“Butter lyrics over hominy grit”†
Comparing audio and psychology-based text features in MIR tasks

Kim, Jaehun; Demetriou, Andrew M.; Manolios, Sandy; Stella Tavella, M.; Liem, Cynthia C.S.

Publication date
2020

Document Version
Final published version

Published in
Proceedings of the 21st International Society for Music Information Retrieval Conference, ISMIR 2020

Citation (APA)
Kim, J., Demetriou, A. M., Manolios, S., Stella Tavella, M., & Liem, C. C. S. (2020). “Butter lyrics over hominy grit”†: Comparing audio and psychology-based text features in MIR tasks. In J. Cumming, J. H. Lee, B. McFee, M. Schedl, J. Devaney, J. Devaney, C. McKay, E. Zangerle, & T. de Reuse (Eds.), *Proceedings of the 21st International Society for Music Information Retrieval Conference, ISMIR 2020* (pp. 150-156). (Proceedings of the 21st International Society for Music Information Retrieval Conference, ISMIR 2020). International Society for Music Information Retrieval (ISMIR).

“BUTTER LYRICS OVER HOMINY GRIT”[†]: COMPARING AUDIO AND PSYCHOLOGY-BASED TEXT FEATURES IN MIR TASKS

Jaehun Kim[1 *] **Andrew M. Demetriou**[1*] **Sandy Manolios**[1]
M. Stella Tavella[2] **Cynthia C. S. Liem**[1]
[1] Delft University of Technology, Netherlands
[2] Musixmatch, Bologna, Italy
J.H.Kim@tudelft.nl
A.M.Demetriou@tudelft.nl

ABSTRACT

Psychology research has shown that song lyrics are a rich source of data, yet they are often overlooked in the field of MIR compared to audio. In this paper, we provide an initial assessment of the usefulness of features drawn from lyrics for various fields, such as MIR and Music Psychology. To do so, we assess the performance of lyric-based text features on 3 MIR tasks, in comparison to audio features. Specifically, we draw sets of text features from the field of Natural Language Processing and Psychology. Further, we estimate their effect on performance while statistically controlling for the effect of audio features, by using a hierarchical regression statistical model. Lyric-based features show a small but statistically significant effect, that anticipates further research. Implications and directions for future studies are discussed.

1. INTRODUCTION

Popular Western music very often contains lyrics. Social science research has shown informative relationships between popular songs and their lyrical content: e.g., country music lyrics rarely include political concepts [1], songs with more typical [2] and more negative [3] lyrics appear to be more successful, and the psychological content of song lyrics appears to correlate with cultural changes in psychological traits [4]. As for music consumption, lyrics have also been shown to be a salient component of music in the minds of listeners [5]. Furthermore, [6] showed that patients are more likely to choose music with lyrics when participating in music-based pain reduction interventions; [7] showed that lyrics enhance self reported emotional responses to music, although melody had an overall larger effect, and [8] showed a number of additional brain regions were active during the listening of sad music with lyrics, vs. sad music without lyrics.

In the Music Information Retrieval (MIR) field, some interest for lyrics and how they can be used to improve MIR tasks has been shown. Popular uses of lyrics for MIR tasks consider mood classification [9–12], genre classification [13, 14] and topic detection for indexing and browsing [15, 16]. [17] also proposed a metric to assess the novelty of lyrics, and suggested that novelty can play a role in music preference.

From these findings, one can conclude that lyrics are a rich data source. Although MIR interests have historically focused more on audio, lyrics information may fruitfully be leveraged for various MIR tasks. Still, there are many possible ways to extract information from lyrics text, and it is an open question what information extraction procedure will turn out most fruitful. To gain more insight into this, we present a study investigating several textual feature sets. In shaping these sets—acknowledging potential value of the topic for social science research—we are inspired by the way text analysis has been performed in the Psychology domain, and draw several of our extractors from prior work in that field. We will assess the performance of these textual feature sets on 3 common MIR tasks, and will statistically control for the effect of each chosen feature set, including an audio feature set for comparison. Our analysis will be performed on a large dataset from the online Musixmatch lyrics catalogue.

In the remainder of the paper, in Section 2, we discuss relevant previous work on text information extraction in the Psychology literature. Section 3 will subsequently explain our research design, after which Section 4 discusses the feature sets we used. Section 5 describes the data collection and pre-processing procedures, after which Section 6 details the experimental design. Section 7 justifies our chosen analytical strategy, followed by a presentation of results in Section 8 and the conclusion in Section 10.

[*] Authors contributed equally to the work.

[†] Quoted words are lyrics written by Clifford Smith, from the song “The What”, by the Notorious B.I.G. featuring Methodman, on the album “Ready to Die”, released in 1994.

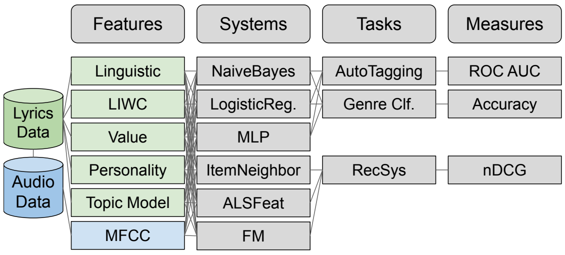

Figure 1. Overview of the experimental pipeline.

2. RELATED WORK

The field of Psychology has long pondered the importance of the words people choose to use, and how this reflects their individual differences [18]. The features we use in present work are primarily inspired by two prior lines of work in which Natural Language Processing (NLP) techniques were applied in psychology research: one employing closed-vocabulary lexicon approaches, the other employing open vocabulary approaches. Firstly, [19] used NLP techniques to derive estimates of personality for music genres. Specifically, they created a lexicon (a meaningful group of words) from psychology research that described personality dimensions, as well as a corpus of lyrics, separated into music genres. They then computed the similarity between the lyrics of music genres and the groups of personality dimension words, and considered this result to be an estimate of the personality dimension represented in the lyrics of each genre. Lexicon-based approaches have generally been popular, also thanks to the release of the Linguistic Inquiry Word Count (LIWC) lexicon-based software [20]; e.g., in the context of lyrics, [21] used it to examine psychological distress in the lyrics of musicians that committed suicide vs. those who had not.

Secondly, [22] demonstrated the usefulness of an open vocabulary approach vs. a lexicon approach while examining personality in the context of online social networks. Although lexicons are carefully curated and meaningful, they are also time-consuming to create and context-specific. In contrast, data-driven techniques can automatically estimate latent topics from groups of words that tend do appear together. [22] showed relationships between personality scores and automatically extracted latent topics. Further, they showed that the open vocabulary approach may have stronger correlations to self-reported personality scores than the closed-vocabulary lexicon approaches.

3. RESEARCH DESIGN

In this study, we seek to examine the relative importance of lyric-based text features—especially features drawn from psychology research— for various popular MIR tasks. We wish to compare this importance to that of conventional audio based features.

An overview of our experimental pipeline is given in Figure 1. Various feature sets will feed into various systems, that are appropriate for various MIR machine learning tasks. We employ a full-factorial experimental design for feature sets, tasks, and the systems attached to each task, which means we research all the possible combinations of those factors. For each combination, we will employ the traditional train-validation-test machine learning setup. Performance results on the test sets will feed into our statistical analysis, where we will explicitly control for the effect of each of the feature sets.

4. FEATURE SETS

In this work, we will consider 5 lyric-based text feature sets and an audio-based feature set. More details are given in the following subsections; a summary of the dimensionalities of all feature sets is given in Table 1.

4.1 Linguistic Features

As baseline textual features for this study, we first extract several simple *linguistic* features:

- *NumWords:* the number of words included in the lyrics text.
- *NumUniqueWords:* the number of unique words in the lyrics text.
- *NumStopWords:* the number of stop words in the lyrics text [1] .
- *NumRareWords* the number of words that appeared in less than 5 lyrics.
- *NumCommonWords* the number of words extremely commonly used within a lyrics corpus. We set the threshold as the 30% percentile of the document frequency of words.

Along with the absolute number, we also compute the ratio over the total number of words for each lyrics text.

4.2 Topic Modeling

As a more advanced feature extraction technique, we employ probabilistic Latent Semantic Analysis (pLSA) [24] for *topic* modeling. We treat each of the lyric texts as a document, and will take the found topic distribution for a given document as the document feature. We chose the number of topics $K = 25$, which maximizes validation log-perplexity. Taking advantage of the unsupervised learning setup, we use the total pool of songs to setup the training-validation-test split.

4.3 LIWC

Linguistic Inquiry Word Count (LIWC) is a software package built on a lexicon that has been validated for text analysis in psychological studies [20]. It uses a curated lexicon, separated into 73 categories (e.g., the category ‘Social Processes’ includes references to family and friends). The software outputs the counts of words in a given text for each of the 73 categories. We employ the latest LIWC, released in 2015.

[1] As we will focus on English lyrics in this study, we used the English stop words corpus from the Natural Language Toolkit (NLTK) [23]

4.4 Psychology Inventory Scores

We will consider two more feature sets, inspired by psychology inventory scores: a feature set focusing on *personality* and a feature set focusing on *values*. In both cases, we will use lexicons from literature. However, rather than performing a word count as was done in LIWC, we will use more contemporary NLP techniques based on word embeddings.

Contemporary personality theory is derived from lexical studies: it has been suggested that meaningful individual psychological differences between people are captured in the adjectives that describe people [25]. Although the number of meaningful clusters of adjectives (called Personality Dimensions) is under debate, the OCEAN or Big-Five model is often used. It is composed of 5 traits : Openness to Experience, Conscientiousness, Extroversion, Agreeableness and Neuroticism [25]. Our *personality* feature set consists of 2 word clusters per dimension, comprised of words representing positive and negative aspects for each personality dimension, derived from prior research [26].

Personal *values* are another important component of identity, though less studied. They are stable over time and represent who people want to be, targeting the most important things for them in life at the most abstract level. The traditional way to obtain people's personal values is through questionnaires, but recent works focused on NLP techniques to extract them from text [27–29]. In our work, we used the value inventory and lexicon from [28].

Both for the *personality* and *values* feature sets, we will exploit the `word2vec` model [30] to approximate distances between lyrics and the various inventory categories in the feature sets. For this, we use the model pre-trained on the Google News dataset[2]. The average distance score $s_{d,c}$ for each lyric text d, and category c is computed by taking the average cosine distance between the words belonging to the lyrics and the categories, respectively:

$$s_{d,c} = \frac{1}{|\mathcal{W}_d||\mathcal{W}_c|} \sum_{n \in \mathcal{W}_d} \sum_{m \in \mathcal{W}_c} \frac{\langle \mathbf{v}_n, \mathbf{v}_m \rangle}{||\mathbf{v}_n|| \cdot ||\mathbf{v}_m||} \qquad (1)$$

where $\mathcal{W}_d$ and $\mathcal{W}_c$ represent the set of words belonging to the lyrics text d and the category c. v_n and v_m denote the pre-trained word vectors corresponding to word n in the lyrics and word m in the category, respectively.

4.5 MFCC

Finally, we employ a set of audio features based on the Mel-Frequency Cepstral Coefficients (MFCC). We include these, such that the effect of the lyric-based text features can be compared to a commonly used feature set from the primary modality of interest in many MIR tasks. Specifically, we adopt the feature computation introduced in [31] with 40 mel bins.

Feature Set	Dimensions
Audio	240
LIWC	73
Values	49
Topics	25
Personality	10
Linguistic	9

Table 1. Number of dimensions per feature set

5. DATA COLLECTION

We analyzed the lyrics contained in the Musixmatch dataset[3], which is the official lyrics metadata selection integrated in the Million Song Dataset (MSD) [32], a collection of relevant data and metadata for one million popular contemporary songs. Musixmatch is a lyrics and music language platform. The Musixmatch community drives the content production by adding, correcting, syncing and translating lyrics of songs. The process of lyrics quality verification involves several steps, including spam detection, formatting, spelling and translation checking. These steps are accomplished by the use of both artificial intelligence and machine learning models. In addition, they are manually verified by more than 2000 Curators worldwide, and a local team of Musixmatch Editors, who are native speakers in different languages.

The data used for the purpose of this project consists of $182,808$ lyrics, plus relevant metadata such as the unique identifier, artist and title. The data encompasses $20,219$ unique artists over various genres of music.

5.1 Preprocessing

For the given lyrics dataset, we consider the following preprocessing steps: the sentence strings are 1) tokenized and 2) lemmatized, followed by 3) stop-words filtering and 4) filtering extremely rare and extremely common words (see Section 4.1). Finally, we filter out non-English lyrics by a filtering process using the topic modeling. More precisely, we fit the topic model to detect whether the topics contain non-English words above a certain threshold. Songs that mostly load on non-English topics are removed.

6. EXPERIMENT

6.1 Tasks & Systems

As shown in Figure 1, to assess the lyrics feature set, we consider 3 popular MIR machine learning tasks; for each of these, we use 3 different commonly used types of systems, and a task-specific performance measure is considered, as detailed below.

6.1.1 Music Genre Classification

Music Genre Classification (MGC) is a multi-class classification problem. Typically, a set of music genres is given as the classes, and music audio content or features are used as the observations. In this study, we examine 3 machine

[2] `https://code.google.com/archive/p/word2vec/`

[3] `http://millionsongdataset.com/musixmatch/`

learning based systems: *Gaussian Naive Bayes* (GNV), *Logistic Regression* (LR) and the *Multi-Layer Perceptron* (MLP). For performance quantification, we opt for *classification accuracy*.

For this task, we use the data in the intersection between our lyrics database and the part of the MSD for which the music genre mapping introduced in [33] can be made. By choosing the intersection with the MSD, our audio features can be extracted from the MSD preview audio excerpts. Due to genre label availability, this leads to $67,719$ songs being used in this task.

6.1.2 *Music Auto-Tagging*

Music Auto-Tagging (MAT) is often formulated as a multi-label classification problem in which multiple positive labels may exist for one input music observation. We used the same set of systems as in the MGC task[4]. Again, we cross-match to the MSD, now also considering MSD's LastFM social tags. Similarly to [31], we choose to focus on the 50 most frequent tags from the dataset. The *Area Under Curve - Receiver Operating Characteristic* (AUC-ROC) is used as the performance measure, which will be referred to as AUC^{song} for the rest of this paper[5]. Due to tagging label availability, $137,095$ songs are used under this task.

6.1.3 *Music Recommendation*

Finally, Music Recommendation (MR) is considered for a user-related retrieval task. In particular, we consider a cold-start scenario, in which a batch of songs is newly introduced to the market, and required to be recommended to users. Due to the lack of previous interaction history, in such a scenario, a model will be maximally dependent on item attributes. As this is a substantially different type of task than the previous classification tasks, a different set of the systems common to the recommender systems field is used. *Item Nearest Neighbor* (INN) is a memory-based collaborative filtering method, which recommends the items closest to those that the user had consumed. We employ the feature vector introduced in Section 4 to compute the distance between entities using the cosine distance. We also use the *Feature-augmented Matrix Factorization* (FMF) [34] method, as well as the *Factorization Machine* [35] (FM). These models are more sophisticated collaborative recommenders, which also are capable of exploiting item attributes. The systems are developed and evaluated using the MSD-Echonest dataset [32]. Due to limits on available computational resources, we exploit a densified subset with $96,551$ users and $66,850$ songs from the initial song pool with the lyrics[6]. Finally, the binarized *normalized Discounted Cumulative Gain* (nDCG) is considered as performance measure, for the top-100 songs recommended.

6.2 Task Simulation Setup

All MIR tasks above are machine learning tasks, but the systems and data we choose to use for them did not yet exist in a real-life system. Therefore, we ran the machine learning procedures to initiate them. For this, for each task, we randomly split the available song data into *training/validation/test* subsets by a ratio of $8:1:1$. Each model is trained using the *training* set and evaluated on the *validation* set to tune the hyper-parameters. Once the optimal hyper-parameters are found, final performance is measured on the the *test* set.

For MLP and FMF, which have more than one hyper-parameter, automatic hyper-parameter tuning is conducted through a Bayesian approach, using the Gaussian Process[7] [8]. Every search process iterates through 50 training-validation procedures to reach the optimal point. For the MGC and MAT tasks, the hyper-parameters are searched at every trial, while in the MR task, the search process runs only once and is used for all the other trials.

7. ANALYTIC STRATEGY

We wish to assess the usefulness of each of the feature sets for the 3 MIR tasks. Therefore, the resulting performance score from each trial run in our experimental setup (see Section 3) forms the measurement that is our outcome variable of interest. We seek to estimate the relative contribution of each feature set, while statistically controlling for the contribution of all other variables in the analysis. In addition, we assess whether feature sets perform better or worse, depending on the task.

Our data has a nested structure. Specifically, we might say that our systems are nested within the tasks: each task is likely to influence the score, as will the underlying systems that were used for each task. Further, not all systems were used in all tasks. To account for this structure, we employed hierarchical regression models which allow for the modeling of variances of nested data.

The typical example for this category of models is the task of modeling the standardized test scores of various students within various schools. Test scores may be due to the performance of the student, but the school itself may also influence the scores. In this case, the students are said to be nested within the school. If we wanted to accurately assess the effect of e.g. a specific teaching technique on the scores of the students, we would want to statistically control for the effect of the nested structure. A hierarchical regression allows for us to estimate the variance in both intercept and slope of the school, to more accurately assess the effect of the teaching technique on the score of the student. For example, the following equations allow us to model the varying intercepts and slopes of each school:

[4] We employ a one-vs-rest strategy for the LR and GNV, which transforms a multi-label classification problem to multiple binary classification problems.

[5] We employed song-wise aggregation for this study

[6] We initially matched the original Echonest dataset to our initial song pool and 30% of randomly sampled users. Consequentially, we apply a filter, such that users who interacted with more than 5 songs remain, and vice versa for songs.

[7] We use the implementation from the *scikit-optimize* package.

[8] We do not search the hyper-parameters for FM and use a manually tuned setup, mostly due to the computational complexity required for this specific model.

$$y_i = a_{j[i]} + \beta x_i + \epsilon_i \tag{2}$$

$$\alpha_j = a_0 + b_0 u_j + \eta_{j1} \tag{3}$$

$$\beta_j = a_1 + b_1 u_j + \eta_{j2} \tag{4}$$

where i refers to the individual students, and $j[i]$ refers to the school that student i attends. The first line is similar to a classic regression, where the x represents a predictor at the level of student, the teaching technique in our example, and the ϵ represents the error term of the main regression. However, equations (3) and (4) allow for the modeling of the intercept and slope respectively, where the u and η expressions are the predictors and error terms at the school levels.

By statistically controlling for these additional variances, hierarchical modeling allows for a more precise estimate of the variables of interest. A more complete discussion can be found in [36].

In our study, we treat the task similarly to the school in our example, and the systems similarly to the students. By controlling for these variances, we estimate the effect of each feature set. From the resulting parameter estimates, we extract 95 % confidence intervals, which we then interpret for our results.

This approach also allows for the comparison of models containing different specifications, where the specifications refer to which specific parameter estimates are computed. As some parameters may not meaningfully contribute to the variance, their effects will be estimated at very close to 0, and may be removed to improve model fit. Indices of fitness, i.e. Akaike and Bayesian Information Criteria (AIC and BIC respectively) give an estimate of model fit, which is penalized by the number of terms. We can therefore arrive at the best-fitting model with the fewest parameters estimated, by systematically removing poorly performing parameter estimates, comparing successive fit indices e.g. with a Likelihood Ratio Test.

Following from our strategy, we examined the usefulness of the inclusion of the various features sets on the 3 considered MIR tasks. Our variables of interest are 1) binary indicators for the inclusion of each of the feature sets: *linguistic*, *topic*, *LIWC*, *personality*, and *values*, as well as the set of audio features, where (0 = not included, 1 = included), 2) a categorical variable representing each of the MIR tasks, 3) a categorical variable representing the systems implemented within each task, and 4) the resulting Measurement scores which were standardized within each task for comparability. We further estimate whether feature sets perform better or worse for certain tasks, by examining interactions between each feature set, and our task variable. Feature sets had differing numbers of sub-dimensions which were not individually analyzed (see Table 1)[9] .

We ran multiple models and compared the results of our feature sets across specifications (see Figure 2). Model specifications varied based on 1) how we accounted for the nested structure (i.e. task and systems), as we can estimate

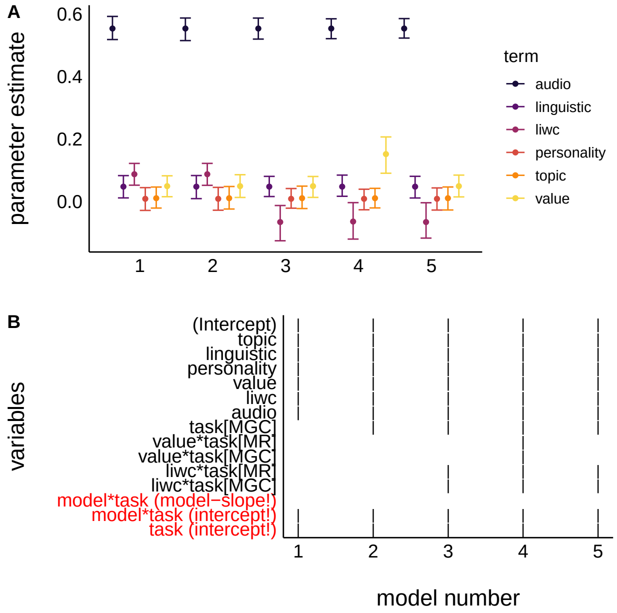

Figure 2. A: Parameter estimates of 5 hierarchical regression models. Error bars are 95% confidence intervals, bootstrapped 500 times. B: Specific parameters that are estimated in the each of the models. Parameters that form the structure of the model are denoted both in red and with a "!" symbol, feature sets of primary interest are denoted in black, and variables for which two terms separated by a "*" are interaction terms.

intercepts for task, for system, for system within task, as well as as slopes for tasks, for systems, and for systems within task, etc., and 2) the interaction terms we specified, i.e. whether we estimated an interaction term for a given feature set and our task variable.

8. RESULTS

We assessed models with two nested structures specified, where the parameters estimated are referred to as "random effects". The first included intercepts for each task, and the system used within task. The second estimated the same intercepts, and additionally estimated a slope for each system. For each of these two random effects structures, we then determined which parameters to estimate, referred to as "fixed effects". Specifically, we estimated parameters for each feature set, and interactions between all feature sets and the tasks. We first specified a "maximal" model, with all features and the task variable, and all two-way interactions among these variables. To remove unnecessary parameters, we ran a protocol which iteratively removed parameter estimates, retaining only those that either 1) significantly decrease model fit if not included, or 2) do not significantly decrease model fit if excluded. The Step function in the *lmerTest* package, was used for this phase [37]. What remained were two interaction terms: the interaction between *values* and task, and between *LIWC* and task. As such, we estimated models with no interaction terms, as well as models with and without each of those inter-

[9] Analyses were conducted on two servers running R 3.6.3. and 3.4.4.

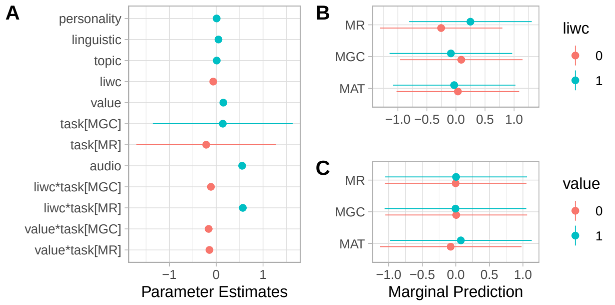

Figure 3. A: Parameter estimates of model 4. Error bars are 95% confidence intervals. Interaction terms are denoted with the "*" symbol. B: Predicted scores for the inclusion of *LIWC* on each of three MIR tasks, where 1 indicates that it was included and 0 indicates that it was not. C: Predicted scores for the inclusion of *values* on each of three MIR tasks, where 1 indicates that it was included and 0 indicates that it was not.

action terms. When we assessed the interaction term, we also included the main effect of task. Thus, we also ran models with and without task included. The 5 models included for interpretation were those that converged without error. Parameter estimates are shown in Figure 2A, and Figure 2B shows which parameters were estimated in each model. For the full specification of our models, we refer the readers to the reproducibility package [10] accompanying this paper.

As is shown in Figure 2A, we observe a consistent, large, positive effect of *audio features* on the score, and no meaningful effects of *topic* and *personality* feature sets. Further, we observe a consistent, small, positive effect of *values* across our specifications. This effect size increases in model 4, where the interaction between values and task was included. Similarly, *LIWC* shows a small but positive effect, that appears to decrease when the interaction term of *LIWC* and task is included. This suggests that *LIWC* and *values* may perform differently, depending on the task.

To clarify if this is the case, we examined the parameter estimates of model 4, which included interaction terms for both *LIWC* and *values* (see Figure 3A). Although both interaction terms were statistically significant, we observe that the confidence intervals for the main effect of task are very wide. This was expected, as 1) we were assessing an interaction effect which might increase the width of a the confidence interval, and 2) we were largely accounting for this variance by standardizing the score within each task, and by including task in the random effect structure. Figures 3A and 3B show the predicted values for both *LIWC* and *values* across tasks. Although the score was higher when *LIWC* was included in the MR task and when *values* was included in the MAT task, the predicted estimates are imprecise, as evidenced by the wide confidence intervals. As such, a more sensitive study design is likely required to obtain estimates of these interaction effects, e.g. analyses on individual dimensions of feature sets, to establish the most informative features, and/or more systems and more MIR tasks. Thus, we conclude that *linguistic* and *values* feature sets show the most consistent positive effects, and that *LIWC* and *values* may vary in performance based on task.

9. LIMITATIONS AND FUTURE WORKS

Several limitations are still present in our current study. Firstly, although our feature sets did show promising yet small effect sizes, we did not assess the performance of individual dimensions. Given that the feature sets vary greatly in both in terms of the number and content of subdimensions (see Table 1), reducing the overall set may result in a more sensitive set of features to examine.

Secondly, we did not consider subgroups of users, or of groups of songs. It may be possible that some users are more sensitive to the content of lyrics than others, and that lyric-sensitive users would benefit far more from lyric features than others. Further, it may be the case that lyrics are very important in some groups of songs vs. others (e.g Hip-Hop music vs. electronic dance music). Further research could examine the potential existence of a lyric-sensitive sub-group of users, lyric-sensitive songs, and how these two may interact.

Thirdly, aspects of our experimental design can be elaborated in future work: 1) Although we strategically sampled a limited number of MIR tasks and a limited number of systems, we did not fully address all possibilities. For instance, future work can include more contemporary systems such as deep learning, thereby increasing generalizability of our results. 2) Certain task metrics could be improved, although we strategically designed our experiment to prevent local noise from skewing our conclusions: e.g. a different performance measure for the genre classification (i.e. AUC-ROC) could deliver a more accurate experimental result, given its skewed class distribution.

Lastly, the reliability of all of our feature sets could be better assessed in the future. This is particularly true of our *personality* features: they contain words that have been shown to describe individuals that have or lack in personality traits, but it is not clear that individuals with those traits use the specific words that describe them.

10. CONCLUSION

Although the *audio* features in our analysis most positively affected performance on various MIR tasks, our lyric-based text features did show some promise. More specifically, *linguistic* and *values* feature sets showed consistent, small effect sizes. Given that the interactions between *LIWC* and task were significant, it may be the case that *LIWC* features are also useful. We can conclude that text-based features drawn from Psychology literature anticipate further research, and that further investigations addressing the current limitations will lead to better data-driven understanding of the role lyrics play in music consumption.

[10] `https://github.com/mmc-tudelft/lyricpsych-ISMIR20`

11. ACKNOWLEDGEMENT

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

12. REFERENCES

[1] R. W. Van Sickel, "A world without citizenship: On (the absence of) politics and ideology in country music lyrics, 1960–2000," *Popular music and society*, vol. 28, no. 3, pp. 313–331, 2005.

[2] A. C. North, A. E. Krause, and D. Ritchie, "The relationship between pop music and lyrics: A computerized content analysis of the United Kingdom's weekly top five singles, 1999–2013," *Psychology of Music*, pp. 1–24, 2020.

[3] C. O. Brand, A. Acerbi, and A. Mesoudi, "Cultural evolution of emotional expression in 50 years of song lyrics," *Evolutionary Human Sciences*, vol. 1, pp. 1–14, 2019.

[4] C. N. DeWall, R. S. Pond Jr, W. K. Campbell, and J. M. Twenge, "Tuning in to psychological change: Linguistic markers of psychological traits and emotions over time in popular us song lyrics." *Psychology of Aesthetics, Creativity, and the Arts*, vol. 5, no. 3, pp. 200–207, 2011.

[5] A. Demetriou, A. Jansson, A. Kumar, and R. M. Bittner, "Vocals in music matter: the relevance of vocals in the minds of listeners," in *Proceedings of the 19th International Society for Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, E. Gómez, X. Hu, E. Humphrey, and E. Benetos, Eds., 2018, pp. 514–520.

[6] C. Howlin and B. Rooney, "Patients choose music with high energy, danceability, and lyrics in analgesic music listening interventions," *Psychology of Music*, vol. 0, no. 0, pp. 1–14, 2020.

[7] S. O. Ali and Z. F. Peynircioğlu, "Songs and emotions: are lyrics and melodies equal partners?" *Psychology of music*, vol. 34, no. 4, pp. 511–534, 2006.

[8] E. Brattico, V. Alluri, B. Bogert, T. Jacobsen, N. Vartiainen, S. K. Nieminen, and M. Tervaniemi, "A functional MRI study of happy and sad emotions in music with and without lyrics," *Frontiers in psychology*, vol. 2, pp. 1–16, 2011.

[9] X. Hu and J. S. Downie, "When lyrics outperform audio for music mood classification: A feature analysis," in *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010, Utrecht, Netherlands, August 9-13, 2010*, J. S. Downie and R. C. Veltkamp, Eds. International Society for Music Information Retrieval, 2010, pp. 619–624.

[10] M. McVicar, T. Freeman, and T. D. Bie, "Mining the correlation between lyrical and audio features and the emergence of mood," in *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011*, A. Klapuri and C. Leider, Eds. University of Miami, 2011, pp. 783–788.

[11] Y. Hu, X. Chen, and D. Yang, "Lyric-based song emotion detection with affective lexicon and fuzzy clustering method," in *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009, Kobe International Conference Center, Kobe, Japan, October 26-30, 2009*, K. Hirata, G. Tzanetakis, and K. Yoshii, Eds. International Society for Music Information Retrieval, 2009, pp. 123–128.

[12] X. Wang, X. Chen, D. Yang, and Y. Wu, "Music emotion classification of Chinese songs based on lyrics using TF*IDF and rhyme," in *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011*, A. Klapuri and C. Leider, Eds. University of Miami, 2011, pp. 765–770.

[13] R. Mayer, R. Neumayer, and A. Rauber, "Rhyme and style features for musical genre classification by song lyrics," in *Proceedings of the 9th International Society for Music Information Retrieval Conference, ISMIR 2008, Drexel University, Philadelphia, PA, USA, September 14-18, 2008*, J. P. Bello, E. Chew, and D. Turnbull, Eds., 2008, pp. 337–342.

[14] A. Tsaptsinos, "Lyrics-based music genre classification using a hierarchical attention network," in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, S. J. Cunningham, Z. Duan, X. Hu, and D. Turnbull, Eds., 2017, pp. 694–701.

[15] F. Kleedorfer, P. Knees, and T. Pohle, "Oh oh oh whoah! Towards automatic topic detection in song lyrics," in *Proceedings of the 9th International Society for Music Information Retrieval Conference, ISMIR 2008, Drexel University, Philadelphia, PA, USA, September 14-18, 2008*, J. P. Bello, E. Chew, and D. Turnbull, Eds., 2008, pp. 287–292.

[16] S. Sasaki, K. Yoshii, T. Nakano, M. Goto, and S. Morishima, "Lyricsradar: A lyrics retrieval system based on latent topics of lyrics," in *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, H. Wang, Y. Yang, and J. H. Lee, Eds., 2014, pp. 585–590.

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

[18] Y. R. Tausczik and J. W. Pennebaker, “The psychological meaning of words: LIWC and computerized text analysis methods,” *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[19] Y. Neuman, L. Perlovsky, Y. Cohen, and D. Livshits, “The personality of music genres,” *Psychology of Music*, vol. 44, no. 5, pp. 1044–1057, 2016.

[20] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, “The development and psychometric properties of LIWC2015,” Tech. Rep., 2015.

[21] D. M. Markowitz and J. T. Hancock, “The 27 Club: Music lyrics reflect psychological distress,” *Communication Reports*, vol. 30, no. 1, pp. 1–13, 2017.

[22] 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*, vol. 8, no. 9, pp. 1–16, 2013.

[23] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O’Reilly, 2009.

[24] T. Hofmann, “Unsupervised learning by probabilistic latent semantic analysis,” *Machine Learning*, vol. 42, no. 1/2, pp. 177–196, 2001.

[25] L. R. Goldberg, “An alternative "description of personality": the Big-Five factor structure.” *Journal of personality and social psychology*, vol. 59, no. 6, pp. 1216–1229, 1990.

[26] G. Saucier and L. R. Goldberg, “Evidence for the Big Five in analyses of familiar english personality adjectives,” *European Journal of Personality*, vol. 10, no. 1, pp. 61–77, 1996.

[27] S. Wilson, R. Mihalcea, R. Boyd, and J. Pennebaker, “Disentangling topic models: A cross-cultural analysis of personal values through words,” in *Proceedings of the First Workshop on NLP and Computational Social Science*, 2016, pp. 143–152.

[28] S. R. Wilson, Y. Shen, and R. Mihalcea, “Building and validating hierarchical lexicons with a case study on personal values,” in *International Conference on Social Informatics*. Springer, 2018, pp. 455–470.

[29] H. Liu, Y. Huang, Z. Wang, K. Liu, X. Hu, and W. Wang, “Personality or value: A comparative study of psychographic segmentation based on an online review enhanced recommender system,” *Applied Sciences*, vol. 9, no. 10, pp. 1–28, 2019.

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, Eds., 2013, pp. 3111–3119.

[31] K. Choi, G. Fazekas, M. B. Sandler, and K. Cho, “Transfer learning for music classification and regression tasks,” in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, S. J. Cunningham, Z. Duan, X. Hu, and D. Turnbull, Eds., 2017, pp. 141–149.

[32] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, “The Million Song Dataset,” in *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28,2011*, A. Klapuri and C. Leider, Eds. University of Miami, 2011, pp. 591–596.

[33] H. Schreiber, “Improving genre annotations for the Million Song Dataset,” in *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015*, M. Müller and F. Wiering, Eds., 2015, pp. 241–247.

[34] D. Liang, M. Zhan, and D. P. W. Ellis, “Content-aware collaborative music recommendation using pre-trained neural networks,” in *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015*, M. Müller and F. Wiering, Eds., 2015, pp. 295–301.

[35] S. Rendle, “Factorization machines,” in *Proceedings of the 10th IEEE International Conference on Data Mining, ICDM 2010, Sydney, Australia, 14-17 December 2010*, G. I. Webb, B. Liu, C. Zhang, D. Gunopulos, and X. Wu, Eds. IEEE Computer Society, 2010, pp. 995–1000.

[36] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge university press, 2006.

[37] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, “lmerTest package: tests in linear mixed effects models,” *Journal of statistical software*, vol. 82, no. 13, 2017.