Ground Truthing is a Field

Andrew M. Demetriou

2025-03-26

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

1 Introduction

This is a book created from markdown and executable code.

See [1] for additional discussion of literate programming.

2 Summary

In summary, this book has no content whatsoever.

3 Vocals in Music Matter: the Relevance of Vocals in the Minds of Listeners

In music information retrieval, we often make assertions about what features of music are important to study, one of which is vocals. While the importance of vocals in music preference is both intuitive and anticipated by psychological theory, we have not found any survey studies that confirm this commonly held assertion. We address two questions: (1) what components of music are most salient to people's musical taste, and (2) how do vocals rank relative to other components of music, in regards to whether people like or dislike a song. Lastly, we explore the aspects of the voice that listeners find important. Two surveys of Spotify users were conducted. The first gathered open-format responses that were then card-sorted into semantic categories by the team of researchers. The second asked respondents to rank the semantic categories derived from the first survey. Responses indicate that vocals were a salient component in the minds of listeners. Further, vocals ranked high as a self-reported factor for a listener liking or disliking a track, among a statistically significant ranking of musical attributes. In addition, we open several new interesting problem areas that have yet to be explored in MIR.

4 Introduction

The Music Information Retrieval (MIR) community has historically focused on content-based understanding of music. The type of content-based analysis studied over time is typically driven by the data available to the task, or the interests of the specific researchers. An alternative motivator could be to study topics that are salient in the minds of listeners, especially with respect to listener's musical preference. Specifically, understanding which attributes of music contribute the most to music preference, and their relative weight, could help guide research efforts. One attribute of music we would expect to be salient in the minds of listeners is the singing voice.

Psychology research anticipates the importance of the human voice as a salient stimulus, and as a component of music in particular. The human ability to communicate exceeds that of any other species studied thus far, with both speech and singing being cultural universals reliant on vocal production. It is theorized that the advanced human ability to communicate, discriminate, and to experience emotional responses in vocalizations has allowed for the emergence of music **juslin2003music?**. Our emotions are often accompanied by involuntary changes in our physiology and nonverbal expressions, such as facial expressions and vocalizations **porges2001polyvagal?**. Our reactions to the emotional content expressed in the vocals in music may have similar effects. As such, much psychological research has focused on the singing voice even more than speech, due to the precision required to execute and process musical vocalizations **hutchins2013linked?**. This makes musical vocals a well-anticipated candidate for study as a feature of music, as we would expect people to have a sophisticated ability to deliver, empathize with, and process vocal communications.

We would therefore expect that the vocals in music would be an especially salient component, if not the most salient. While a complete review is beyond the scope of this paper, some research is particularly worth noting. For example, it has been shown that both adults **weiss2012something?** and children **weiss2015enhanced?** recall melodies more correctly when sung with the voice than when played with instruments. Hutchins and Moreno **hutchins2013linked?** review literature that shows relatively precise perception of pitch in the human voice, yet fewer noticeable pitch errors in the voice relative to musical instruments or synthesized voices **hutchins2012frog?**. Neuroscience studies show specific areas of the brain involved in processing human voices **belin2000voice?**. Although similar regions of the brain are involved in processing both music and voices, there is differential processing of the human voice relative to music **armony2015specificity?**. As such, the human voice may be processed as a uniquely significant sound.

However, while prior research suggests that vocals would be especially relevant to music preference, no study to our knowledge has assessed the importance of the voice in music, relative to other musical components. To address this gap, we test the hypothesis that the voice is as or more important than other musical components across implicit and explicit datasets, using traditional social science techniques, as well as data mining techniques. First, we mine data available from Spotify, including playlist titles, search data and artist biographies, to test whether terms related to vocals are prevalent. However, we show that the results of the data mining are inconclusive as to whether or not vocals are salient in the minds of listeners. Specifically, it is not clear whether the vocals can be disentangled from other factors in playlist titles and search queries, such as genre. For more conclusive results, we gather data from users explicitly. To this aim we conduct two online survey studies: the first gathered subjective data on the salient components of music directly from listener reports, which were separated into semantic categories using card-sorting. The second asked participants to rank the semantic categories from the first study in terms of importance to their musical preference. We conclude that two aspects related to the voice are especially salient, namely the voice itself, and the lyrics of the song. Furthermore, we highlight the importance of gathering explicit data to complement implicit techniques, in situations where factors may not be easily disentangled.

5 Vocals in Semantic Data

Prior research has shown that semantic descriptors of music may be an appropriate means for users to query music databases **lesaffre2008potential?**. Given the large amount of semantic data available to Spotify such as playlist titles, search results, and artist biographies, one might hypothesize that terms describing the vocals would commonly appear in this implicit data.

5.1 Playlist Tags and Search Queries

Non-common words or groups of words and emojis appearing in the titles of a large number of Spotify's user-generated playlists were aggregated to create a list of the 1000 most frequently occurring *tags*. Each of these 1000 tags was assigned a category by a professional curator based on the tag itself and information from the tracks most frequently associated with the tag. The categories, determined by the curator, were Genre (e.g. "K-Pop"), Mood (e.g. "sad"), Activity (e.g. "gym"), Popularity (e.g. "Today's hits"), Artist (e.g. "Justin Timberlake"), Era (e.g. "70's"), Culture (e.g. "Latin"), Lyrics (e.g. "clean"), Rhythm (e.g. "groove"), Instrument (e.g. "guitar"), Tempo (e.g. "slow"), Voice (e.g. "female singers"), or Other (e.g. "favorites", "Jenna", "hi"). The percentage of playlists containing each of these tag categories is displayed in Figure 1, top.

Surprisingly, we see that tags explicitly related to vocals are not at all common compared to other types of tags, with the most common tags being related to genre, mood, or activity. Playlist titles can be viewed as labels for groups of music, and this analysis suggests that people do not often label groups of music based on explicit characteristics of the vocals. However, specific vocal characteristics (as well as many other musical attributes) may be implicit in many of the other tag categories, particularly for genre, mood, and artist. As vocal delivery style and genre are closely related, emotions communicated by the voice and the mood of the collection of songs may be related, and as each artist has a unique voice, we conclude that the relative weight of vocals may not have been disentangled from other factors.

(Top) Percentage of Spotify playlists containing one of the top 1000 tags corresponding to each category. (Middle) Percentage of descriptive search queries corresponding to each tag category, sampled from one day of search data. (Bottom) tf-idf for each term category in artist biographies compared with Wikipedia term frequencies.

We perform a similar analysis on descriptive terms from one day's worth of Spotify search queries, and obtained similarly inconclusive results, shown in Figure 1, middle.

5.2 Artist Biographies

Finally, we analyze descriptive terms that occur in 100,000 professionally authored artist biographies on Spotify. We use TF-IDF **sparck1972statistical?** to retrieve terms that are distinctive to music writers, by comparing the frequency of terms in artist biographies to the frequency of the same terms in Wikipedia. The 100 most distinctive terms, grouped into semantic categories, are displayed in Figure 1, bottom. While many terms are much more frequent in music text (e.g. "bassist", "jazz", "songwriter"), vocals specifically were not more frequently mentioned than other musical aspects. One can hypothesize that the TF-IDF method is insufficient for this particular task, due to vocals being commonly discussed outside the context of music, and thus a relatively more common word in Wikipedia.

5.3 Conclusions

Our results thus far do not show support for our general hypothesis. It may be the case that the intuitive notion of the relevance of vocals to user preference is misleading. On the other hand, it may also be the case that the importance of vocals is implicit in this data, as certain vocal styles are indicative of genre or mood. As such, the overlap between the voice and a number of the tags and descriptors analyzed prevents us from disentangling the unique effect of the voice from other musical components.

6 Vocals in Survey Data

In order to disentangle the unique effect of the voice among other components, we gathered explicit data from users. Specifically, we conducted two online survey studies in order to collect self-reported data on 1) the salient components of music, and 2) their relative ranking. Unlike prior surveys, such as **lesaffre2008potential?** that presented users with short musical excerpts and groupings of adjectives to rate, we allowed the users to freely enter their responses to the question "When you listen to music, what things about the music do you notice?". This allowed us to assess whether vocals would emerge as a salient component of music. In addition, we explored what aspects of the voice users report as being important to their musical taste.

6.1 Survey 1: Semantic Components of Music

The aim of our first survey was to establish an unranked set of self-reported salient components of music. While our hypothesis was that the vocals would be prominent, it was crucial to avoid biasing respondents as the data collected were explicit. As such, our first survey asked participants what they notice when listening to music that might make them like or dislike a song. We deliberately did not specify anything further, such as the type of music, or that we were interested in components of music, nor were participants asked to listen to musical excerpts so as not to bias responses. As an exploratory measure, we then asked participants to describe what about vocals specifically might make them like or dislike a song *after* the previous open ended questions, so as not to bias responses. Responses to these two open-response questions were manually sorted into semantic categories by the researchers.

6.1.1 Recruitment

A random sample of 50,000 people was drawn from the database of Spotify's Monthly Active Users (MUAs), divided approximately equally between the United States and Canada. 860 individuals responded to the survey, however 224 did not respond to any questions beyond the consent form, and 9 were removed for giving nonsensical responses. 626 individuals — 338 women (average age 33.6 years with a standard deviation of 16.1); 288 men (average age 30.6 years with a standard deviation of 15.5) — completed the survey in its entirety.

6.1.2 Survey

An online consent form was first presented to respondents. We then asked:

> *Q1:* When you listen to music, what things about the music do you notice? Please list as many as you can think of here:

The respondents were shown a screen with open-response format fields to complete, in which they could complete up to seven fields. On the following screen, respondents were presented with a list of their responses in random order, and asked:

> *Q2:* Please rank how important the aspects you listed are to your musical preference, where 1 is the most important.

They were then asked the following two questions about the items they ranked from 1 to 3:

> *Q3:* (a) What about _____ would make you like a song? (b) What about _____ would make you dislike a song?

Lastly, to explore what aspects of vocals may be relevant, participants responded to the following:

> *Q4:* (Please ignore these questions if you've already mentioned the vocals, the voice, the singer/rapper etc.) (a) When would vocals make you like a song? (b) When would vocals make you dislike a song?

They were then given the opportunity to comment on the survey, and were shown a final debriefing screen.

6.1.3 Semantic Categorization

A number of partially completed surveys contained responses sufficiently complete for card sorting. 317 sufficient responses — 262 from the completed surveys as well as 55 sufficiently complete partial – were then card-sorted by a team of researchers. Card-sorting is a common technique used in social sciences and elsewhere to discover clusters of related concepts **miller1969psychological?**. Traditionally, individuals are presented with physical paper "cards" that have terms and/or descriptions printed on them, printed pictures, or a group of objects. They are then asked to group items in a way that makes sense, given the research question. Here, we apply card-sorting to derive semantically meaningful groupings of musical components from the freely entered words and phrases that participants entered in each field.

Participant responses to *Q1* (i.e. "When you listen to music, what things do you notice?") were printed twice, once next to their response to *Q3a* ("What about _ would make you like a song?"), and again next to the response to *Q3b* ("What about _ would make you dislike a

song?"). As such, researchers had respondents' top 3 terms printed out twice, once next to the positive descriptive aspects of the term, and once next to the negative descriptive aspects. A term (e.g. "the lyrics") and its descriptor (e.g. "when they have meaning") comprised a card. Figure 2 shows examples of positive and negative cards that were used in card sorting.

Survey 1 sample answers for Q3. (Top) Card for an answer to Q3a. (Bottom) Card for an answer to Q3b.

As some responses were unclear (e.g. "the melody" was mentioned, but the descriptor clearly focused on the quality of the singer's voice), the research team was instructed to look at both the term and its descriptor when determining its semantic category. The researchers then reviewed the cards a second time, and defined sub-categories where necessary.

6.1.4 Results

The output of this study was two sets of semantic categories: broad semantic categories of music, and vocal-specific semantic categories. Statistical testing was not possible, given the intentionally imprecise nature of the responses. However, out of the 626 responses to the first question, 186 (29.7%) mentioned the vocals, the voice, or the singer, 348 (55.6%) mentioned the lyrics, or the words, and 101 (16.1%) mentioned both. While this is no indication of relative importance, it does demonstrate that the voice and the lyrics were salient musical components to our respondents.

The broad semantic categories determined by the researchers are presented in the left column of Table [tab:mus-attributes] (note that the other results in Table [tab:mus-attributes] are from Study 2). The category of *Emotion/mood* referred to the ability of a song to evoke emotion, whether the emotion was a match or a mismatch to the current or desired mood or current activity, whether the emotion was desirable or undesirable, and nostalgia. *Voice* included genre related terms (e.g. mumble rap, metal, auto-tune, speechiness/rapping), descriptions of how the voice is used (e.g. unique/novel, screaming, pitch/pitch range, presence or absence of effects, intensity/effort/power, emotionality, authenticity, whininess/nasality, melodic-ness), skill, the innate qualities of the voice, liking/disliking, and the mix/blend. The *Lyrics* category represented items that indicated whether or not lyrics were present, their intelligibility, the presence of profanity, how "well" crafted they were, the "message", the meaning behind them or general lyrical content and how relatable they are. *Beat/Rhythm* referred to whether it was liked/disliked, whether it "fit" the song, danceability, and uniqueness. The *Structure/complexity* of songs included liking or disliking the hook or chorus, and the song length. Instrumentation referred to drums, bass, and guitar. *Sound* referred to audio quality and related concerns. Self-explanatory categories included *Tempo/BPM*, the mention of a *Specific Artist*, *Genre*, *Harmony*, *Chords*, *Musicianship*, *Melody*, and *Popularity/Novelty*.

6.2 Survey 2: Component Ranking

While the first study aimed at determining what attributes of music were salient in the minds of listeners, the aim of the second survey was to determine the relative importance of each of the components. Specifically, we explored whether the voice would be ranked highest among a list of musical attributes. To accomplish this, participants were asked to rank a list of attributes derived from the results of our first survey, thus allowing an assessment of whether or not vocals rank above other components.

6.2.1 Recruitment

A randomized sampling method was employed among the database of Spotify's Monthly Active Users (MAUs) that had not opted-out of email correspondence. An email with a link to an online survey was sent to 50,000 potential respondents, approximately equally divided among the United States and Canada.

A total of 531 respondents — 263 of which were women (average age 31.8 years, with a standard deviation of 16.5); 268 were men (average age 34.2 years, with a standard deviation of 14.8) — completed the survey in its entirety. 429 participants completed the first half of the survey (broad semantic categories), whereas 360 participants completed the second half (vocal semantic categories).

6.2.2 Survey

An online consent form was first presented to respondents. The derived semantic categories were rephrased to be more easily understood (see Table [tab:mus-attributes], Description). Participants were presented with the new list of descriptions in random order, and asked to "Please click all the items below that would make you like or dislike a song." They were then presented with a list of all the items they had clicked, also in random order, and asked to rank them.

Broad Semantic Category	**Description**	**Borda score**	p**-value**
Emotion/mood	How it makes you feel - the emotions/mood	4641	**<0.001**
Voice	Voice/vocals	3688	**<0.001**
Lyrics	Lyrics	3656	**<0.001**
Beat/rhythm	Beat/rhythm	3460	**<0.001**
Structure/Complexity	How it's composed, the hook, the structure	2677	1.000
Musicianship	Skill of the musicians, musicianship	2583	1.000
Melody	The main melody	2577	1.000
Sound	The "sound", or the recording quality	2406	1.000
Specific Artist	The specific artist	2349	1.000

Broad Semantic Category	Description	Borda score	p-value
Genre	The specific genre	2293	1.000
Instrumentation	The musical instruments (e.g. drums, bass, guitar)	2084	1.000
Tempo/BPM	How fast or slow the song is	1828	1.000
Harmony	Harmony	1763	1.000
Chords	The chords	1086	1.000
Popularity/Novelty	How popular or unique it is	777	1.000

As a continuation of our exploratory study of vocal characteristics, a second list was then presented, comprised of terms derived from the vocal and lyrics semantic categories. For clarity, the terms were rephrased as they appear in Table [tab:voc-attributes].

6.2.3 Analytic Strategy

Responses were subjected to Borda counting **Borda1781?** and Robust Rank Aggregation **kolde2012robust?**. Borda counting is a simple procedure for aggregating votes by summing ranks. The Borda score B_i for an item i is computed as $B_i = \sum_{p=0}^{N} \left(|r_p| - r_{p,i}\right)$ where N is the number of participants, $r_{p,i}$ is participant p's rank of item i, starting at zero, and $|r_p|$ is the number of items ranked by p. The Borda method does not naturally extend to partial lists**dwork2001rank?** — we have chosen to award higher scores to preferred items in long lists.

To verify the statistical significance of our findings we supplement the Borda count with Robust Rank Aggregation (RRA), in which we compare our survey results to a null hypothesis. Each item receives a score based on its observed position, compared to an expected random ordering. Upper bounds to p-values are computed using Bonferroni correction, with values of 1.0 indicating null findings. In this work we used the implementation provided by the RobustRankAggreg package[1].

6.2.4 Results and Conclusion

Results can be found in Tables [tab:mus-attributes] and [tab:voc-attributes], with categories ordered by descending Borda count. We are able to show statistical significance of both the most salient broad and vocal semantic categories. Importantly, our results show that the Vocals and Lyrics ranked second and third among the list of components (Borda scores and RRA agree on the order of the first four broad categories). This indicates that, relative to other musical components, respondents overall indicated the importance of the vocals and lyrics.

[1] https://www.musicbusinessworldwide.com/files/2022/12/f23d5bc086957241e6177f054507e67b.png

	Vocal Semantic Categories	**Borda score**	***p*-value**
A	Singing skill	3423	**<0.001**
B	How well the voice fits or matches the rest of the music	3380	**<0.001**
C	Lyrical skill / cleverness / wit	3145	**<0.001**
D	The meaning, or the "message" of the words	3038	**0.048**
E	Authenticity / "realness"	2884	**<0.001**
F	Uniqueness	2780	**<0.001**
G	If the voice is emotional	2771	**0.006**
H	Voice strength / intensity / effort	2721	1.000
I	If the voice sounds natural	2480	1.000
J	Being able to relate	2256	1.000
K	If the voice is melodic	2202	1.000
L	Whether or not you can understand the lyrics	2056	1.000
M	If it's whiny or nasal	1801	1.000
N	Whether or not there's screaming	1771	1.000
O	The overall pitch, or the range of the pitch	1400	1.000
P	Whether or not there are lyrics	1250	1.000
Q	Whether it has production effects on it, like autotune	1230	1.000
R	Profanity, explicit lyrics	1086	1.000
S	Whether or not there is rapping	909	1.000

7 New Avenues for Research

While the musical attributes related to the broad musical categories (Table [tab:mus-attributes]) are well studied in MIR, the attributes related to vocals (Table [tab:voc-attributes]) present a number of exciting and unexplored research directions. A limiting factor to studying some of these problems, as is often the case, is the availability of data, and we encourage researchers to focus data collection efforts in these areas as well. A further limiting factor is that users of online musical platforms may come from a specific demographic, e.g. regular internet users typically younger than 35, who engage in music related activities in about one third of the online time, have had at least some musical education, and have a preference for pop, rock and classical music **lesaffre2008potential?**. In addition, our sample was derived from the U.S. and Canada. As such, a cross-cultural sample may differ in their relative preference for vocals.

Our exploratory data suggest that there is a vast space of research in tagging and measuring different qualities of the singing voice, such as whether a singing voice is authentic, powerful, natural, melodic, nasal, or emotional (Table [tab:voc-attributes], rows E, H, I, K, M and G). In addition to these categories, determined by untrained listeners, there are a number of other more specific categories such as modes of phonation that could be explored. Further, in addition to vocal qualities, there are genre-centric vocal styles, such as identifying rap or screaming (Table [tab:voc-attributes], rows S and O).

Another interesting and (as far as we are aware) unexplored research area is to measure whether a voice fits or blends well with the background music (Table [tab:voc-attributes], row B). This is somewhat related to the problem of determining "mashability" in automatic-mashup generation. This is a broad problem that is likely based on many factors, such as the style of the vocalist compared to the background, the way the song is mixed, and the overall expectations of the musical genre. We suspect this could be most easily studied when isolated vocals/backgrounds are available in order to automatically generate examples of vocals that do not match the background by blending random combinations.

The problem of identifying whether a voice is "unique" is likely challenging (Table [tab:voc-attributes], row F), as it is not necessarily a quality that can be determined in isolation, but rather relative to many other voices. One possible approach to this problem would be to treat the problem as one of outlier detection.

Production effects applied to the singing voice are increasingly common, especially different types of distortion or the infamous auto-tune (Table [tab:voc-attributes], row Q). Automatic identification of these production effects presents an interesting challenge, and one where data

could be automatically generated with the help of plugins for generating effects and databases with isolated vocals with corresponding backgrounds.

Measuring the relatability (Table [tab:voc-attributes], row J) of a singer is a quality that is relative to the listener, rather than absolute. Factors that could affect a singer's relatability could include the age, gender, culture or language of the singer relative to the listener, which might require automatic identification of each of these attributes of the singer.

Lyric intelligibility (Table [tab:voc-attributes], row L) has not been well studied, and also presents a novel challenge **ibrahim2017intelligibility?**. This problem does not necessarily directly require lyric transcription, and may be able to be determined from qualities of the audio. Similarly, determining whether a singing voice contains lyrics or is wordless has not been studied (Table [tab:voc-attributes], row P).

Automatic lyric transcription has been studied **mcvicar2014leveraging?**, **kruspe2016retrieval?** but is not yet solved, and would power the automatic estimation of many of these vocal attributes. For lyric-related terms, given textual lyrics, while some attributes would be relatively simple to estimate (e.g. whether or not there is profanity), others present interesting NLP challenges, such as estimating whether the lyrics are "clever" or are "meaningful" (Table [tab:voc-attributes], rows R, C, and D).

8 Discussion and Conclusions

While our analyses of playlist titles and search queries were inconclusive, we show evidence that English-speaking respondents from the U.S. and Canada clearly indicated that the voice is a salient component of music. Specifically, Spotify users were asked what they notice about music while listening. Despite the unassuming nature of the question, our results showed that the voice was indeed salient among the group of reported musical attributes. Furthermore, users ranked the voice as the second most important component to their musical preference, after emotions.

Our results have a number of implications. With regards to MIR research specifically, our results suggest that the voice and lyrics are indeed relevant attributes that warrant further study. While individuals may not necessarily want or know how to describe vocals themselves, i.e. in their playlists or search queries, surveying listeners directly does indicate that they find vocals to be important. As such, clarifying how the voice relates to music preference is an important topic for future research.

Secondly, users indicated that the ability of a song to evoke emotions was the most important factor. This confirms findings in prior research of the relevance of emotional content in music, and how it is linked to musical preference, e.g. **krumhansl2017listening?**. Therefore, examining how music affects the emotions of listeners remains an important theme. Interestingly, while genre was the most frequent term used to label playlists or search for music, respondents did not rank the specific genre as important relative to the other attributes. Understanding why this is the case warrants further study.

More relevant to our hypothesis, is that the vocals and the lyrics of a song were ranked second and third by respondents who were directly asked what components of music are important to their preferences. Therefore the link between emotions perceived in the voice and lyrics, and the emotions felt in listeners, is very relevant to questions of music preference. Clarification of these links was out of scope in these studies, and could be addressed in future research.

Lastly, we show the relevance of explicitly collected data that might guide future research. While we showed inconclusive findings regarding the prevalence of vocals in implicit data, we did show that the unique effect of vocals on music preference may be observed using survey data. As such, explicit data-gathering techniques often found in the social sciences, as well as collaborations with social scientists, may be of great use to MIR researchers.

9 "Butter Lyrics Over Hominy Grit": Comparing Audio and Psychology-Based Text Features in MIR Tasks

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.

9.1 Introduction

[1]

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 **van2005world?**, songs with more typical **north2020relationship?** and more negative **brand2019cultural?** lyrics appear to be more successful, and the psychological content of song lyrics appears to correlate with cultural changes in psychological traits **dewall2011tuning?**. As for music consumption, lyrics have also been shown to be a salient component of music in the minds of listeners **demetriou2018vocals?**. Furthermore, **howlin2020patients?** showed that patients are more likely to choose music with lyrics when participating in music-based pain reduction interventions; **ali2006songs?** showed that lyrics enhance self reported emotional responses to music, although melody had an overall larger effect, and **brattico2011functional?** showed

[1] https://www.musicbusinessworldwide.com/files/2022/12/f23d5bc086957241e6177f054507e67b.png

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 **hu2010lyrics?**, **mcvicar2011mining?**, **hu2009lyric?**, **wang2011music?**, genre classification **mayer2008rhyme?**, **tsaptsinos2017lyrics?** and topic detection for indexing and browsing **kleedorfer2008oh?**, **sasaki2014lyricsradar?**. **ellis2015quantifying?** 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 Section2, we discuss relevant previous work on text information extraction in the Psychology literature. Section3 will subsequently explain our research design, after which Section4 discusses the feature sets we used. Section5 describes the data collection and pre-processing procedures, after which Section6 details the experimental design. Section7 justifies our chosen analytical strategy, followed by a presentation of results in Section8 and the conclusion in Section10.

10 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 **tausczik2010psychological?**. 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, **neuman2016personality?** 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 **pennebaker2015development?**; e.g., in the context of lyrics, **markowitz201727?** used it to examine psychological distress in the lyrics of musicians that committed suicide vs. those who had not.

Secondly, **schwartz2013personality?** 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. **schwartz2013personality?** 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.

11 Research Design

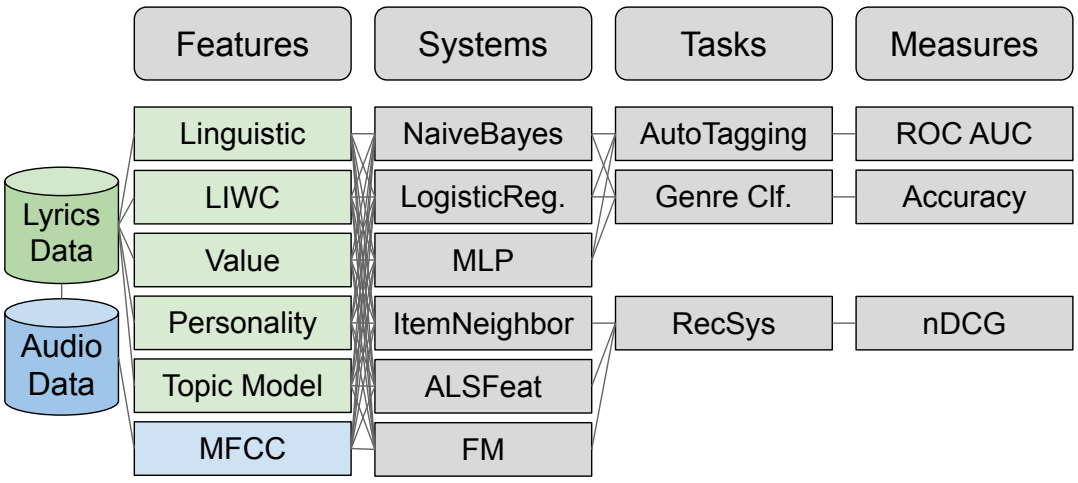

Figure 11.1: Overview of the experimental pipeline.

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.

12 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.

12.1 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.

12.2 Topic Modeling

As a more advanced feature extraction technique, we employ probabilistic Latent Semantic Analysis (pLSA) **DBLP:journals/ml/Hofmann01?** 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.

[1] https://www.gwi.com/reports/music-streaming-around-the-world

12.3 LIWC

Linguistic Inquiry Word Count (LIWC) is a software package built on a lexicon that has been validated for text analysis in psychological studies **pennebaker2015development?**. 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.

12.4 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 **goldberg1990alternative?**. 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 **goldberg1990alternative?**. 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 **saucier1996evidence?**.

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 **wilson2016disentangling?**, **wilson2018building?**, **liu2019personality?**. In our work, we used the value inventory and lexicon from **wilson2018building?**.

Both for the *personality* and *values* feature sets, we will exploit the `word2vec` model **DBLP:conf/nips/Mikolov** 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:

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

[2] https://en.wikipedia.org/wiki/List_of_instrumental_number_ones_on_the_UK_Singles_Chart

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.

12.5 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 **DBLP:conf/ismir/ChoiFSC17?** with 40 mel bins.

Table 12.1: Number of dimensions per feature set

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

13 Data Collection

We analyzed the lyrics contained in the Musixmatch dataset[1], which is the official lyrics metadata selection integrated in the Million Song Dataset (MSD) **DBLP:conf/ismir/Bertin-MahieuxEWL11?**, 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.

13.1 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.

[1] https://research.atspotify.com/2020/09/the-million-playlist-dataset-remastered/

14 Experiment

14.1 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.

14.1.1 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 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 **DBLP:conf/ismir/Schreiber15?** 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.

14.1.2 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[1]. Again, we cross-match to the MSD, now also considering MSD's LastFM social tags. Similarly to **DBLP:conf/ismir/ChoiFSC17?**, 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[2]. Due to tagging label availability, $137,095$ songs are used under this task.

[1] https://www.musixmatch.com/

[2] Each member independently screened each lyric and the screening process overall was discussed at length.

14.1.3 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) **DBLP:conf/ismir/LiangZE15?** method, as well as the *Factorization Machine* **DBLP:conf/icdm/Rendle10?** (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 **DBLP:conf/ismir/Bertin-MahieuxEWL11?**. 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[3]. Finally, the binarized *normalized Discounted Cumulative Gain* (nDCG) is considered as performance measure, for the top-100 songs recommended.

14.2 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[4][5]. 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.

[3]e.g., Spotify reports over 100 million songs in its cataloguehttps://newsroom.spotify.com/company-info/

[4]Full code of our sampling procedure is at https://anonymous.4open.science/r/lyrics-value-estimators-CE33/1_stimulus_sampling/stratified_sampling.py

[5]https://prolific.co

15 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:

$$y_i = a_{j[i]} + \beta x_i + \epsilon_i \tag{15.1}$$

$$\alpha_j = a_0 + b_0 u_j + \eta_{j1} \tag{15.2}$$

$$\beta_j = a_1 + b_1 u_j + \eta_{j2} \tag{15.3}$$

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 ([eq:2]) and ([eq:3]) 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 **gelman2006data?**.

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)[1].

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 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.

[1] https://qualtrics.com

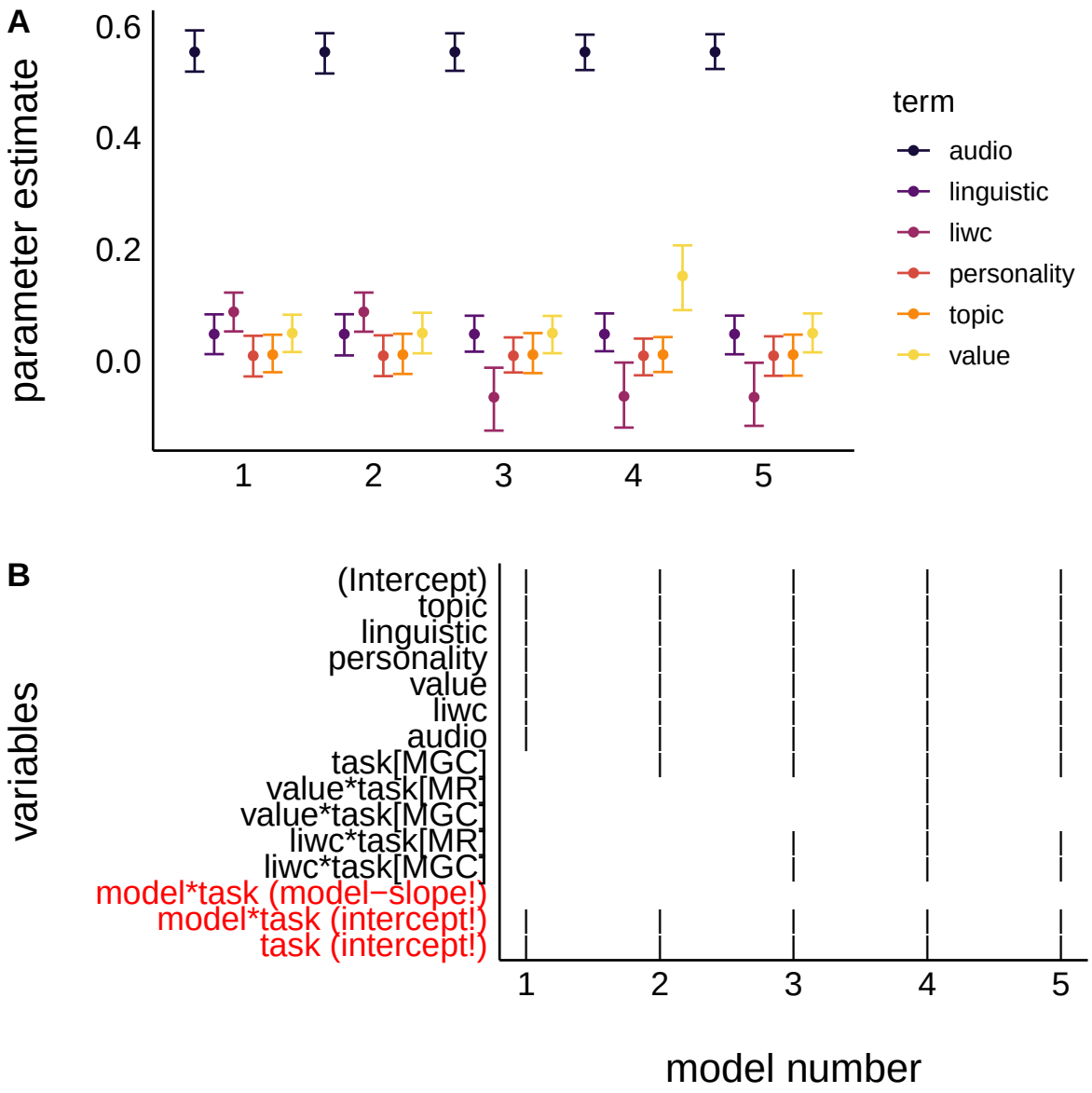

Figure 15.1: A: Parameter estimates of 5 hierarchical regression models. Error bars are 95% confidence intervals, bootstrapped 500 times. B: Specific parameters that are estimated in the each of the models. Parameters that form the structure of the model are denoted both in red and with a "!" symbol, feature sets of primary interest are denoted in black, and variables for which two terms separated by a "*" are interaction terms.

16 Results

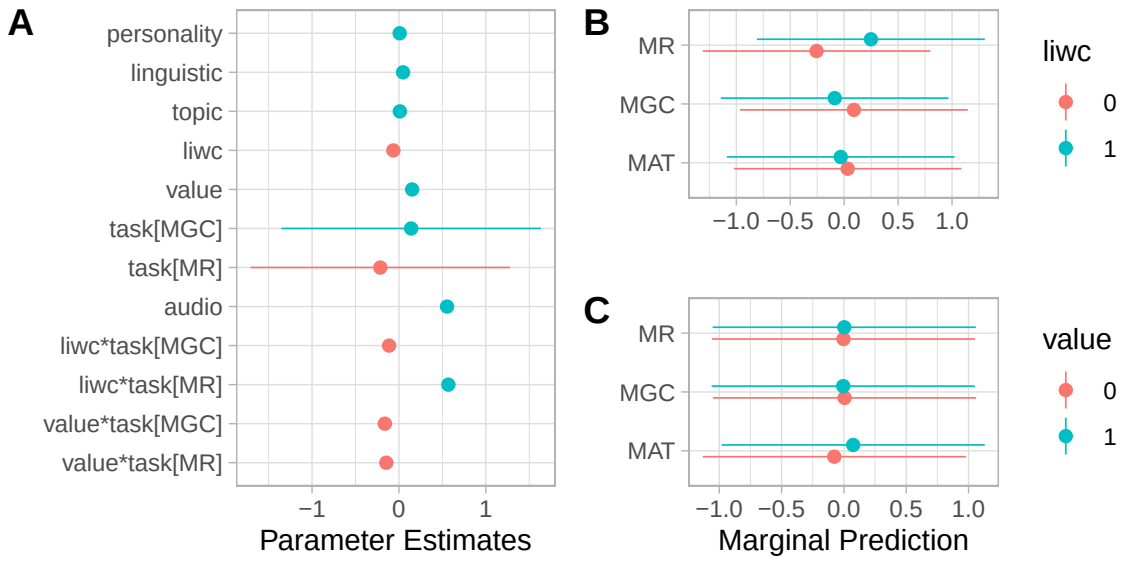

Figure 16.1: A: Parameter estimates of model 4. Error bars are 95% confidence intervals. Interaction terms are denoted with the "*" symbol. B: Predicted scores for the inclusion of *LIWC* on each of three MIR tasks, where 1 indicates that it was included and 0 indicates that it was not. C: Predicted 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.

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 **kuznetsova2017lmertest?**. 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 interaction 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 readers to the reproducibility package[1] 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. Figures3A and3B 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.

[1]It has shown correlations ranging from .45-.70 per value with longer more established procedures, test-retest reliability, as well as the typical values structure shown in Figure 2

17 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 sub-dimensions (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.

18 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.

19 Acknowledgement

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

20 Towards Automated Personal Value Estimation in Song Lyrics

Most music widely consumed in Western Countries contains song lyrics, with U.S. samples reporting almost all of their song libraries contain lyrics. In parallel, social science theory suggests that personal values - the abstract goals that guide our decisions and behaviors - play an important role in communication: we share what is important to us to coordinate efforts, solve problems and meet challenges. Thus, the values communicated in song lyrics may be similar or different to those of the listener, and by extension affect the listener's reaction to the song. This suggests that working towards automated estimation of values in lyrics may assist in downstream MIR tasks, in particular, personalization. However, as highly subjective text, song lyrics present a challenge in terms of sampling songs to be annotated, annotation methods, and in choosing a method for aggregation. In this project, we take a perspectivist approach, guided by social science theory, to gathering annotations, estimating their quality, and aggregating them. We then compare aggregated ratings to estimates based on pre-trained sentence/word embedding models by employing a validated value dictionary. We discuss conceptually 'fuzzy' solutions to sampling and annotation challenges, promising initial results in annotation quality and in automated estimations, and future directions.

21 Introduction

Popular music in Western countries almost always contains lyrics, making song lyrics a widely, repeatedly consumed **conrad2019extreme?** form of text. Over 616 million people subscribe to streaming services worldwide[1], many of whom stream more than an hour of music every day[2]. Lyrics have been shown to be a salient component of music **demetriou2018vocals?**, and out of over 1400 number-1 singles in the UK charts, only 30 were instrumental[3]. The two representative US population samples that were our annotators indicate a median 90% of songs in their libraries contain lyrics (Figure 1).

Distribution of self-reported percentage of music library containing lyrics from two representative US samples, n=505 and n=600 respectively.

It is thus not surprising that informative relationships between popular songs and their lyrical content have been shown: e.g., country music lyrics rarely include political concepts **van2005world?**, and songs with more typical **north2020relationship?** and more negative **brand2019cultural?** lyrics appear to be more successful. **howlin2020patients?** showed that patients are more likely to choose music with lyrics when participating in music-based pain reduction interventions, although melody had an overall larger effect **ali2006songs?** showed that lyrics enhance self reported emotional responses to music, and **brattico2011functional?** showed a number of additional brain regions were active during the listening of sad music with lyrics, vs. sad music without lyrics. In fields closer to MIR, **kim2020butter?** show that estimating psychological concepts from lyrics showed a small benefit in a number of MIR tasks, and **preniqi2022more?** showed a correlation between moral principles estimated from song lyrics and music preferences.

A connection between music lyrics and music preferences anticipated by theory involves the personal values perceived in the lyrics by listeners. Prior work has shown correlations between an individual's values, and the music they listen to **manolios2019influence?**, **gardikiotis2012rock?**, **swami2013metalheads?**, **preniqi2022more?**, suggesting that we seek music in line with our principles. Yet we have not seen an attempt to measure perceived personal values expressed in the lyrics themselves via human annotation or automated methods.

In this work we take a first step towards the automated estimating the values perceived in song lyrics. As artistic and expressive language, lyrics are ambiguous text: they contain different forms of analogy and wordplay **sandri2023don?**. Thus we take a perspectivist approach

[1] https://www.musicbusinessworldwide.com/files/2022/12/f23d5bc086957241e6177f054507e67b.png
[2] https://www.gwi.com/reports/music-streaming-around-the-world
[3] https://en.wikipedia.org/wiki/List_of_instrumental_number_ones_on_the_UK_Singles_Chart

to the annotations: because we expect that perceptions will vary substantially more than in other annotation tasks, we aim to represent the general perceptions of only one population. We account for the subjectivity by gathering a large number of ratings (median 27) per song from a targeted population sample (U.S.), of 360 carefully sampled song lyrics, using a psychometric questionnaire that we adjust for this purpose. We treat values in line with theory: as ranked lists, using Robust Ranking Aggregation (RRA) to arrive at our 'ground truth'. We then gather estimates from word embedding models, by measuring semantic similarity between the lyrics and a validated dictionary. We show that ranked lists from estimates correlate moderately with annotation aggregates. We then discuss the implications of our results, the limitations of this project, and anticipated future work.

22 Personal Values

The modern study of human values spans over 500 samples in nearly 100 countries over the past 30 years, and has shown a relatively stable structure **sagiv2022personal?**, as illustrated in Figure 2. Personal values are a component of personality, defined as the hierarchy of principles that guide a person's thoughts, behaviors, and the way they evaluate events **schwartz1987toward?**, **schwartz2012overview?**. Basic human values can be used to describe people or groups: social science theory suggests that each person uses a hierarchical list of values as life-guiding principles **rokeach1973nature?**, such that we prioritize some values over others as we make decisions. Schwartz's theory is the most widely used in social and cultural psychology, and has shown correlations with important behaviors, ranging from political affiliation to personal preferences **sagiv2022personal?**.

We communicate our values in order to gain cooperation and coordinate our efforts, according to Schwartz **schwartz1992universals?**. Thus our values will manifest in the words that we use **boyd2017language?**. Although personal values are traditionally measured by having individual people complete validated psychological questionnaires, it has been argued that values may be clearly expressed in the speech and text that we produce **boyd2017language?**.

Visualization of the Schwartz 10-value inventory from used in this paper, such that more abstract values of Conservation, vs. Openness to Change, and Self-transcendence vs. Self-enhancement form 4 higher-order abstract values. Illustration adapted from .

A common approach to measuring psychological aspects in text is to validate dictionaries: curated sets of words, with subsets aimed at measuring each component of the psychological aspect in question **pennebaker2015development?**, **graham2009liberals?**, **holtrop2022exploring?**, **ponizovskiy2020development?**. Some work estimating the values of the authors of text has been conducted on individuals who have written personal essays and social media posts e.g. **maheshwari2017societal?**, **ponizovskiy2020development?**, and in arguments abstracted from various forms of public facing text **kiesel2022identifying?**. However, we have not seen work aimed at measuring values *perceived* in text, measuring them along a scale as in prior work **schwartz1992universals?**, or ultimately treating them as a hierarchical list in line with theory **rokeach1973nature?**.

23 Primary Lyrics Data

We aim to collect a sample of lyric data where the lyrics are as accurate as possible, and our sample is as representative as possible. We sampled from the population of songs in the Million Playlist Dataset (MPD)[1] as it is large and recent compared to other similar datasets. The lyrics themselves were obtained through the API of Musixmatch[2], a lyrics and music language platform. Musixmatch lyrics are crowdsourced by users who add, correct, sync, and translate them. Musixmatch then engages in several steps to verify quality of content, including spam detection, formatting, spelling and translation checking, as well as manual verification by over 2000 community curators, and a local team of Musixmatch editors. Via their API, Musixmatch provided us with an estimated first 30% of the lyrics of each song.

Using the 'fuzzy' stratified sampling method described below, we sampled 2200 songs. Three members of the research team manually screened approximately 600 of the 2200 songs for inclusion. Each set of lyrics was confirmed to be a match to the actual song, and for suitability[3]. Lyrics were unsuitable if they were: 1) not English, 2) completely onomatopoetic, 3) repetitions of single words or phrases, 4) too few words to estimate values present or, 5) were not a match to the meta-data of the song, e.g. artist title, song name. This resulted in 380 songs, 20 of which were used in a pilot study to determine the number of ratings to gather per song, and 360 were used for annotation.

23.1 Fuzzy Stratified Sampling

An initial challenge is determining how to represent a corpus. In our case, the population of songs is known to be very large[4]. An ideal scenario would be one in which we aim for a known number of songs, randomly sampled from within clearly defined strata, i.e. relevant subgroups, also known as *stratified random sampling* **groves2009survey?**. However, for music, we do not know how many songs we would need to sample in order to reach saturation, what the relevant strata to randomly sample within should be, and how to measure relevant parameters from each stratum.

Some measurable strata that affect the use of language in song lyrics are clear: e.g., the year of release, which may reflect different events or time-specific colloquial slang. Others are less

[1] https://research.atspotify.com/2020/09/the-million-playlist-dataset-remastered/
[2] https://www.musixmatch.com/
[3] Each member independently screened each lyric and the screening process overall was discussed at length.
[4] e.g., Spotify reports over 100 million songs in its cataloguehttps://newsroom.spotify.com/company-info/

clear: e.g., there is no single metric of popularity for music, although it can be estimated from various sources such as hit charts. Some may be very subjective, such as genre, for which there may be some overlap of human labelling, but no clear taxonomy exists in the eyes of musicological domain experts **Liem2012MusicGap?**.

Based upon these considerations, we aimed for a stratified random sampling procedure, based on strata that we acknowledge to be justifiable given our purpose, yet in some cases conceptually 'fuzzy': (1) release date; (2) popularity, operationalized as artist playlist frequency from the MPD **DBLP:conf/recsys/ChenLSZ18?**; (3) genre, estimated from topic modeling on Million Song Dataset artist tags **schindler2012facilitating?**; (4) lyric topic, through a bag-of-words representation of the lyrics data. Popularity and Release date were divided into equally spaced bins; e.g. we divided release year into decades (60s, 70s, 80s, and so on), and genre and lyric topic were divided into categories.

Release date was quantized into 14 bins in 10-year increments from 1890-2030. Popularity was exponentially distributed, and thus manually binned, to make the quantiles per each of the 7 bins as similar as possible. Thus, the first bin contained the lowest 40% of the population in terms of popularity, while the 7th bin contained the highest 9%. Topic modelling was applied on a bag-of-word representation of the lyrics data and artist-tag data to yield 25 estimated genres and 9 lyrics topic strata, respectively.

We observed a skewness of data concentration with regard to several of our strata, e.g., songs that are recent and widely popular are most likely be drawn. To correct for this and thus get a more representative sample of an overall song catalogue, we oversample from less populated bins. For this, we use the maximum-a-posteriori (MAP) estimate of the categorical distribution of each stratum. The oversampling is controlled by concentration parameter a of the symmetric Dirichlet distribution. We heuristically set this parameter such that songs in underpopulated bins still will make up up 5-10 % of our overall pool[5]. Through this method, we subsampled our initial 2200 songs lyrics.

MDS plots derived from the correlation plot reported in , and our participant responses as confidence-weighted means

[5]Full code of our sampling procedure is at https://anonymous.4open.science/r/lyrics-value-estimators-CE33/1_stimulus_sampling/stratified_sampling.py

24 Ground-Truthing Procedure

We chose to obtain our annotations from samples of the US population, representative in terms of self-reported sex, ethnicity and age, through the Prolific[1] platform. Annotator pools comprised of two samples, the first n=505 wave participated in a pilot study to estimate the number of ratings per song needed on average, and the second n=600 wave comprised our main data collection. Participants completed the survey on the Qualtrics [2] platform.

We clearly differentiate between the Author and the Speaker of lyrics by explaining to participants that the Author of song lyrics may write from the perspective of someone or something else (the Speaker). 17 randomly selected sets of lyrics were then shown to each participant along with instructions to annotate each with the values of the Speaker. We adapted the 10-item questionnaire used in **lindeman2005measuring?** for the value annotations, as it is the shortest questionnaire for assessing personal values whose validity and reliability have been assessed[3]. As in **lindeman2005measuring?**, each questionnaire item is a specific value along with additional descriptive words e.g. POWER (social power, authority, wealth). We adjusted it by asking participants to indicate the values of the Speaker of the lyrics, and by having them indicate on a bar with -100 (opposed to their principles) on one end, and +100 (of supreme importance) on the other end instead of a likert scale. In addition, we asked participants to indicate how confident they were in their ratings, on a scale of 0 (not at all confident) to 100 (extremely confident), inspired by work that has shown that self-reported confidence in ratings can be used to estimate the accuracy of individual ratings **cabitza2020if?**.

We used a procedure similar to **DeBruine_Jones_2018?** in order to determine the number of raters. Specifically, we recruited a representative 500+ participant sample of the US using the Prolific platform, who completed our survey for 20 songs. We then computed canonical mean ratings of each of the 10 values per song, and inter-rater reliability using Cronbach's Alpha. We then estimated Cronbach's alpha for a range of subsample sizes (5 to 50 participants in increments of 5), for each of the 10 values. We repeated this procedure 10 times per increment, separately for each of the 10 values, and examined the distribution of Cronbach's Alpha. We specifically looked for the sample size with which Alpha exceeded .7 [4]. We arrived at a conservative estimate of 25 ratings per set of lyrics, with songs receiving a median 27 ratings (range 22-30).

[1] https://prolific.co
[2] https://qualtrics.com
[3] It has shown correlations ranging from .45-.70 per value with longer more established procedures, test-retest reliability, as well as the typical values structure shown in Figure 2
[4] .7 is a commonly considered an acceptable level of reliability in the form of internal consistency

24.1 Reliability, Agreement and Initial Validation

The rater reliability was estimated via intra-class correlation for each personal value, (type 2k: see **koo2016guideline?**) using the '`psych`' package in R **Psychcitation?**, all of which exceeded .9 (excellent reliability). As an initial validation, we compare data simulated from values in the upper triangle of a correlation matrix reported in **schwartz2001extending?** to those derived from our study. To aggregate our participants rankings for this purpose, we compute confidence-weighted means inspired by **cabitza2020if?**: we estimate confidence-weights by dividing participant's self-reported confidence of a given rating by the highest possible response (100), and then compute aggregated means weighted by these. For both the simulated data and confidence-weighted mean scores, we generate a multi-dimensional scaling plot (MDS) **davison2000multidimensional?** for visual comparison, which has previously been used as method to assess measurements conform to theory **ponizovskiy2020development?**, **lindeman2005measuring?**. Note: the interpretation is to observe whether each of the values appears next to expected neighboring values, and not each value's orientation. From these plots (Figure 3), in as little as our 360 annotated lyrics, we surprisingly see similar clusters and relative positioning relations emerging as those obtained from a formal cross-cultural study.

We coerced the annotated scores to ranked lists of values, such that the highest scoring value was at the top. We derived ranked lists per participant per song, and then used Robust Ranking Aggregation (RRA) to extract a single ranked list per song. Aggregation was conducted using R version 4.2.2.**Rcitation?**, and the `RobustRankAggreg` package **RRAcitation?**. Briefly, RRA produces a ranked list by comparing the probability of the observed ranking of items to rankings from a uniform distribution. Essentially, scores are determined by comparing the height of an item on a set of lists to where it would appear if its rank were randomly distributed across lists. These scores are then subjected to statistical tests, where the resulting p value is Bonferronni corrected by the number of input lists**kolde2012robust?**. Thus, when an item appears in different positions on a list, the resulting p value is high, as its position appears randomly distributed.

As lyrics are ambiguous, we expect that some songs' values are completely subjective. We operationalize these as randomly distributed rankings for all personal values for completely subjective songs, i.e. p values above .05 for all 10 items on the ranked list. Results from the RRA show 62 songs with p values above .05 for all 10 values, and 96 songs with only 1 value ranked. At most, 5 values were ranked, which occurred for 35 songs. Thus, we confirm that although there was correspondence in the scores that participants assigned per value per song, ranked lists did not always agree.

Rank correlations between NLP systems / word counts and Robust Ranking Aggregation lists, by normalization scheme.

25 Automated Scoring

For automated scoring, we use a dictionary of words associated with the 10 Schwartz values **ponizovskiy2020development?**. With this dictionary as reference, we computationally estimate the degree to which each value is reflected in the lyrics text according to traditional word counting **ponizovskiy2020development?**, as well as by assessing cosine similarity between dictionary words and lyrics texts using four classes of pre-trained word embeddings: `word2vec`, a generic English word embedding trained on Google News dataset **DBLP:conf/nips/MikolovSCCD13?**; `glove`, another generic English word embedding trained on Common Crawl dataset **DBLP:conf/emnlp/PenningtonSM14?**; `mxm-far-[1∼10]`, trained on the collected initial lyrics candidate pool, employing the Glove model **DBLP:conf/emnlp/PenningtonSM14?** (using ten models populated from ten cross-validation folds, whose parameters are tuned based on English word similarity judgement data **DBLP:conf/acl/FaruquiD14?**.); `mxm-cv-[1∼10]`, ten variants of lyrics based word-embeddings from cross-validation folds selected by Glove loss values on the validation set; and finally, `sent-bert`, a transformer model that encodes sentence into a embedding vector, fine-tuning of a generic self-supervised language model called MPNet, which is trained on a large scale English corpus **DBLP:conf/emnlp/ReimersG19?**. Our process thus resulted in 24 sets of scores: 5 from models and one from word-counting, normalized using four methods.

We take the perspective from theory that that value assessments should be seen as ranked lists, and thus coerce scores to ranked lists per model per song. We then compute rank correlations between ranked lists derived from model scores and RRA lists from participants. As RRA lists assess lack of consensus on rankings, personal values with high p values received tied rankings, at the bottom of the list. Correlations were computed using Kendall's τ which is robust to ties (Figure 4).

In earlier work **richard2003one?**, **ponizovskiy2020development?**, Pearson correlations of 0.1-0.2 were considered as moderate evidence of the validity of a proposed dictionary in relation to a psychometrically validated instrument. Although we are using a different metric, we observe several models whose mean rank correlations exceed the .10 mark. The mean Kendall's τ values were highest for the word2vec, sent-bert, and wordcount models with null normalization (SD=.24, .30, and .34 respectively). We further observe that 76% of the rank correlations for word2vec exceed the .10 mark, followed by 56.1% from sent-bert, and 47.8% from wordcounts. Although none of these models had been thoroughly optimized and thus this cannot be interpreted as a thorough benchmark, we do see evidence of higher than expected correlations.

We also explored whether our fuzzy strata might hint towards more or less automatically scorable lyrics. We found most strata to be uninformative. However, when examining the rank correlations for our overall best performing model, word2vec, we did observe higher mean correlations for some artist tag topics than others (Figure 5). In particular, topics 10 (which included the tags: 'jazz', 'chillout', 'lounge', 'trip-hop', 'downtempo'), 11 (which included the tags like: 'metal', 'celtic', 'thrash metal', 'dutch', 'seen live'), and 16 (which included tags like: 'country', 'Soundtrack', 'americana', 'danish', 'Disney'). Although speculative, we do expect that certain genres are more difficult to interpret than others, in particular for people who are generally unfamiliar with such music.

Rank correlations between word2vec scores Robust Ranking Aggregation lists, per genre grouping operationalized as Artist Tag Topic.

26 Descriptive Analyses

We conduct a further exploratory data analysis by examining the gathered value annotations with respect to the song strata introduced in Section 3.1. To better understand the overall patterns of value rankings in songs we visualize the average ranking of each value for each level of each stratum. To reflect the uncertainty of aggregated ranking from RRA, we employ 'truncated' rankings: the values within each aggregated ranked list are considered ties if their p-values higher than the threshold ($p = 0.05$), hence with high uncertainty in their ranking positions.[1]

In all results, we observe that there is a tendency of overall value ranking: 1) a generally strong presence of HEDONISM in higher ranks in all cases, followed by STIMULATION and SELF (SELF-DIRECTION). 2) ACHIEVEMENT and POWER generally follow next across all figures, and 3) the rest of the values, including BENEVOLENCE, UNIVERSALISM, SECURITY, CONFORMITY, and TRADITION overall rank lower, but show higher variability across strata. We refer to these three groups of values as *Group1* (HEDONISM, STIMULATION and SELF), *Group2* (ACHIEVEMENT and POWER), and *Group3* (the rest) for the rest of the section.

Average value ranking from 'release year' (A) and 'artist-playlist frequency' (B). x and y axis represent the strata and average ranking measure from RRA, respectively. Each point in different point shapes and vertical bars denote the average ranking value and its confidence interval (at 95% level). For visual convenience, we connected the same values with lines.

Zooming in each to stratum, in Figure 6, we observe that the 'release year' (sub-figure **A**) strata show the most consistent and visible trend especially for Group3, which generally declines over time. Such a trend is not as obvious in Group1 and only partially observed in Group2. The low presence of Group3 is especially noticeable in the 1990s, although it regained its presence to some degree, a pattern which the SELF value from Group1 partially shares. Such visible movements suggest that the rank of specific values may evolve over time. In sub-figure **B**, we observe the most flat response across all strata considered: beyond the fluctuation pattern that is shared by all groups, there is no substantial variability among groups, which implies that popularity might not be as correlated as the 'release year'.

Average value ranking from 'artist-tag topic' (C) and 'lyrics topic' (D).

[1] We assume that the adjusted exact p-value from RRA monotonically decreases as the rank position ascends (i.e., the lower the p-value is, the higher the ranking position is).

Moving onto Figure 7, we discuss the value presence pattern in two 'topic' strata. First, in sub-figure **C**, we observe that Group3 values show overall higher variability than 'artist playlist frequency'. It is notable that there are a few distinct topics in which Group3 values show a significant difference; the sixth, seventh and fourteenth topics, which correspond to the 'under 2000 listeners/musical', 'folk/singer-songwriter', and 'Hip-Hop/rap' topics when represented in primary topic terms. Specifically, we see that first two topics show a high presence of Group3 values, while the latter topics show the least presence of Group3 values. It suggests that the artists in these styles/genres were perceived on average to present clearly different sets of values through their lyrics, distinguished by the inclusion/exclusion of values such as BENEVOLENCE or UNIVERSALISM.

Finally, considering sub-figure **D**, we observe a similar pattern as 'artist playlist frequency' in 6, albeit with relatively more variability in Group3 values. Notably, the 'rap/hip-hop' lyrics topic shows the least presence of Group3 values, which aligns to the observation from previous sub-figure. The 'sad/romantic1' topic, on the other hand, shows the highest ranking of Group3 values. Another remarkable topic is 'gospel/reggae' topic, where HEDONISM value is least present, which semantically aligns well with the typical lyrical theme of those songs.

27 Limitations and Future Work

In this work we attempt to ground-truth perceptions of ambiguous song lyrics for perceived human values. We adopt a validated questionnaire from the social sciences for this purpose, in addition to a purposeful, if conceptually 'fuzzy', stratified sampling strategy, and estimate the average number of ratings needed to estimate the average perception of values in a song. We acknowledge our current sample of 360 lyrics is small and may need expansion for more typical work, and that, while we had a representative population sample, not every member of the sample rated every song. We thus did gather diverse opinions, but cannot claim they fully represent the target population. In addition, the small sample of songs allowed for only limited observation of patterns that might emerge in larger samples with relation to our defined strata, and indefinite conclusions given the overall massive population of songs in existence. We also did not assess whether variations on the annotation instrument might result in substantial differences in the annotations we received **kern2023annotation?**, nor did we repeat our procedure **inel2023collect?**. In addition, we acknowledge that participants from different groups will perceive and thus annotate corpora differently **homan2022annotator?**, **prabhakaran2023framework?**. Thus, we expect that lyrics may be especially sensitive to varying perceptions, which we did not explore in this work. Finally, we only provide a preliminary comparison to automated scoring methods, and did not leverage the most contemporary tools for this purpose (e.g. Large Language Models). All of these are rich and promising avenues for future work.

The most interesting avenues are potential relationships that could be revealed with more annotated songs, and eventual automated scoring methods. In particular, we see potential in understanding music consumption more broadly from patterns revealed in the dominant value hierarchies in specific music genres, popularity segments, lyrical topics, and even release year. And for understanding music consumption more narrowly, from patterns revealed in an individual's music preferences, and the degree to which they conform with their own value hierarchy.

28 Conclusion

Song lyrics remain a widely and repeatedly consumed, yet ambiguous form of text, and thus a promising and challenging avenue for research into better understanding the people that consume them. We observe promising initial results for the annotation of personal values in songs, despite our limitations. MDS plots of aggregated ratings showed the beginnings of the expected structure of values, conforming more closely than might be expected from as little as 360 songs. We also observed high inter-rater reliability in the raw scores, suggesting a sufficiently reliable annotation procedure with 25 ratings. Thus, we see promise on our method for ground-truthing lyrics despite their ambiguity. A post-hoc procedure revealed that 15 ratings may be enough on average: we repeatedly subsampled 5, 10, 15 and 20 ratings for each value within each song, and calculated pearson correlations between subsample means and canonical means. From this, we see Pearson correlations to the canonical mean exceed 0.9 for all values from 15 subsampled ratings. Further lyric annotation may thus require fewer annotations per song than what was gathered in this work. In addition, we observe promising rank correlations between ranked rater scores and our automated methods, with over 75% of the rankings in our best performing model above a minimal threshold of .10. Despite inherent challenges in the task, our method shows initial promise, and multiple fruitful avenues for future work.

29 Ethics Statement

Our study includes data gathered from people, and was approved by the Human Research Ethics board of our university. We follow Prolific's guidelines on fair compensation to set our compensation rates. Survey design and data handling were pre-discussed with our institutional data management and research ethics advisors, we obtained formal data management plan and human research ethics approval. Participants gave informed consent before proceeding with the survey, which informed them of the intentions of use for their data, and that it could be withdrawn at any time.

[1] D. E. Knuth, "Literate programming," *Comput. J.*, vol. 27, no. 2, pp. 97–111, May 1984, doi: 10.1093/comjnl/27.2.97.