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 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. religion, 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 (National Center for Education Statistics), 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 [28]. 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 will feel psychologically threatened, leading to changes in affect, physiology, and behavior[16, 31, 24, 5].

While these insidious effects lead to outcomes that are extraordinary in their injustice, the description of a psychological process which 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 [27].

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 [12, 23]. There is also evidence that these affirmations reduce disruptive or aggressive behavior in the classroom [29, 30].

In short, research has definitively shown that values affirmations can reduce the achievement gap. 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 [25]), these efforts have been

[*] tar2119@columbia.edu; Corresponding Author

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 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 must 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.

2. DATA

Our data comes from a series of studies conducted on the effectiveness of values affirmations. For the datasets which have resulted in publications, detailed descriptions of the subjects and procedures can be found in those publications [4, 5, 24, 25]. The unpublished data follow nearly identical procedures with respect to the essay generation.

The values affirmation writing procedure asks participants to read a list of values (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), and select one, two, or three which are particularly important to them. Following this, participants are instructed to think about the importance of these values and to describe how and why the selected values are important to them. These essays were written by participants in a series of lab and field studies. Across all studies, students 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. In the control version, participants view the same list of values, but are asked to select those values which are least important to them and then to write about why those values might be important to someone else. 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 psychological process behind the generation of a control essay is fundamentally different from the process when 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 different nature.

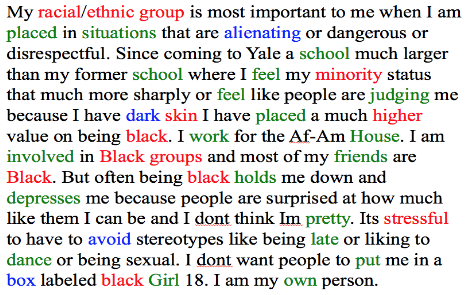
Including only the essays which met the above criteria resulted in a final dataset of 3,097 essays. Given that some individuals wrote up to 7 essays over the period of their participation, the 3,097 essays came from 1,255 writers (425 Black, 473 White, 41 Asian, 174 Latino, 9 other, 83 unrecorded; 657 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.

The length of the essays vary in length (median number of words = 39, mean = 44.83, SD = 35.85), some essays being 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 [19]. This allows words which 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 provided 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 essays written by the college student we gave 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 model 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 ($\tilde{2}$-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 use both in psychology and computational linguistics as a method for content analysis. LIWC lexicon consists of a set of 64 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 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 Lati-

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

nos 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 identify as 'other' or declined to state) for our between-group differences analyses and classification experiments.

4.1 Interpreting Topic Models

We first describe the results of using LDA to see whether we can detect topics which feature strong and theoretically interesting between-group differences. Accurately interpreting the meaning of learned topics is not an easy process [13] 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. 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 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 spend together and hanging out (Topic22). Finally, Topic33 show and 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 our analysis, we identified topics that show theoretically interesting between-group differences (e.g., with respect to gender or ethnicity). To perform this analysis we looked at the differences in the distributions of each topic by group. 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 the differences in the distribution by group. An ANOVA showed on this topic revealed a main effect of *gender* type F(1, 2,393) = 5.02, p = .03, with females devoting a greater proportion to the topic (M = .12, SD = .27) than males (M = .09, SD = .24).

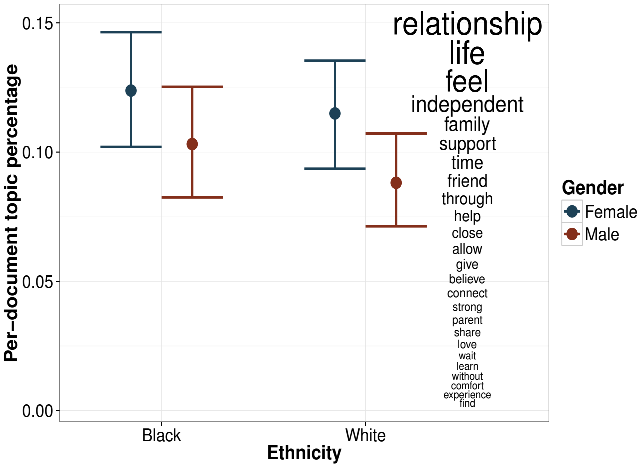


Figure 2: Topic3: effect of gender. Topic distribution as a function of gender and ethnicity. Error bars represent bootstrapped 95% confidence intervals. Word size represents weighting in the topic

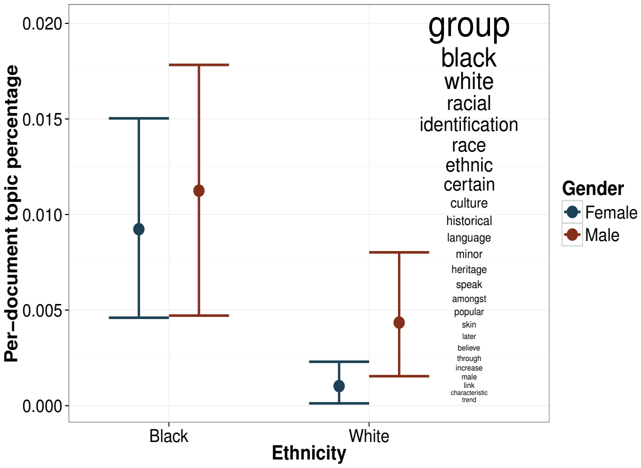


Figure 3: Topic33: effect of ethnicity. Topic distribution as a function of gender and ethnicity. Error bars represent bootstrapped 95% confidence intervals. Word size represents weighting in the topic

There were also topics which strongly discriminated between ethnicities. Figure 3 presents findings from one such topic (Topic33; ethnic group). An ANOVA investigating the effect of gender and ethnicity on this topic revealed the expected main effect of *ethnicity* F(1, 2,393) = 12.17, $p < .01$, with Black writers devoting a greater proportion of their writing to the topic (M = .01, SD = .07) than White writers (M = .003, SD = .03).

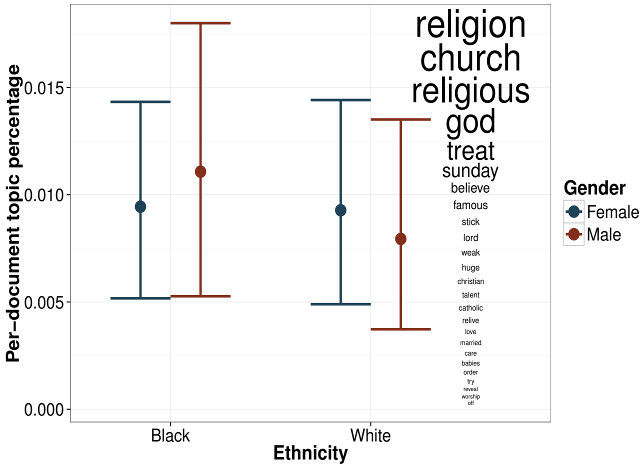


Figure 4: Topic43: Topic distribution as a function of gender and ethnicity. Error bars represent bootstrapped 95% confidence intervals. Word size represents weighting in the topic

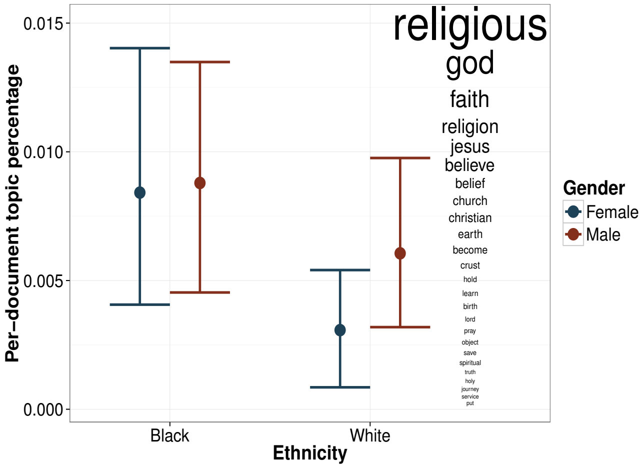


Figure 5: Topic47: Topic distribution as a function of gender and ethnicity. Error bars represent bootstrapped 95% confidence intervals. Word size represents weighting in the topic

We also looked for topics which seemed to cover similar content, such as Topic43 and Topic47 related to religion. For Topic43 which is related to religion as it pertains to elements of the institution (Figure 4), an ANOVA revealed no effects of ethnicity, gender, nor their interaction (p's $> .05$). In contrast, for Topic47 which is focused on the content of the faith itself (Figure 5), an ANOVA revealed a main effect of *ethnicity* F(1, 2,393) = 4.29, $p = .04$, with Black writers writing more about this topic ($M = .009$, $SD = .06$) than White writers ($M = .005$, $SD = .04$).

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 on essay based on its 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 1) TF-IDF (words weighted by their TF-IDF values) [1]; LDA (topic distributions are used as additional features); WTMF (the 50 dimensional latent vector used as 50 additional features); LIWC categories.

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).

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 is presented in Table 3. There are several interesting observations when analyzing these results. First, LIWC features seems to be more prominent in the Gender classification task and rather absent from the other two classification tasks, while LDA topics seems to useful distinguishing features in the Ethnicity and Gender-Ethnicity classifications. In terms of content, we notice that this analysis supports some of the findings from our topic model analysis. For example, Topic33 (ethnic group) is one of the most discriminative features for ethnicity (Blacks tend to write more about this topic). What is interesting, is that when looking at Gender-Ethnicity, we notice that Topic33 only appears in the Black Female vs. White Female and not for Black Male vs. White Male. The topic model results (Figure 3) also show a stronger effect w.r.t to Black Female vs White Female. Another aspect that seems to support the findings using LDA, is that females tend to write more about Family (LIWC-Family category) and relations with others (LIWC-Inclusive category which contains words such as ”along”, ”both”, ”open”, ”with”) than males do. However, if we are looking at Gender-Ethnicity we notice that Topic22 (time spend with family, Table 1) seems to be one of the discriminative feature for Black Males. Interestingly enough, male seem to use more positive emotions and words related to affect in general (LIWC-Positive Emotion and LIWC-Affect category) than females. Even though both female and male write about sports, female tend to write about sports such as swimming, and horse riding, while males write about running and football.

[1] 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
LIWC-Inclusive	LIWC-Affect	Topic33-ethnic group	Topic15- relationship, creative
LIWC-Conjunction	LIWC-Positive Emotion	barely	younger
softball	verry	romantic	Topic25-music, play, enjoy
LIWC-2nd Person Pronoun	LIWC-Biological Processes	distance	family
jump	LIWC-Health	africa	less
LIWC-Family	LIWC-Filler	avoid	weird
swim	avail	heaven	LIWC-Affect
horse	score	dig	hockei
doctor	run	doubl	lazy
happier	footbal	result	lager

Females		Males	
Black	White	Black	White
Topic33-ethnic group	decorate	Topic22-time,hang,family	Topic2-reply, already, old
above	guitar	head	Topic25-music, play, enjoy
double	rock	apart	Topic17-humor, sense, sport
combine	peer	avoid	larger
ill	grandparent	nap	sit
south	saxaphon	phone	golf
option	crowd	race	rock
design	handl	motiv	handy
race	horse	award	holiday
factor	stronger	famous	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.[25] 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, 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 which 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 which is also true for the topics we have not discussed in this paper.

One other finding of interest concerns the discriminative ability of LIWC. In our data, several LIWC categories emerged as strong discriminators of gender. There are many other studies which also show gender differences in LIWC categories [22, 18, 21, 15], to say nothing of the broader literature on differences in language use between men and women [14, 11]. However, there is far less consistent evidence for differences in LIWC categories as a function of ethnicity [17]. That our results indicate features from LDA are more discriminative for ethnicity suggests the utility of a bottom-up approach for distinguishing these groups. However, it should be noted that, in general, classification performance on ethnicity was not as good as classification on gender.

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 which 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 [26]. 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 [20]. 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 which 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. 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: 2013. special report nsf 13-304. Technical Report Special 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] R. T. Lakoff. *Language and woman's place: Text and commentaries*, volume 3. Oxford University Press, 2004.

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

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

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