TOWARDS AUTOMATED PERSONAL VALUE ESTIMATION IN SONG LYRICS

Andrew M. Demetriou[1] **Jaehun Kim**[2] **Sandy Manolios**[1] **Cynthia C. S. Liem**[1]

[1] Delft University of Technology
[2] SiriusXM/Pandora

a.m.demetriou@tudelft.nl / c.c.s.liem@tudelft.nl

ABSTRACT

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.

1. INTRODUCTION

Popular music in Western countries almost always contains lyrics, making song lyrics a widely, repeatedly consumed [1] 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 [2], 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).

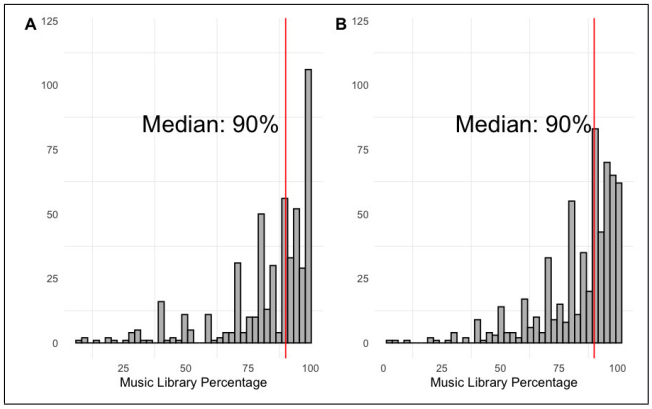

Figure 1. Distribution of self-reported percentage of music library containing lyrics from two representative US samples, n=505 and n=600 respectively.

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 [3], and songs with more typical [4] and more negative [5] lyrics appear to be more successful. [6] 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 [7] showed that lyrics enhance self reported emotional responses to music, and [8] showed a number of additional brain regions were active during the listening of sad music with lyrics, vs. sad music without lyrics. In fields closer to MIR, [9] show that estimating psychological concepts from lyrics showed a small benefit in a number of MIR tasks, and [10] 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 [10–13], suggesting that we seek music in

[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

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 [14]. Thus we take a perspectivist approach 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.

2. 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 [15], 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 [16, 17]. 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 [18], 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 [15].

We communicate our values in order to gain cooperation and coordinate our efforts, according to Schwartz [19]. Thus our values will manifest in the words that we use [20]. 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 [20].

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 [22–25]. 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. [25, 26], and in arguments abstracted from various forms of public facing text [27]. However, we have not seen work aimed at measuring values *perceived* in text, measuring them along a scale as in prior work [19], or ultimately treating them as a hierarchical list in line with theory [18].

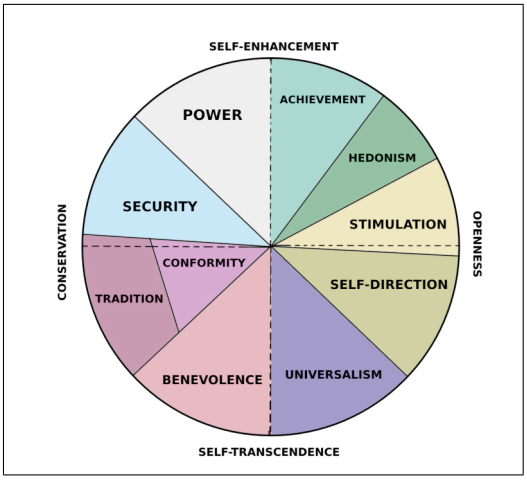

Figure 2. Visualization of the Schwartz 10-value inventory from [19] used in this paper, such that more abstract values of Conservation, vs. Openness to Change, and Self-transcendence vs. Self-enhancement form 4 higher-order abstract values. Illustration adapted from [21].

3. 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) [4] as it is large and recent compared to other similar datasets. The lyrics themselves were obtained through the API of Musixmatch [5], 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 [6]. 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.

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

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

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

3.1 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 [7]. 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* [28]. 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 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 [29].

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 [30]; (3) genre, estimated from topic modeling on Million Song Dataset artist tags [31]; (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 [8]. Through this method, we subsampled our initial 2200 songs lyrics.

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

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

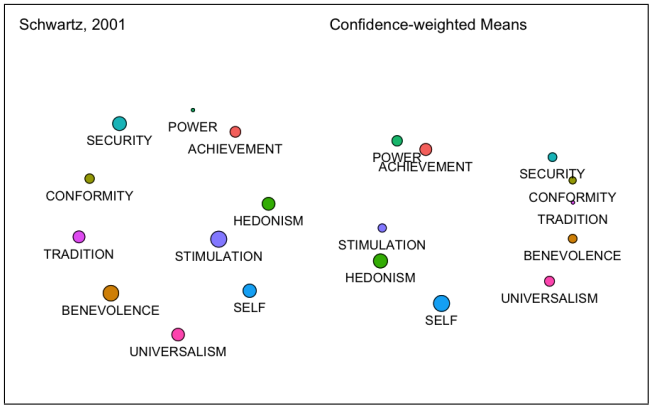

Figure 3. MDS plots derived from the correlation plot reported in [32], and our participant responses as confidence-weighted means

4. 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 [9] 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 [10] 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 [33] for the value annotations, as it is the shortest questionnaire for assessing personal values whose validity and reliability have been assessed [11]. As in [33], 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 [34].

We used a procedure similar to [35] 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

[9] https://prolific.co

[10] https://qualtrics.com

[11] 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

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 [12]. We arrived at a conservative estimate of 25 ratings per set of lyrics, with songs receiving a median 27 ratings (range 22-30).

4.1 Reliability, Agreement and Initial Validation

The rater reliability was estimated via intra-class correlation for each personal value, (type 2k: see [36]) using the 'psych' package in R [37], 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 [32] to those derived from our study. To aggregate our participants rankings for this purpose, we compute confidence-weighted means inspired by [34]: 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) [38] for visual comparison, which has previously been used as method to assess measurements conform to theory [25, 33]. 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. [39], and the `RobustRankAggreg` package [40]. 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 [41]. 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.

[12] .7 is a commonly considered an acceptable level of reliability in the form of internal consistency

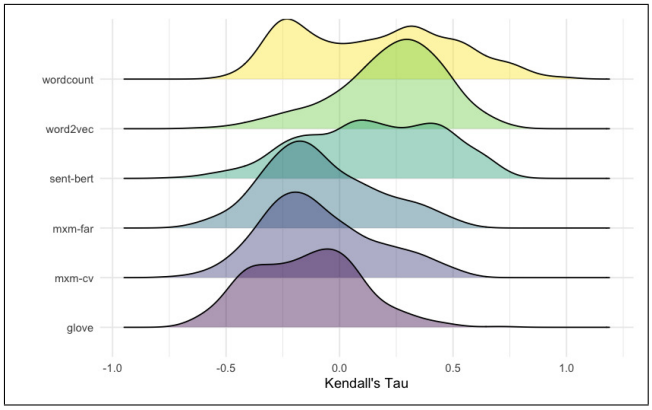

Figure 4. Rank correlations between NLP systems / word counts and Robust Ranking Aggregation lists, by normalization scheme.

5. AUTOMATED SCORING

For automated scoring, we use a dictionary of words associated with the 10 Schwartz values [25]. 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 [25], 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 [42]; `glove`, another generic English word embedding trained on Common Crawl dataset [43]; `mxm-far-[1∼10]`, trained on the collected initial lyrics candidate pool, employing the Glove model [43] (using ten models populated from ten cross-validation folds, whose parameters are tuned based on English word similarity judgement data [44].); `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 [45]. 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).

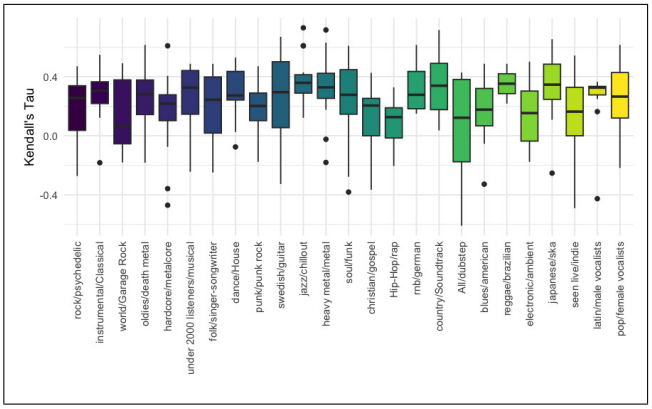

Figure 5. Rank correlations between word2vec scores Robust Ranking Aggregation lists, per genre grouping operationalized as Artist Tag Topic.

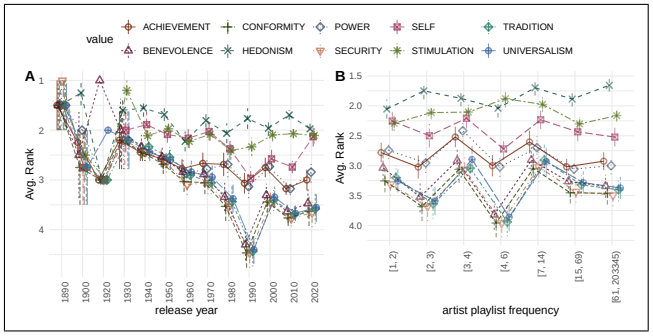

Figure 6. Average value ranking from 'release year' (A) and 'artist-playlist frequency' (B). x and y axis represent the strata and average ranking measure from RRA, respectively. Each point in different point shapes and vertical bars denote the average ranking value and its confidence interval (at 95% level). For visual convenience, we connected the same values with lines.

In earlier work [25, 46], 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.

6. 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. [13]

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.

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'.

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

[13] 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).

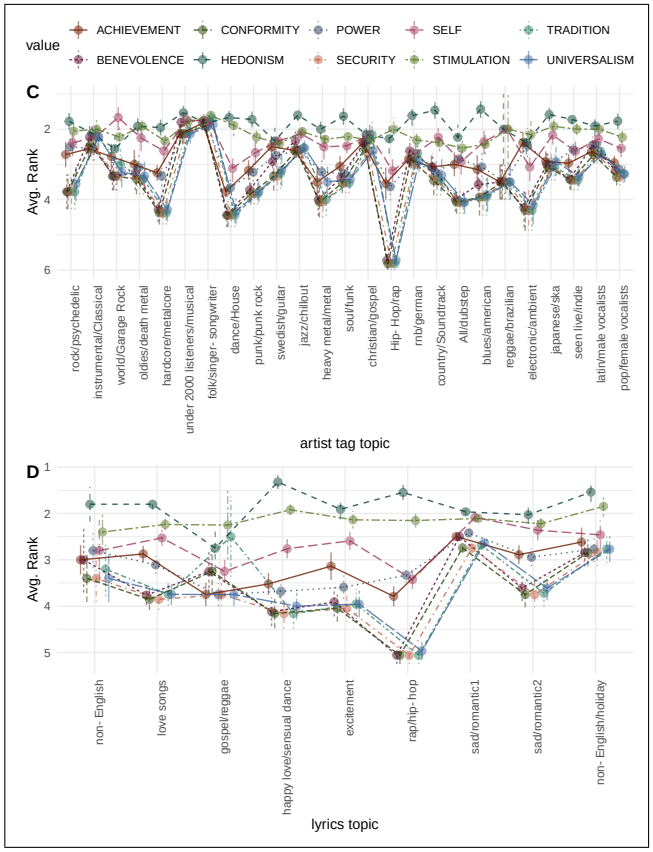

Figure 7. Average value ranking from 'artist-tag topic' (C) and 'lyrics topic' (D).

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.

7. 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 [47], nor did we repeat our procedure [48]. In addition, we acknowledge that participants from different groups will perceive and thus annotate corpora differently [49, 50]. 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.

8. 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.

9. 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.

10. REFERENCES

[1] F. Conrad, J. Corey, S. Goldstein, J. Ostrow, and M. Sadowsky, "Extreme re-listening: Songs people love... and continue to love," *Psychology of Music*, vol. 47, no. 2, pp. 158–172, 2019.

[2] A. Demetriou, A. Jansson, A. Kumar, and R. M. Bittner, "Vocals in music matter: the relevance of vocals in the minds of listeners." in *ISMIR*, 2018, pp. 514–520.

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

[4] A. C. North, A. E. Krause, and D. Ritchie, "The relationship between pop music and lyrics: A computerized content analysis of the united kingdom's weekly top five singles, 1999–2013," *Psychology of Music*, p. 0305735619896409, 2020.

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

[6] C. Howlin and B. Rooney, "Patients choose music with high energy, danceability, and lyrics in analgesic music listening interventions," *Psychology of Music*, p. 0305735620907155, 2020.

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

[8] E. Brattico, V. Alluri, B. Bogert, T. Jacobsen, N. Vartiainen, S. K. Nieminen, and M. Tervaniemi, "A functional mri study of happy and sad emotions in music with and without lyrics," *Frontiers in psychology*, vol. 2, p. 308, 2011.

[9] J. Kim, A. M. Demetriou, S. Manolios, M. S. Tavella, and C. C. Liem, "Butter lyrics over hominy grit: Comparing audio and psychology-based text features in mir tasks." in *ISMIR*, 2020, pp. 861–868.

[10] V. Preniqi, K. Kalimeri, and C. Saitis, "" more than words": Linking music preferences and moral values through lyrics," *arXiv preprint arXiv:2209.01169*, 2022.

[11] S. Manolios, A. Hanjalic, and C. C. Liem, "The influence of personal values on music taste: towards value-based music recommendations," in *Proceedings of the 13th ACM Conference on Recommender Systems*, 2019, pp. 501–505.

[12] A. Gardikiotis and A. Baltzis, "'rock music for myself and justice to the world!': Musical identity, values, and music preferences," *Psychology of Music*, vol. 40, no. 2, pp. 143–163, 2012.

[13] V. Swami, F. Malpass, D. Havard, K. Benford, A. Costescu, A. Sofitiki, and D. Taylor, "Metalheads: The influence of personality and individual differences on preference for heavy metal." *Psychology of Aesthetics, Creativity, and the Arts*, vol. 7, no. 4, p. 377, 2013.

[14] M. Sandri, E. Leonardelli, S. Tonelli, and E. Ježek, "Why don't you do it right? analysing annotators' disagreement in subjective tasks," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 2420–2433.

[15] L. Sagiv and S. H. Schwartz, "Personal values across cultures," *Annual review of psychology*, vol. 73, pp. 517–546, 2022.

[16] S. H. Schwartz and W. Bilsky, "Toward a universal psychological structure of human values." *Journal of personality and social psychology*, vol. 53, no. 3, p. 550, 1987.

[17] S. H. Schwartz, "An overview of the schwartz theory of basic values," *Online readings in Psychology and Culture*, vol. 2, no. 1, p. 11, 2012.

[18] M. Rokeach, *The nature of human values.* Free press, 1973.

[19] S. H. Schwartz, "Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries," in *Advances in experimental social psychology*. Elsevier, 1992, vol. 25, pp. 1–65.

[20] R. L. Boyd and J. W. Pennebaker, "Language-based personality: A new approach to personality in a digital world," *Current opinion in behavioral sciences*, vol. 18, pp. 63–68, 2017.

[21] G. R. Maio, "Mental representations of social values," in *Advances in experimental social psychology*. Elsevier, 2010, vol. 42, pp. 1–43.

[22] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," Tech. Rep., 2015.

[23] J. Graham, J. Haidt, and B. A. Nosek, "Liberals and conservatives rely on different sets of moral foundations." *Journal of personality and social psychology*, vol. 96, no. 5, p. 1029, 2009.

[24] D. Holtrop, J. K. Oostrom, W. R. J. van Breda, A. Koutsoumpis, and R. E. de Vries, "Exploring the application of a text-to-personality technique in job interviews," *European Journal of Work and Organizational Psychology*, vol. 31, no. 6, pp. 799–816, 2022.

[25] V. Ponizovskiy, M. Ardag, L. Grigoryan, R. Boyd, H. Dobewall, and P. Holtz, "Development and validation of the personal values dictionary: A theory–driven tool for investigating references to basic human values in text," *European Journal of Personality*, vol. 34, no. 5, pp. 885–902, 2020.

[26] T. Maheshwari, A. N. Reganti, S. Gupta, A. Jamatia, U. Kumar, B. Gambäck, and A. Das, "A societal sentiment analysis: Predicting the values and ethics of individuals by analysing social media content," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, 2017, pp. 731–741.

[27] J. Kiesel, M. Alshomary, N. Handke, X. Cai, H. Wachsmuth, and B. Stein, "Identifying the human values behind arguments," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022, pp. 4459–4471.

[28] R. M. Groves, F. J. Fowler Jr, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau, *Survey methodology*. John Wiley & Sons, 2009, vol. 561.

[29] C. C. S. Liem, A. Rauber, T. Lidy, R. Lewis, C. Raphael, J. D. Reiss, T. Crawford, and A. Hanjalic, "Music Information Technology and Professional Stakeholder Audiences: Mind the Adoption Gap," in *Dagstuhl Follow-Ups*. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2012, vol. 3. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2012/3475/

[30] C. Chen, P. Lamere, M. Schedl, and H. Zamani, "Recsys challenge 2018: automatic music playlist continuation," in *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, S. Pera, M. D. Ekstrand, X. Amatriain, and J. O'Donovan, Eds. ACM, 2018, pp. 527–528. [Online]. Available: https://doi.org/10.1145/3240323.3240342

[31] A. Schindler, R. Mayer, and A. Rauber, "Facilitating comprehensive benchmarking experiments on the million song dataset." in *ISMIR*. International Society for Music Information Retrieval, 2012, pp. 469–474.

[32] S. H. Schwartz, G. Melech, A. Lehmann, S. Burgess, M. Harris, and V. Owens, "Extending the cross-cultural validity of the theory of basic human values with a different method of measurement," *Journal of cross-cultural psychology*, vol. 32, no. 5, pp. 519–542, 2001.

[33] M. Lindeman and M. Verkasalo, "Measuring values with the short schwartz's value survey," *Journal of personality assessment*, vol. 85, no. 2, pp. 170–178, 2005.

[34] F. Cabitza, A. Campagner, and L. M. Sconfienza, "As if sand were stone. new concepts and metrics to probe the ground on which to build trustable ai," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, pp. 1–21, 2020.

[35] L. M. DeBruine and B. C. Jones, "Determining the number of raters for reliable mean ratings," Aug 2018. [Online]. Available: osf.io/x7fus

[36] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of chiropractic medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[37] William Revelle, *psych: Procedures for Psychological, Psychometric, and Personality Research*, Northwestern University, Evanston, Illinois, 2023, r package version 2.3.9. [Online]. Available: https://CRAN.R-project.org/package=psych

[38] M. L. Davison and S. G. Sireci, "Multidimensional scaling," in *Handbook of applied multivariate statistics and mathematical modeling*. Elsevier, 2000, pp. 323–352.

[39] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[40] R. Kolde, *RobustRankAggreg: Methods for Robust Rank Aggregation*, 2022, r package version 1.2.1. [Online]. Available: https://CRAN.R-project.org/package=RobustRankAggreg

[41] R. Kolde, S. Laur, P. Adler, and J. Vilo, "Robust rank aggregation for gene list integration and meta-analysis," *Bioinformatics*, vol. 28, no. 4, pp. 573–580, 2012.

[42] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, Eds., 2013, pp. 3111–3119. [Online]. Available: https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

[43] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, A. Moschitti, B. Pang, and W. Daelemans, Eds. ACL, 2014, pp. 1532–1543. [Online]. Available: https://doi.org/10.3115/v1/d14-1162

[44] M. Faruqui and C. Dyer, "Community evaluation and exchange of word vectors at wordvectors.org," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*. The Association for Computer Linguistics, 2014, pp. 19–24. [Online]. Available: https://doi.org/10.3115/v1/p14-5004

[45] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, 2019, pp. 3980–3990. [Online]. Available: https://doi.org/10.18653/v1/D19-1410

[46] F. D. Richard, C. F. Bond Jr, and J. J. Stokes-Zoota, "One hundred years of social psychology quantitatively described," *Review of general psychology*, vol. 7, no. 4, pp. 331–363, 2003.

[47] C. Kern, S. Eckman, J. Beck, R. Chew, B. Ma, and F. Kreuter, "Annotation sensitivity: Training data collection methods affect model performance," *arXiv preprint arXiv:2311.14212*, 2023.

[48] O. Inel, T. Draws, and L. Aroyo, "Collect, measure, repeat: Reliability factors for responsible ai data collection," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 11, no. 1, 2023, pp. 51–64.

[49] C. Homan, T. C. Weerasooriya, L. Aroyo, and C. Welty, "Annotator response distributions as a sampling frame," in *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, 2022, pp. 56–65.

[50] V. Prabhakaran, C. Homan, L. Aroyo, A. Parrish, A. Taylor, M. Díaz, and D. Wang, "A framework to assess (dis) agreement among diverse rater groups," *arXiv preprint arXiv:2311.05074*, 2023.