to be created containing an equal amount of females and males, in order to prevent overfitting as a result of under-representation of females. A balanced dataset containing an equal number of instances for each gender was created by taking all female users (2545 instances) and extending this set with a user set containing an equal amount of randomly selected male users. Instead of 5-fold-classification, we chose to use 10-fold classification for the balanced gender subset to increase the available instances to train the classifiers on.

5.1.2 Song selection

Due to the absence of the computational resources required to gather lyrics for all songs listened to by the subset of users, features were extracted only for the top 10,000 they most commonly listened to. An additional top 5,000 songs were identified for listeners that had less than 10 of the previously gathered 10,000 songs in their listening history, to increase the number of usable listening events for these users.

5.1.3 Lyrics

Next, several datasets containing lyrics for a significant number of songs are required. As we need the lyrics to be complete and unprocessed in order to be able to detect some of the proposed stylefeatures (rhyme and alliteration, song title mentions, sentiment polarity), datasets containing bag-of-word representations such as the Million Song Dataset (MSD) [3] or the LyricFind dataset [12] were of no use to us. A dataset containing unprocessed lyrics for half a million songs can be found on machine learning platform Kaggle[5] (from hereon referred to as Kaggle Lyrics 1, or KL 1). A second Kaggle Lyrics dataset (KL 2)[6] was added to both speed up the lyric scraping process and to increase the odds of finding lyrics for each song. This dataset is assumed to speed up the process as it is a smaller dataset, which is expected to contain mostly popular songs and thus iterating over this set first should increase the speed of lyric gathering. If lyrics were not found in KL 1 or 2 for songs that were nonetheless popular among the selected last.fm users, these were obtained by webscraping Lyricfinder[7]. Songs for which less than 65% of the unique words were in the English Wordnet dictionary were judged to be non-English. A manual check of 1,000 randomly selected songs revealed that setting this threshold at 65% resulted in 0 false positives (songs that

[4] http://www.cp.jku.at/datasets/LFM-1b/

[5] https://www.kaggle.com/artimous/every-song-you-have-heard-almost

[6] https://www.kaggle.com/mousehead/songlyrics

[7] www.lyricfinder.org

were judged to be English, but actually were not). Lowering the percentage further (e.g. 60%), resulted in an increase in the number of false positives. No features were extracted for non-English lyrics, since both topics (due to stopwords being English) and style features depend on language-specific properties. Some words may exist across different cultures and be used in different contexts or hold different meanings, whose inclusion would introduce noise in the topic model. Concerning the style-features, a song that is only partially English can be expected to have a high unique word ratio, there is no way to accuractely determine the function word ratio or sentiment if the language is unknown. Moreover, the pronunciation used for detecting rhyme and alliteration only works for English words, as each language is spoken using different phonemes. Ultimately, lyrics were extracted for 9,372 songs, which resulted in a set of style-features for 9197 songs. This means that 175 songs were judged to be non-English.

5.2 Data preprocessing

Porter stemming was implemented to improve the LDA topic modeling, as there are some genres (e.g. hiphop) where the same words are spelled in many different ways, often cutting off, or changing the end of the word (e.g. killing becomes killin'). The standard NLTK stopword set was used for stopword removal.

6. RESULTS

In this section, the experimental results are shown that are most relevant for answering the research questions. First, the results from T-tests between several demographically separable groups of users are presented. Next, the results from the experiments on gender and nationality classification, as well as age regression, are reported. The Weka models that were used are published in the GitHub repository following the link in Appendix C for reproducability purposes.

6.1 T-tests

The complete table containing T-tests can be found in Appendix B.

Firstly, the results for the Russian versus non-Russian group mean comparison per feature, reveal that 13 feature means are different at a significance level of P <= 0.05. This suggests that users with the Russian nationality have music tastes that significantly differ from the group of users that are not Russian for these features (See the boldprinted features in Table 9). For example, the most significant difference is found in the sentiment polarity value, which is much lower for Russians than for the overall population, while the Absolute Sentiment Polarity is not significantly different. Moreover, the fraction of unique Rhyme- and unique Alliteration words is lower for Russians, while on the other hand they listen more to songs about topic 2 (Religion) and less to songs about topic 5 (Urban/Hiphop).

Secondly, there is the gender feature mean comparison. As can be seen in Table 10, significant differences were found for all but three features (Absolute Sentiment Polarity, the standard deviation in Rhymes following the AABB scheme, and topic 3). The most significant differences can be found in topics 1 (Love), 2 (Religion) and 5 (Urban/Hiphop). Apparently, male users listen less to songs about love, and significantly more to songs about religion and urban life. Also notably different, are the group mean values for the Unique Word Ratio, the Function Word Ratio, Song Title Mentions and the Song Polarity Value. Male users apparently listen, on average, to songs that contain more unique words and to less function words. On the other hand, the average spread of male users (standard deviation) for these two features is significantly bigger than that of female users too. Female users appear to prefer songs with a higher number of song title mentions and a higher sentiment polarity value. Finally, they score significantly higher on 19 out of 24 rhyme/alliteration features. In line with the higher UWR preference, males do score higher on the Fraction of Unique Rhyme/Alliteration words.

Thirdly, we ran T-tests on the means of users aged under 25 (adolescents) versus users aged 25 and older (adults), which can be found in Table 11. Interestingly, although adolescents listen to songs that have significantly more rhyme and alliteration schemes in them than adults do, the percentage of lines that contain rhyme is significantly lower for adolescents. This suggests that adolescents listen to songs that have substantially longer lyrics that rhyme relatively little compared to those listened to by adults. Furthermore, adolescents listen to songs that score low on SPV, but high on ASP. Moreover, they prefer songs with a lower unique word ratio and a higher function word ratio. Finally, adolescents listen more to songs about love (T1), and less to christmas songs (T3), songs in foreign languages (T4) and songs about Urban life (T5).

6.2 Gender classification

The classification of gender using the regular dataset proved to be difficult, due to the unbalanced dataset in favor of male users. For every 4 male users, there is only one female user in the dataset. Training on such an unbalanced dataset results in classifiers overfitting hugely on males, misclassifying the majority of female users. The results (Table 4) reflect this overfitting, since the highest accuracy with its 80.47%(Achieved through Adaboost using random forests) is only 1.36% above the majority vote-baseline. Plotting a confusion matrix (See Table 5) reveals that the best performing classification algorithm still has a very low recall when it comes to females, as it predicts the label male for 11,703 instances out of a total 12,181.

Classifier	**Style**	**Topic**	**LFM-1B**	**All**
Adaboost (Random Forest)	80.14%	76.64%	79.10%	80.47%
Multilayer Perceptron	79.53%	79.11%	79.11%	80.38%
Logistic Regression	78.40%	78.38%	79.16%	80.18%
Simple Logistic	79.09%	79.11%	79.13%	80.09%
SMO (RBF)	79.11%	79.11%	79.11%	79.11%
SMO (Poly)	79.11%	79.11%	79.11%	79.11%
ZeroR	79.11%	79.11%	79.11%	79.11%

Table 4: Gender classification accuracy(unbalanced)

The results in Table 4 are quite poor and uninformative about the usefulness of the featuresets due to overfitting. The best performing algorithm on the balanced gender classification task is the simple logistic regression, which accurately predicts the gender of 70.79% of the instances (See

Female	Male	Correct Label
322	2223	**Female**
156	9480	**Male**

Table 5: Unbalanced gender classification - confusion matrix

Table 6). For gender classification, the style-features carry most predictive value of the three stand-alone featuresets, followed by the topics and LFM-1B features which perform roughly equal.

Classifier	Style	Topic	LFM-1B	All
Simple Logistic	67.39%	64.72%	64.79%	70.79%
Multilayer Perceptron	67.70%	64.68%	64.52%	70.08%
SMO (Poly)	66.80%	64.05%	63.99%	70.00%
Logistic regression	68.10%	64.66%	64.58%	69.94%
Adaboost (Random Forest)	69.00%	60.73%	62.81%	69.73%
SMO (RBF)	61.04%	54.26%	61.53%	66.33%
ZeroR	50.00%	50.00%	50.00%	50.00%

Table 6: Gender classification accuracy(balanced)

6.3 Nationality classification

The classification of user nationality proved to be difficult, partially due to the presence of 61 different nationalities, for the other part due to the unbalanced distribution of nationalities. The two most common nationalities (US and Russia) represent 38.75% of the dataset, while there are 28 nationalities which are represented by less than 50 users each. The best performing algorithm (Multilayer Perceptron) predicts the accurate label 38.15% of the cases, which is just higher than the baseline of 27.30%, and much lower than the results obtained in the State-of-the-Art. However, the confusion matrix shows that the unbalance in this dataset (27.3% American users vs. 19% in the State-of-the-Art) negatively impacts the learning by creating a model that overfits on the American nationality. The LFM-1B features appear to be more useful than the style- or topic-features for country prediction. The latter features do seem to add some extra predictive value, as the performance is always best when training on all features. Adaboost Random Forest could not be included as too much memory was required to handle the 61 classes.

Classifier	Style	Topic	LFM-1B	All
Multilayer Perceptron	27.94%	27.12%	37.90%	38.15%
Simple logistic	27.08%	27.30%	35.94%	37.01%
SMO (Poly)	27.30%	27.30%	34.82%	35.58%
ZeroR	27.30%	27.30%	27.30%	27.30%

Table 7: Nationality classification accuracy

6.4 Age regression

The baseline for age regression for the used user subset is a Mean Absolute Error (MAE) of 3.85. As the results in Table 8 show, the multilayer perceptron scores a much lower mean absolute error than the others do. If only one featureset would be available for the task of age regression, one would be wise to choose the LFM-1B features, disregardless of the used regression algorithm.

Classifier	Style	Topic	LFM-1B	All
Multilayer Perceptron	3.59	3.82	3.29	3.12
SMOreg (Poly)	3.62	3.77	3.47	3.34
SMOreg (RBF)	3.66	3.78	3.45	3.34
Simple Linear Regression	3.81	3.84	3.68	3.68
ZeroR	3.85	3.85	3.85	3.85

Table 8: Age regression (MAE)

7. DISCUSSION

The T-tests revealed that different user groups listen to songs about different topics. Firstly, male users listen less to songs about love, and significantly more to songs about religion and urban life. Comparing users of Russian nationality with the rest, revealed that they listen more to songs about religion and less to songs about urban life. Finally, adolescents listen more to songs about love, and less to Christmas songs, songs in foreign languages or songs about urban life.

On average, male users listen to songs that contain more unique words and less function words. They also score higher on the fraction of unique rhyme/alliteration words. Female users appear to prefer songs with a higher number of song title mentions, a higher sentiment polarity value and absolute sentiment polarity, and to songs that rhyme and alliterate more, disregardless of the rhyme scheme it appears in. The sentiment polarity value is much lower for Russians than for the overall population, while the absolute sentiment polarity is not significantly different. For adolescents, similar results were found, although their absolute sentiment polarity was even smaller than that of adults. This suggests that both Russians and adolescents, listen more to songs of which the lyrics carry a negative sentiment. Finally, adolescents prefer songs with a lower unique word ratio and a higher function word ratio.

The release of a different dataset than the one that was actually used by the authors, hurts the reproducibility of the State-of-the-Art experiments [22], as well as the comparability of their results to the ones obtained in this study. The impact of training and testing machine learning algorithms on unbalanced data, as seen when comparing Tables 4 and 6, is big. As the published user subset is significantly more unbalanced than the one reported by the authors, the results obtained in our experiments can be expected to suffer more from overfitting on the most prevalent value of the specific target variable. Therefore, the obtained results should be compared to the majority vote-baseline. Unfortunately, this means that it is impossible to compare the information gain of the features that were newly proposed in this paper, to the ones used in the State-of-the-Art. However, the results in Tables 4, 6, 7 and 8 all show that on average, classifiers that use both the newly suggested features and the LFM-1B features, perform best. The LFM-1B features seem to be more informative for country prediction and age regression, while gender prediction benefits most from the style-features.

7.1 Conclusion

It was already known that demographic user traits can be predicted using features extracted from their listening history (e.g. mainstreaminess or the time of day the users listens to music the most [35]). In this paper we extracted features about the writing style or the topics found in lyrics of the songs in a user its listening history. T-tests revealed several significant differences in the lyrical writing- and topic preferences of user groups, which can be used by classifiers to separate these groups. These features are especially useful for the prediction of gender, but can also be used to improve prediction accuracy for nationality and age. Another interesting result is the structural high performance of the multilayer perceptron, suggesting that the results obtained in the State-of-the-Art might be improved upon simply by exploring the use of neural networks for prediction.

7.2 Future work

Several features capturing text-style were not implemented in this study even though they might contain useful information. First of all, there are the text statistics such as the relative length of choruses or verses to the song length [26] or the number of words per second [27]. The first could not be gathered as we considered the process of extracting the choruses and verses non-reproducible, whereas for the latter the information on the song duration was missing. The structure of natural text can also be captured by reporting the ratios of Part Of Speech (POS)-tags in each song [27]. Since it is unknown which POS-tags are most informative in this context, the reported tags should be counts for the number of common nouns, proper nouns, verbs, pronouns, prepositions, adverbs, determiners, modals, and adjectives. In order to take into account different document lengths, all of these values should then be divided by the number of words in the song and represented as ratio in the range of $< 0 : 1 >$. These features had to be excluded due to the limited resources available for this project in combination with the high computational demand attached to the extraction of POS-tags. On top of that, imperfect rhyme and alliteration detection might be useful for improving the rhyme and alliteration extraction deployed in this study [15]. It would also be interesting to see whether the unique word ratio and the function word ratio could be replaced by extracting the Lexical Novelty Score [12], and evaluating whether this score differs between user groups. Finally, the results achieved by the multilayer perceptron are highly competitive in comparison to older classification and regression algorithms. Therefore, it would be interesting to expand these experiments with newer types of neural networks (e.g. convolutional-, recurrent neural networks or Hierarchical attention networks), which are already proven effective in music genre [41] and mood[8] prediction. On a critical sidenote, the effectiveness of style-features in predicting a user its gender means that style- and topic-features derived from songlyrics can also compromise demographic information without user consent. This calls for extensions to previous work on guarding user privacy (e.g. through listener anonymization recommenders [42]) to take lyric-based features into account in the future as well.

8. REFERENCES

[1] R. Arun, V. Suresh, C. V. Madhavan, and M. N. Murthy. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 391–402. Springer, 2010.

[2] C. Bauer and M. Schedl. On the importance of considering country-specific aspects on the online-market: An example of music recommendation considering country-specific mainstream. 2018.

[3] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Ismir*, volume 2, page 10, 2011.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[5] E. Cambria, S. Poria, R. Bajpai, and B. Schuller. Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2666–2677, 2016.

[6] Z. Cheng, J. Shen, L. Nie, T.-S. Chua, and M. Kankanhalli. Exploring user-specific information in music retrieval. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 655–664. ACM, 2017.

[7] K. Choi and J. S. Downie. Exploratory investigation of word embedding in song lyric topic classification: Promising preliminary results. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, pages 327–328. ACM, 2018.

[8] K. Choi, G. Fazekas, M. Sandler, and K. Cho. Convolutional recurrent neural networks for music classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2392–2396. IEEE, 2017.

[9] K. Choi, J. H. Lee, X. Hu, and J. S. Downie. Music subject classification based on lyrics and user interpretations. *Proceedings of the Association for Information Science and Technology*, 53(1):1–10, 2016.

[10] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[11] J. S. Downie. Music information retrieval. *Annual review of information science and technology*, 37(1):295–340, 2003.

[12] R. J. Ellis, Z. Xing, J. Fang, and Y. Wang. Quantifying lexical novelty in song lyrics. In *ISMIR*, pages 694–700, 2015.

[13] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, and L. Trigg. Weka-a machine learning workbench for data mining. In *Data mining and knowledge discovery handbook*, pages 1269–1277. Springer, 2009.

[14] G. Häubl and V. Trifts. Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing science*, 19(1):4–21, 2000.

[15] H. Hirjee and D. G. Brown. Automatic detection of internal and imperfect rhymes in rap lyrics. In *ISMIR*, pages 711–716, 2009.

[16] M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.

[17] X. Hu and J. S. Downie. When lyrics outperform audio for music mood classification: A feature analysis. In *ISMIR*, 2010.

[18] X. Hu, J. S. Downie, and A. F. Ehmann. Lyric text mining in music mood classification. In *ISMIR*, 2009.

[19] D. A. Huffaker and S. L. Calvert. Gender, identity, and language use in teenage blogs. *Journal of computer-mediated communication*, 10(2):JCMC10211, 2005.

[20] M. L. Kern, J. C. Eichstaedt, H. A. Schwartz, G. Park, L. H. Ungar, D. J. Stillwell, M. Kosinski, L. Dziurzynski, and M. E. Seligman. From âĂIJsooo excited!!!âĂİ to âĂIJso proudâĂİ: Using language to study development. *Developmental psychology*, 50(1):178, 2014.

[21] F. Kleedorfer, P. Knees, and T. Pohle. Oh oh oh whoah! towards automatic topic detection in song lyrics. In *Ismir*, pages 287–292, 2008.

[22] T. Krismayer, M. Schedl, P. Knees, and R. Rabiser. Predicting user demographics from music listening information. *Multimedia Tools and Applications*, pages 1–24, 2018.

[23] V. Kumar and S. Minz. Mood classifiaction of lyrics using sentiwordnet. In *Computer Communication and Informatics (ICCCI), 2013 International Conference on*, pages 1–5. IEEE, 2013.

[24] C. Laurier, J. Grivolla, and P. Herrera. Multimodal music mood classification using audio and lyrics. In *Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on*, pages 688–693. IEEE, 2008.

[25] J.-Y. Liu and Y.-H. Yang. Inferring personal traits from music listening history. In *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 31–36. ACM, 2012.

[26] J. P. Mahedero, Á. MartÍnez, P. Cano, M. Koppenberger, and F. Gouyon. Natural language processing of lyrics. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 475–478. ACM, 2005.

[27] R. Mayer, R. Neumayer, and A. Rauber. Rhyme and style features for musical genre classification by song lyrics. In *Ismir*, pages 337–342, 2008.

[28] B. McFee, L. Barrington, and G. Lanckriet. Learning content similarity for music recommendation. *IEEE transactions on audio, speech, and language processing*, 20(8):2207–2218, 2012.

[29] R. Mihalcea and C. Strapparava. Lyrics, music, and emotions. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 590–599. Association for Computational Linguistics, 2012.

[30] M. Montaner, B. López, and J. L. De La Rosa. A taxonomy of recommender agents on the internet. *Artificial intelligence review*, 19(4):285–330, 2003.

[31] R. Neumayer and A. Rauber. Integration of text and audio features for genre classification in music information retrieval. In *European Conference on Information Retrieval*, pages 724–727. Springer, 2007.

[32] F. Pachet and J.-J. Aucouturier. Improving timbre similarity: How high is the sky. *Journal of negative results in speech and audio sciences*, 1(1):1–13, 2004.

[33] J. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.

[34] F. Ricci, L. Rokach, and B. Shapira. Recommender systems: introduction and challenges. In *Recommender systems handbook*, pages 1–34. Springer, 2015.

[35] M. Schedl. The lfm-1b dataset for music retrieval and recommendation. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 103–110. ACM, 2016.

[36] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[37] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy. Improvements to the smo algorithm for svm regression. *IEEE transactions on neural networks*, 11(5):1188–1193, 2000.

[38] M. Slaney, K. Weinberger, and W. White. Learning a metric for music similarity. In *International Symposium on Music Information Retrieval (ISMIR)*, 2008.

[39] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[40] J. Špeh, A. Muhic, and J. Rupnik. Parameter estimation for the latent dirichlet allocation. 2013.

[41] A. Tsaptsinos. Lyrics-based music genre classification using a hierarchical attention network. *arXiv preprint arXiv:1707.04678*, 2017.

[42] K. Tsukuda, S. Fukayama, and M. Goto. Listener anonymizer: Camouflaging play logs to preserve userâĂŹs demographic anonymity. *ISMIR*, 2018.

[43] A. Van den Oord, S. Dieleman, and B. Schrauwen. Deep content-based music recommendation. In *Advances in neural information processing systems*, pages 2643–2651, 2013.

[44] A. Van den Oord, S. Dieleman, and B. Schrauwen. Deep content-based music recommendation. In *Advances in neural information processing systems*, pages 2643–2651, 2013.

[45] M.-J. Wu, J.-S. R. Jang, and C.-H. Lu. Gender identification and age estimation of users based on music metadata. In *ISMIR*, pages 555–560, 2014.

[46] Y. Xia, L. Wang, and K.-F. Wong. Sentiment vector space model for lyric-based song sentiment classification. *International Journal of Computer Processing Of Languages*, 21(04):309–330, 2008.

[47] Z.-Q. Zeng, H.-B. Yu, H.-R. Xu, Y.-Q. Xie, and J. Gao. Fast training support vector machines using parallel sequential minimal optimization. In *Intelligent*

System and Knowledge Engineering, 2008. ISKE 2008. 3rd International Conference on, volume 1, pages 997–1001. IEEE, 2008.

APPENDIX

A. LDA TOPICS

Topic 1 —> 0.007*”love” + 0.005*”babi” + 0.005*”know” + 0.005*”want” + 0.004*”feel” + 0.004*”time” + 0.004*”never” + 0.004*”let” + 0.004*”heart” + 0.004*”thi”

Topic 2 —> 0.004*”lord” + 0.004*”jesu” + 0.004*”god” + 0.004*”sing” + 0.003*”prais” + 0.003*”holi” + 0.003*”king” + 0.002*”sky” + 0.002*”glori” + 0.002*”earth”

Topic 3 —> 0.021*”christma” + 0.005*”santa” + 0.005*”merri” + 0.004*”snow” + 0.003*”bell” + 0.003*”clau” + 0.002*”sleigh” + 0.002*”jingl” + 0.002*”carol” + 0.002*”mistleto”

Topic 4 —> 0.003*”ang” + 0.001*”vie” + 0.001*”que” + 0.001*”bop” + 0.001*”ikaw” + 0.001*”nah” + 0.001*”ako” + 0.001*”che” + 0.001*”kong” + 0.001*”frosti”

Topic 5 —> 0.003*”got” + 0.003*”get” + 0.003*”like” + 0.003*”nigga” + 0.003*”man” + 0.002*”yeah” + 0.002*”gon” + 0.002*”fuck” + 0.002*”girl” + 0.002*”hey”

B. T-TEST RESULTS

Feature	**T-statistic**	**P-value**
UWR mean	-1.6331	0.1025
UWR sd	0.7144	0.4750
FWR mean	0.1533	0.8782
FWR sd	-0.0173	0.9862
STM mean	-1.5088	0.1314
STM sd	-1.5827	0.1135
SPV mean	-8.1356	0.0000
SPV sd	-3.0191	0.0025
ASP mean	-1.1569	0.2473
ASP sd	-2.6711	0.0076
FUR mean	-2.6179	0.0089
FUR sd	-0.4407	0.6594
FUA mean	-3.0446	0.0023
FUA sd	-3.4959	0.0005
PLR mean	-1.5490	0.1214
PLR sd	-1.1661	0.2436
PLA mean	1.7671	0.0772
PLA sd	1.5040	0.1326
RAA mean	0.4653	0.6417
RAA sd	0.4576	0.6472
RAABB mean	-2.9917	0.0028
RAABB sd	-2.9388	0.0033
RABBA mean	3.0674	0.0022
RABBA sd	1.3754	0.1690
RABAB mean	1.5343	0.1250
RABAB sd	-1.3126	0.1893
AAA mean	0.5110	0.6093
AAA sd	-0.4941	0.6212
AAABB mean	1.1374	0.2554
AAABB sd	-1.0772	0.2814
AABBA mean	-0.9865	0.3239
AABBA sd	-0.6563	0.5116
AABAB mean	0.9142	0.3607
AABAB sd	-2.0124	0.0442
T1 mean	1.6943	0.0902
T2 mean	3.5450	0.0004
T3 mean	0.2730	0.7848
T4 mean	-2.0934	0.0363
T5 mean	-4.0126	0.0001

Table 9: Russian/Non-Russian T-test results

C. GITHUB REPOSITORY

Follow this link for the user set containing the style-, topic- and LFM-1B features, as well as the Weka models used for classification and regression. The balanced gender dataset is also provided here:

https://github.com/svenvdbeukel/Separating-Music-Consumers-Through-Lyrics

Feature	**T-statistic**	**P-value**
UWR mean	14.7766	0.0000
UWR sd	11.0091	0.0000
FWR mean	-9.8700	0.0000
FWR sd	10.1510	0.0000
STM mean	-11.7971	0.0000
STM sd	-7.3730	0.0000
SPV mean	-10.5702	0.0000
SPV sd	-6.0263	0.0000
ASP mean	-4.8990	0.0000
ASP sd	1.4122	0.1579
FUR mean	8.3694	0.0000
FUR sd	7.9227	0.0000
FUA mean	6.0435	0.0000
FUA sd	8.7278	0.0000
PLR mean	-8.4439	0.0000
PLR sd	-2.1298	0.0332
PLA mean	-11.8410	0.0000
PLA sd	-3.4280	0.0006
RAA mean	-7.4777	0.0000
RAA sd	-3.6266	0.0003
RAABB mean	-3.5756	0.0004
RAABB sd	1.6633	0.0963
RABBA mean	-5.6460	0.0000
RABBA sd	-9.6724	0.0000
RABAB mean	-6.6129	0.0000
RABAB sd	-8.7738	0.0000
AAA mean	-9.4871	0.0000
AAA sd	-7.7526	0.0000
AAABB mean	-7.2976	0.0000
AAABB sd	-5.2715	0.0000
AABBA mean	-5.7809	0.0000
AABBA sd	-8.4451	0.0000
AABAB mean	-8.9883	0.0000
AABAB sd	-9.0379	0.0000
T1 mean	-17.2468	0.0000
T2 mean	16.2208	0.0000
T3 mean	1.1865	0.2355
T4 mean	2.1858	0.0289
T5 mean	14.5764	0.0000

Table 10: Male/Female T-test results

Feature	**T-statistic**	**P-value**
UWR mean	-4.8948	0.0000
UWR sd	-1.1758	0.2397
FWR mean	4.5708	0.0000
FWR sd	-1.1281	0.2593
STM mean	2.1119	0.0347
STM sd	9.0227	0.0000
SPV mean	-3.8613	0.0001
SPV sd	5.5735	0.0000
ASP mean	13.2386	0.0000
ASP sd	7.6280	0.0000
FUR mean	-0.8177	0.4135
FUR sd	-1.0862	0.2774
FUA mean	0.1036	0.9175
FUA sd	0.3115	0.7554
PLR mean	-2.1208	0.0340
PLR sd	-2.5503	0.0108
PLA mean	1.7454	0.0809
PLA sd	1.6533	0.0983
RAA mean	13.6579	0.0000
RAA sd	14.5812	0.0000
RAABB mean	3.0884	0.0020
RAABB sd	4.5682	0.0000
RABBA mean	7.5931	0.0000
RABBA sd	10.7523	0.0000
RABAB mean	3.5084	0.0005
RABAB sd	4.4818	0.0000
AAA mean	10.5738	0.0000
AAA sd	13.9052	0.0000
AAABB mean	5.9109	0.0000
AAABB sd	6.8595	0.0000
AABBA mean	5.3707	0.0000
AABBA sd	12.9044	0.0000
AABAB mean	5.3709	0.0000
AABAB sd	5.2247	0.0000
T1 mean	4.9360	0.0000
T2 mean	1.9401	0.0524
T3 mean	-8.0942	0.0000
T4 mean	-6.8823	0.0000
T5 mean	-2.7015	0.0069

Table 11: Adolescent/Adult T-test results