Towards Algorithmic Fidelity: Mental Health Representation across Demographics in Synthetic vs. Human-generated Data

Shinka Mori, Oana Ignat, Andrew Lee, Rada Mihalcea
University of Michigan, Ann Arbor
Ann Arbor, MI, USA
{shinkamo, oignat, ajyl, mihalcea}@umich.edu

Abstract

Synthetic data generation has the potential to impact applications and domains with scarce data. However, before such data is used for sensitive tasks such as mental health, we need an understanding of how different demographics are represented in it. In our paper, we analyze the potential of producing synthetic data using GPT-3 by exploring the various stressors it attributes to different race and gender combinations, to provide insight for future researchers looking into using LLMs for data generation. Using GPT-3, we develop HEADROOM, a synthetic dataset of 3,120 posts about depression-triggering stressors, by controlling for race, gender, and time frame (before and after COVID-19). Using this dataset, we conduct semantic and lexical analyses to (1) identify the predominant stressors for each demographic group; and (2) compare our synthetic data to a human-generated dataset. We present the procedures to generate queries to develop depression data using GPT-3, and conduct analyzes to uncover the types of stressors it assigns to demographic groups, which could be used to test the limitations of LLMs for synthetic data generation for depression data. Our findings show that synthetic data mimics some of the human-generated data distribution for the predominant depression stressors across diverse demographics.

Keywords: Synthetic Data Generation, LLMs, Bias in LLMs, Mental Health Datasets

1. Introduction

The emergence of Large Language Models (LLMs) poses many exciting use cases in various applications (Bang et al., 2023). In particular, synthetic data generation (Tang et al., 2023) has great potential to impact domains such as mental health, where it can scale hard-to-acquire data to improve medical information extraction (Agrawal et al., 2022), provide clinical decision support (Shen et al., 2023), and enhance patient-doctor communication (Kreimeyer et al., 2017).

Prompt Template	Sample Output
"I want you to act like a {race} {gender} who is feeling depressed. Write a *blog post* to describe the main source of stress in your life'	*.. I'm **not good enough**, like I'm not doing enough. I'm struggling to **make ends meet** and I'm constantly worried about **money**. I'm worried about **my family** and their **safety** ...*

Table 1: Example of prompt templates used for HEADROOM, as well as sample outputs.

However, before using synthetic data, we need to understand the potential biases across demographics within the underlying models generating such data. Otherwise, a subsequent model trained on biased synthetic data can have undesirable consequences, such as misrepresentations of minority voices or, specifically in mental health, a misdiagnosis (Potts et al., 1991; Call and Shafer, 2018).

To address this need, we conduct extensive analyses to understand the similarities and differences between synthetic and human-generated data. We focus on mental health data, specifically depression stressors across races and genders. We study if GPT-3 accurately captures depression stressors across demographics and if the stressors map closely to those found in human-generated data. Argyle et al. (2022) coin the term "algorithmic fidelity" to describe the degree to which models mimic the real-life distributions for a particular group. Inspired by this, we aim to measure how accurately GPT-3 represents depression stressors for different demographics with the following research questions:

RQ1: What are the depression stressors identified by GPT-3 for different demographic groups and does it capture demographic biases?

RQ2: How does synthetic data about depression stressors compare to human-generated data across demographics?

We closely follow the analyses done by Aguirre et al. (2022) on human-generated data to discover patterns of depression stressors among demographics. Namely, we generate a similar dataset by prompting GPT-3 to produce outputs representative of diverse demographics and compare our findings to theirs.

Our work makes the following contributions.

LREC-COLING 2024, pages 16378–16391
20-25 May, 2024.

First, we develop and publish HEADROOM: a syntHEtic dAtaset of Depression-triggering stRessOrs acrOss deMographics, using GPT-3 while controlling for race, gender, context, and time phase – before and after COVID-19. **Second**, we identify the most predominant depression stressors for each demographic group. **Third**, we conduct semantic and syntactic analyses to compare our synthetic data to a human-generated dataset. Our findings show that GPT-3 exhibits some degree of "algorithmic fidelity" – the generated data mimics some real-life data distributions for the most prevalent depression stressors among diverse demographics.

2. Related Work

LLMs for Generating Mental Health Datasets Across Demographics. Psychological studies show that depression affects racial and gender groups differently (Brody et al., 2018). Despite this, there are still discrepancies where minority groups are often overlooked for depression diagnoses (Stockdale et al., 2008). While demographic information is a key aspect to consider when conducting mental health studies, obtaining such data in the mental health domain is challenging due to safety and privacy regulations (Mattern et al., 2022). As a result, researchers often turn to alternative methods of obtaining demographic labels, such as using automated classifiers, keywords, or lists of names (Wang and Jurgens, 2018). However, as presented in Field et al. (2021), such methods fail to account for the multidimensionality of race due to simplifications inherent in classification models: i.e., classifiers predicting demographics in tweets perform poorly on Asian and Hispanic samples (Wood-Doughty et al., 2021). Furthermore, commonly used mental health datasets, such as CLPsych (Coppersmith et al., 2015) and Multitask (Benton et al., 2017), underrepresent specific demographics such as men and Hispanic individuals (Aguirre et al., 2021).

An alternative to predicting demographic labels using machine learning is to *generate* demographic data using LLMs. Argyle et al. (2022) show that GPT-3 can generate political stances regarding recent elections in the United States that strongly correlate with real-life voter distributions. Møller et al. (2023) compare the performance of classifiers trained on human-generated versus LLM-generated data, demonstrating that classifiers trained on synthetic data can perform well on tasks such as social dimensions.

Considering the inherent risks of applying these approaches to mental health tasks, measuring the *algorithmic fidelity* of LLMs in the mental health domain is essential. For instance, Lin et al. (2022) demonstrate that LLMs carry different mental health stigmas for men and women. In our work, we also study demographic biases in LLMs by generating synthetic data using GPT-3 and analyzing it against human-generated data.

Depression-triggering Stressors Across Demographics Analysis. Depression stressors can vary greatly depending on the demographic, due to systemic racism, racial dynamics, gender discrimination, immigration status, and other factors such as COVID-19 (McKnight-Eily et al., 2021). Specifically, Loveys et al. (2018) analyze data from self-reported depression users in an online peer support community.[1] Similar to our work, Loveys et al. (2018) use Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007, 2015), a lexical analysis tool, to compare the stressors between racial groups, and found some critical differences in stressor patterns between demographics.

3. Datasets

To answer our research questions, we generate data using GPT-3 while controlling for race, gender, and time (before and after COVID-19), to simulate human-generated data. We then conduct semantic and lexical analyses to find patterns in the synthetic data. Next, we compare our findings to those based on The University of Maryland - OurDataHelps dataset (UMD-ODH) (Kelly et al., 2020, 2021), a demographically diverse human-generated dataset about depression stressors.

3.1. UMD-ODH: Human-generated Data

UMD-ODH (Kelly et al., 2020, 2021) contains open-ended responses from patients clinically diagnosed with depression and psychosis. Patients were asked: *"Describe the biggest source of stress in your life at the moment. What things have you done to deal with it?"*

Aguirre et al. (2022) further process this data by selecting the survey responses that have demographic data available, resulting in $2,607$ survey responses. The resulting demographic information is shown in Table 2.

3.2. HEADROOM: GPT-3 generated Data

We generate our synthetic dataset with GPT-3.[2] We use GPT-3 because it is one of the largest LLMs available, and has been demonstrated to effectively

[1] https://www.7cups.com/

[2] Text-Davinci-003, https://platform.openai.com/docs/models/gpt-3-5

	RACE					GENDER			COVID-19	
Responses	White	Black	Asian	Latinx	Other	Male	Female	Other	Before Pandemic	After Pandemic
UMD-ODH	1,761	221	246	277	102	1,857	659	94	890	1,717
HEADROOM	780	780	780	780	–	1,500	1,500	–	1,440	1,680

Table 2: Demographic statistics from UMD-ODH and HEADROOM.

Topic	UMD-ODH	HEADROOM	Similarity
Family	family, focus, year, planning, friends	can, stress, deal, lot, person	0.78
Work	work, lot, hard, week, balance	job, stress, work, sourc, life, work	0.90
Health	like, feel, surgery, found, productive	constant, feel, take, health, mental	0.85
Finance	money, bills, pay, sleep, lack	job, struggl, find, make, end	0.94
Relationship	day, stressed, relationship, tried, think	feel, thing, make, depress, like	0.92
School	problems, friends, program, plans, dissertation	asian, succeed, expect, pressur, fall	0.90
News, Social media	help, people, social, use, avoid	like, climat, current, stress, polit	0.84
Unemployment	job, new, finding, lost, looking	look, lost, time, get, month	0.87

Table 3: The keywords corresponding to each overarching topic in UMD-ODH and HEADROOM, together with the cosine similarity between the averaged GloVe embeddings (Pennington et al., 2014) of the keywords corresponding to each topic. A cosine similarity between two sets of random words gives us a baseline of 0.75.

emulate human texts (Argyle et al., 2022), but our study can be done with any LLM.[3]

Prompt Tuning. To simulate human-generated data, we paraphrase the prompt that Kelly et al. (2020, 2021) use for their *human survey*.

We find that we can obtain more detailed responses with additional context in our prompts. Therefore, we provide additional context such as *writing a blog post*, *posting on Reddit*,[4] or *talking to a therapist*. We use three contexts to obtain more diverse responses. For each prompt, we also specify the user's *gender* (women and men), *race* (Asian, African American, Hispanic, and White), and *context* (blog post, Reddit post, and therapy session). We produce outputs for each *race*, *gender*, and *context* combination. The prompts and example outputs are displayed in Table 1.

The human-generated data also contains samples collected after the start of COVID-19, which may affect the stressor patterns. Therefore we also control for time by indicating the year (2020, 2021) in the prompts while preserving the data distribution. Prompts that do not indicate the year are assumed to represent pre-COVID-19 samples.

[3]We also attempted to use ChatGPT, but due to its content filters, the prompt had to be heavily engineered, which may add confounding variables.

[4]r\Depression

For the *blog post* context, we generate 720 samples, 30 samples per demographic group before COVID-19 and 60 samples after COVID-19. For the other two contexts, we generate $2,400$ total samples, 150 samples per demographic group. The data statistics are summarized in Table 2.

4. Dataset Analysis Methods

Following Aguirre et al. (2022), we conduct two analyses on our synthetic dataset: (1) Semantic analyses using Structural Topic Model (STM), and (2) Lexical analyses using log-odds-ratio with Latent Dirichlet prior.

Semantic Analyses. STM is a variant of Latent Dirichlet allocation (LDA) that also allows the addition of covariates, or metadata, to accompany the textual features. Unlike LDA, which calculates topic prevalence and content from Dirichlet distributions whose parameters are set in advance, STM uses metadata to find the topic prevalence and content. Following Aguirre et al. (2022) who annotated and filtered their topics to 25, we use gender, race, and time (before and after COVID-19) and generate 25 topics. Two annotators labeled the topics based on their most prevalent keywords, while filtering out unclear topics. The annotators obtained a Fleiss' kappa score of $k = 0.52$, which shows a *moderate* agreement (Fleiss, 1971, 1973).

	GENDER			
	Category	Ratio	Category	Ratio
(a)	Women (+)		Men (-)	
	female	3.95	male	-4.06
	i	2.83	we	-1.88
	pro1	2.60	verb	-1.76
	ppron	1.32	tentat	-1.21
	home	1.10	money	-1.02
	ETHNICITY			
	Category	Ratio	Category	Ratio
(b)	Asian (+)		African American (-)	
	work	4.06	see	-8.41
	nonflu	3.11	percept	-4.71
	home	2.31	health	-3.88
	reward	1.57	and	-1.76
	achiev	1.53	compare	-1.68
(c)	Asian (+)		White (-)	
	leisure	2.73	anx	-2.60
	work	2.24	focusfuture	-1.50
	reward	2.14	tentat	-1.37
	achiev	1.62	ingest	-1.32
	negate	1.34	relativ	-0.99
(d)	Hispanic (+)		White (-)	
	home	7.48	insight	-3.36
	leisure	7.37	percept	-3.17
	family	7.18	cogproc	-2.95
	affiliation	5.30	see	-2.62
	social	2.51	compare	-1.83
(e)	African American (+)		White (-)	
	see	5.88	insight	-2.28
	bio	4.01	adverb	-2.25
	percept	3.78	tentat	-2.06
	health	3.40	you	-1.80
	body	2.78	space	-1.78
(f)	Hispanic (+)		African American (-)	
	home	7.33	see	-8.49
	family	6.79	percept	-6.96
	leisure	6.72	bio	-5.24
	affiliation	3.67	health	-4.67
	social	3.51	feel	-4.40
(g)	Hispanic (+)		Asian (-)	
	affiliation	5.13	cogproc	-3.00
	home	5.04	i	-2.75
	leisure	4.70	reward	-2.60
	family	4.61	negate	-2.39
	social	2.73	certain	-2.23

Table 4: Highlights of LIWC log-odds ratio analysis on HEADROOM showing LIWC categories related to predominant stressors when comparing between genders and demographic group pairs. For the (+) group, higher score indicates higher prominence; for the (-) group, lower score indicates higher prominence

We obtain 23 fine-grained topics that we manually cluster into eleven overarching topics. The fine-grained and corresponding overarching topics can be seen in the Appendix (Table 5). Of the eleven overarching topics, eight match those in the human-generated data. The matching topics and their keywords are shown in Table 3. Topics from UMD-ODH that did not have a match with our data include `school/ grad school` and `daily stress`. UMD-ODH has two topics related to `school` – the first relates to school in general, and the second relates to graduate school. While our dataset has a topic for `general stress`, we concluded that none of the keywords are similar enough to be considered a match.[5]

The four overarching topics that appear in the synthetic data and not in the human-generated data are: `general stress`, `racism and police violence`, `immigration status` and `pandemic`. We conduct a pairwise analysis between each gender and race pair using these topics. Effectively, to find the difference between topic proportions for each demographic pair, we estimate a regression to find the topic proportion with the added covariate information. This is then used to extract the prevalence of a topic (topic distribution) for each demographic pair. We show and discuss the difference in the topic prevalence for each demographic pair in Figure 2 and Figure 3.

Lexical Analyses. LIWC (Pennebaker et al., 2007, 2015) includes dictionaries of English words related to human cognitive processes. Specifically, we use the LIWC 2015 dictionary, which contains 6,400 word stems. Each word stem is assigned to multiple categories, e.g., *father* is assigned to: `male`, `family` and `social`.

Aguirre et al. (2022) apply log-odds-ratio with Latent Dirichlet prior, based on the work of Monroe et al. (2017), which aims to capture how a demographic group uses a specific LIWC category compared to another demographic group. For example, to compare the proportions in which one group uses the `negemo` (negative emotion) LIWC category compared to another group, we calculate the log-odds-ratio to get the odds of `negemo` being used in the first group compared to the latter. To calculate the Dirichlet prior, we use the LIWC category counts in the CLPsych dataset (Coppersmith et al., 2015). Note that we do not normalize the results with a control text unrelated to depression, to preserve comparison fidelity with Aguirre et al. (2022), who also do not normalize.

[5] `feeling stuck`, `staying strong`, `uncertainty`, `comparing to others`, `helplessness`, `stress and anxiety`, `loneliness`, `perfectionism`

We show the top five words that have a high log-odds-ratio in Table 4 and highlight the LIWC categories also present in the pairwise lexical analysis from Aguirre et al. (2022) in the Appendix Tables 6, 7, 8, 9, 10, 11, and 12. Insights from the data analysis are presented in Section 5.

5. Research Questions

RQ1. What are the depression stressors identified by GPT-3 for different demographic groups, and do they capture demographic biases? The topic proportion between different demographics, and lexical analyses indicate demographic differences regarding stressors. Refer to Figure 2, Figure 3 and Table 4 for the figures.

Gender. Between genders, women have more mentions of first-person pronouns (`pro1` and `ppron`) (Table 4 (a)). Also, we find the following prevalent stressors: `health`, `news and social media`, `news and politics`, `family`, and `relationship`. See Figure 2 (g), Figure 3 (g).

In contrast, men tend to mention stressors regarding `finances and unemployment`, and `school` more than women. Furthermore, topics regarding `racism` and `police brutality` are much more prominent in men than women.

Both women and men mention stressors related to `work`, but for different reasons: women about `work2/ work-pressure` and men about `work1/ work-fatigue`. The two types of work-related stressors are defined in Appendix Table 5.

We acknowledge that we are excluding other gender identities by only comparing between two genders, *women* and *men*. We take this decision because the comparison data used in Aguirre et al. (2021) is primarily from binary genders, *women, men*, and very few from *other*.

Race. We conduct a pairwise analysis for each race group.

African American. The African American group tends to mention words related to `health`, `body`, `perception` and `family` (LIWC categories: `bio`, `health`, `body`, `percept`, and `see`). Topics relating to `racism` and `police brutality` are also more likely compared to other groups. See Table 4 (b, e, f), Figure 2 (a, d, e), and Figure 3 (a, d, e)

Asian. For the Asian group, the topics `perfectionism` and `comparing to others` are significant stressors. The Asian group also tend to be more concerned with `work`, `school` and `reward` (LIWC categories: `work`, `reward` and `achiev`). See Table 4 (b, c, g), Figure 2 (a, b, c), and Figure 3 (a, b, c).

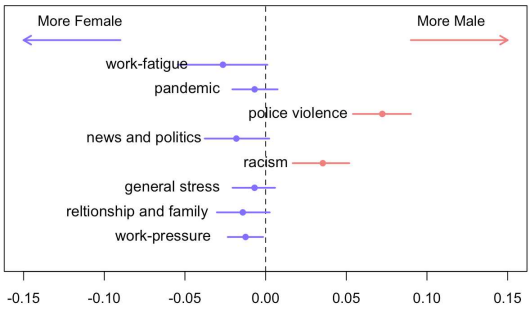

Figure 1: Topic Modeling: topic proportions between race and gender intersectionality – African American women vs. African American men. The bars represent confidence intervals. The closer to the graph extremities, the more prevalent the topics are for the corresponding demographics

Hispanic. The Hispanic group has more stressors related to `immigration`. Other stressors include `family` and `social interactions`, while other prominent topics include `finances`, `news`, `work`. See Table 4 (d, f, g), Figure 2 (b, e, f), and Figure 3 (b, e, f).

White. In the White group, the most prominent stressors are: `general stress`, `news and social media`, `relationships` and `uncertainty`. See Table 4 (c, d, e), Figure 2 (c, d, f), and Figure 3 (c, d, f).

Race and Gender Intersectionality. We also analyze the intersectionality of race and gender in HEADROOM, and provide an excerpt of the experiment to demonstrate how it could be used to study the data further. Analyzing all possible demographic combinations would be too expansive, hence we only provide an excerpt to demonstrate its use case. Focusing on only one demographic category, such as race or gender, can overlook the fine-grained inequalities in demographic groups. Field et al. (2021) give the example that only looking at African American Group emphasizes the more gender-privileged group (African American men), and similarly, only looking at gender may lead to over-representing the race-privileged group (White women). We fit an STM model on the race-gender metadata to find stressor patterns comparing **African American women** to **African American men**. In Figure 1, we show that `police violence` and `racism` are more prevalent stressors for African American men. African American women are more concerned about `work`, `pandemic`, `news and politics`, `general stress`, and `relationship and family`.

In future work, if we can access a demographically labeled depression dataset, we can compare the intersectional stressor patterns to real-life data,

as the data from Aguirre et al. (2022) does not include analyses on intersectionality.

RQ2. How does synthetic data about depression stressors compare to human-generated data across demographics? We compare the analysis findings from our synthetic dataset with the findings from UMD-ODH (Aguirre et al., 2022): (1) between the *keywords* extracted from the aggregated data, not split by demographics,[6] and (2) between the stressors obtained for each demographic group. When comparing the stressors across demographics, we also compare them with other findings related to stressor patterns (McKnight-Eily et al., 2021; Aguirre et al., 2022; Loveys et al., 2018). Our analysis results as depicted in Figure 2, and Table 4 show that the most prevalent depression stressors across demographics are comparable between the human-generated and the synthetic datasets. At the same time, GPT-3 also identifies other stressors not present in the real-life data, as shown in Figure 3.

Topic Similarity. From Aguirre et al. (2022), we obtain the top 30 most prevalent keywords for each topic from UMD-ODH. We then compare them with the keywords from our topics to measure how closely they match each other.

For each topic pair, we convert all keywords in each topic into word embeddings using GloVe (Pennington et al., 2014).[7] We then average the embeddings within each topic, and calculate the cosine similarity of the averaged embeddings. Topics with one-to-many matches (e.g., `work1/ work-fatigue` and `work2/ work-pressure`) are consolidated into one topic. Table 3 shows the cosine similarity scores between each topic pair.

Gender. Comparing gender-related stress patterns, the findings in Aguirre et al. (2022) show that stressors related to `finances`, `relationships`, and `health` are more prevalent for women than men. In contrast, stressors about `social interactions` (LIWC categories: `home`, `leisure`, `social`, `affiliation`, `we` and `family`) are more prevalent in men.

Our findings largely support this pattern and show that the prevalence of `health` and `relationships` stressors are more dominant in women.

However, different from Aguirre et al. (2022), we find that in HEADROOM, stressors related to `finance` are more dominant in men than in women (LIWC category: `money`). See Table 4 (a), Figure 2 (g), and Figure 3 (g).

African American Group. Real-life depression patterns indicate that African Americans are more likely to discuss `health` and use more `social` terms compared to other groups (Loveys et al., 2018). Our synthetic data also supports this: The stressor `health` is more prevalent for African American groups than for Asian, White, or Hispanic groups (LIWC categories: `health`, `body`, and `bio`).

In our synthetic data, we also find that stressors for `police brutality`, `police violence`, and `racism` are more prominent in African American groups despite these topics not being present in Aguirre et al. (2022). However, Alang et al. (2021) showed that hostile police encounters significantly affect African American and Hispanic individuals and are associated with depressed mood and anxiety. See Table 4 (b, e, f), Figure 2 (a, d, e), and Figure 3 (a, d, e).

Asian Group. In our synthetic data, the Asian Group demonstrates a more substantial prevalence of `school` stressors, matching prior findings from Aguirre et al. (2021); Loveys et al. (2018). Surprisingly, topics relating to the impacts of COVID-19 are more commonly associated with the Asian group, despite prior findings showing that Hispanic groups were severely affected by it due to lack of housing and basic needs (McKnight-Eily et al., 2021). See Table 4 (b, c, g), and Figure 2 (a, b, c).

Hispanic Group. Aguirre et al. (2022); McKnight-Eily et al. (2021) showed that stressors `education`, `finance`, `government`, and `family` are prevalent in the Hispanic group. These findings align with ours: `finance`, `school`, `family`, and `politics` are more prevalent for the Hispanic group than other groups

However, Loveys et al. (2018) found that the Hispanic group tends to make fewer mentions of social terms than African Americans, which contradicts our findings. We find that LIWC categories `social` and `affiliation` are more common in Hispanic groups. See Table 4 (d, f, g), Figure 2 (b, e, f), and Figure 3 (b, e, f).

White Group. Similar to Aguirre et al. (2022) and McKnight-Eily et al. (2021), we find that for the White Group, the `finance` stressor is less prevalent than in other racial groups (Figure 2 (c, d, f)).

Different from previous findings (McKnight-Eily et al., 2021), when comparing White and African American groups, we find that `family` stressors are more prevalent in the African American group (see Figure 2 (d)).

[6] We compute over the aggregated data as we could not obtain keywords split by demographic from Aguirre et al. (2022).

[7] We use `glove.6B.300d` from `https://nlp.stanford.edu/projects/glove/`

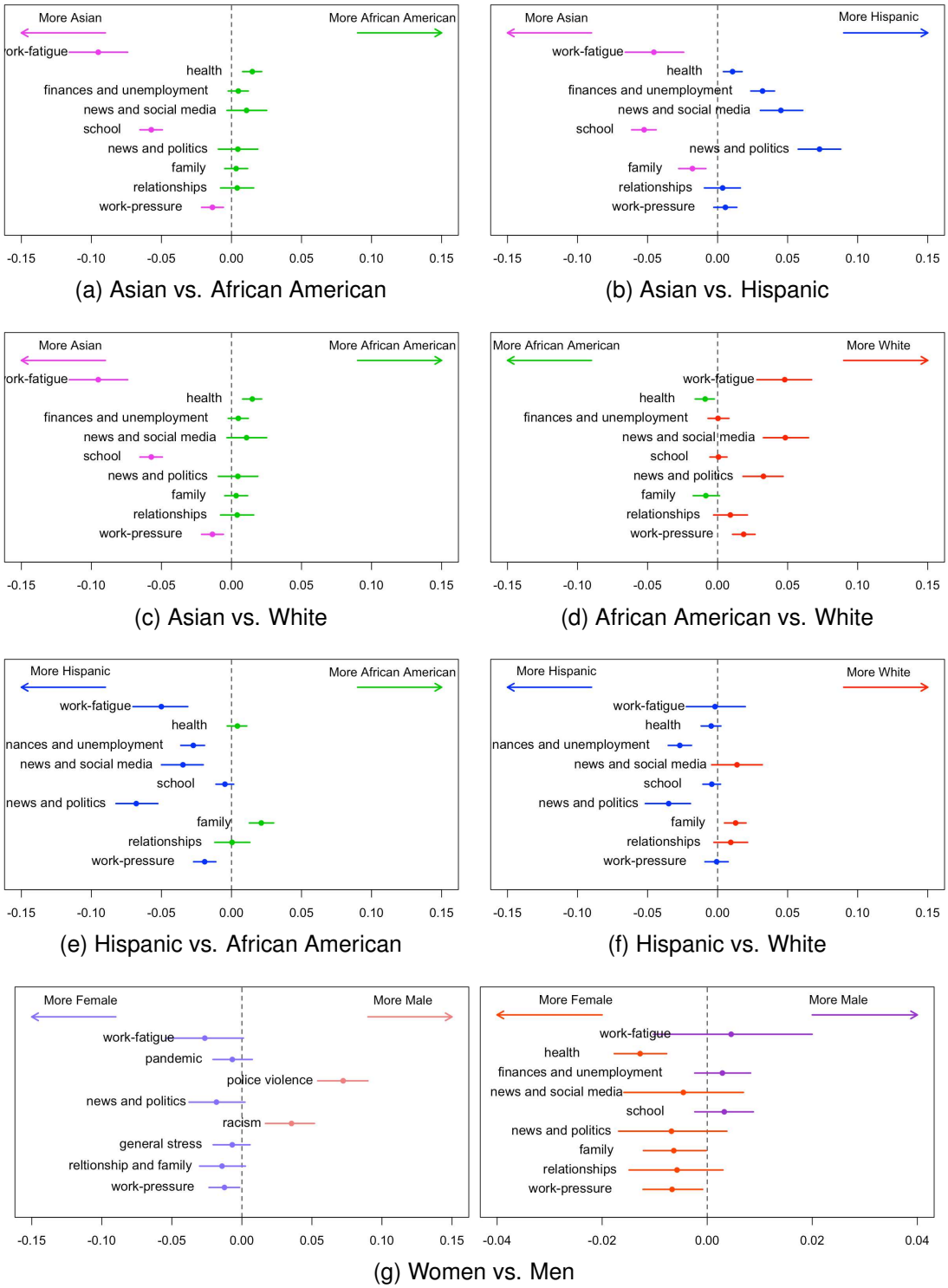

Figure 2: Topic Modeling: topic proportion between different demographics, as detected in GPT-generated data and in real-life data. Colors represent different races and genders: Men – purple, Women – orange, Asian – magenta, African American – green, Hispanic – blue, and White – red. The bars represent confidence intervals. The closer to the graph extremities, the more prevalent the topics for the corresponding demographics. For example, graph (a) Asian vs. African American shows that stressors such as `work1/ work-fatigue`, `work2/ work-pressure` and `school` are more prevalent for Asian than for African American. *Best viewed in color.*

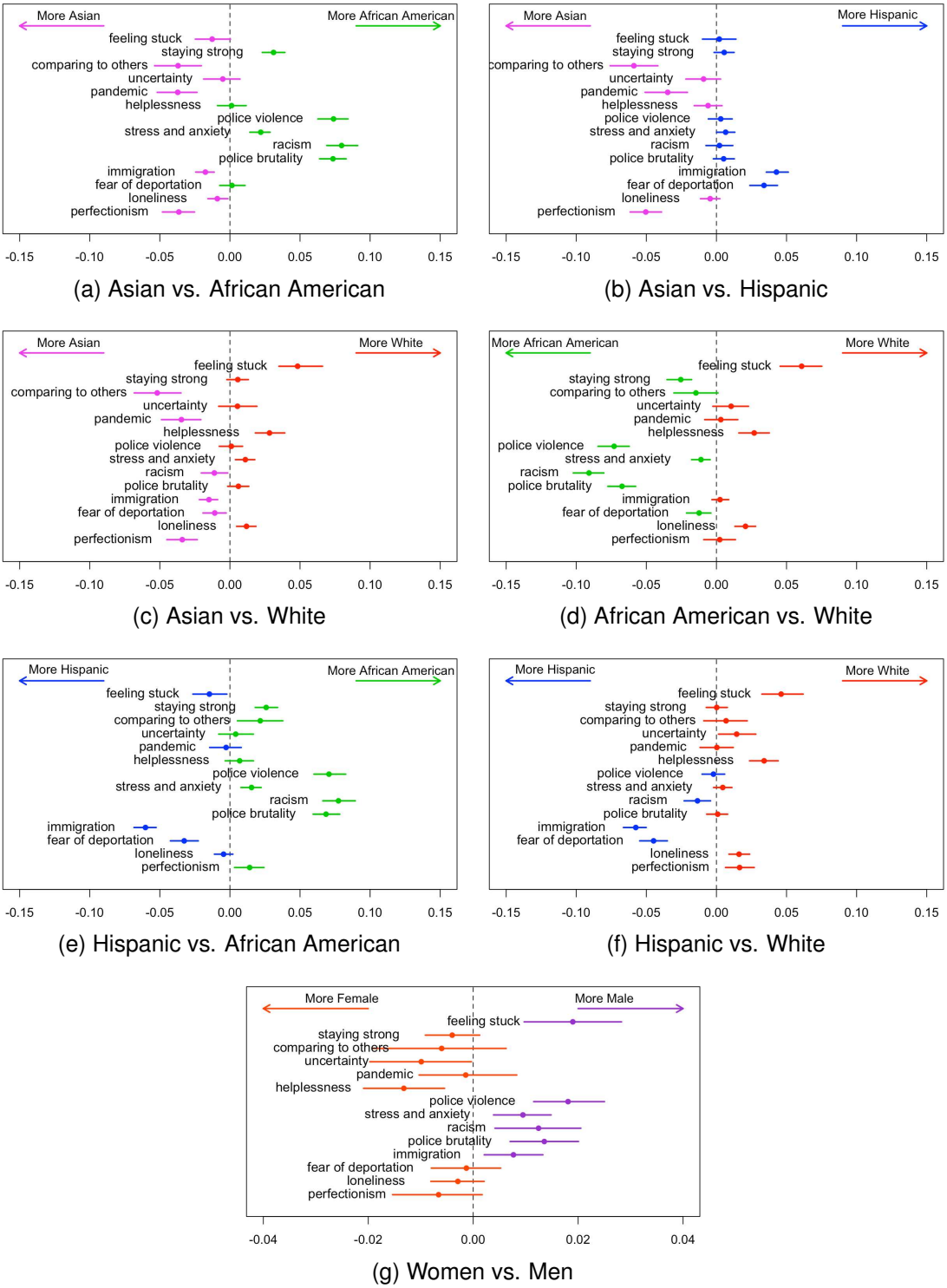

Figure 3: Topic Modeling: topic proportion between different demographics, as detected in GPT-generated data and *not* in real-life data. Colors represent different races and genders: Men – purple, Women – orange, Asian – magenta, African American – green, Hispanic – blue, and White – red. The bars represent confidence intervals. The closer to the graph extremities, the more prevalent the topics for the corresponding demographics. *Best viewed in color.*

6. Conclusion

In this paper, we developed a procedure to produce depression data using GPT-3, which could be applied to other LLMs to test their capability for creating synthetic mental health data. We perform semantic and lexical analyses on this dataset to understand how GPT-3 represents depression stressors across demographics. Our findings show the differences in the types of depression stressors GPT-3 attributes to different demographics, and that some prominent stressors across demographics are similar to those in real-life data from UMD-ODH. Our synthetic data and code is for research purposes only and is made available at `https://github.com/MichiganNLP/depression_synthetic_data`.

7. Acknowledgments

We thank the anonymous reviewers for their constructive feedback. This project was partially funded by a National Science Foundation award (#2306372). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Philip Resnnik and Carlos Aguirre for providing us with their topic model generation code as well as the data.

Ethical Statement and Limitations

7.1. Ethical Statement

We clarify that the intent of our research nor our dataset is not a proxy for creating mental health datasets. We see our paper as a way to discover the biases that LLMs have for different demographics and compare them with available human data. We do not believe this data should be used for supplementing current human data because it can enforce biases. Instead, we propose to use our data for research, to investigate the biases of current LLMs in mental health and how they compare to human data. Our dataset is created using only GPT-3, and according to the IRB of our institution, does not classify as a human subjects research. It is also difficult to explain why GPT-3, or LLMs in general, speculate these stressors, and lack of explainability of the outputs should be considered when following our methods. However, despite the lack of explainability, the recent evolution in the quality of LLM output is driving researchers to consider its application in synthetic data generation such as hate-speech data(Møller et al., 2023). Creating procedures to analyze the algorithmic fidelity of these datasets encourage researchers to use LLMs for synthetic data generation with caution by developing a series of methods to understand its underlying behaviors and potential risks.

Limitations

Gender Representation. We are aware that the gender and race categories we explored are exclusionary and do not capture the full spectrum of gender identity, sexuality, race, and ethnicity; our choice in race and gender groups were predominantly decided by the availability of existing, human-generated datasets.

Location Representation. The authors who collected the UMD-ODH dataset did not mention the location of the patients, so we also do not mention it in the GPT-3 prompts (Kelly et al., 2021, 2020). The data it generates is probably not comprehensive of the whole world, and the findings do not represent all cultures.

Sensitive Information. Across all groups, we note that mentions of *suicide* or *self-harm* are not included in our synthetic data. At the same time, they tend to be mentioned in real-life depression texts (Aguirre et al., 2022). This difference may be a result of model restrictions.

Dataset Size. The size of the dataset is based on the UMD-ODH dataset used by (Aguirre et al., 2022) which consisted of 2607 samples; we also keep our synthetic dataset size small while balancing for demographic groups to conduct a fair comparison to their results.

Using Real-life Depression Data. Due to the difficulty of obtaining demographically-labeled depression datasets, we could not conduct fine-grained analyses between our data and human-generated depression data. While we conduct some quantitative analyses based on the topic keywords provided by the authors of Aguirre et al. (2022), having access to a human-generated dataset would have allowed us to obtain more detailed observations. We also produced relatively short samples, and we do not know whether these stressor patterns hold for longer text sequences.

Model Variability. At the moment, OpenAI does not mention updating `text-davinci-003`; however, we do not know if this will remain true. Possible changes to the model may alter our findings. Additionally, the model is only trained with data up to June 2021, and cannot predict relevant stressors beyond that time frame. The prompts used here are flexible in that the time context can be replaced easily to target a specific time frame, and can be used with other LLMs that has similar capabilities. One could use another LLM model using our prompt to explore its potential to be applied in depression analysis.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Carlos Aguirre, Mark Dredze, and Philip Resnik. 2022. Using Open-Ended Stressor Responses to Predict Depressive Symptoms across Demographics. ArXiv:2211.07932 [cs].

Carlos Aguirre, Keith Harrigian, and Mark Dredze. 2021. Gender and Racial Fairness in Depression Research using Social Media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2932–2949, Online. Association for Computational Linguistics.

Sirry Alang, Donna McAlpine, and Malcolm McClain. 2021. Police Encounters as Stressors: Associations with Depression and Anxiety across Race. *Socius*, 7:2378023121998128. Publisher: SAGE Publications.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2022. Out of One, Many: Using Language Models to Simulate Human Samples. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 819–862. ArXiv:2209.06899 [cs].

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multi-Task Learning for Mental Health using Social Media Text. ArXiv:1712.03538 [cs].

Debra J. Brody, Laura A. Pratt, and Jeffery P. Hughes. 2018. Prevalence of depression among adults aged 20 and over: United states, 2013-2016. *NCHS data brief*, (303):1–8.

Jarrod B. Call and Kevin Shafer. 2018. Gendered manifestations of depression and help seeking among men. *American Journal of Men's Health*, 12(1):41–51.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Joseph L. Fleiss. 1973. Statistical methods for rates and proportions.

Deanna L. Kelly, Max Spaderna, Vedrana Hodzic, Glen Coppersmith, Shuo Chen, and Philip Resnik. 2021. Can language use in social media help in the treatment of severe mental illness? *Current research in psychiatry*, 1(1):1–4.

Deanna L. Kelly, Max Spaderna, Vedrana Hodzic, Suraj Nair, Christopher Kitchen, Anne E. Werkheiser, Megan M. Powell, Fang Liu, Glen Coppersmith, Shuo Chen, and Philip Resnik. 2020. Blinded Clinical Ratings of Social Media Data are Correlated with In-Person Clinical Ratings in Participants Diagnosed with Either Depression, Schizophrenia, or Healthy Controls. *Psychiatry Research*, 294:113496.

Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F. Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. 2017. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics*, 73:14–29.

Inna Lin, Lucille Njoo, Anjalie Field, Ashish Sharma, Katharina Reinecke, Tim Althoff, and Yulia Tsvetkov. 2022. Gendered Mental Health Stigma in Masked Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2152–2170, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 78–87, New Orleans, LA. Association for Computational Linguistics.

Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf, and Mrinmaya Sachan. 2022. Differentially private language models for secure data sharing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4873, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lela R. McKnight-Eily, Catherine A. Okoro, Tara W. Strine, Jorge Verlenden, NaTasha D. Hollis, Rashid Njai, Elizabeth W. Mitchell, Amy Board, Richard Puddy, and Craig Thomas. 2021. Racial and Ethnic Disparities in the Prevalence of Stress and Worry, Mental Health Conditions, and Increased Substance Use Among Adults During the COVID-19 Pandemic — United States, April and May 2020. *Morbidity and Mortality Weekly Report*, 70(5):162–166.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2017. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis*, 16(4):372–403. Publisher: Cambridge University Press.

Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2023. Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks. ArXiv:2304.13861 [physics].

James W. Pennebaker, Roger John Booth, and Martha E. Francis. 2007. Linguistic inquiry and word count (LIWC2007).

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate G. Blackburn. 2015. The development and psychometric properties of LIWC2015.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Marilyn K. Potts, M. Audrey Burnam, and Kenneth B. Wells. 1991. Gender differences in depression detection: A comparison of clinician diagnosis and standardized assessment. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, 3:609–615.

Yiqiu Shen, Laura Heacock, Jonathan Elias, Keith D. Hentel, Beatriu Reig, George Shih, and Linda Moy. 2023. ChatGPT and Other Large Language Models Are Double-edged Swords. *Radiology*, 307(2):e230163. Publisher: Radiological Society of North America.

Susan E. Stockdale, Isabel T. Lagomasino, Juned Siddique, Thomas McGuire, and Jeanne Miranda. 2008. Racial and ethnic disparities in detection and treatment of depression and anxiety among psychiatric and primary health care visits, 1995-2005. *Medical Care*, 46(7):668–677.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of LLMs help clinical text mining? ArXiv:2303.04360 [cs].

Zijian Wang and David Jurgens. 2018. It's going to be okay: Measuring Access to Support in Online Communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 33–45, Brussels, Belgium. Association for Computational Linguistics.

Zach Wood-Doughty, Paiheng Xu, Xiao Liu, and Mark Dredze. 2021. Using Noisy Self-Reports to Predict Twitter User Demographics. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 123–137, Online. Association for Computational Linguistics.

A. Appendix

Overarching Topics	Fine-grained Topic
Work	work-fatigue (work 1) work-pressure (work 2)
Racism/ police brutality	Fear of police and violence Racism Police brutality
General stress	Feeling stuck Staying strong Uncertainty Comparing to others Helplessness Stress and anxiety Loneliness Perfectionism
Immigration status	Fear of deportation Life as an immigrant
News	News and social media Politics and economy
Finances	Finances and unemployment
Pandemic	Pandemic
Family	Family
Relationships	Relationships
Health	Health
School	School

Table 5: All topics from our synthetic data. The overarching topics that also match the topics in the UMD-ODH data are highlighted in **bold**.

Gender			
Women(+)		Men(-)	
category	ratio	category	ratio
female	3.95	**male**	-4.06
adverb	3.16	**see**	-2.45
i	2.83	we	-1.88
pro1	2.56	verb	-1.76
feel	1.81	ipron	-1.70
anx	1.52	auxverb	-1.35
ppron	1.32	tentat	-1.21
affect	1.27	body	-1.21
posemo	1.26	article	-1.05
home	1.10	money	-1.02
insight	1.06	interrog	-1.01
leisure	1.04	health	-0.95
sad	1.02	compare	-0.89
friend	1.00	focuspast	-0.86
conj	0.99	discrep	-0.84

Table 6: Lexical analysis of our data between Women and Men. LIWC categories that also matches the topics in the UMD-ODH data are highlighted in **bold**

Ethnicity			
Asian(+)		White	
category	ratio	category	ratio
leisure	2.73	**anx**	-2.60
family	2.62	adverb	-2.57
certain	2.60	see	-2.54
pro1	2.57	motion	-1.84
home	2.50	**negemo**	-1.55
work	2.24	focusfuture	-1.49
i	2.20	interrog	-1.44
reward	2.14	tentat	-1.37
achiev	1.62	ingest	-1.32
drives	1.59	insight	-1.16
posemo	1.47	space	-1.02
negate	1.34	relativ	-0.99
auxverb	1.14	anger	-0.98
focuspast	1.09	percept	-0.92
nonflu	1.08	adj	-0.86

Table 7: Lexical analysis of our data between Asian and White group. LIWC categories that also matches the topics in the UMD-ODH data are highlighted in **bold**

Hispanic(+)		White(-)	
category	ratio	category	ratio
home	7.48	insight	-3.36
leisure	7.37	percept	-3.17
family	7.18	cogproc	-2.95
affiliation	5.30	see	-2.62
we	4.36	feel	-2.62
drives	2.63	tentat	-2.39
social	2.51	compare	-1.83
focuspast	2.28	differ	-1.74
money	2.25	ipron	-1.48
anx	2.01	**health**	-1.28
achiev	1.75	**bio**	-1.23
auxverb	1.53	space	-1.22
cause	1.44	power	-1.12
number	1.30	prep	-1.10
pro1	1.07	negate	-1.05

Table 8: Lexical analysis of our data between Hispanic and White group. LIWC categories that also matches the topics in the UMD-ODH data are highlighted in **bold**

African American(+)		White(-)	
category	ratio	category	ratio
see	5.88	insight	-2.28
bio	4.01	adverb	-2.25
percept	3.78	tentat	-2.07
certain	3.74	nonflu	-2.04
we	3.68	work	-1.81
number	3.44	informal	-1.81
health	3.40	you	-1.80
body	2.78	space	-1.78
time	2.18	ppron	-1.67
adj	1.86	power	-1.55
feel	1.77	differ	-1.45
compare	1.75	cogproc	-1.41
prep	1.66	shehe	-1.26
affiliation	1.66	discrep	-1.25
money	1.62	focusfuture	-1.18

Table 9: Lexical analysis of our data between African American and White group. LIWC categories that also matches the topics in the UMD-ODH data are highlighted in **bold**

Hispanic(+)		African American(-)	
category	ratio	category	ratio
home	7.33	see	-8.49
family	6.79	percept	-6.96
leisure	6.72	bio	-5.24
affiliation	3.67	health	-4.68
social	3.51	feel	-4.40
anx	2.88	compare	-3.59
focuspast	2.81	certain	-3.36
ppron	2.47	**prep**	-2.77
work	2.05	body	-2.74
you	1.85	adj	-2.56
drives	1.76	**number**	-2.14
focusfuture	1.74	ipron	-2.01
achiev	1.66	cogproc	-1.56
nonflu	1.58	time	-1.50
informal	1.47	quant	-1.17

Table 10: Lexical analysis of our data between Hispanic and African American group. LIWC categories that also matches the topics in the UMD-ODH data are highlighted in **bold**

Hispanic(+)		Asian(-)	
category	ratio	category	ratio
affiliation	5.13	cogproc	-3.00
we	5.06	feel	-2.99
home	5.04	i	-2.75
leisure	4.70	reward	-2.60
family	4.61	negate	-2.39
anx	4.61	percept	-2.26
social	2.73	certain	-2.23
negemo	2.44	insight	-2.21
focusfuture	2.05	work	-2.00
adverb	1.69	posemo	-1.94
money	1.57	**compare**	-1.91
motion	1.22	nonflu	-1.53
interrog	1.19	pro1	-1.50
focuspast	1.19	power	-1.34
verb	1.14	differ	-1.09

Table 11: Lexical analysis of our data between Hispanic and Asian group. LIWC categories that also matches the topics in the UMD-ODH data are highlighted in **bold**

Asian (+)		African American (-)	
category	ratio	category	ratio
work	4.06	see	-8.41
i	3.17	bio	-4.79
nonflu	3.11	percept	-4.71
ppron	2.64	we	-4.38
home	2.31	health	-3.89
family	2.19	body	-3.16
leisure	2.03	**number**	-2.90
power	1.76	adj	-2.73
negate	1.74	**prep**	-2.14
informal	1.66	risk	-1.93
focuspast	1.62	article	-1.91
reward	1.57	anx	-1.76
achiev	1.53	compare	-1.68
pronoun	1.48	ingest	-1.52
pro1	1.46	affiliation	-1.48

Table 12: Lexical Analysis on our synthetic data: Log-odds-ratio of LIWC categories between Asians and African Americans. LIWC categories that also matches the topics in the UMD-ODH data are highlighted in **bold**