Mining a Written Values Affirmation Intervention to Identify the Unique Linguistic Features of Stigmatized Groups

Travis Riddle [*], Sowmya Sree Bhagavatula[1], Weiwei Guo[1], Smaranda Muresan[1], Geoff Cohen[2], Jonathan E. Cook[3], and Valerie Purdie-Vaughns[1]

[1]Columbia University
[2]Stanford University
[3]Pennsylvania State University

ABSTRACT

Social identity threat refers to the process through which an individual underperforms in some domain due to their concern with confirming a negative stereotype held about their group. Psychological research has identified this as one contributor to the underperformance and underrepresentation of women, Blacks, and Latinos in STEM fields. Over the last decade, a brief writing intervention known as a values affirmation, has been demonstrated to reduce these performance deficits. Presenting a novel dataset of affirmation essays, we address two questions. First, what linguistic features discriminate gender and race? Second, can topic models highlight distinguishing patterns of interest between these groups? Our data suggest that participants who have different identities tend to write about some values (e.g., social groups) in fundamentally different ways. These results hold promise for future investigations addressing the linguistic mechanism responsible for the effectiveness of values affirmation interventions.

Keywords

Interventions, Natural Language Processing, Achievement Gap

1. INTRODUCTION

In the American education system, achievement gaps between Black and White students and between male and female students persist despite recent narrowing. This is true in STEM fields in particular, with the underachievement leading in turn to problems with underemployment and underrepresentation more generally. Women, for example, make up a scant 28% of the STEM workforce [1].

While we acknowledge that the reasons for underachievement and underrepresentation are numerous and complex, *social identity threat* has consistently been shown to be one factor which contributes to these problems and features a psychological basis [32]. Social identity threat refers to the phenomenon in which an individual experiences stress due to concerns about confirming a negative stereotype held about his or her social group. For instance, Black students are stereotyped to be less capable in academic settings than White students. Therefore, a Black student who is aware of this stereotype may feel psychologically threatened, leading to changes in affect, physiology, and behavior[17, 35, 27, 5].

The description of a psychological process that partly accounts for these achievement gaps opens the door to possible psychological interventions. Indeed, a brief, relatively simple intervention derived from self-affirmation theory known as a *values affirmation* has been shown to diminish these achievement gaps - especially when delivered at key transitional moments, such as the beginning of an academic year [6, 4]. The values-affirmation intervention instructs students to choose from a series of values, and then reflect on why this value might be important to them. The intervention draws on self-affirmation theory, which predicts that a fundamental motivation for people is to maintain self-integrity, defined as being a good and capable individual who behaves in accordance with a set of moral values [31].

Accumulating evidence indicates that this intervention is effective in reducing the achievement gap. For instance, students who complete the intervention have shown a blunted stress response [8] and improved academic outcomes longitudinally [4], as well as in the lab [13, 26]. There is also evidence that these affirmations reduce disruptive or aggressive behavior in the classroom [33, 34].

In short, research has definitively shown that values affirmations can reduce achievement gaps. However, the content of the essays themselves has not been as thoroughly examined. While some studies have examined the content of expressive writing for instances of spontaneous affirmations [7], or examined affirmations for instances of certain pre-defined themes (e.g., social belonging [28]), these efforts have been on a relatively small scale, and have been limited by the usual constraints associated with hand-annotating (e.g., experimenter expectations, annotator bias, or excessive time

[*]tar2119@columbia.edu; Corresponding Author

requirements).

The goal of this paper is to explore the *content of values affirmation essays* using *data mining techniques*. We explore the differences in the content of affirmation essays as a function of ethnic group membership and gender. We are motivated to address these questions because ethnicity and gender, in the context of academic underperformance and the affirmation intervention, are categorical distinctions of particular interest. Identifying as Black or as a woman means that one is likely to contend with negative stereotypes about intelligence, which in turn puts the individual at risk of experiencing the negative effects of social identity threat. The content of the essays produced by individuals under these different circumstances could lead to insights on the structure of threat or the psychological process of affirmation. Additionally, we hope to eventually use information from this initial study to create affirmation prompts which are tailored to individual differences. That is, it may be beneficial to structure the values-affirmation in different ways depending on the particular threatening context or identity of the writer.

We will explore these issues from two different perspectives. First, we investigate the latent topics of essays using Latent Dirichlet Allocation (LDA) [2], which is a generative model that uncovers the thematic structure of a document collection. Using the distribution of topics in each essay, we will present examples of topics which feature strong and theoretically interesting between-group differences. Second, we approach the question of between-group differences in text as a classification problem. For instance, given certain content-based features of the essays (e.g., topics, n-grams, lexicon-based words), how well can we predict whether an essay was produced by a Black or White student? This approach also allows us to examine those features which are the most strongly discriminative between groups of writers. Finally, classification will allow us to closely compare the relative strength of each model's features with respect to differences between groups.

2. DATA

Our data come from a series of studies conducted on the effectiveness of values affirmations. For the datasets that have resulted in publications, detailed descriptions of the subjects and procedures can be found in those publications [4, 5, 27, 28]. The unpublished data follow nearly identical procedures with respect to the essay generation.

As an illustrative example of the essay generation process, we describe the methods from Cohen et. al [4]. This study, conducted with seventh-graders, featured a roughly equal number of Black and White students who were randomly assigned to either the affirmation condition or a control condition. The affirmation intervention was administered in the student's classrooms, by teachers who were blind to condition and hypothesis. Near the beginning of the fall semester, students received closed envelopes from their teachers, who presented the work as a regular classroom exercise. Written instructions inside the envelope guided students in the affirmation condition to chooose their most important values (or, in study 2, their top two or three most important values) from a list (athletic ability, being good at art, being smart or getting good grades, creativity, independence, living in the moment, membership in a social group, music, politics, relationships with friends or family, religious values, and sense of humor), while control students were instructed to select their least important value (two or three least important values in study 2). Students in the affirmation condition then wrote about why their selected value(s) are important to them, while students in the control condition wrote about why their selected values might be important to someone else. All students quietly completed the material on their own.

The other samples in our data include both lab and field studies and feature methods largely similar to those just described. Across all studies, participants completing the affirmation essays are compared with students who do not suffer from social identity threat as well as students who complete a control version of the affirmation. Our datasets feature students of college age, as well as middle school students. Below we show two examples of *affirmation essays* (one from a college student and one from a middle school student) and a *control essay* (middle school student):

> **Affirmation Essay (college student):** My racial/ethnic group is most important to me when I am placed in situations that are alienating or dangerous or disrespectful. Since coming to Yale a school much larger than my former school where I feel my minority status that much more sharply or feel like people are judging me because I have dark skin I have placed a much higher value on being black. I work for the Af-Am House. I am involved in Black groups and most of my friends are Black. But often being black holds me down and depresses me because people are surprised at how much like them I can be and I dont think Im pretty. Its stressful to have to avoid stereotypes like being late or liking to dance or being sexual. I dont want people to put me in a box labeled black Girl 18. I am my own person.

> **Affirmation Essay (middle school student:)** Being smart and getting good grades is important to me because it is my path to having a succesful life. Independence is also important because I don't want to be like everybody else. I want to be special in my own way. I want to be different.

> **Control Essay:** I think that being good in art can be important to someone else who likes and enjoys art more than I do. I also think this because there are people who can relate and talk about art by drawing and stuff like that but I don't.

In total, we were able to obtain 6,704 essays. Of these, our analyses included all essays which met the following criteria:

1. The essay was an *affirmation* essay (not control). We opted to exclude control essays because the psycholog-

ical process behind the generation of a control essay is fundamentally different from the process that generates an affirmation essay. We are interested in the *affirmation* process, and including control essays in a topic model, for instance, would only add noise to the signal we are interested in exploring.

2. The writing prompt did not deviate (or deviated only slightly) from the writing prompt most widely used across various studies [4]. For example, most of the essays used prompts mentioned above (e.g., athletic ability, religious values, independence). We excluded prompts such as reflection on President Obama's election, since they are of a different nature.

Including only the essays which met the above criteria resulted in a final dataset of 3,097 essays. Given that some individuals wrote up to 7 essays over the period of their participation, the 3,097 essays came from 1,255 writers (425 Black, 473 White, 41 Asian, 174 Latino, 9 other, 83 unrecorded; 657 females, 556 males, 42 unrecorded). The majority of these writers (n = 655) were from a field study in which 8 cohorts of middle school students were followed over the course of their middle school years. The remainder were from several lab-based studies conducted with samples of college students. Before modeling, all essays were preprocessed by removing stop words and words with frequency counts under four. We also tokenized, lemmatized, and automatically corrected spelling using the jazzy spellchecker [11].

The essays varied in length (median number of words = 39, mean = 44.83, SD = 35.85). Some essays are very short (e.g., 2 sentences). As we describe in the next section, this posed some interesting opportunities to test different methods of modeling these essays, especially with regard to using topic models.

3. MODELS FOR CONTENT ANALYSIS

To explore the differences in the content of affirmation essays as a function of ethnic group membership and gender we used several methods to model essay content.

Latent Dirichlet Allocation (LDA). Graphical topic models such as LDA [2] have seen wide application in computational linguistics for modeling document content. Such topic models assume that words are distributed according to a mixture of topics and that a document is generated by selecting a topic with some mixture weight, generating a word from the topic's word distribution, and then repeating the process. LDA specifies a probabilistic procedure by which *essays* can be generated: the writer chooses a topic z_n at random according to a multinomial distribution (θ), and draws a word w_n from $p(w_n|z_n, \beta)$, which is a multinomial probability conditioned on the topic z_n ($\theta \sim Dir(\alpha)$). The topic distribution θ describes the portion of each topic in a document. One drawback of the current LDA framework is that it assumes equal contribution of each word to the topic distribution of a document θ. Since many of our writers tended toward using repetitive language (e.g., miming the essay prompt), we used a modified version of LDA to model our essays, which uses a tf-idf matrix instead of the standard word-count matrix [21]. This allows words that are more unique in their usage to take on greater weight in the topic model. We settled on a model with 50 topics, as this provided a good fit to our data, and topics with good subjective interpretability. Given that a primary goal of our analysis was to investigate the topics, we prioritized interpretable topics over statistical fit when necessary. Figure 1 shows the affirmation essay written by the college student given in Section 2, where words are highlighted to show their topic assignments. This example includes three topics, one of which is clearly related to ethnic group (red text), while the other two are somewhat more ambiguous. Section 4 shows some of the learned topics, an analysis of the topic distributions as a function of gender and race, and the results of using the topic distributions as additional features for classification experiments (gender, ethnicity, and gender-ethnicity).

My racial/ethnic group is most important to me when I am placed in situations that are alienating or dangerous or disrespectful. Since coming to Yale a school much larger than my former school where I feel my minority status that much more sharply or feel like people are judging me because I have dark skin I have placed a much higher value on being black. I work for the Af-Am House. I am involved in Black groups and most of my friends are Black. But often being black holds me down and depresses me because people are surprised at how much like them I can be and I dont think Im pretty. Its stressful to have to avoid stereotypes like being late or liking to dance or being sexual. I dont want people to put me in a box labeled black Girl 18. I am my own person.

Figure 1: An example essay from a college-aged writer. Words have been highlighted to show their topic assignments

Weighted Textual Matrix Factorization (WTMF). Topic models such as LDA [2] have been successfully applied to relatively lengthy documents such as articles, web documents, and books. However, when modeling short documents (e.g., tweets) other models such as Weighted Textual Matrix Factorization (WTMF) [10] are often more appropriate. Since most of our essays are relatively short (2-3 sentences), we use WTMF as an additional method to model essay content. The intuition behind WTMF is that it is very hard to learn the topic distribution only based on the limited observed words in a short text. Hence Guo and Diab [10] include unobserved words that provide thousands more features for a short text. This produces more robust low dimensional latent vector for documents. However, while WTMF is developed to model latent dimensions (i.e., topics) in a text, a method for investigating the most frequent words of these latent dimensions is not apparent (unlike LDA). We therefore use this content analysis method only for the classification tasks (gender, ethnicity, gender-ethnicity), with the induced 50 dimensional latent vector as 50 additional features in classification (Section 4).

Linguistic Inquiry and Word Count (LIWC). Pennebaker et al.'s LIWC (2007) dictionary has been widely used both in psychology and computational linguistics as a method for content analysis. The LIWC lexicon consists of a set of 64

Table 1: Top 10 words from select LDA topics

Topic3	Topic22	Topic33	Topic43	Topic47
relationship	time	group	religion	religious
life	spring	black	church	god
feel	play	white	religious	faith
independent	hang	racial	god	religion
family	talk	identify	treat	jesus
support	help	race	sunday	believe
time	friend	ethnic	believe	belief
friend	family	certain	famous	church
through	homework	culture	stick	christian
help	school	history	lord	earth

word categories grouped into four general classes organized hierarchically: 1) Linguistic Processes (LP) [e.g., Adverbs, Pronouns, Past Tense, Negation]; 2) Psychological Processes (PP) [e.g., Affective Processes [Positive Emotions, Negative Emotions [Anxiety, Anger, Sadness]], Perceptual Processes [See, Hear, Feel], Social Processes, etc]; 3) Personal Concerns (PC) [e.g., Work, Achievement, Leisure]; and 4) Spoken Categories (SC) [Assent, Nonfluencies, Fillers]. LIWC's dictionary contains around 4,500 words and word stems. In our analysis we used LIWC's 64 categories as lexicon-based features in the classification experiments (Section 4).

4. RESULTS

One of our primary questions of interest is whether we can discover between-group differences in the content of the essays. In order to examine this idea in a straightforward way, we limit the analyses to only those individuals who identified as Black or White (2,392 essays from 897 writers). While there are stereotypes suggesting that Asians and Latinos should perform well and poorly in academic domains, respectively, many individuals in our samples who identify with these groups are born in other countries, where the nature of prevailing stereotypes may be different. This is not true to the same extent of individuals who identify as Black or White. We thus exclude Asians and Latinos (as well as those who identified as "other" or declined to answer) for our between-group differences analyses and classification experiments. Inferential analyses were conducted using R [20], and figures were generated using the ggplot2 package [36].

4.1 Interpreting Topic Models

We first describe the results of using LDA to see whether we can detect topics that feature strong and theoretically interesting between-group differences. Accurately interpreting the meaning of learned topics is not an easy process [14] and more formal methods are needed to qualitatively evaluate these topics. However, our initial investigation suggests that participants use common writing prompts to write about values in different ways, depending on the group to which they belong.

Table 1 provides the top 10 words from several learned LDA topics[1]. Manually inspecting the topics, we noticed that LDA not only learned topics related to the values given, but it seemed to be able to learn various aspects related to these

[1]As noted in section 3, we are unable to investigate WTMF models in the same fashion.

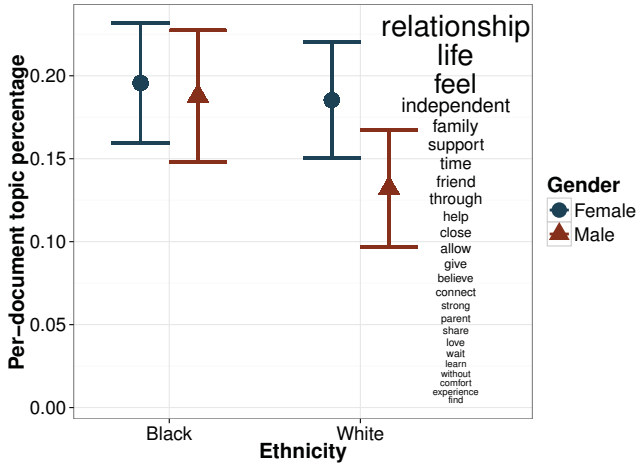

Figure 2: Topic3: Most prominent topic. Points represent fixed effect estimates. Error bars represent represent +/- 1.96 standard errors. Word size represents weighting in the topic

values. For example, Topic43 and Topic47 both relate to religious values but Topic43 refers to religion as it pertains to elements of the institution (including words such as church, sunday, and catholic), while Topic47 seems to focus more on the content of faith itself (indicated by words such as faith, jesus, and belief). A similar interpretation can be given to Topic3 and Topic22 — they both refer to relationship with family and friends, but one focuses on the support and help aspect (Topic3), while the other seems to refer to time spent together and hanging out (Topic22). Finally, Topic33 shows an example where the topic learned is about ethnic group, even if ethnicity was not a specific value given as a prompt (rather the more general value of 'membership in a social group' was given). Figure 1 shows an example of an essay and the word-topic assignments, where Topic33 is one of the topics (ethnic group, shown in red).

In order to identify interesting between-group differences in topic distributions, we fit a series of mixed-effects linear regressions, with each of the 50 topics as the outcomes of interest. For each model, we estimated effects for gender, ethnicity, and the interaction between the two. For the random effects component, we allowed the intercept to vary by writer. Across the 50 models and excluding the intercept, we estimated a total of 150 effects of interest. Of these, 23 reached the threshold for statistical significance. This proportion is greater than would be expected by chance ($p < .01$). Having established that there are real and meaningful between-groups differences, we more closely examined topics which had theoretically interesting insights.

For example, Figure 2 shows the most frequent words from the most prominent topic (Topic3; relationships with family and friends as basis of support/help) across all essays, along with differences between groups. The model for this topic yielded marginal effects of gender ($B = .02$, $SE = .01$, $p = .08$), with female writers devoting a greater proportion of their writing to the topic ($M = .12$, $SD = .27$) than males ($M = .09$, $SD = .24$). There was also a marginal effect of

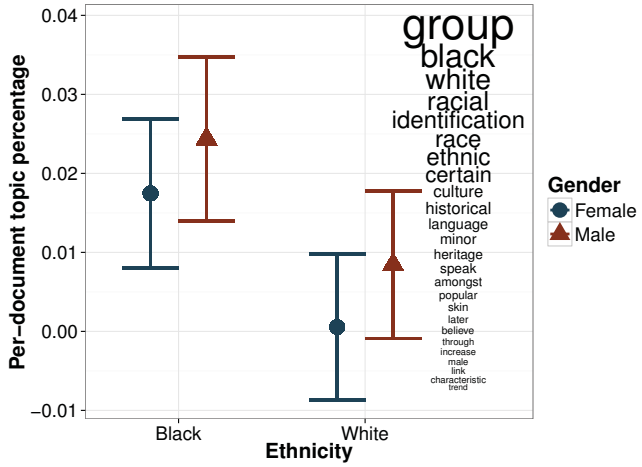

Figure 3: Topic33: effect of ethnicity. Points represent fixed effect estimates. Error bars represent represent +/- 1.96 standard errors. Word size represents weighting in the topic

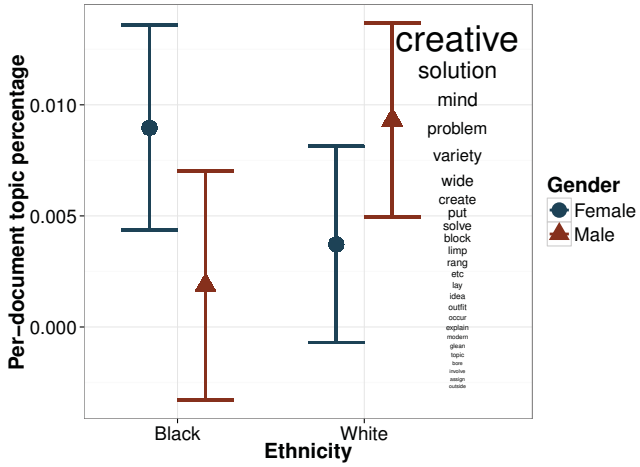

Figure 4: Topic23: Interaction between Gender and Ethnicity. Points represent fixed effect estimates. Error bars represent represent +/- 1.96 standard errors. Word size represents weighting in the topic

ethnicity, ($B = .02$, $SE = .01$, $p = .10$), with black writers ($M = .11$, $SD = .26$) devoting more of their writing to the topic than white ($M = .10$, $SD = .25$) writers.

There were also topics which strongly discriminated between ethnicities. Figure 3 presents findings from one such topic (Topic33; ethnic group). The model for this topic revealed the expected main effect of ethnicity ($B = .008$, $SE = .02$, $p < .01$), with black writers devoting a greater proportion of their writing to the topic ($M = .01$, $SD = .07$) than white writers ($M = .003$, $SD = .03$).

The LDA model also estimated topics that were utilized differently by black and white writers, depending on if they happened to be males or females. For instance, Figure 4 presents a topic which is related to problem-solving. Modeling this topic showed that the interaction between gender and ethnicity was significant ($B = .003$, $SE = .01$, $p < .01$). Specifically, for black writers, women wrote more about this topic ($M = .009$, $SD = .07$) than males did ($M = .001$, $SD = .02$, $p < .05$). For white writers, the difference is in the opposite direction, and marginally significant, with males using more of their writing on this topic ($M = .009$, $SD = .08$) than women ($M = .004$, $SD = .03$, $p = .08$). Similarly, the difference for black and white males is statistically significant ($p < .05$), whereas the difference is reversed and marginal for black and white females ($p = .11$).

The findings from the LDA topic modeling show that there are between-group differences emerging from the affirmation essays. To investigate further, in the next section we present the results of a study where we approach the question of between-group differences as a classification problem.

4.2 Classification:Gender, Ethnicity, Gender-Ethnicity

Given certain content-based features of the essays (e.g., distribution of topics, LIWC categories, n-grams), these experiments aim to classify essays based on the writer's ethnicity and/or gender: Black vs. White (Ethnicity classification), Female vs. Male (Gender classification), and Black-Male vs White-Male and Black-Female vs. White-Female (Ethnicity-Gender classification). In all classification experiments we use a linear Support Vector Machine (SVM) classifier implemented in Weka (LibLINEAR) [9]. We ran 10-fold cross validation and for all results we report weighted F-1 score. As features we used TF-IDF (words weighted by their TF-IDF values)[2]; LDA (topic distributions are used as additional features); WTMF (the 50 dimensional latent vector used as 50 additional features) and LIWC (LIWC's 64 word categories are used as features).

The classification results are displayed in Table 2. We notice that all features give similar performance per classification task. In general, the results were better for the gender classification task (best results 74.09 F1 measure), while the worse results seems to be for the ethnicity classification (best result 66.37 F1). None of the classification tasks showed significant differences as a function of the included features ($p > .05$).

However, the aspect we were more interested in was to analyze the most discriminative features for each classification task with the hope of discovering interesting patterns for between-groups differences. The top 10 discriminating features from each classification type on the TF + LDA + LIWC features are presented in Table 3. There are several interesting observations when analyzing these results. First, supporting the results of the classification experiment, we see that unigrams feature prominently. We also note that LIWC features are largely missing from the top ten, with the only exception being the 10th feature for males in the gender classification. LDA topics, on the other hand, appear as strongly distinguishing in 3 of the 4 classification tasks. Further, in terms of content, the discriminative features sup-

[2]We experimented with presence of n-grams but using TF-IDF gives better results.

Table 2: SVM Results - cell contents are number of P/R/F1

Features	Classification			
	Gender	Ethnicity	Bl vs Wh Female	Bl vs Wh Male
TF-IDF	73.38/73.38/73.33	71.34/67.91/65.13	73.43/69.70/67.97	75.26/70.76/67.29
TF-IDF + LDA	73.48/73.46/73.40	**70.54/68.41/66.37**	73.29/69.62/67.90	74.72/70.85/67.63
TF-IDF + WTMF	73.52/73.46/73.37	71.72/68.00/65.11	**73.11/70.02/68.55**	74.62/70.59/67.23
TF-IDF+LIWC	74.07/74.0/73.92	72.07/68.08/65.10	73.49/69.78/68.07	75.20/70.85/67.45
TF-IDF+LDA+LIWC	**74.09/74.09/74.04**	71.38/68.58/66.24	73.49/69.78/68.07	**74.98/71.02/67.82**

Table 3: Most discriminative features from classifiers with TF-IDF+LDA+LIWC as features

Gender		**Ethnicity**	
Female	Male	Black	White
softball	verry	race	Topic15-relationship, creative
jump	available	result	Topic25-music, play, enjoy
swim	football	heaven	younger
happier	Topic26-play, soccer	barely	less
horse	score	disappoint	weird
cheerleader	language	romantic	Topic17-humor, sense, laugh
doctor	lazy	NBA	larger
Topic14-music, relax	moreover	outdoor	rock
boyfriend	baseball	africa	tease
reason	LIWC27-affect	double (game double dutch)	heavy
Females		**Males**	
Black	White	Black	White
double (game double dutch)	decorate	Topic22-spring, hangout	Topic25-music, play, enjoy
above	rock	NBA	Topic17-humor, sense, laugh
ill	guitar	race	Topic2-reply, already, told
race	peer	head	larger
thick	horse	motive	sit
south	handle	health	cheer
option	grandparents	apart	rock
lord	saxaphone	phone	skate
result	crowd	award	handy
york	less	famous	holiday

port some of the results from the topic model analysis. For instance, topic 33 (ethnic group) is the most discriminative, non-unigram feature for ethnicity, and is the 56th most strongly associated feature with Black writers overall. It is also the most discriminative, non-unigram feature for the female-ethnicity classification, as the 44th most strongly associated feature with Black female writers. However, this topic does not show up for the Black vs White male classification. The topic results (Figure 3) also indicate a somewhat stronger relationship for Black vs. White Females.

We also notice that there are strong effects related to sports. In particular, some of the most discriminative features are consistent with social expectations regarding participation in various types of sports. Females, for instance, are more likely to write about softball, swimming, and jumping rope, whereas males are more likely to write about football and baseball. Similar differences can be seen for ethnicity (NBA, double dutch), and gender-ethnicity classifications (females: double dutch, horse; males: NBA, skate).

5. RELATED WORK

As mentioned in the introduction, there have been some smaller-scale investigations into the content of affirmation essays. For instance, Shnabel et al.[28] hand-annotated a subset of the data presented here for presence of social belonging themes. They defined social belonging as writing about an activity done with others, feeling like part of a group because of a shared value or activity, or any other reference to social affiliation or acceptance. Their results indicate that the affirmation essays were more likely to contain such themes than control essays, and that Black students who wrote about belonging themes in their affirmation essays had improved GPAs relative to those who did not write about social belonging. A subsequent lab experiment confirmed this basic effect and strengthened the hypothesized causal claim. The data here are consistent with the idea that social themes are a dominant topic in these essays. Indeed, the most prominent topic (Topic3) seems to be a topic that directly corresponds to social support (see Table 1). Further, even a cursory glance at the topics we have included here will show that references to other people feature prominently - a pattern that is also true for the topics we have not discussed in this paper.

One other finding of interest concerns the discriminative ability of LIWC. Only for the gender classification did LIWC categories appear among the discriminative features. There

are many studies that show gender differences in LIWC categories [25, 19, 24, 16], to say nothing of the broader literature on differences in language use between men and women [15, 12]. However, there is far less consistent evidence for differences in LIWC categories as a function of ethnicity [18]. That our results indicate features from LDA are more discriminative for ethnicity suggests the utility of a bottom-up approach for distinguishing between these groups. However, it should be noted that, in general, classification performance on ethnicity was not as good as classification on gender.

Finally, we also note that this is one of a small, but growing number of studies directly contrasting LIWC and LDA as text modeling tools [30, 22, 25]. While this other work tends to find that LDA provides additional information which results in improvements to classification performance in comparison to LIWC, our do not display this pattern. It is not clear why this may be, although we suspect that frequent misspellings present in our data could lead to some of the discrepancy.

6. CONCLUSIONS

We used data mining techniques to explore the content of a written intervention known as a *values affirmation*. In particular, we applied LDA to examine latent topics that appeared in students' essays, and how these topics differed as a function of whether the group to which the student belonged (i.e., gender, ethnicity) was subject to social identity threat. We also investigated between-groups differences in a series of classification studies. Our results indicate that there are indeed differences in what different groups choose to write about. This is apparent from the differences in topic distributions, as well as the classifier experiments where we analyzed discriminative features for gender, ethnicity and gender-ethnicity.

Why might individuals coping with social identity threat write about different topics than those who are not? Some literature shows that racial and gender identity can be seen as a positive for groups contending with stigma [29]. The model of optimal distinctiveness actually suggests that a certain degree of uniqueness leads to positive outcomes [3]. This suggests that if an individual from a stigmatized group perceives their identity to be unique, it may be a source of pride. In the current context, this could be reflected in an increase of writing devoted to the unique social group students are a part of (i.e., African American). On the other hand, there is some evidence that individuals downplay or conceal identities they perceive to be devalued by others [23]. This work would suggest that students in our data would choose to write about what they have in common with others. Our work here seems to provide some support for the former, but we have not addressed these questions directly, and so cannot make any strong claims.

Looking forward, we intend to investigate the relationship between essay content and academic outcomes. Do stigmatized students who write about their stigmatized group experience more benefit from the affirmation, as would be suggested by the optimal distinctiveness model? This work could provide data that speak to this issue. Furthermore, we hope to model the trajectory of how the writing of an individual changes over time, especially as a function of whether they completed the affirmation or control essays. Given that values affirmations have been shown to have long-term effects, and our data include some individuals who completed multiple essays, exploration of longitudinal questions about the affirmation are especially intriguing. We also intend to model the essays using supervised-LDA, which would allow us to jointly model the topics with the grouping information. Last but not least we plan to investigate whether there are differences between the middle school students and the college-level students.

7. ACKNOWLEDGMENTS

We would like to thank Robert Backer and David Watkins for assistance with this project. This work was supported in part by the NSF under grant DRL-1420446 and by Columbia University's Data Science Institute through a Research Opportunities and Approaches to Data Science (ROADS) grant.

8. REFERENCES

[1] Women, minorities, and persons with disabilities in science and engineering. Technical Report NSF 13-304, National Science Foundation, National Center for Science and Engineering Statistics, Arlington, VA., 2013.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] M. B. Brewer. The social self: On being the same and different at the same time. *Personality and Social Psychology Bulletin*, 17(5):475–482, 1991.

[4] G. L. Cohen, J. Garcia, N. Apfel, and A. Master. Reducing the racial achievement gap: A social-psychological intervention. *Science*, 313(5791):1307–1310, 2006.

[5] G. L. Cohen, J. Garcia, V. Purdie-Vaughns, N. Apfel, and P. Brzustoski. Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science*, 324(5925):400–403, 2009.

[6] J. E. Cook, V. Purdie-Vaughns, J. Garcia, and G. L. Cohen. Chronic threat and contingent belonging: Protective benefits of values affirmation on identity development. *Journal of Personality and Social Psychology*, 102(3):479, 2012.

[7] J. D. Creswell, S. Lam, A. L. Stanton, S. E. Taylor, J. E. Bower, and D. K. Sherman. Does self-affirmation, cognitive processing, or discovery of meaning explain cancer-related health benefits of expressive writing? *Personality and Social Psychology Bulletin*, 33(2):238–250, 2007.

[8] J. D. Creswell, W. T. Welch, S. E. Taylor, D. K. Sherman, T. L. Gruenewald, and T. Mann. Affirmation of personal values buffers neuroendocrine and psychological stress responses. *Psychological Science*, 16(11):846–851, 2005.

[9] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear - a library for large linear classification, 2008. The Weka classifier works with version 1.33 of LIBLINEAR.

[10] W. Guo and M. Diab. Modeling sentences in the latent space. In *Proceedings of the 50th Annual Meeting of the Association for Computational*

Linguistics: Long Papers-Volume 1, pages 864–872. Association for Computational Linguistics, 2012.

[11] M. Idzelis. Jazzy: The java open source spell checker, 2005.

[12] R. T. Lakoff. *Language and woman's place: Text and commentaries*, volume 3. Oxford University Press, 2004.

[13] A. Martens, M. Johns, J. Greenberg, and J. Schimel. Combating stereotype threat: The effect of self-affirmation on women's intellectual performance. *Journal of Experimental Social Psychology*, 42(2):236–243, 2006.

[14] Q. Mei, X. Shen, and C. Zhai. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 490–499. ACM, 2007.

[15] A. Mulac, J. J. Bradac, and P. Gibbons. Empirical support for the gender-as-culture hypothesis. *Human Communication Research*, 27(1):121–152, 2001.

[16] M. L. Newman, C. J. Groom, L. D. Handelman, and J. W. Pennebaker. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211–236, 2008.

[17] H.-H. D. Nguyen and A. M. Ryan. Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, 93(6):1314, 2008.

[18] M. Pasupathi, R. M. Henry, and L. L. Carstensen. Age and ethnicity differences in storytelling to young children: Emotionality, relationality and socialization. *Psychology and Aging*, 17(4):610, 2002.

[19] J. W. Pennebaker and L. A. King. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296, 1999.

[20] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[21] D. Ramage, S. T. Dumais, and D. J. Liebling. Characterizing microblogs with topic models. *ICWSM*, 5(4):130–137, 2010.

[22] P. Resnik, A. Garron, and R. Resnik. Using topic modeling to improve prediction of neuroticism and depression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1348–1353. Association for Computational Linguistics, 2013.

[23] L. M. Roberts. Changing faces: Professional image construction in diverse organizational settings. *Academy of Management Review*, 30(4):685–711, 2005.

[24] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205, 2006.

[25] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[26] J. R. Shapiro, A. M. Williams, and M. Hambarchyan. Are all interventions created equal? A multi-threat approach to tailoring stereotype threat interventions. *Journal of Personality and Social Psychology*, 104(2):277, 2013.

[27] D. K. Sherman, K. A. Hartson, K. R. Binning, V. Purdie-Vaughns, J. Garcia, S. Taborsky-Barba, S. Tomassetti, A. D. Nussbaum, and G. L. Cohen. Deflecting the trajectory and changing the narrative: How self-affirmation affects academic performance and motivation under identity threat. *Journal of Personality and Social Psychology*, 104(4):591, 2013.

[28] N. Shnabel, V. Purdie-Vaughns, J. E. Cook, J. Garcia, and G. L. Cohen. Demystifying values-affirmation interventions writing about social belonging is a key to buffering against identity threat. *Personality and Social Psychology Bulletin*, 39(5):663–676, 2013.

[29] T. B. Smith and L. Silva. Ethnic identity and personal well-being of people of color: a meta-analysis. *Journal of Counseling Psychology*, 58(1):42, 2011.

[30] A. Stark, I. Shafran, and J. Kaye. Hello, who is calling?: Can words reveal the social nature of conversations? In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 112–119. Association for Computational Linguistics, 2012.

[31] C. M. Steele. The psychology of self-affirmation: Sustaining the integrity of the self. *Advances in Experimental Social Psychology*, 21:261–302, 1988.

[32] C. M. Steele, S. J. Spencer, and J. Aronson. Contending with group image: The psychology of stereotype and social identity threat. *Advances in Experimental Social Psychology*, 34:379–440, 2002.

[33] S. Thomaes, B. J. Bushman, B. O. de Castro, G. L. Cohen, and J. J. Denissen. Reducing narcissistic aggression by buttressing self-esteem: An experimental field study. *Psychological Science*, 20(12):1536–1542, 2009.

[34] S. Thomaes, B. J. Bushman, B. O. de Castro, and A. Reijntjes. Arousing âĂIJgentle passionsâĂİ in young adolescents: Sustained experimental effects of value affirmations on prosocial feelings and behaviors. *Developmental Psychology*, 48(1):103, 2012.

[35] G. M. Walton and G. L. Cohen. Stereotype lift. *Journal of Experimental Social Psychology*, 39(5):456–467, 2003.

[36] H. Wickham. *ggplot2: Elegant graphics for data analysis*. Springer New York, 2009.