Human and Machine: Analyzing Language Trends in Descriptions of Academic Philosophy

Sherri Lynn Conklin[1], Alex Dayer[2], Michael Nekrasov[3], Carolyn Dicey Jennings[4]

[1] Washington State University [2] University of California, Merced/University of Arkansas at Little Rock [3] Independent [4] University of California, Merced

Abstract

Advances in machine learning hold promise for corpus analysis: they have the potential to allow for more efficient and less biased analyses of text. This would be a boon for qualitative research, such as the survey research conducted by Academic Philosophy Data and Analysis. In this paper we examine the utility of automated machine learning for select survey questions, with a focus on LDA (statistical topic modeling) and VADER (rules-based sentiment analysis). We thus compare human and machine coding on the question of whether underrepresented philosophers are more likely to respond negatively to questions concerning diversity and inclusivity in academic philosophy. Our study has mixed results, revealing the potential to utilize automated machine learning for high-level classification and sentiment while emphasizing the need for more traditional hand-coding techniques to delve into nuance.

Human and Machine: Analyzing Language Trends in Descriptions of Academic Philosophy

1 Introduction

> "In my opinion and experience, I find that graduate students are often more enthusiastic and open towards ideas, whilst academics at some point often get caught within institutional and administrational webs that prevent them from fully immersing themselves in their research. Students are often an antidote to such a climate, and know more than they are given credit for, so I think they should definitely be taken more seriously, and be more involved in departmental activity." (an anonymous response from the 2021 APDA survey)

As observed above, graduate students can offer invaluable insights into the future and health of academic philosophy, which is why the Academic Philosophy Data and Analysis (APDA) project has been collecting surveys from PhD students and recent graduates since 2016. In response to these surveys, thousands of past and current philosophy PhD students have described their graduate program, its climate, and its support for nonacademic employment, among other topics. In 2018 and 2021 the survey included questions concerning issues of diversity and inclusivity in academic philosophy in order to obtain insights on these topics from the perspective of current and recently-minted philosophy PhDs.[1] These surveys found that underrepresentation occurs in a surprising spread of categories, presenting a challenge for those who aim to address issues of diversity and inclusion across these different groups.[2] Yet, they also yielded data that may illuminate both particular and shared areas of concern.

---

[1] While such questions were also on the 2023 survey, this chapter was written before that survey took place.

[2] "Women, people of color, first-generation college students, and veterans are underrepresented relative to the United States population and doctoral graduates, but so are those with conservative political leanings" (Jennings & Dayer, 2022).

However, accessing these insights, using either quantitative or qualitative measures, can prove difficult (Gelo, Braakmann, & Benetka, 2008). While both methodologies afford advantages, quantitative research is fairly straightforward, lending itself to reproducibility and the appearance of objectivity, and is therefore often favored by researchers. Yet, quantitative measures commonly used in surveys, such as the Likert scale, risk reducing insights to a limited range of numeric scores tallied on a set of statements pre-identified by the researcher.[3] Survey measures of this sort both lack important details pertaining to participant insights and risk missing them altogether when the researcher frames the topic differently than the participant. This occurs because quantitative measures tend to adopt the framing of the researcher.[4]

Qualitative research, on the other hand, allows researchers to detect such differences—qualitative methodologies are more sensitive to the perspective of the participant in that they allow participants to use their own words and topic framing. Yet, qualitative measures can be difficult to analyze and assess. Where analyses of quantitative measures can be largely automated using "off the shelf" (OTS) software, qualitative measures often require more time, effort, and funding to analyze, precisely because the greater richness and detail included in the responses typically necessitates hand coding, which is open to bias on the part of the coder.[5]

The development of new machine learning tools can help to close the gap between qualitative and quantitative research. One recent paper, for example, uses automated tools

---

[3] For example, "cultural differences emerging from Likert scale data do not always concur with differences predicted by cultural experts" (Ogden & Lo, 2012, 351) and Likert scale responses can diverge substantially from text-based ones: "Such inconsistencies between different forms of data may reflect measurement error and the psychometric limitations of Likert scales" (Ogden & Lo, 2012, 358).

[4] As one paper puts it, "while quantitative approaches are usually deductive and theory-driven (i.e. they observe specific phenomena on the base of specific theories of reference), qualitative ones are inductive and data-driven (i.e. they start from the observation of phenomena in order to build up theories about those phenomena)" (Gelo et al., 2008, 272).

[5] For qualitative research, "an attempt is usually made to understand a small number of participants' own frames of reference or worldviews, rather than trying to test hypotheses on a large sample" (Gelo et al., 2008, 268).

to analyze text-based responses to a survey on related topics: the study proposes "a novel automatic analysis framework that can be used to automatically mine the text responses to open-ended questions in student feedback questionnaires" (Nawaz et al., 2022, 2). This is, essentially, a quantitative analysis of a qualitative measure, potentially allowing the researcher to perform more efficient analyses.

The current chapter applies mixed methods corpus linguistics to some free-response questions from APDA's 2018 and 2021 surveys (Jennings & Dayer, 2022; Jennings et al., 2019). For our purposes, "mixed methods" refers to projects implementing both quantitative and qualitative analytical approaches (Creswell, 1999). Using these two different approaches, we identify the main topics discussed in the responses as well as the general sentiment (positive, negative, or neutral). The first approach is the more typical "qualitative" method of hand-coding the responses using human judgment to identify topics and sentiment.[6] The second approach is an automated "quantitative" approach, which uses Latent Dirichlet Allocation (LDA) Topic Modeling (Blei, Ng, & Jordan, 2003) to identify response topics and Valence Aware Dictionary and sEntiment Reasoner (VADER) (Hutto & Gilbert, 2014) to evaluate response sentiment. We then compare the results of the two approaches to identify the best way to apply machine learning approaches to textual analysis of survey responses in the context of academic philosophy.

We decided to incorporate response data from four ADPA survey questions concerning issues of diversity and inclusivity in philosophy. Based on previous work (Jennings et al., 2019), of particular interest are differences between demographic groups and whether underrepresented philosophers tend to use more negative language in assessing their program and philosophy as a whole. This follows the finding that "underrepresented graduates are less likely to recommend their program to others than those who are not underrepresented" (Jennings et al., 2019). Further, those with one or more

---

[6] We are very grateful to Alisha Nesslage, who worked as an undergraduate research assistant to hand code all of the survey responses with the relevant topics, their associated valence, and overall sentiment.

underrepresented factors[7] both rate themselves as less comfortable in philosophy and rate philosophy as less welcoming to those who are underrepresented than those with no underrepresented factors,[8] a difference that is statistically significant (Jennings et al., 2019). Text-based responses following these questions leaned negative, according to a coder of the 2018 survey (Jennings et al., 2019). Focusing on these questions in the course of comparing analytical methods will allow us to obtain more detailed, yet quantifiable, insights into the perspective of our research participants on issues of diversity in philosophy and will hopefully demonstrate the value of these methods for experimental philosophers working on other projects.[9]

1.1 Overview

The analyses conducted in this paper are largely exploratory. In broad strokes, we are investigating differences between automated, quantitative tools and hand coded, qualitative tools for tagging topics and sentiments from select free-response questions of ADPA's 2018 and 2021 surveys. Across all comparisons, we generally expect to find that manual methods for coding these data generate results that better correspond to participant self-reports than automated methods (as measured through comparisons to Likert data). We expect this because we are using untrained, OTS machine learning approaches on a small set of corpus data, which means that the automated approaches, especially LDA topic modeling, will most likely struggle to identify the sort of nuances that might be easily accessible to a human reader. A recent paper contrasting automated machine learning with

---

[7] This includes those who identify as women, non-binary, transgender, LGBQA, Asian/Pacific Islander, Black/African American, Chicanx/Latinx/Hispanic, two or more races, first generation, military/veteran, or one or more disabilities.

[8] This includes straight, White men who are neither transgender nor a first generation college student and have neither military status nor known disability.

[9] While there are many categories of underrepresentation in philosophy, as mentioned above, we decided to focus on gender and race/ethnicity in this study. One reason for this focus has to do with the available tools: we do not have demographic information for the full data set and NamSor can only categorize names in terms of gender and race/ethnicity.

both supervised machine learning and hand coding concluded that automated machine learning has only limited utility when used on its own to analyze texts with complex concepts (Nelson, Burk, Knudsen, & McCall, 2021). Further, while it is possible to successfully use automated machine learning techniques on small datasets, even in the case of linguistic data, it is widely considered to present significant challenges (Hudon et al., 2021). Because hand coding is resource intensive, we nonetheless hope that automated methods could offer some comparable or otherwise meaningful insights into the data. Our findings are somewhat mixed.

Our study comprises two parts. The first part involves a descriptive comparison between the topics and sentiments tagged using qualitative and quantitative methods to understand the extent to which the two methods converge on similar results. The second part investigates how the results of these methods impact real world scientific inquiry, in other words, whether the outputs of automated techniques lead to the same conclusions as traditional hand coding techniques. We apply our methodology to examine the difference in topic and sentiment between the responses of "underrepresented" and all other participants, as defined in Section 2.2.

In part one, we examine the extent to which the topics identified by hand coding correlate with those identified through the LDA model. We hypothesize that there will be relatively little correlation between the topics identified through the two methods. Next, we compare both qualitative (hand coding) and quantitative (VADER) sentiment identification techniques to self-reported Likert values, which were included in 3 of the survey questions. We hypothesize that there will be an overall correlation between the sentiment values assigned through the two methods to each response but that the hand-coded sentiment will more closely approximate the self-reported sentiment. Our findings are consistent with both hypotheses, but the automated technologies did surprise us.

In part two, we investigate whether each of the 4 diversity-related questions from the ADPA survey elicit different sentiment responses for underrepresented and all other

participants.[10] As Haslanger put it in 2008, "In my experience it is very hard to find a place in philosophy that isn't actively hostile towards women and minorities" (Haslanger, 2008). As a result, we hypothesize that the responses of underrepresented participants will correlate more with negative sentiment, per question, than those of all other participants. Our findings on this hypothesis are mixed and shed light on differences between the two coding techniques, which we discuss in 4.

2 Methods

In this section, we describe the methods used in preparing, coding, and analyzing the data through the two approaches described above.

2.1 ADPA Surveys

APDA has examined employment and other trends in philosophy for over 10 years, using both quantitative and qualitative methods. The methods used for collecting ADPA data, including the survey data, have been discussed at length in numerous other publications and reports, many of which are publicly accessible online through the website (Jennings, 2015; Jennings et al., 2016; Jennings, Cobb, Pablo, & Kyrilov, 2017; Jennings & Dayer, 2022; Jennings et al., 2019, 2015).[11] For detailed descriptions of how the original data were collected and previously analyzed, please see these publications. Here, we characterize the data analyzed in this study by presenting a very general overview of the ADPA survey, its methods, and questions used from the 2018 and 2021 surveys.

APDA's database contains nearly 20,000 current PhD students and recent graduates, with information about both graduation and employment for these graduates. Early surveys were sent to everyone in the database with known contact information (61% of the

[10] In this case "underrepresented" includes women and people of color, but not other categories of underrepresentation, as mentioned in a footnote above.

[11] https://philosophydata.org/about

database), but more recent surveys have been restricted to current students and graduates of the past 10 years.[12] APDA has so far run 5 separate surveys: in 2016, 2017, 2018, 2021, and 2023.[13] Around 1,000 people take part in any given survey and in 2021 it was found that 25% of those contacted had taken part in at least one survey.[14] While there is substantial overlap in the survey questions, each survey also includes unique questions intended to target a specific focus of the researchers running that survey. These foci have included, for example, nonacademic employment (Jennings et al., 2017) and diversity and inclusivity (Jennings et al., 2019).

For this paper we are utilizing a corpus of 4 questions from the 2018 and 2021 surveys that focus on diversity and inclusion: asking participants about the climate in their graduate program (question 4), their level of comfort in philosophy (6), how welcoming philosophy is to those from underrepresented groups (7), and how philosophy might become more inclusive (8). A list of questions along with number of responses per question is presented in Appendix A. Note that three of the questions are broken into two parts, a five point Likert-scaled numerical response and a free response. Our corpus comprises responses from 992 participants. Each participant answered between one and four of the free-response questions, totaling 2,062 responses. From these responses, we started with 5,931 total words in our corpus.

For the text analysis we use the free-response portion of these questions. We also compare the Likert-scaled responses to the results of our sentiment analysis. For this, we convert the [1,5] Likert response into a [-1,1] negative to positive sentiment range by subtracting 3 and dividing by two. This was chosen to match the scoring of the automated sentiment methods.

---

[12] APDA has international coverage, but has been able to determine its level of completeness only for the U.S.: it includes around 90% of the philosophy PhD graduates counted by the National Science Foundation's Survey of Earned Doctorates over the past 10 years.

[13] The 2025 survey is currently underway.

[14] Our new survey platform, which adds a further layer of anonymity for participants, makes it impossible to determine this number going forward.

2.2 Demographic Information

As a use case for our methods we identified underrepresented demographics as an area of focus. We used self-reported demographic information from the APDA surveys to determine if a survey respondent fell into one of two underrepresented groups: woman or person of color. In the absence of available survey information we used NamSor, a name-based tool for gender and race/ethnicity determination (Namsor, 2021). We found that 35% of participants identified as women, whereas 21% identified as (at least one of) American Indian or Alaskan Native, Asian or Asian American, Black or African American, Chicanx/Latinx/Hispanic, Other (e.g. MENA), or Pacific Islander. In contrast, 32% of these participants were coded as women, and 30% were coded as either Black, Asian, or Hispanic by NamSor (the three options other than White, not Latino). When self-report information is available, the two methods agree in 95% of cases with respect to gender, and 77% of cases with respect to race/ethnicity.

For this analysis we group responses into originating from either an "underrepresented" or "other" participant. In this context, we define "underrepresented" to mean that if the participant selected either "man" or "woman" in the gender question or one of the available options in the race/ethnicity question(s) then we look to see if the participant identified as either a woman or a person of color (defined as anyone but White, non-Hispanic). If the participant identified as either of these they count as underrepresented. For all other participants (i.e. those who did not answer the questions as defined above), we use NamSor to see if the name is categorized as belonging to a woman or person of color. In either case the person is additionally counted as underrepresented. Anyone that did not fall into one of these categories (woman by survey or NamSor, person of color by survey or NamSor) is counted as "other."

Given the difference between self-report and NamSor mentioned above, this methodology is likely to include more participants as underrepresented than if we had used self-report alone.

2.3 Hand-Coded Responses

Hand coding occurred in parallel to machine learning. That is, the authors split into two groups, with each pair tackling either hand coding or machine learning. They did not share material or findings during this period. The research assistant (RA) worked exclusively on the hand coding side while the authors worked in parallel. We used this setup to better allow for comparison across the two methods.

The first step in hand coding responses is to determine the topics. For this step two authors read all of the available survey responses for two of the questions each and created a list of topics they encountered in those questions.[15] These authors then met to consolidate their topics, with the goal of having a single list for all four questions. The authors decided to group these topics under headings to ease the coding process. An explicit goal of the list was to cover as many topics raised in the responses as possible with minimal overlap between them, while also keeping to a manageable total number for the coder. This process led to the creation of 18 topics, divided into four sections: time, space, resources; individuals; interactions; action/reaction (see Appendix E). These two authors then created a coding document for the RA with each topic, positive and negative examples of the topic, and whether the topic was likely to come up for that question. For example, the "resilient" topic was associated with the positive example of "culture is healthy and can bounce back" versus the negative example of "culture is toxic and hard to improve," and was noted as primarily associated with the questions on climate and improving inclusivity.[16]

The second step was to train the RA over multiple sessions. The RA was recruited on the basis of her high level of knowledge about philosophy, both in terms of content and as

---

[15] The two authors described here are Alex Dayer and Carolyn Jennings. This division of labor was driven partly by IRB restrictions that prevented the other two authors, Sherri Conklin and Michael Nekrasov, from accessing non-public survey responses, and partly by the greater skill of those authors with respect to machine learning.

[16] We had hoped to look at variations in use of topics by underrepresented and all other participants, as well as differences in sentiment on topics, but were forced to remove that section due to space considerations. If we had included that section, we would have noted that underrepresented participants are more likely to use the topic of resilience than other participants, but in a negative way.

an academic discipline.[17] In this step the two authors met with the RA to go through sample responses. All three would individually code the sample responses and then discuss their choices. This process was continued until all three were satisfied. Following this training, the RA carefully read all of the survey responses for the four questions identified. For each response, the RA hand coded one of three variables: 1 (positive valence), 0 (neutral), or -1 (negative valence) for the overall sentiment, while also using these variables to tag the relevant topics contained in each response. In this context, positive valence meant that the topic was brought up in a positive manner (e.g. for the topic of accessibility/inclusion, the response might say that all groups are included), whereas negative valence meant that the topic was brought up in a negative manner (e.g. exclusion of low SES, disability, etc.). Finally, neutral meant that the topic was brought up without a clearly associated valence.

2.4 LDA Topic Modeling

As mentioned above, the other two authors worked in parallel to determine topics using machine learning. Our first automated analysis employs LDA Topic Modeling to generate the main topics discussed in the selected free-response questions of the ADPA surveys. LDA is an automated method for analyzing text, which groups words based on the statistical likelihood that they will co-occur in a corpus. These groups of words are called a "topic." When LDA is applied to the individual documents within a corpus, which, in our case, are the individual survey responses, each document is weighted based on the likelihood of its engagement with each topic discussed in the corpus (Blei et al., 2003).

2.4.1 Pre-Processing. To prepare the responses for analysis, we first tokenized the responses into individual words, all lower case, with symbols and punctuation stripped.

Second, we excluded stop words from the corpus. Stop words are words, such as "the" or "and" in English, that appear so frequently in a language that they dilute the results. In

[17] As mentioned above, the RA described in this section is Alisha Nesslage, who is now a graduate student at UC Irvine.

(a) Word cloud of raw responses

(b) Word cloud after stemming and lemmatization, removal of stop words and mapping of terms

Figure 1. Word cloud showing normalized frequency of terms corresponding to size. Larger terms in graph are more frequent.

addition to common stop words, we removed words such as “philosopher,” “academic,” and “department,” since the ADPA free response questions analyzed in this study dealt with diversity-related issues in the respondent’s academic philosophy department or philosophy at large, and these concepts were contextually assumed and appeared in the results for nearly every topic. We identified 237 unique stop words, listed in Appendix C.

Third, we formed bi-grams and tri-grams of words that co-occur at least 5 times. For this we employed the Gensim Phrases Model (Gensim, 2022) with a threshold of 10. This allows us to capture word pairs and triplets such as “working class” in addition to individual words. The threshold value determines the required association between words to be considered a pair or triplet; we settled on a value of 10 from experimentation on our sample data, as this struck a balance of capturing common word groupings while minimizing spurious pairing.

Fourth, we lemmatized the words in the analyzable data, using spaCy (spaCy, 2022). Lemmatization is the process of reducing inflected words, such as “marginalized” or “humiliates,” or the grammatical variants of words to the base word (i.e., “marginalize” or

"humiliate"). Lemmatization ensures that the model better captures the relative significance of words in the responses because the process of condensing the grammatical variations of words into a single term can increase the frequency of that word appearing in the corpus, and, correspondingly, its relative weight in the statistical model.

Fifth, we applied term mapping to create common terms for different words with the same or similar meanings. Term mapping fixed common spelling discrepancies. For example, word spellings sometimes differ between American and British English (e.g., color and colour); we standardized terms to American English. Similarly, we applied mapping between common acronyms and shortenings, such as "TA" and "grad," to the expanded counterparts (e.g., "teaching assistant" and "graduate"). Given the small size of our corpus, mapping was particularly important for terms referring to minority groups that have infrequent individual occurrence but, on aggregate, amount to a significant frequency. For example, terms like "LGB," "LGBT," "LGBTQU," and "LGBTQIA" were mapped to a common "LGBTQ" term. While this limits the specificity of the discussion, it prevents marginalization of the topic in automated analysis due to diversity of the language that philosophy PhDs use to discuss these critical topics. A list of mappings is provided in Appendix D.

Finally, we performed a second pass of stop word removal, to remove words that may have been re-introduced due to lemmatization and mapping steps. Figure 1 shows word clouds of the term frequencies in the raw data verses the resulting term frequencies after stop word removal, mapping, and lemmatization. The total process cut an initial list of 5,931 unique words to 4,314 words (including the introduction of new word pairs via n-gram creation).

2.4.2 LDA Topics Selection. To this set of words, we trained an LDA Topic Model using Gensim (Gensim, 2022). The output of the LDA is, put simply, bags-of-words. To create a good model, we need to know how many bags (i.e., topics) are needed for accurately describing the corpus and how many words should be included in each bag. To

choose the number of topics, we iterated over different number of topics and evaluated the model's perplexity and coherence scores using the C_v Measure (Röder, Both, & Hinneburg, 2015), shown in Figure 2. The perplexity score is a measurement of the predictive capacity of a model (i.e., how well the model performs when new data are introduced). A low perplexity score means that the model performs well when new data are introduced. The perplexity score increases with each new topic, but we wanted to have as many topics as possible without losing human-interpretability or coherence. The model's coherence score is a measurement of the semantic similarity between the topic's words, which helps to distinguish between statistical artefacts and human-interpretable results. A high coherence score means that a human is likely able to interpret the results. The best available model has a good balance of perplexity score relative to the coherence score. For our model, we chose 19 topics at a coherence score of 0.378 and a perplexity score of -8.162. This was a local maximum of coherence with a comparatively high perplexity.

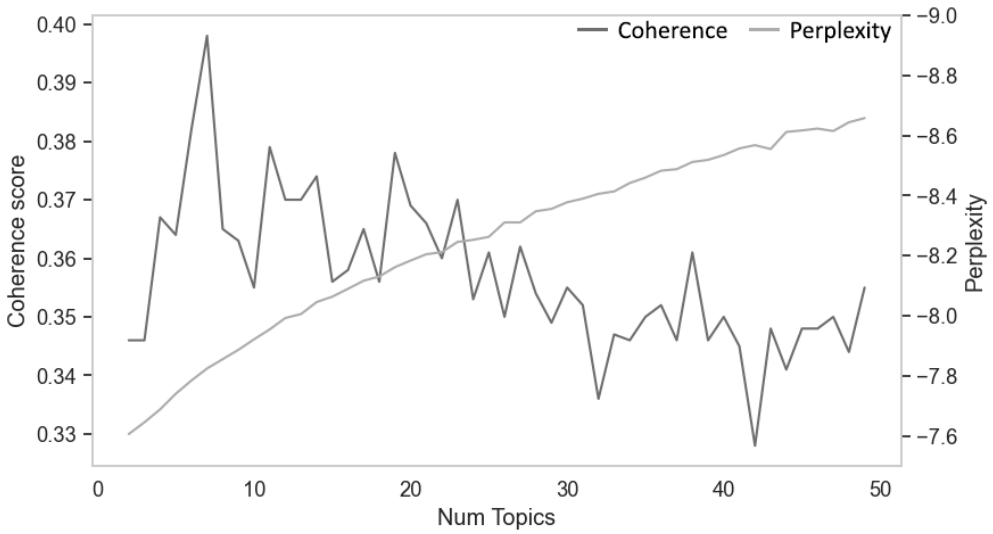

Figure 2. Topic Coherence (dark gray, left y-axis) is shown alongside Topic Perplexity (light gray, inverted right y-axis) for varying numbers of topics (x-axis)

We built a 19-topic LDA model from the cleaned corpus. The resulting topics, numbered 0 through 18, are provided in Appendix E. We then applied the model to each of the 2,062 responses in the data set, recording the prevalence of each topic in each response

(note that any given response can have terms from multiple topics represented simultaneously). We then applied a cutoff of 0.33 (chosen through experimentation) to determine if a particular response fits with a given topic, leading responses to contain at most 2 topics with a median of 1 topic. 14 responses were not tagged with any topics.

2.5 Sentiment Analysis

Our second automated approach employs an open-source sentiment analysis tool called VADER (Hutto & Gilbert, 2014; NLTK-VADER, 2022). We selected VADER because it is designed to analyze short-form documents, such as those found on social media, which are similar in length to the free responses in the ADPA surveys. VADER maps the words used in a document (in this use case a document consists of one response to a survey question) to a library of words with pre-calculated emotional intensities. VADER analyzes the words in relation to other words in the document to calculate the sentiment score of the entire document based on how words modulate sentiment in an embedded context. For example, the statement "I am angry" might generate a high negative sentiment score whereas the statement "I am a little angry" might reduce the intensity of the negative sentiment score because the word "little" is a common diminutive modifier.

For this analysis we use the raw corpus with no stop word removal, mapping or lemmatization. Unlike hand coding, VADER only gives an overall sentiment score for the entire response, not differentiated by a particular topic.

2.6 Statistical Analysis

For this paper we rely on the Chi^2 test for independence and the Spearman correlation test. We briefly remind the reader on the interpretation of these tests.

The Chi^2 test is used on categorical variables and hypothesizes variable independence. If the resulting p-values are less than 0.05 we reject that hypothesis and

deduce that there is an interaction between the two variables.[18] As this test expects a categorical variable, when comparing results of VADER sentiment, we first convert the score into a categorical variable according to standard convention: negative [-1,-0.05], neutral (-0.05,0.05), and positive values [0.05,1].[19]

Spearman correlation test is used on ordinal variables and examines the relationship between these variables ranging from -1 to 1. The magnitude of the Spearman correlation coefficient indicates the strength of the association between the variables. The sign of the coefficient indicates the direction of their relationship, with negative signs indicating an inverse relationship.

3 Results

In this section we present a comparison of the qualitative and quantitative methods we used for the evaluation of the survey response data followed by a short analysis of the data based on these methods. For a high-level summary of the breakdown of topics to questions, please reference Appendix B, which shows the percentage of responses that fit a given hand-coded topic followed by the percentage of responses that fit a given LDA topic.

3.1 Comparing Hand Coded Topics to LDA Model

First, we examine the extent to which the 19 topics identified by the LDA Model correlate with the 18 topics identified through the qualitative coding. We hypothesized above that there would be relatively little correlation between the topics identified through the two methods. In order to test this, we run a Spearman correlation and Chi^2 tests for the co-occurrence of hand-coded topics against the LDA topics in Figure 3.

The findings are consistent with our hypothesis. Overall, we observed little correlation between the majority of the topics. Notably, we do observe that LDA topic 0

[18] As this paper is largely exploratory we do not use Bonferroni correction when we report significance. One can assume that around 5% of statistically significant results are due to chance alone. Given the large

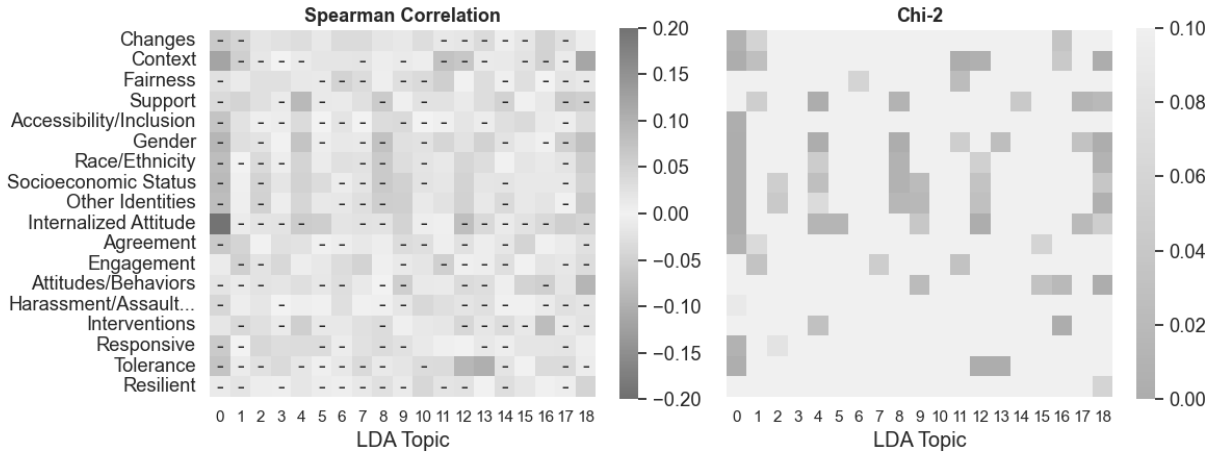

Figure 3. Correlation between LDA and hand coded *topics*. The left-hand chart shows Spearman correlation coefficients, and the right-hand chart shows the result of a Chi^2 test.

(defined by terms such as: *comfortable, depend, conference, interact, professional, work, awkward, program, interaction, good*) positively correlates to the Internalized Attitude topic (with examples "confident, relaxed" vs. "anxiety, stereotype threat, imposter syndrome") and weakly to Context, while having negative correlation with other topics. It also shows (uncorrected) statistically significant interactions with hand-coded topics from the Chi^2 test. This suggests that LDA 0 may be related to Internalized Attitude (and less strongly Context) but not the other topics.

3.1.1 Comparing Hand Coded Groups to LDA. As noted in Section 2.3 and Appendix B, each of the 18 qualitatively identified topics was also assigned to one of 4 higher-order groups. Since we observed little relationship between LDA and hand-coded topics, we run Spearman correlation and Chi^2 tests for the co-occurrence of the LDA topics against the 4 condensed hand-coded groups in Figure 4.

---

number of "tests" reported in the charts, many of the resulting cases of statistical significance will be due to chance.

[19] Parentheses are used here to signify a range excluding the endpoints, whereas brackets signify including the endpoints.

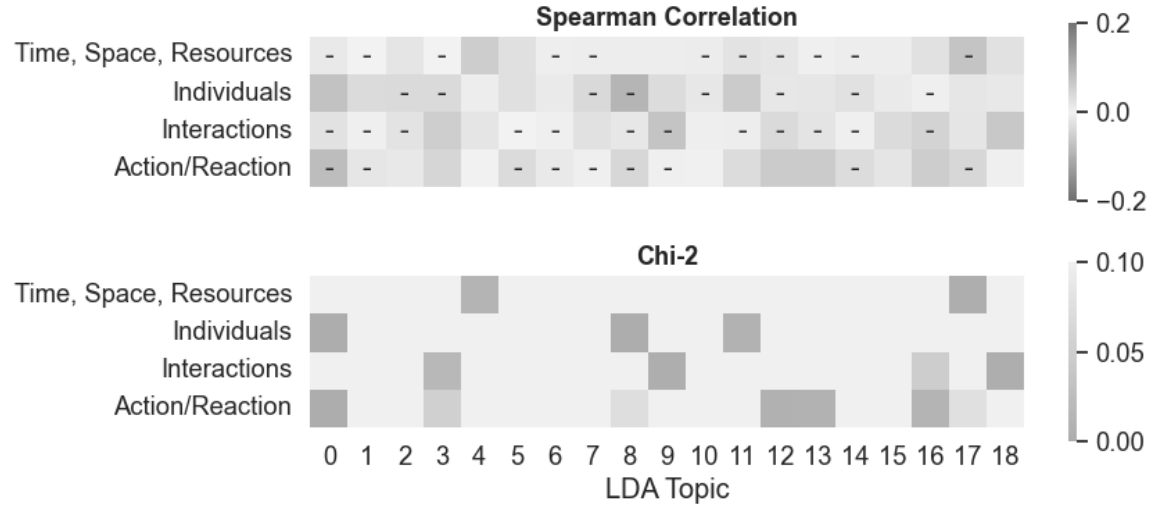

Figure 4. Correlation between LDA and hand-coded *groups*. Top row shows Spearman correlation coefficients, and the second row shows the result of a Chi^2 test.

While Chi^2 shows some interactions between topics and groups, the Spearman correlation coefficients are all quite low (less than 0.1). This indicates a low overall alignment between topics and any particular group. If we accept that the correlation is giving us a loose indication of group affiliation, we observe that topics 4 and 17 may relate to Time, Space, Resources, 0 and 11 may relate to Individuals, 12, 13 and 16 may relate to Action/Reaction, and 18 to may relate to Interactions. The observed correlations are, however, too small to draw any definite conclusions.

3.2 Assessing the interoperability of LDA topics vs hand coding

In this study we decided to have minimal interaction between the hand coding and machine coding techniques to provide a crisp distinction between them. Others have found supervised machine learning to perform similarly to hand-coding techniques, while nonetheless arguing that machine learning should not be used on its own, but only as a complement to hand coding: “the evidence is mixed as to whether they can fully replace traditional approaches” (Nelson et al., 2021, 226). This is in part, as they note, due to the complexity of concepts that are typically of interest to researchers (inequality in their

study), which are difficult to capture with pure machine learning techniques. We had hoped that after the first step of developing topic lists in parallel, the two methods would reveal gaps and oversights in each other. In reality, we found the topics identified by machine learning to be difficult to interpret, even when we restricted them to a much smaller number. There were some points of insight—Topic 0, for example, seems related to experience, with associated words like "comfortable" and "awkward," whereas Topic 6 alone mentions both "white" and "man"—but not enough for us to meaningfully compare and contrast the two approaches.

3.2.1 Comparing Sentiment Techniques. Second, we compare sentiment identification techniques. Only 3 of the survey questions examined in this study included self-reported sentiment data obtained from a Likert scale.[20] While there is a slight difference in intent between the Likert-scaled value, which is a direct evaluation by the survey participant, and post-facto coding, which is a third-party evaluation of the free-response follow-up, we propose to use the Likert-scale results as the baseline against which we can compare the hand-coded sentiment and VADER machine-coded sentiment. We hypothesized above that there would be a correlation between the sentiment scores assigned through the two methods but that the hand-coded sentiment would more closely approximate the self-reported sentiment. To test this, we run a Spearman correlation between the Likert score, the hand-coded sentiment score, and the VADER sentiment score for each of the 3 questions and overall in Figure 5.

The initial findings are consistent with our hypothesis. Chi^2 shows (uncorrected) statistically significant interaction between the three methods and we observed a strong Spearman correlation (0.63 overall) between participant self-reported Likert score and hand coding. The machine-based coding evidenced a weaker but still significant correlation to the other techniques (0.32 and r=0.33 overall).

---

[20] No self reported score was asked for question 8.

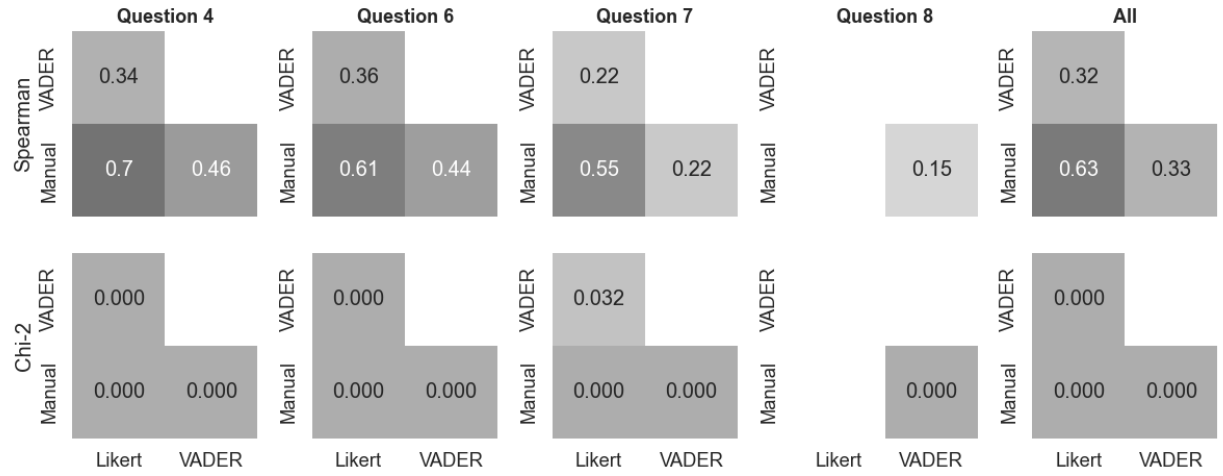

Figure 5. Correlation between sentiment techniques by question. Top row shows Spearman correlation coefficients, with darker shades of gray for stronger correlations, and the second row shows the result of a Chi^2 test, with darker shades of gray for smaller p-values.

In Figure 6, we present the overall score distributions for each method. As is expected, the Likert distribution is heavily quantized on the five whole values with a mean of overall positive sentiment. The hand-coded sentiment score is restricted to the extreme (negative, neutral, positive) with a mean of negative overall sentiment. Unlike the other methods, VADER provides a more nuanced measure of sentiment, with a continuous range of scores whose mean (and median) are of an overall positive sentiment.

While Likert and hand-coded sentiment scores exhibit a stronger correlation, VADER scores more closely track the overall positive Likert sentiment values than the hand-coded scores when examining the sway of absolute values and the means of those values. Hand coding seems to detect a stronger negative sentiment than is self-reported or that VADER reports. This nuance somewhat contradicts our hypothesis, and we do not currently have a full explanation for this difference. Yet, we have at least two potential explanations, which reveal the advantages and disadvantages of each approach. The breakdown by question in section 3.4 indicates that the difference in overall sentiment may be due to the greater sensitivity of the RA to complex concepts and the negativity conveyed by them. Alternatively, given our findings on the interaction between representation and sentiment

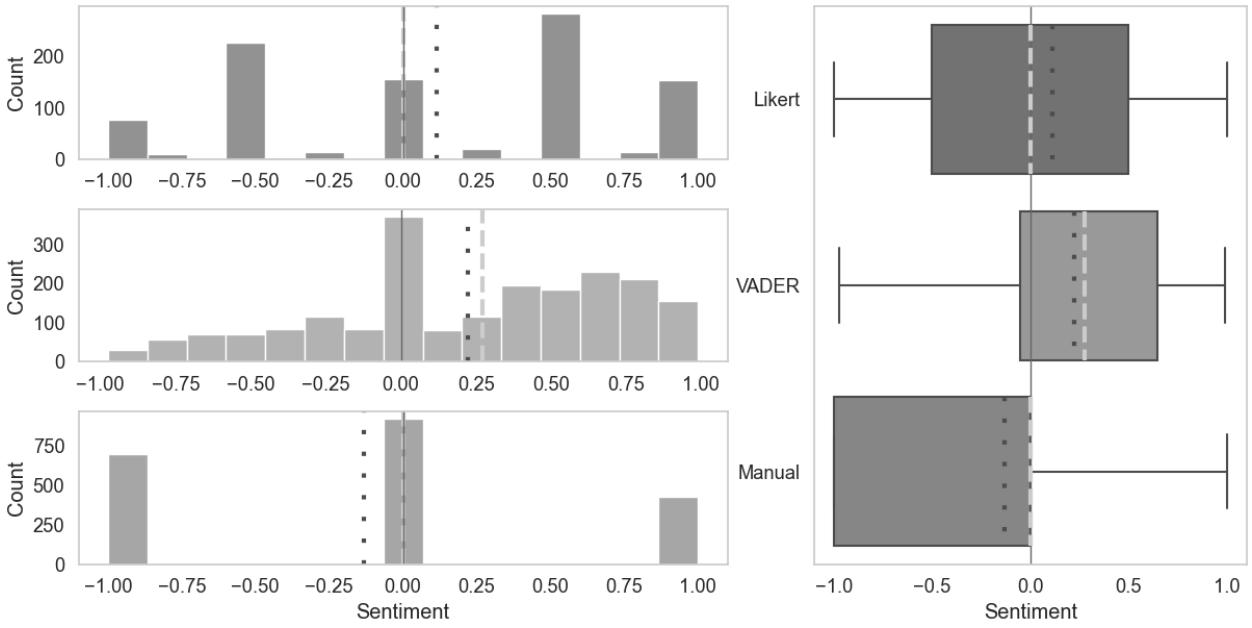

Figure 6. Distribution of values for normalized self reported Likert, scaled assessment, and VADER sentiment analysis. Counts are provided on the left, comparative boxplots on the right. Medians are provided with a dashed light gray line, while means are provided as a sparsely dotted black line.

and the fact that the RA is herself from an underrepresented group, it may be that hand coding of sentiment has a negative bias.

3.3 Examining Representation

To examine how these methods behaved in practice, we apply our methodology to examine the difference in topic and sentiment between the underrepresented participants and all other participants, as defined in Section 2.2. Based on this determination, we find a roughly even split in our corpus with 46% underrepresented participants and 47% underrepresented responses. As noted above in Section 2.2, this likely overestimates the proportion of underrepresented participants.

3.4 Representation By Question

First, we examine whether each of the 4 questions elicit different sentiment responses between underrepresented and all other participants. Given the substantial body of

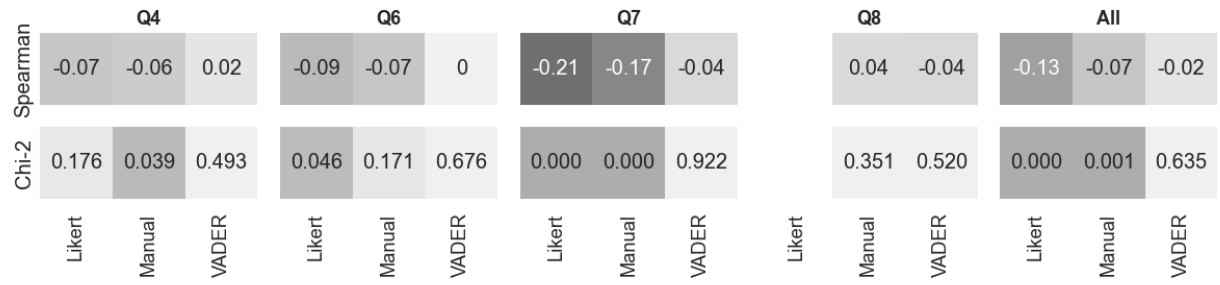

Figure 7. Correlation between sentiment and under-representation by question. Top row shows Spearman correlation coefficients, and the second row shows the result of a Chi^2 test, with darker grays corresponding to stronger correlations (top row) and smaller p-values (bottom row).

evidence that philosophy can be hostile towards members of underrepresented groups, at both the level of individual departments and in the discipline more generally, we hypothesized above that underrepresented participants will have more negative sentiment, per question, when compared to all other participants.[21] We further hypothesized that hand-coded sentiment would most closely correlate with the Likert-scale values self-reported by underrepresented participants.

To test the first hypothesis, we run a Spearman correlation using underrepresented as a binary categorical value and compare it to the sentiment values of each of the three techniques. The results are presented in Figure 7. For this analysis, a positive value of the Spearman correlation coefficient indicates that the underrepresented group has more positive sentiment relative to the other group. A negative value indicates an inverse correlation, meaning that the underrepresented group has a more negative sentiment relative to the other group. Values close to zero indicate that the two groups do not have have a clear delineation of sentiment.

We observed that Question 7 (How welcoming do you find academic philosophy to be toward underrepresented groups) has a weak negative correlation for both Likert and hand-coded sentiment. This means that underrepresented participants are likely more

[21] See, for example, (Hassoun, Conklin, Nekrasov, & West, 2022), (Conklin, Artamonova, & Hassoun, 2019), (Wilhelm, Conklin, & Hassoun, 2018)

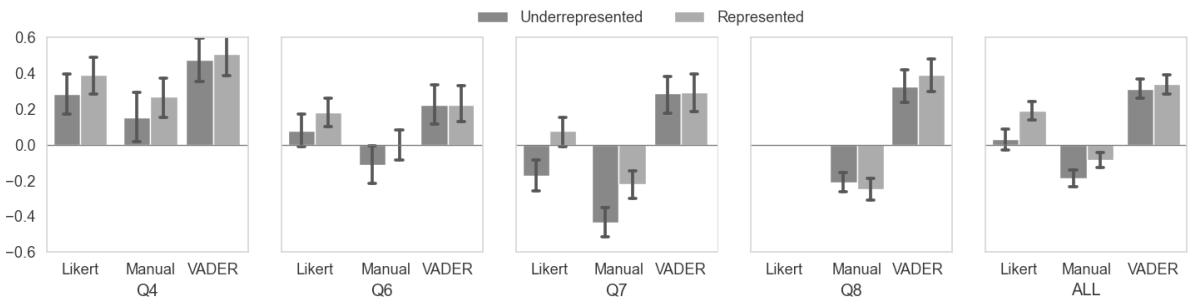

Figure 8. Comparing mean sentiment by question, with underrepresented participants in dark gray and all other participants in light gray.

negative than other participants in response to this question (in keeping with Jennings et al. (2019)).

To test the second hypothesis, we run a Chi^2 test between the two groups of participants and the three sentiment methods. For Question 7, participant group has a statistically significant ($p < 0.01$) interaction with the Likert and hand-coded sentiment. In none of the questions was there any statistically significant interactions between VADER and participant group.

In Figure 8, we present the mean values of sentiment across questions. Prior to this point, hand coding seemed to most closely track Likert scoring, while VADER performed worse on every measure. However, the mean sentiment scores indicate considerable disagreement between the methods, which muddies this picture. Where all three methods agreed on Question 4, hand-coded sentiment for Question 6 evidenced a slight negative sentiment (statistically significant for underrepresented participants), while the other two methods evidenced a slight positive sentiment. For Question 7, where we observed a statistically significant separation for underrepresented participants, we also observe a strong negative sentiment value on hand coding, a small negative sentiment value by Likert scoring, and a small positive sentiment value from VADER. Further, where we have no comparative baseline with self-reported Likert scores on Question 8, we see a statistically significant disagreement between VADER and hand coding.

We think these differences may be driven by the complexity of underrepresentation in academic philosophy, such that negativity can be conveyed in complex ways. For instance, the statements "you would have to be a man to think the discipline is ok" and "there are minimal provisions for child-rearing adults at conferences and it would be great if this could change" are both clearly negative from the standpoint of someone in the discipline, but not clearly negative from a machine learning standpoint. Complex concepts are unlikely to be picked up by automated machine learning techniques like those used in this paper. This view is supported, for instance, by the fact that responses associated with the hand coded topic of "harassment/assault/abuse of power" were marked as having the most negative mean sentiment by the hand coder, but only a slightly negative sentiment by VADER (see B). Yet, the comparison with VADER raises the possibility of bias in the different methods: nearly every topic is associated with negative mean sentiment according to hand coding (except "engagement" and "interventions"), but positive mean sentiment according to VADER (except "harassment/assault/abuse of power"). It is unclear whether this means that the hand coder has a negativity bias, VADER has a positivity bias, or the two methods are just tracking different aspects of the responses.

4 Discussion

We compared qualitative (i.e., hand-coded) and quantitative (i.e., LDA and VADER) approaches for identifying topic and sentiment in a small corpus, comprising four free-response questions from the 2018 and 2021 APDA surveys (Jennings & Dayer, 2022; Jennings et al., 2019). We further observed the behavior of these methods while investigating differences between underrepresented participants and all other participants for questions relating to diversity and inclusivity. For the most part, we found little correlation between the topics identified by hand-coding and LDA topic modeling. We found greater correlation between the sentiments identified by hand-coding and VADER

and found that both methods correlated with self-reported Likert values. When comparing how these methods behave in practice, we observed conflicting outcomes. Where we had no Likert data, the two different methods generated opposing results. Even so, hand-coded sentiments correlated with Likert values on the one question where we found a statistically significant difference between underrepresented participants and all other participants (Q7). In light of our findings, here are our takeaways.

It seems likely that both approaches are tracking topics actually present in the responses but that they track different kinds of topics. For instance, in machine learning topic 1a, where the top ten words are: philosopher, philosophy, people, group, department, faculty, student, think, make, and work (in no particular order), we do observe something akin to a human interpretable topic contained in this list—"what people, interpreted broadly, do in philosophy departments" (or, more likely, "what is done to people in philosophy departments"). Call this the "meta-topic." There should be no wonder that these words would deluge the LDA model and that topics would group around these words because the meta-topic is, at a general level, what the questions and corresponding responses are indeed about.

This highlights a key difference in the methodological approaches, as well as their limitations. Humans can be biased about what we are looking for when interpreting a text. The authors chose to analyze questions about diversity and inclusion in philosophy departments, and we were likely specifically identifying topics relating to this. It is possible to have excluded or ignored some topics because they were either not obvious to us or irrelevant to our research agenda. After all, we excluded the meta-topic both in the LDA model and in hand-coding because we already know about the general context. This is not necessarily a bad thing, since we often conduct research in order to answer highly specific questions and cull extraneous information in order to conduct focused analyses. An OTS LDA model does not know to do this without training. Yet, we hold that an LDA model

may help qualitative researchers recognize their methodological biases, perhaps alerting them to questions the researchers are unaware are being answered.

The findings of the sentiment analysis offer a little more insight into the utility of using hand coding versus automated methods, such as VADER. Both methods analyze arrangements of and relations between words in a document to determine sentiment. Perhaps VADER is unable to capture, at the level of hand coding, the nuanced sentiment of each individual response. Yet, VADER can capture the sway of sentiment in a body of text, which can be useful when we want to compare groups of responses corresponding to variables of interest, such as demographics. VADER might thus be a useful tool for making general observations about sentiment when Likert, or other standard measures, are not included. For example, it seems fairly common to include non-directed questions at the end of a survey (e.g., asking for additional comments), for which Likert measures would be inappropriate and impossible to interpret. Using VADER could save resources in comparison to hand coding of these responses.

In addition to the main methodologies, it is worth noting the clear utility of NamSor, which we used to automatically categorize participants by gender and race/ethnicity. This was especially effective for gender, which had a high level of agreement with self-identified gender. Race and ethnicity had somewhat lower agreement, but could be used to get rough estimates for a study lacking self-identifying information on these categories.[22]

Finally, we want to address a shortcoming of the paper: the limited data set prompted us to condense complex and highly nuanced diversity-related terms into a flat category. On the one hand, we were able to glean a bit more about diversity-related issues in philosophy by condensing such terms. On the other hand, such practices are criticized for erasing the unique characteristics of different marginalized groups. Because we could

---

[22] While using NamSor may raise ethical concerns that should be considered by researchers (thanks to a reviewer for pointing this out), it is worth mentioning that they offer a paid service in which NamSor signs a non-disclosure agreement and considers the names confidential information that is destroyed following the analysis; we used this service.

not conduct the sort of analysis we were after without engaging in such practices, we caution that some automated research tools may simply be inappropriate for conducting research on marginalization in philosophy because they risk promoting further marginalization when data are so limited. This does not mean that we should stop trying to use these tools for this sort of research, but it does highlight the onus on philosophy researchers to collect more nuanced data in these areas and to use such tools ethically.

5 Conclusion

This paper explores two different approaches to corpus analysis: human and machine. What we find fits publications from other disciplines: automated machine learning techniques can be used to complement hand coding, but have limited utility on their own. The most significant insight gained from using automated machine learning techniques in this paper was related to sentiment analysis: hand coding indicated mean negative sentiment, whereas Likert-scaled questions and machine learning indicated mean positive sentiment. This difference has multiple explanations that require further study. One such explanation reveals the advantage of hand coding: it is able to pick up on more nuance. Another explanation reveals the advantage of machine coding: it is less likely to be biased in terms of sentiment. As it is, we continue to see machine learning as a good approach to corpus analysis, but with significant drawbacks for small data sets like our own.

Appendix A

Survey Question Reference

Question	Text	Responses
4A	Rate your satisfaction with this program's efforts to foster a healthy, respectful academic culture or climate. *[very unsatisfied, somewhat unsatisfied, neutral, somewhat satisfied, very satisfied]*	
4B	Please elaborate on your previous answer.	350
6A	When you interact with other philosophers in professional and social settings, how comfortable do you find yourself? *[very uncomfortable, somewhat uncomfortable, neither comfortable nor uncomfortable, somewhat comfortable, very comfortable]*	
6B	Please elaborate on your previous answer.	543
7A	How welcoming do you find academic philosophy to be toward students who are members of underrepresented groups, e.g., women, racial or ethnic minorities, members of the LGBTQ community, people with low socio-economic status, veterans and members of the military, and people with disabilities? *[very unwelcoming, somewhat unwelcoming, neither welcoming nor unwelcoming, somewhat welcoming, very welcoming]*	
7B	Please elaborate on your previous answer.	552
8	What steps should philosophy take to become more inclusive, if any?	617
		2062

Appendix B

Group	Topic	% Occurrence 4B	6B	7B	8	Mean Sentiment Likert	VADER	Manual
Time, Space, Resources	Changes	9	4	9	4	0.1	0.3	-0.1
	Context	6	11	5	0	0.1	0.2	-0.1
	Fairness	3	0	2	3	0.1	0.3	-0.4
	Support	9	4	3	12	0.3	0.3	-0.1
	Accessibility/Inclusion	4	3	8	8	0.1	0.2	-0.3
Individuals	Gender	8	5	13	12	-0.1	0.2	-0.3
	Race/Ethnicity	5	3	12	11	-0.1	0.2	-0.3
	Socioeconomic Status	3	3	10	10	-0.1	0.2	-0.3
	Other Identities	4	5	11	11	-0.1	0.2	-0.3
	Internalized Attitude	1	28	3	1	0.1	0.2	-0.1
Interactions	Agreement	5	2	1	1	0.2	0.3	-0.2
	Engagement	8	8	0	1	0.4	0.4	0.2
	Attitudes/Behaviors	18	18	14	6	0.1	0.3	-0.1
	Harassment/Assault/ Abuse of Power	3	1	1	2	-0.3	-0.1	-0.6
Action/ Reaction	Interventions	3	0	1	1	0.2	0.4	0.0
	Responsive	5	1	2	5	0.1	0.3	-0.2
	Tolerance	4	5	3	12	0.0	0.3	-0.3
	Resilient	2	0	1	0	-0.4	0.1	-0.5
LDA	0	8	19	4	4	0.2	0.2	0.1
	1	4	3	4	5	0.1	0.2	-0.2
	2	4	4	3	5	-0.0	0.2	-0.3
	3	5	5	6	7	0.1	0.3	-0.1
	4	7	2	8	7	0.2	0.2	-0.2
	5	2	6	5	4	-0.0	0.1	-0.4
	6	3	6	5	2	-0.1	0.2	-0.1
	7	11	4	3	5	0.3	0.2	0.0
	8	4	5	5	5	0.2	0.2	0.1
	9	6	8	5	7	0.1	0.2	-0.1
	10	3	3	4	2	0.1	0.3	-0.0
	11	7	6	4	8	0.2	0.1	-0.2
	12	3	3	4	8	0.2	0.3	-0.1
	13	5	5	6	9	0.2	0.2	-0.2
	14	2	4	4	2	-0.1	0.1	-0.3
	15	8	4	4	4	0.2	0.4	-0.1
	16	7	3	4	7	0.3	0.3	-0.0
	17	3	6	5	3	0.1	0.2	-0.2
	18	8	6	17	5	0.0	0.2	-0.2

Appendix C

Stop Words

A.1 Standard Stop Words from NLTK set

a, about, above, after, again, against, all, am, an, and, any, are, as, at, be, because, been, before, being, below, between, both, but, by, can, did, do, does, doing, don, don't, down, during, each, few, for, from, further, had, has, have, having, he, her, here, hers, herself, him, himself, his, how,i, if, in, into, is, it, it's, its, itself, just, me, more, most, my, myself, no, nor, not, now, of, off, on, once, only, or, other, our, ours, ourselves, out, over, own, s, same, she, she's, should, should've, so, some, such, t, than, that, that'll, the, their, theirs, them, themselves, then, there, these, they, this, those, through, to, too, under, until, up, very, was, we, were, what, when, where, which, while, who, whom, why, will, with, you, you'd, you'll, you're, you've, your, yours, yourself, yourselves

A.2 Additional Stop Words

ain, also, aren, aren't, could, couldn, couldn't, d, didn, didn't, doesn, doesn't, hadn, hadn't, hasn, hasn't, haven, haven't, isn, isn't, ll, m, ma, mightn, mightn't, mustn, mustn't, needn, needn't, o, re, shan, shan't, shouldn, shouldn't, though, ve, wasn, wasn't, weren, weren't, whose, won, won't, would, wouldn, wouldn't, wouldnt, y, youre

A.3 Content Specific Stop Words

academic, although, around, become, department, departmental, departments, etc, etc, even, faculty, false, feel, find, general, generally, go, group, however, like, lot, make, many, may, might, much, obvious, often, people, phil, philosopher, philosophic, philosophical, philosophising, philosophy, put, rather, seem, seemed, since, still, student, thing, things, think, true, way, well, within

Appendix D

Term Mapping

Diversity		Academic	
From	To	From	To
lgbtqu	lgbtq	gen	generation
lgb	lgbtq	philosophically	philosophical
lgbt	lgbtq	philsophical	philosophical
lgbtqia	lgbtq	philosophic	philosophical
gay	lgbtq	profs	faculty
genderqueer	lgbtq	prof	faculty
gender_queer	lgbtq	professor	faculty
genderfluid	lgbtq	ta	teaching_assistant
gender_fluid	lgbtq	ma	master
queer	lgbtq	ba	bachelor
lesbian	lgbtq	post_doc	postdoc
bisexual	lgbtq	phds	phd
trans	lgbtq	grad	graduate
transgender	lgbtq		
non_binary	lgbtq		
black	bipoc		
brown	bipoc		
indigenous	bipoc		
non_white	bipoc		
nonwhite	bipoc		
students_color	bipoc		
poc	bipoc		
female	woman		
male	man		
white_male	white_man		
white_males	white_man		
cisgender	cis		
cisgendere	cis		

Additionally map British spellings to American counterparts ex: colour -> color

Appendix E

LDA Topics

Topic	Terms
0	comfortable, depend, conference, interact, professional, work, awkward, program, interaction, good
1	woman, member, person, include, support, welcome, teach, especially, program, funding
2	need, woman, conference, less, field, say, program, graduate, man, take
3	need, field, experience, help, try, university, profession, self, work, support
4	program, experience, woman, issue, good, graduate, member, face, members group, effort
5	conference, discipline, sometimes, effort, say, welcome, profession, minority, especially, long
6	field, work, man, culture, know, graduate, experience, white, social, person
7	year, issue, supportive, climate, see, talk, work, aware, problem, professional
8	inclusive, discipline, answer, question, quite, change, work, respectful, open, welcome
9	comfortable, get, diverse, work, course, woman, know, change, minority, experience
10	profession, welcome, need, tradition, experience, little, include, know, effort, work
11	woman, need, culture, work, good, tenure, school, man, program, problem
12	discipline, work, diversity, need, idea, area, canon, inclusive, teach, see
13	woman, know, work, teach, program, interest, hire, conference, member, term
14	study, welcome, first, work, example, bipoc, take, area, change, crowd
15	time, experience, work, program, different, support, issue, inclusive, woman, place
16	diversity, include, program, perspective, member, respect, course, work, good, support
17	person, science, witness, time, sometimes, institution, discipline, social, member, study
18	welcome, woman, discipline, experience, member, especially, really, issue, unwelcome, conference

Resulting first 10 words (in sequence of significance) for the 19 identified topics.

A References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Conklin, S. L., Artamonova, I., & Hassoun, N. (2019). The state of the discipline: New data on women faculty in philosophy. *Ergo: An Open Access Journal of Philosophy*, *6*.

Creswell, J. W. (1999). Mixed-method research: Introduction and application. In *Handbook of educational policy* (pp. 455–472). Elsevier.

Gelo, O., Braakmann, D., & Benetka, G. (2008). Quantitative and qualitative research: Beyond the debate. *Integrative psychological and behavioral science*, *42*(3), 266–290.

Gensim. (2022). *Gensim: Topic modeling for humans.* `https://radimrehurek.com/gensim`. (Accessed on June 01 2022)

Haslanger, S. (2008). Changing the ideology and culture of philosophy: Not by reason (alone). *Hypatia*, *23*(2), 210–223.

Hassoun, N., Conklin, S. L., Nekrasov, M., & West, J. (2022). The past 110 years: Historical data on the underrepresentation of women in philosophy journals. *Ethics*.

Hudon, A., Beaudoin, M., Phraxayavong, K., Dellazizzo, L., Potvin, S., Dumais, A., et al. (2021). Use of automated thematic annotations for small data sets in a psychotherapeutic context: Systematic review of machine learning algorithms. *JMIR mental health*, *8*(10), e22651.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international aaai conference on web and social media* (Vol. 8, pp. 216–225).

Jennings, C. D. (2015). An empirical look at gender and research specialization. In *Invited colloquium presentation for the metaphilosophy & diversity colloquium at boston university: March.*

Jennings, C. D., Cobb, P., Kerster, B., Gordon, C., Kyrilov, A., Montes, E., ... Vlasits, J. (2016). *Academic placement data and analysis: 2016 final report.*

Jennings, C. D., Cobb, P., Pablo, C., & Kyrilov, A. (2017). *Academic placement data and analysis: 2017 final report* (Tech. Rep.). Academic Placement Data and Analysis Project.

Jennings, C. D., & Dayer, A. (2022). Academic placement data and analysis (apda) 2021 survey of philosophy ph. d. students and recent graduates: Demographic data, program ratings, academic job placement, and nonacademic careers. *Metaphilosophy*, *53*(1), 100–133.

Jennings, C. D., Fronda, R., Hunter, M., Johnson King, Z., Spivey, A., & Wilson, S. (2019). *The diversity and inclusivity survey: Final report* (Tech. Rep.). Academic Placement Data and Analysis Project.

Jennings, C. D., Kyrilov, A., Cobb, P., Vlasits, J., Vinson, D. W., Montes, E., & Franco, C. (2015). *Academic placement data and analysis: 2015 final report* (Tech. Rep.). Academic Placement Data and Analysis Project.

Namsor. (2021). *Namsor: Name ethnicity and gender classifier api.* `https://https://www.namsor.com/`. (Accessed on Jan 05 2022)

Nawaz, R., Sun, Q., Shardlow, M., Kontonatsios, G., Aljohani, N. R., Visvizi, A., & Hassan, S.-U. (2022). Leveraging ai and machine learning for national student survey: Actionable insights from textual feedback to enhance quality of teaching and learning in uk's higher education. *Applied Sciences*, *12*(1), 514.

Nelson, L. K., Burk, D., Knudsen, M., & McCall, L. (2021). The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, *50*(1), 202–237.

NLTK-VADER. (2022). *Natural language toolkit: vader.* `https://www.nltk.org/_modules/nltk/sentiment/vader.html`. (Accessed on June 01 2022)

Ogden, J., & Lo, J. (2012). How meaningful are data from likert scales? an evaluation of how ratings are made and the role of the response shift in the socially disadvantaged. *Journal of health psychology*, *17*(3), 350–361.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth acm international conference on web search and data mining* (pp. 399–408).

spaCy. (2022). *spacy: Industry-strength natural language processing.* `https://spacy.io/`. (Accessed on June 01 2022)

Wilhelm, I., Conklin, S. L., & Hassoun, N. (2018). New data on the representation of women in philosophy journals: 2004–2015. *Philosophical Studies*, *175*(6), 1441–1464.