Ground-truthing perspectives on highly subjective text: basic human values perceived in song lyrics

Anonymous ACL submission

Abstract

We present an interdisciplinary approach to gathering a dataset on a highly subjective text annotation task. The task thus requires explicit insight into broad human annotator perceptions, and conscious curation of what will be annotated. With strong inspiration from best practices in the social sciences, we add to emerging and increasing calls for greater accountability with regard to data and its quality. For our task, we choose the annotation of human values as they are perceived in song lyrics. We present our strategy to select song lyrics for annotation, draw annotators from a representative US sample, estimate number of annotators needed, and assess data quality. We obtain a dataset of 360 richly annotated lyrics, and substantiate the benefits of our approach, which can be adapted to many domains and tasks. Finally, we give a first illustration of how our data can be employed in connection to applied machine learning approaches.

1 Introduction

With growing interest in AI and the rising popularity of Large Language Models, AI advances appear to require larger datasets to train models, which ideally need few human annotations. At the same time, language is a cultural phenomenon, in which human interpretation plays a key role in transmission and understanding.

In broader applications where machine learning techniques may automate and scale up actions that formerly relied on human perception and judgement, the question of what makes for good data and 'ground truth' to depart from has been less articulated and appreciated than the promise of generalizability and scalability (Birhane et al., 2022; Sambasivan et al., 2021). However, calls for data-centric AI have recently emerged[1], as has recognition that human annotator disagreement can be a meaningful signal, rather than noise suggesting unreliable annotation (Aroyo and Welty, 2015). In parallel, awareness of the need for more explicit data documentation is rising (Gebru et al., 2021; Mitchell et al., 2019; Geiger et al., 2020), with institutional efforts to encourage responsible practices visible in the *ACL communities (Rogers et al., 2021) with mandatory checklists accompanying manuscript submissions.

Computational researchers historically have not been trained to be aware of data quality considerations. Standardized checklists, forms and best practice 'rules of thumb' help lowering the threshold to report and discuss these, yet may not stimulate critical reflection on current practices. For example, in response to the question, "How many annotators would be needed for NLP corpus ground truth?", a well-cited book on natural language annotation for machine learning (Pustejovsky and Stubbs, 2013) suggests to "have your corpus annotated by at least two people (more is preferable, but not always practical)" before being ready to move on to gold standard data. This is a remarkably low number, without clear substantiation of whether this indeed would be sufficient.

Typically for a much longer time than the computational domain, other disciplines have been building expertise on how to best curate data, and capture aspects of the data that may not trivially be measurable. For example, both in archives and museums, long-standing traditions of purposeful and well-documented curation exist (Jo and Gebru, 2020; Huang and Liem, 2022). In quantiative psychology, practices exist to gather responses from a sample people that represent a population, and reliabily measure *constructs*, i.e., phenomena that cannot directly physically be quantified. For this, the basics of psychometrics (Furr and Bacharach, 2014) and survey science (Groves et al., 2009) are often taught as entry-level courses to social science students, which inform robust sampling

[1] see https://datacentricai.org/

and survey design for gathering human responses. While this expertise has been referred to in several works targeting computationally oriented communities (Welty et al., 2019; Jacobs and Wallach, 2021; Kern et al., 2023), institutionalized uptake of the expertise remains rare.

Present work resulted from an interdisciplinary team with backgrounds in the computational and social sciences. Our aim was to take steps towards ground-truthing procedures for (highly) subjective annotation tasks using perceived human values in song lyrics as a case study. We approach this task by merging the idea that variance in annotation is signal rather than noise (Aroyo and Welty, 2015), with data-collection principles from the social sciences, increasing the odds of representative measurements by 1) being purposeful about strata in data sampling and the 2) population of annotators, 3) considering distributions of annotations as the ground truth, and therefore 4) investigating the impact of high numbers of annotations per item. We explicitly rely on evidence-backed theory of basic human values, and employ 5) measurement methods that were shown in prior work to be reliable and valid. We relate our results to established prior literature showing initial promise in our methods, and suggest ways that our approach can be applied to multiple domains where variation in the annotations is expected to be meaningful.

2 Background

2.1 Perspectivist Ground Truthing

Automated systems often rely on manually annotated reference data for training and evaluation. Multiple labels from multiple annotators are gathered for reasons associated with the annotators, e.g. a lack of trust in crowdsourcing or annotations from non-experts, or because there is an expectation that people will vary in their responses to the phenomenon of interest (Cabitza et al., 2023; Basile et al., 2020). These annotations are then aggregated to produce a single label that is used to train and/or evaluate systems, as it is often incumbent on automated systems to produce a singular response.

Thus, most problems are treated as 'classification' problems. Variance in reported annotations is removed, usually by taking the label chosen most often by the annotators. Even in 'objective' problems where annotators are medical experts (Kompa et al., 2021), variance is often treated as an error even when a case can be made that there are indeed multiple ways to interpret the phenomenon of interest (Aroyo and Welty, 2015), e.g., when different groups of annotators reliably label media differently (Prabhakaran et al., 2023; Homan et al., 2022), or when the task itself is ambiguous (Artstein and Poesio, 2008).

A growing movement in the field of ground-truthing has taken to viewing this variation in some instances as being a part of the 'ground truth'[2]. It is argued that annotation projects occur on a continuum: on one end are objective phenomena whose interpretation is not expected to vary based on the perspective of the annotator, and on the other are phenomena where it is indeed expected to vary based on the lived experience, feelings etc. of the annotator (Cabitza et al., 2023). In some instances, the expectation is that there will be multiple valid labels for an item, based on the social group of the annotator e.g. (Prabhakaran et al., 2023) or because the text itself is ambiguous (Sandri et al., 2023a). Thus, variance in annotator characteristics may lead to a distribution of annotations.

Although determining the degree of subjectivity of a task is a challenge, and research is ongoing in terms of appropriate methods and metrics to extract, the Perspectivist approach advocates creating and reporting disaggregated data (Cabitza et al., 2023), so that more appropriate methods can be applied as they are developed, thus allowing for a continuous update as to knowledge on the dataset (Liem and Demetriou, 2023).

2.2 Human values

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 (Rokeach, 1973). Schwartz's theory is the most widely used in social and cultural psychology, and broadly defines basic human values as abstract goals that guide and motivation actions towards them, across contexts (Sagiv and Schwartz, 2022).

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 (Sagiv and Schwartz, 2022), as illustrated in Figure 1. This structure been observed across cultures in terms of the specific values present, and which values are prioritized together. Obtained

[2]Referred to as the Perspectivist manifesto.

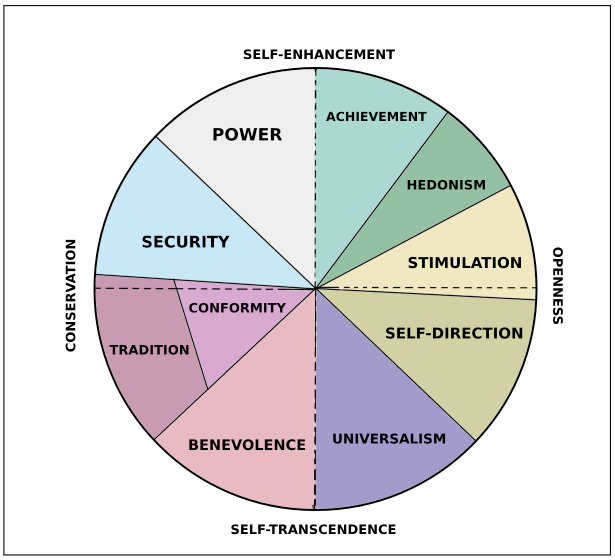

Figure 1: Visualization of the Schwartz 10-value inventory from (Schwartz, 1992) used in this paper, such that more abstract values of Conservation, vs. Openness to Change, and Self-transcendence vs. Self-enhancement form 4 higher-order abstract values. Illustration adapted from (Maio, 2010).

scores across cultures also correlate with a broad range of impactful phenomena. Cultures valuing conservation and conforming to authority tend towards religiosity and away from openness and self-direction. Altruistic behavior correlates with self-transcendent values like benevolence and universalism, and competitiveness and unethical behavior correlate with self-enhancement goals like achievement and power. Right-wing political ideology correlates with tradition, conformity and security, where universalism correlates with left-wing ideology (Sagiv and Schwartz, 2022).

As such, the structure can be used to understand what individuals use to guide their actions, but also what entire populations prioritize when representative samples are aggregated. In addition, the relative stability of the structure allows for a convenient method to estimate the reliability and validity of measurements in novel contexts: new measurement methods should, in principle, show similar structure.

2.3 Human Values in Text

We communicate our values in order to gain cooperation and coordinate our efforts, according to Schwartz (Schwartz, 1992), which will manifest in the form of words in speech and text (Boyd and Pennebaker, 2017). A vast amount of text and speech is produced and consumed: every minute in 2022 an estimated 1 million hours of content were streamed, and over 350,000 tweets were shared [3].

However, researchers rarely, if ever, have access to the 'ground truth' values of people that produce the most impactful text or speech, e.g. politicians, famous artists, authors, and other influencers. Further, how language is perceived may vary substantially depending on what group is perceiving it: e.g. perceptions have been shown to vary widely by group in terms of what language is harmful (Solaiman et al., 2023; Prabhakaran et al., 2023), and how emotions are described even when there is a common structure (Jackson et al., 2019). Thus, measuring how this text is perceived by those who widely consume and are effected by it is relevant from a social sciences perspective (in order to understand the behavior and values of social groups), and from a computational perspective (estimating how text is perceived by the large number of people consuming it vs. the small number that produce it).

Some work estimating the values of the authors of text has been conducted to measure the values of individuals who have written personal essays and social media posts e.g. (Maheshwari et al., 2017; Ponizovskiy et al., 2020), and in arguments abstracted from various forms of public facing text (Kiesel et al., 2022). However, we have not observed work on how to measure values perceived in text, estimate them along a scale as in prior work (Schwartz, 1992), or ultimately treats them as a hierarchical list in line with theory (Rokeach, 1973).

2.4 Music Lyrics

Music listening is an extremely popular activity. Over 616 million people subscribe to streaming services worldwide[4], and out of the music industry's reported 31.2 billion USD[5] revenue, more than 17 billion comes from music streaming[6]. Out of over 1400 number-1 singles in the UK charts, only 30 were instrumental[7]. Lyrics were shown to be a salient component of music (Demetriou et al., 2018), and thus are likely to be a widely consumed form of text, and of importance to a broad audi-

[3] https://web-assets.domo.com/miyagi/images/product/product-feature-22-data-never-sleeps-10.png

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

[5] https://midiaresearch.com/blog/recorded-music-market-2022-reality-bites

[6] https://cms.globalmusicreport.ifpi.org/uploads/Global_Music_Report_State_of_The_Industry_5650fff4fa.pdf

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

ence. As reported in Appendix A.1, the responses of our annotation participants, who were drawn from representative samples of the US population, quantitatively confirm the prevalence and importance of lyrics to them as music listeners.

Importantly, the annotation of lyrics is a challenging task as we expect substantial variance in annotator responses. As with harmful speech, different social groups may perceive the values in lyrics differently (Solaiman et al., 2023; Prabhakaran et al., 2023). Further, as artistic and expressive language lyrics are ambiguous text: they contain different forms of analogy and wordplay (Sandri et al., 2023b). Thus, the steps taken towards a method for the annotation of values in music lyrics are likely to be applicable to other domains in which perceptions are of interest, or in which the text is subjective, or both.

3 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[8]. 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* (Groves et al., 2009). 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. We expect this problem will be similar in other related tasks (e.g., perceptions of values in other corpora like political speeches and podcasts, or of other phenomena like personality or morality in the same corpora).

Some measurable strata that affect the use of language are clear, in the song lyrics as in other domains (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 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 (Liem et al., 2012). Based upon these considerations, we advocate for a stratified random sampling procedure, based on strata that we acknowledge to be justifiable given our purpose, yet in some cases conceptually 'fuzzy'. In our case these include: (1) release date; (2) popularity, as estimated via artist playlist frequency from the MPD (Chen et al., 2018); (3) genre, estimated from topic modeling on Million Song Dataset artist tags (Schindler et al., 2012); (4) topic, through a bag-of-words representation of the lyrics data.

3.1 Primary Lyric 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)[9] as it is large and recent compared to other similar datasets. The lyrics themselves were obtained through the API of Musixmatch[10], 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.

To draw an initial subpopulation of songs, we first uniformly subsampled 60,000 out of 300,000 artists from the Million Playlist Dataset (MPD) (Chen et al., 2018). We then queried the Musixmatch API to determine if the lyrics for each of the songs of the 60,000 sample of artists was available.

We expect that our dataset will require a bias correction. Specifically, we observe 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 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[11]. Through this method, we

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

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

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

[11] Full code of our sampling procedure is at https://anonymous.

Figure 2: Visualization of the annotation interface on Qualtrics for two of ten annotated values

subsampled 2200 songs with lyrics.

3.2 Inclusion Criteria

As the annotation of highly subjective perceived values in lyrics has not been studied yet, it is unclear whether any valid and reliable annotations can be obtained. As such, together with the ambition to investigate many annotations from a representative population sample, it may be unwise to immediately annotate thousands of songs, but rather focus on rich insights on smaller well-curated data. For this, the following screening procedure was followed: three members of the research team manually screened several hundreds of songs randomly sampled from our 2200 songs. They verified the match of songs to lyrics, the available metadata, and rejected songs that had words that were not English, contained very few words, were only onomatopoetic, or were only repetitions. As a consequence, we finally kept 380 songs: 20 for a pilot study, 360 for our main study.

4 Survey Measures

To obtain the perceptions of human values in song lyrics, we design a survey by adapting an existing psychometric instrument, i.e., a validated procedure for measuring psychological constructs (in this case: a series of questions designed to measure human values). Although it may appear as merely a set of questions, designing a psychometric survey is an elaborate, multi-step process that often involves repeated sampling to demonstrate reliability, and correlation with observable behaviors and other established instruments to demonstrate validity (Furr, 2011).

To gain further measurable evidence on the degree to which song lyrics are important yet subjective to a representative population sample, our survey starts with 16 general questions about song lyric preferences. Furthermore, after participants performed their annotations, we also ask them to rate how subjective they considered the task to be (see Appendix A.1).

4.1 Participants

With many lyrics being in the English language, we choose to obtain our annotations from samples of the US population, representative in terms of self-reported sex, ethnicity and age. Such samples can be obtained through the Prolific[12] platform. 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, and participants gave informed consent before proceeding with the survey. 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.

4.2 Short Schwartz Values Survey

Our primary annotations involve impressions of the values expressed in song lyrics. To this end, we adapted the Short Schwartz Values Survey (SSVS) (Lindeman and Verkasalo, 2005) to determine the wording of the questions, as it is the shortest instrument that has shown adequate reliability. The original wording of the questionnaire displays the name of the value being rated, followed by a number of words to describe it e.g. "POWER (social power, authority, wealth)"[13]. Original instructions can be found in Appendix A.2.

We made three adaptations to this questionnaire. First, we adjust the question text to ask not for ratings of life-guiding principles for the individual

4open.science/r/lyrics-value-estimators-CE33/1_stimulus_sampling/stratified_sampling.py

[12] https://prolific.co

[13] Actual wording of items was retrieved from https://blogs.helsinki.fi/everyday-thinking/files/2015/11/The-Short-Schwartzs-Value-Survey.docx.

responding to the survey, but rather for the respondent's impressions of the 'speaker' of the lyrics. This 'speaker' is the someone or something whose perspective is reflected in the lyrics, and may not be the author or artist expressing the lyrics. For example, the speaker in the song 'I gave you power' by the artist Nas is a gun, and the speaker in 'Rosetta Stoned' by the rock band Tool is a person hallucinating from psychedelics. In other words, the creator may use a persona in the writing of song lyrics for artistic purposes, which may not directly represent their values. As such, an annotator's impression of the creator may differ from their impression of the speaker reflected in the lyrics. As we are interested in the values perceived in the text, we explicitly ask participants to respond with the perspective of the speaker in mind, and not the author. Further illustration of our explicit instructions is given in Appendix A.2.

Secondly, the original SSVS uses a 9-point Likert Scale where 0="Opposed to My Principles", 1="Not Important", and 8="Of Supreme Importance". In order to gather continuous, symmetrical measurements, we aimed for a continuous scale where 0 essentially indicates that the value is either not discussed or otherwise not estimatable from the lyric text. Next to this, we balance the scale to have maximum opposition to a given value be at -100, where maximum importance will be at 100. As such, in contrast to the original SSVS, our scale is symmetric with a rating of 0 indicating neutrality (see Figure 2).

Thirdly, (Cabitza et al., 2020) suggested that a rater's confidence is an indication of intra-rater reliability. Thus, we also asked participants "How confident are you in your ratings of these lyrics?", to which they responded on a scale of 0 (Not at all Confident) to 100 (Completely Confident).

4.3 Annotation Interface

The survey was implemented on an instance of the Qualtrics[14] platform. The annotations were collected using the response format shown as illustrated in Figure 2, following explicit instructions as discussed in Appendix A.2. More specifically, a set of lyrics are displayed, with a clickable interface below them. The interface contains brief descriptions of each of the 10 Schwartz values, followed by a vertical bar on which participants can indicate a continuous response, as described in Section 4.2.

[14] https://www.qualtrics.com/

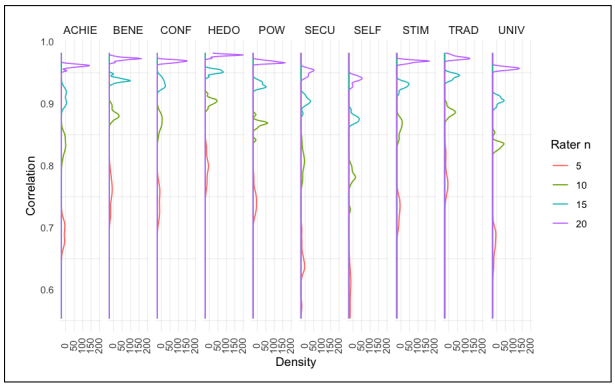

Figure 3: Rotated scaled density plots of Pearson correlations between canonical mean and subsample means, from a mean 27 ratings per each of the 360 songs.

An option to select "Not Applicable" was also available for each value. We considered that "0" and "Not Applicable" responses both indicate that the importance of that given value to the speaker based on the lyrics could not be determined by the participant (i.e., they were either not discussed in the song lyrics, or were otherwise unclear). As we expect that not all songs will discuss all values, and most songs may discuss very few values, we initialize the rating bar at "0".

With this, we now have the setup to gather annotation data. In the remainder of this paper, we discuss how this was done to research three questions: (1) How many annotator ratings are needed for stable annotations to emerge? (2) Do our obtained value perception annotations relate to existing validated knowledge on stable structures among values? and (3) Can our refined annotations be used in computational NLP setups?

5 Determining the Number of Ratings

Our procedure to determine the number of ratings to gather was inspired by (DeBruine and Jones, 2018). Specifically, we first recruited a representative pilot sample (n=505), in which respondents used our interface to annotate perceived values for a fixed set of 20 songs. From these annotations, we computed canonical mean ratings per value, per song. For each of the values, we then estimated Cronbach's α for a range of subsample sizes (5 to 50 participants, in increments of 5), repeating this procedure 10 times per increment. Following this, we visually examined density plots of the distribution of Cronbach's α (Figure 8). In the social sciences, an $\alpha >= 0.7$ is commonly considered an acceptable level of reliability. Taking a con-

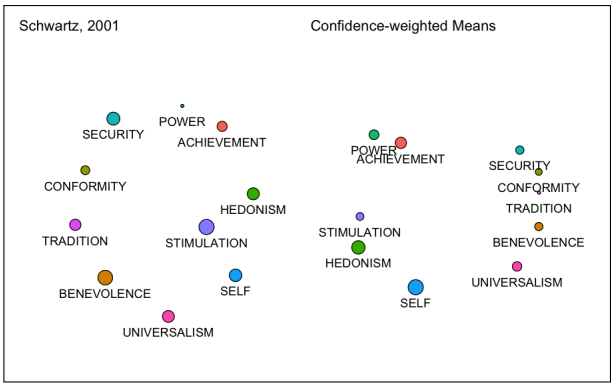

Figure 4: MDS plots derived from the correlation plot reported in (Schwartz et al., 2001), and our participant responses as confidence-weighted means[15].

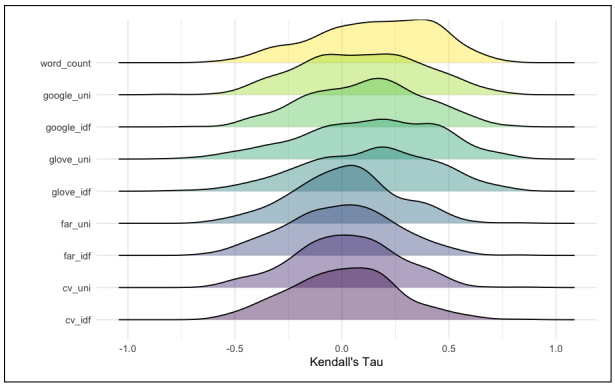

Figure 5: Rank correlations between NLP systems / word counts and confidence-weighted participant means transformed to rankings

servative estimate, we chose to obtain 25 ratings per song lyric in our main study; for that amount of ratings, Cronbach's α in our pilot data would comfortably exceed 0.8.

From this, we perform our main study data collection. We recruit a new representative US population sample (n=600), where each participant goes through our survey questions, and receives 18 randomly selected song lyrics to annotate for perceived values. As a result, we obtained 22-30 annotations per song, with an average of 27.

From these, checking for the reliability of our annotations from this sample, we repeatedly subsampled 5, 10, 15 and 20 ratings for each value within each song, and calculated and visualized intra-class correlations as well as Person correlations between subsample means and canonical means (Figure 3). From this, we see higher Pearson correlations and more stable ICC estimates from higher numbers of ratings. The Pearson correlation to the canonical mean already exceeds 0.9 for all values from 15 subsampled ratings. Further details are given in Appendix A.3.

For further analysis, we must aggregate the subjective labels. Being unaware of a single ideal method to achieve this, we report results using an aggregation method inspired by (Cabitza et al., 2020). Specifically, 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. However, we also provide disaggregated data so that better techniques can be applied as they are researched.

6 Structural comparison

As a first attempt to assess the relative validity of our procedure, we depart from the earlier observations that a cross-cultural stable structure was found based on what values are likely to cluster together. We compare distances as derived from the upper triangle of a correlation matrix reported in (Schwartz et al., 2001) to those derived from proximity in ratings obtained in our study. For both, we generate a multi-dimensional scaling plot (MDS) (Davison and Sireci, 2000) for visual comparison, which has previously been used as method to assess confirmation of earlier theory (Ponizovskiy et al., 2020). From these plots (Figure 4), 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.

7 Comparisons with NLP models

Finally, as first step towards ways in which our data may be connected to computational NLP methods, we perform a preliminary comparison on how computational NLP-based value assessment of lyrics data compares to the way in which our annotators annotated perceived human values.

We again depart from a validated instrument: in this case, a dictionary of words associated with the 10 Schwartz values (Ponizovskiy et al., 2020). 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 (Ponizovskiy et al., 2020), as well as by assessing cosine similarity between dictionary words and lyrics texts using four classes of pre-trained word embeddings: `word2vec-google-new`,

a generic English word embedding trained on Google News dataset (Mikolov et al., 2013); `glove-common-crawl`, another generic English word embedding trained on Common Crawl dataset (Pennington et al., 2014); `faruqui-mxm-[1∼10]`, trained on the collected initial lyrics candidate pool, employing the Glove model (Pennington et al., 2014) (using ten models populated from ten cross-validation folds, whose parameters are tuned based on English word similarity judgement data (Faruqui and Dyer, 2014).); and `cv-mxm-[1∼10]`, ten variants of lyrics based word-embeddings from cross-validation folds selected by Glove loss values on the validation set.

We weigh terms in the lyrics texts in two different ways: uniformly and weighted by Inverse Document Frequency (IDF). Then, we compare value assessments from these computational methods to the ones obtained from our annotators.We take the perspective from theory that that value assessments should be seen as ranked lists, and we consider rank correlations between the machine and human value assessments based on Kendall's τ (Figure 5).

In earlier work (Richard et al., 2003), Pearson correlations of 0.1-0.2 were considered as moderate evidence of the validity of a proposed dictionary in relation to a psychometrically valid instrument. Only the more generic Glove and Google news embeddings seem to reach those levels of correlation (see Appendix A.5). From a rank correlation perspective, the word count methods and these two embedding models hint at slightly positive rank correlations. This may be promising in terms of the degree of specialization needed to assess values; at the same time, neither of the methods presented here have thoroughly been optimized, and as such, these results should not be seen as strong benchmarking evidence. Future work will be needed to more deeply connect computational NLP techniques with our data.

8 Limitations and future work

In this paper, we described our procedure for ground-truthing perceptions of highly subjective text. By paying attention to grounding in social sciences theory and purposeful sampling strategies, the discussion of how to get to 'good data' has been much more extensive than is typical in computational domains. With this, we hope to have illustrated how beyond (welcome) completion of checklists and data sheets, being purposeful about data can pro-actively shape annotation design.

Our procedure may be adapted to a broad range of other annotation tasks. Specifically, our adaptation of the Short Schwartz Value Survey (Lindeman and Verkasalo, 2005) may be used as a tool to design annotation interfaces and surveys for the perceptions of other forms of text (e.g. political speeches, podcasts, tweets etc.). Where other psychological features are of interest, researchers may similarly seek out validated instruments and consider similar adaptations to those we employed. Fuzzy stratified sampling approaches may be employed to determine which texts within a corpus receive annotations. When meaningful disagreement is expected between annotators, the optimal number of annotations to collect may be determined by a procedure similarly inspired by (DeBruine and Jones, 2018). And when that disagreement may be linked to specific populations or social groups, we encourage researchers to thoughtfully sample annotators from within those groups. Those interested in ground-truthing perceived basic human values in lyrics may aim at an average 15 ratings per song.

As for limitations to our work, while we are committed to open science practices, we cannot share the primary lyric data due to copyright prohibitions. However, we do release metadata of the songs of interest, together with our participant annotations, and the code used for the analyses and plots in our paper[16]. 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. We also did not assess whether variations on the annotation instrument might result in substantial differences in the annotations we received (Kern et al., 2023), nor did we repeat our procedure (Inel et al., 2023). In addition, we can further connect our work to related research on examining how participants from different groups will annotate corpora (Homan et al., 2022; Prabhakaran et al., 2023). Finally, while we only provide a preliminary comparison to computational NLP methods, it will be worthwhile to use our data in the context of more sophisticated state-of-the-art NLP systems.

[16] https://anonymous.4open.science/r/values_in_lyrics-8F3F/

References

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Valerio Basile et al. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *CEUR WORKSHOP PROCEEDINGS*, volume 2776, pages 31–40. CEUR-WS.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184.

Ryan L Boyd and James W Pennebaker. 2017. Language-based personality: A new approach to personality in a digital world. *Current opinion in behavioral sciences*, 18:63–68.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Federico Cabitza, Andrea Campagner, and Luca Maria Sconfienza. 2020. 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*, 20(1):1–21.

Ching-Wei Chen, Paul Lamere, Markus Schedl, and Hamed Zamani. 2018. Recsys challenge 2018: Automatic music playlist continuation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 527–528.

Mark L Davison and Stephen G Sireci. 2000. Multidimensional scaling. In *Handbook of applied multivariate statistics and mathematical modeling*, pages 323–352. Elsevier.

Lisa M DeBruine and Benedict C Jones. 2018. Determining the number of raters for reliable mean ratings.

Andrew Demetriou, Andreas Jansson, Aparna Kumar, and Rachel M Bittner. 2018. Vocals in music matter: the relevance of vocals in the minds of listeners. In *ISMIR*, pages 514–520.

Manaal Faruqui and Chris Dyer. 2014. 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*, pages 19–24. The Association for Computer Linguistics.

Mike Furr. 2011. Scale construction and psychometrics for social and personality psychology. *Scale Construction and Psychometrics for Social and Personality Psychology*, pages 1–160.

R. Michael. Furr and Verne R. Bacharach. 2014. *Psychometrics : an introduction*, second edition edition. SAGE Publications.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 325–336.

Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey methodology*, volume 561. John Wiley & Sons.

Christopher Homan, Tharindu Cyril Weerasooriya, Lora Aroyo, and Chris Welty. 2022. Annotator response distributions as a sampling frame. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 56–65.

Han-Yin Huang and Cynthia C. S. Liem. 2022. Social inclusion in curated contexts: Insights from museum practices. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Oana Inel, Tim Draws, and Lora Aroyo. 2023. Collect, measure, repeat: Reliability factors for responsible ai data collection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, pages 51–64.

Joshua Conrad Jackson, Joseph Watts, Teague R Henry, Johann-Mattis List, Robert Forkel, Peter J Mucha, Simon J Greenhill, Russell D Gray, and Kristen A Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Conference on Fairness, Accountability and Transparency (FAT '20) January 27-30 2020, Barcelona, Spain*. ACM, New York, NY, USA.

Christoph Kern, Stephanie Eckman, Jacob Beck, Rob Chew, Bolei Ma, and Frauke Kreuter. 2023. Annotation sensitivity: Training data collection methods affect model performance. *arXiv preprint arXiv:2311.14212*.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471. Association for Computational Linguistics.

Benjamin Kompa, Jasper Snoek, and Andrew L Beam. 2021. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4.

Cynthia C. S. Liem, Andreas Rauber, Thomas Lidy, Richard Lewis, Christopher Raphael, Joshua D. Reiss, Tim Crawford, and Alan Hanjalic. 2012. Music Information Technology and Professional Stakeholder Audiences: Mind the Adoption Gap. In *Dagstuhl Follow-Ups*, volume 3. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik.

Cynthia CS Liem and Andrew M Demetriou. 2023. Treat societally impactful scientific insights as open-source software artifacts. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, pages 150–156. IEEE.

Marjaana Lindeman and Markku Verkasalo. 2005. Measuring values with the short schwartz's value survey. *Journal of personality assessment*, 85(2):170–178.

Tushar Maheshwari, Aishwarya N Reganti, Samiksha Gupta, Anupam Jamatia, Upendra Kumar, Björn Gambäck, and Amitava Das. 2017. 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*, pages 731–741. Association for Computational Linguistics.

Gregory R Maio. 2010. Mental representations of social values. In *Advances in experimental social psychology*, volume 42, pages 1–43. Elsevier.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. 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*, pages 3111–3119.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. 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*, pages 1532–1543. ACL.

Vladimir Ponizovskiy, Murat Ardag, Lusine Grigoryan, Ryan Boyd, Henrik Dobewall, and Peter Holtz. 2020. 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*, 34(5):885–902.

Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Alicia Parrish, Alex Taylor, Mark Díaz, and Ding Wang. 2023. A framework to assess (dis) agreement among diverse rater groups. *arXiv preprint arXiv:2311.05074*.

James Pustejovsky and Amber Stubbs. 2013. *Natural language annotation for machine learning* , first edition edition. O'Reilly Media.

F Dan Richard, Charles F Bond Jr, and Juli J Stokes-Zoota. 2003. One hundred years of social psychology quantitatively described. *Review of general psychology*, 7(4):331–363.

Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. 'just what do you think you're doing, dave?' a checklist for responsible data use in nlp. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics.

Milton Rokeach. 1973. *The nature of human values.* Free press.

Lilach Sagiv and Shalom H Schwartz. 2022. Personal values across cultures. *Annual review of psychology*, 73:517–546.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023a. 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*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Ježek. 2023b. 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*, pages 2420–2433.

Thomas Schäfer and Peter Sedlmeier. 2009. From the functions of music to music preference. *Psychology of Music*, 37(3):279–300.

Alexander Schindler, Rudolf Mayer, and Andreas Rauber. 2012. Facilitating comprehensive benchmarking experiments on the million song dataset. In *ISMIR*, pages 469–474. International Society for Music Information Retrieval.

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.

Shalom H Schwartz, Gila Melech, Arielle Lehmann, Steven Burgess, Mari Harris, and Vicki Owens. 2001. Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of cross-cultural psychology*, 32(5):519–542.

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, et al. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. *arXiv preprint arXiv:2306.05949*.

Chris Welty, Praveen K. Paritosh, and Lora Aroyo. 2019. Metrology for AI: from benchmarks to instruments. *arXiv preprint arXiv:1911.01875*.

A Appendix

A.1 Lyrics affinity and subjectivity perception

Our data collection protocols allowed us to gather self-reports on the importance of lyrics on a sample representative of the US. Our initial pool of questions was inspired by the Preference Intensity scale in (Schäfer and Sedlmeier, 2009), and consisted of Likert-type questions. We turn these into 16 question statements on the participants' relation to music lyrics by also adding our own suggested questions. Participants respond to the questions using a 5-point Likert scale, which included the points "Strongly Disagree", "Somewhat Disagree", "Neither Agree nor Disagree", "Somewhat Agree", and "Strongly Agree". Percentages in the table below indicate the proportion of respondents that indicated either "Somewhat Agree" and "Strongly Agree".

We currently report on responses given to these questions from our two data collection rounds: our pilot study (n=505), whose primary aim was the estimation of the number of ratings needed per song lyric, and our actual annotation collection study (n=600) on which the main outcomes in our paper are reported.

In Figure 6, we visualize self-reported percentages of respondents' music libraries containing lyrics, for both our respondent samples. Here, we

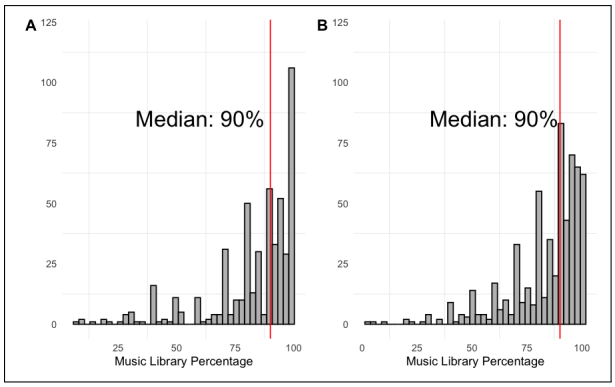

Figure 6: Distribution of self-reported percentage of music library containing lyrics from two representative US samples, n=505 and n=600 respectively.

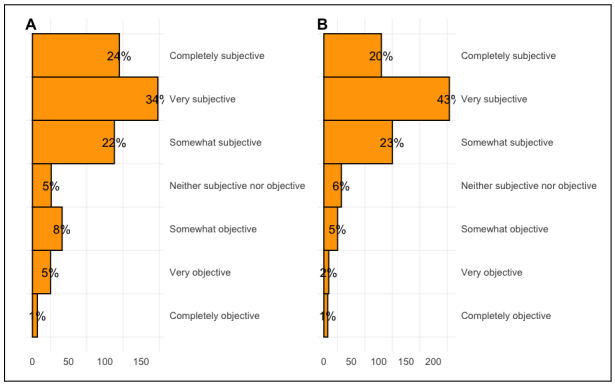

Figure 7: Distribution of self-reported subjectivity of lyric annotation task, n=505 and n=532 respectively.

see that respondents' music overwhelmingly contains lyrics, with a median of 90%. Furthermore, in Table 1, for both samples of respondents, we indicate the percentages of users that indicated to somewhat or strongly agree with given statements. From this, we again observe a strong preference for songs with lyrics (>70% on many of the statements).

Finally, at the end of our survey, we also ask participants to self-report a rating of the subjectivity of the lyrics annotation task we gave to them. Distributions are visualized in Figure 7. From these, we see confirmed the task indeed is perceived as highly subjective in the eyes of our sample population.

As gaining a general understanding of music lyrics affinity is not our current main goal, we chose not to iteratively validate and refine our questions as a formal psychometric instrument at this stage (for this, more explicit iterative analysis would be needed on the instrument being capable of distinguishing between different types of users by making use of the full scale). However, we did start an-

Question	Pilot	Main
I prefer music that contains lyrics, as opposed to music that does not	72%	72%
I always pay attention to the lyrics of a song, if the song has them	70%	72%
If a song has lyrics that I don't like for any reason, I don't listen to it	49%	43%
If I am not sure about the lyrics of a song, I search them on the internet	76%	77%
I memorize the lyrics to the songs I listen to	70%	75%

Table 1: Question wording, and proportion of respondents rounded to the nearest whole number, that indicated either 'somewhat agree' or 'strongly agree' in two surveys, n=505, and n=600 respectively.

alyzing to what extent the current questions may be used as an instrument, or at least as a way to further characterize subpopulations of human respondents. Here, given the large preference towards music that contains lyrics, asking for lyrics vs. non-lyrics music preference will not allow for us to be able to distinguish between respondents. At the same time, responses to the degree to which a respondent pro-actively engages with lyrics (e.g. by actively searching for them, writing about them, or writing lyrics themselves) may yield interpretable factors on which respondents can be distinguished. However, we leave a deeper analysis of this for future work.

A.2 Adjusted Short Schwartz Value Survey

The original Schort Schwartz Value survey appears in (Lindeman and Verkasalo, 2005). The original question wording[17] was:

"Please, rate the importance of the following values as a life-guiding principle for you. Use the 8-point scale in which 0 indicates that the value is opposed to your principles, 1 indicates that the values is not important for you, 4 indicates that the values is important, and 8 indicates that the value is of supreme importance for you."

- POWER (social power, authority, wealth)
- ACHIEVEMENT (success, capability, ambition, influence on people and events)
- HEDONISM (gratification of desires, enjoyment in life, self-indulgence)
- STIMULATION (daring, a varied and challenging life, an exciting life)
- SELF-DIRECTION (creativity, freedom, curiosity, independence, choosing one's own goals)
- UNIVERSALISM (broad-mindedness, beauty of nature and arts, social justice, a world at peace, equality, wisdom, unity with nature, environmental protection)
- BENEVOLENCE (helpfulness, honesty, forgiveness, loyalty, responsibility)
- TRADITION (respect for tradition, humbleness, accepting one's portion in life, devotion, modesty)
- CONFORMITY (obedience, honoring parents and elders, self-discipline, politeness)
- SECURITY (national security, family security, social order, cleanliness, reciprocation of favors)

In our survey, participants were initially shown a set of instructions designed to explain how to use the instrument, and explain our working definitions of 'artist' as separate from the 'speaker' of the lyrics, see(Figure 12). We then presented our adjusted question wording:

"Between the quotation marks below are some song lyrics. Please take a moment to read them and think about the SPEAKER the lyrics. Please remember that this SPEAKER might be a the AUTHOR themselves, or someone or something else:", after which lyrics were displayed, along with the annotation instrument.

A.3 Estimating the Number of Raters

Our initial estimate of approximately 25 raters per song lyric on average was derived in our pilot study. We gathered perceptions of 505 annotators recruited to be representative of the US population, and had them complete the Adjusted Short Schwartz Survey for 20 songs.

From 505 ratings, we subsampled as described in Section 5. We computed the canonical mean and Cronbach's α for all of the ratings. We then

[17] retrieved from https://blogs.helsinki.fi/everyday-thinking/files/2015/11/The-Short-Schwartzs-Value-Survey.docx.

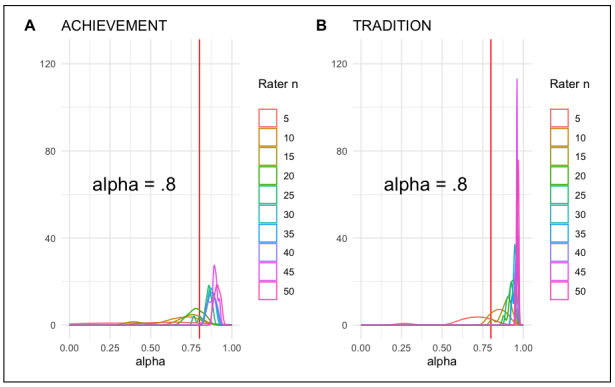

Figure 8: Distribution of Cronbach's α from a representative US Sample (n=505) rating 20 songs, for the values Achievement and Tradition. Vertical line represents the α threshold for comparison.

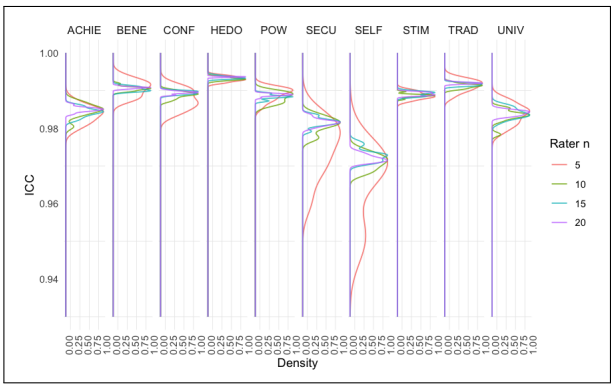

Figure 9: Rotated scaled density plots of ICC for subsamples from annotations on the 360 songs.

computed the mean and alpha for each of the increments. Examples for two values are shown in 8. Distributions of α indicated 25 ratings per song as an initial estimate.

Our main study involved responses from 600 participants, where each song received a median 27 ratings. For a more accurate estimate of the needed ratings per song, we once again computed the mean from all ratings as well as a measure of internal reliability, this time using Intraclass Correlation set to estimate the means of *k* raters, as per the R *psych* package recommendations. We subsampled in increments of 5, ranging from 5 to 25 ten times, and computed the ICC and mean of each increment. As per 9, the ICCs are generally higher, with a narrower spread with higher numbers of ratings. We based our estimate of 15 ratings per song on average from the plot of correlations, 5.

A.4 Distribution of Ratings

Distributions of confidence weighted means by value are shown in Figure 10. Although certainly

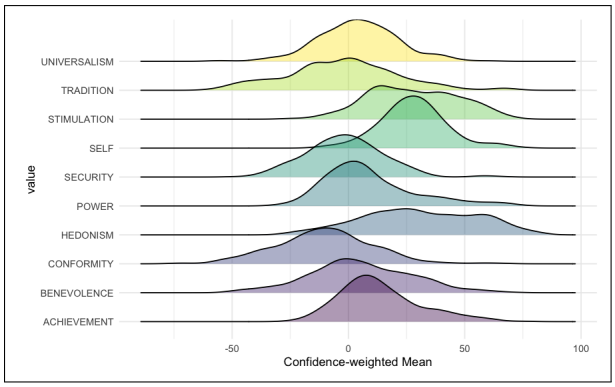

Figure 10: Distributions of confidence-weighted means across songs, by value.

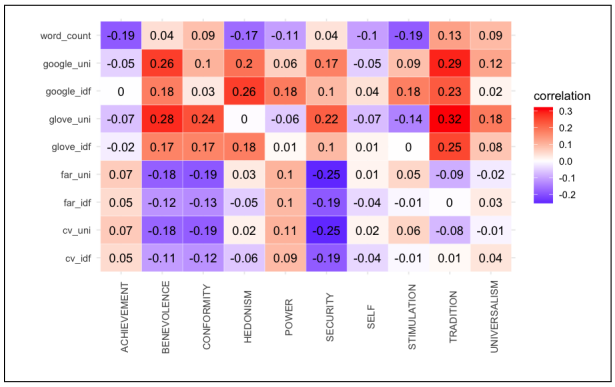

Figure 11: Pearson correlations between NLP systems / word counts, and participant ratings of songs, by value.

an insufficient sample to comment on such a distribution across all songs, some interesting patterns emerge. Universalism, Security, Power, and Achievement have peaks very close to 0, indicated that participants did not perceive them in the lyrics often. However, we see a strong indication that the Speakers of songs were seen as valuing Self-Enhancement, Hedonism and Stimulation, and varied greatly in the degree to which they were perceived as Traditional.

A.5 Correlations between Ratings and NLP systems

Although we chose to interpret rank correlations in line with theory, we report here a more intuitive table of Pearson correlations, shown in Figure 11.

Similarly to the rank correlations, the strongest positive correlations are between participant rankings and pre-trained models. We see weakest correlations between all models and participant ratings on the ACHIEVEMENT dimension, and conversely, relatively strong correlations on the TRADITION and BENEVOLENCE.

Thanks!

You will now be shown parts of song lyrics from 18 songs, and asked to complete some questions about how you perceive them.

IMPORTANT: Lyrics can be written from different perspectives, some of which are not the same as the writer of the lyrics. In other words, the **AUTHOR** of the lyrics may choose a **SPEAKER** for their lyrics that is not themselves.

The SPEAKER of the lyrics could be could be a fictional character, a real person from history or the present, or even an imaginary object. And of course it could be the AUTHOR themselves. Please answer the questions while thinking about the **SPEAKER**.

WARNING: These lyrics are drawn from popular music, some of which use offensive language or describe offensive situations.

Click on the bar (green arrow) to indicate how important you think the value is to the SPEAKER.

Clicking further to the right indicates that you think it is a supremely important guiding principle in their life.
Further to the left indicates that you think it is the opposite of that guiding principle of their life.
Click on 'Not Applicable' (red arrow) or 0 if you don't think the SPEAKER has not expressed anything about that value. They indicate the same thing.

Please click on the arrow below when you're ready to proceed.

Figure 12: Visualization of the instructions page of the annotation interface on Qualtrics.