Values from Lyrics: Pre-Registration

Andrew M. Demetriou

Table of contents

Study Information

Contributors: Andrew M. Demetriou, Jaehun Kim, Sandy Manolios, Cynthia C.S. Liem

Overview

We aim to extend work that used natural language processing (NLP) to estimate psychological values (e.g. (Schwartz et al. 2001)) in social-media text (Ponizovskiy et al. 2020). Our aim is to explore the potential to automate the rating of perceived values in song lyrics.

We will gather ratings of song lyrics from participants using an online survey. Participants will respond to a psychometric instrument that we have adapted for this purpose. We will then use NLP models to estimate scores for each of the personal values. Lastly, we will examine how well the scores from the NLP resemble human ratings.

Contributions

We extend existing work in two ways: Firstly, we compare semantic distance (the degree to which words are related) measured using NLP models to the results of word counting. Prior studies have counted the number of times specific words from a fixed lexicon were used in a given body of text as a means of measuring psychological constructs: (Ponizovskiy et al. 2020) validated such a lexicon of words for measuring a set of 10 personal values in social-media text. However, song lyrics may not contain those exact words, and may instead use synonymous or otherwise meaningfully similar words, or even slang and metaphors. Our method allows for more word coverage: rather than count words from a fixed lexicon, we will estimate the semantic distance between the words in the lexicon that represent each personal value, and the song lyrics in order to derive a score for each value.

Secondly, we linearly combine the output of multiple NLP models into a single latent variable, to represent the shared variance of the machine ratings: as each NLP model is developed using 1) an algorithm trained on 2) a corpus, each algorithm/corpus combination will estimate the semantic distance between two words differently. This loosely parallels how human participants may rate each set of lyrics differently. (Beaty and Johnson 2021) showed that this latent variable of semantic distance estimations resulted in overlap with a latent variable of human ratings as high as r = .9, albeit in a different domain. This approach further allows us to estimate the contribution of each algorithm / corpus setup to the shared variance.

As in (Ponizovskiy et al. 2020), we estimate convergent validity of the grouped NLP models by estimating correlations with related constructs measured using the Linguistic Inquiry Word Count (LIWC) dictionary (see (Boyd et al. 2022) for psychometric details).

Hypotheses

As this is an initial study aimed at examining the feasibility of our method, our hypotheses are not severe:

Primary Hypothesis: Grouped NLP models show a statistically significant correlation with grouped **a)** participant ratings across all 10 personal values, and **b)** with related LIWC constructs - in the same or greater magnitude as shown in (Ponizovskiy et al. 2020).

Null Hypothesis: Grouped NLP models show no evidence of a correlation with participant ratings across all 10 personal values.

Design Plan

Survey Implementation:

Our primary measure is the presence of personal values in song lyrics. Song lyrics may be written from the perspective of the author, but also from the perspective of someone or something else - sometimes referred to as the 'speaker'. As we are measuring the presence of values as suggested in the lyrics themselves, we explicitly ask participants to respond with the perspective of the speaker in mind, and not the author.

The main survey will be implemented a version of the formR.org hosted on the servers of Delft University of Technology to ensure GDPR compliance. The `.csv` survey files used as input to formR were constructed in R. The main component of the survey involves showing participants the lyrics to a number of songs, one at a time. For each song they are asked to respond to set of questions designed to assess the presence of values in the lyrics.

We experienced issues testing our surveys when the number of lyric stimuli in the survey was greater than 60. Thus, our stimulus set will be separated into otherwise identical survey files on the formR server, with no more than 60 lyrics in each survey file. Participants will be randomly assigned to one of the surveys, which will in turn randomly select a subset of stimuli to have rated by participants.

The majority of items require a Likert-type response. In order to gather a more continuous measure, we used a sort of slider with no obvious starting point: the `rating button` option in formR shows a horizontal gray bar with two labeled poles (e.g. agree - disagree). Participants can then be instructed to indicate on the bar the degree to which they agree or disagree, as they might with a slider. However, the gray bar has no visible slider, thus no starting value. The gray bar shows no divisions on it and appears continuous, although it contains 20 subdivisions.

Survey Meausures:

Personal Values:

Prior research (e.g. (Schwartz et al. 2001)) has shown evidence for the presence of personal values as guiding principles in the lives of people. Participants will indicate the degree to which they think 10 values are present for each set of lyrics that they are shown. We chose to use the Short Schwarz's Value Survey (Lindeman and Verkasalo 2005) as it is the briefest instrument whose reliability and validity has been shown to be adequate, to our knowledge. The original instrument displays a brief definition of each of the ten values in the Schwartz inventory, (e.g. "POWER (social power, authority, wealth)") and asks participants to indicate on a Likert scale (0= Opposed to my principles, 8 = Of supreme importance) the degree of importance of the value to them. In our version, participants will indicate on a solid gray bar as described above. As our participants will be rating a stimulus that is not themselves, we

adjusted the wording slightly: e.g. “Please, rate the importance of the following values as a life-guiding principle for the SPEAKER of the lyrics.”

Lyric Preferences:

To assess whether a preference for lyrical content has an effect on the ratings given, we created an ad-hoc scale consisting of 16 items, partially inspired by the Preference Intensity scale in (Schäfer and Sedlmeier 2009). Our original ad-hoc scale consisted of 11 items. Participants in our pilot studies were asked to respond to the 10 items, and to an additional ‘open response’ format item that asked: “Can you think of any other activities or indications that someone has an affinity for song lyrics? If so, please enter them here:”, from which we drew an additional 5 items.

Study Type:

Observational study - Data is collected from study participants that are not randomly assigned to a treatment.

Blinding:

No blinding is involved in this study

Appendix

a) number of ratings per stimulus

We conducted two pilot studies to estimate the number of ratings needed for each song lyric.

Pilot 1

Our first pilot study aimed to gather an tentative estimate of the time it would take participants to complete components of the survey using a small convenience sample. We recruited participants first on reddit.com and then from within the lab of the research team. Participants were shown four lyric stimuli and asked to complete our adapted personal values questionnaire for each song lyric. we used the Qualtrics platform to create and host the survey (qualtrics.com).

Table 1: Time in minutes to complete task.

statistic	lyric preferences	song 1	song 2	song 3	song 4

mean	1.4460848	2.019053	1.626731	1.374369	1.072906
sd	0.7674326	1.572428	1.742973	1.480763	1.053221

Note: Time in minutes determined by subtracting time at the first click in a block of questions from the last click, and dividing by 60.

On average participants took 1.52 minutes per song. However, Table 1 shows that the time to complete the items per song decreased as participants progressed through the questionnaires. Thus we estimated that questions for 20 stimuli could be completed in approximately 30 minutes.

Pilot 2

Following Section , we aimed to gather responses from 500 participants on 20 lyric stimuli to estimate the number of raters necessary to achieve sufficient inter-rater reliability. We recruited a representative sample of the United States on the Prolific recruitment platform (prolific.co) https://www.prolific.co.

b) hypothesized magnitudes

LIWC

liwc insight with self-direction, .43

liwc sexuality with hedonism, .13

liwc achievement with achievement .47

liwc power with power, .19

liwc power with conformity, .16

liwc risk with security, .32

liwc religion with traditionalism, .79

liwc family with benevolence, .57

Beaty, Roger E, and Dan R Johnson. 2021. "Automating Creativity Assessment with SemDis: An Open Platform for Computing Semantic Distance." *Behavior Research Methods* 53 (2): 757–80.

Boyd, Ryan L, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. "The Development and Psychometric Properties of LIWC-22." *Austin, TX: University of Texas at Austin*.

Lindeman, Marjaana, and Markku Verkasalo. 2005. "Measuring Values with the Short Schwartz's Value Survey." *Journal of Personality Assessment* 85 (2): 170–78.

Ponizovskiy, Vladimir, 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.

Schäfer, Thomas, and Peter Sedlmeier. 2009. "From the Functions of Music to Music Preference." *Psychology of Music* 37 (3): 279–300.

Schwartz, Shalom H, 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): 519–42.