Studying Racial Heterogeneity in Language Markers of Depression

Keywords: Mental Health, Depression, Racial Heterogeneity, Language Markers

Extended Abstract

1. Introduction

The manifestation of mental health disorders, including depression, can vary depending on one's race. As a result, mental health models trained on Twitter data having a majority of white population, tend to underperform on persons of color (Aguirre et al. 2021). Moreover, first person Pronouns (*I, we, us*), widely associated with depression, do not indicate depression for this group. Aguirre and Dredze (2021) ascribed the mislabeling to depression class to temporal topics (*TV Award shows*) and demographic specific topics (*AAVE*[1]). In addition to the concerns around the "one size fits all" approach to mental health classifiers, there are psychologically interesting questions such as the trend of usage for linguistic expressions indicating depression that remain unanswered. In this abstract, we investigate the heterogenous language of depression across a cohort of Black and White Americans who consented to share access to their Facebook timeline, self-reported their demographic details, and responded to the PHQ9 inventory.

2. Methods

2.1 Identifying linguistic markers associated with depression and race: Using Differential language analysis Toolkit[2], we extracted linguistic markers (LIWC features and open-vocabulary topics using LDA) from Facebook updates posted by 2121 individuals (20.6% Black). We performed regression to identify the linguistic markers significantly associated with individuals' PHQ9 score after controlling for age and gender. The p-values are corrected using Benjamini-Hochberg correction procedure for multiple hypothesis testing. We repeat the regression with race as an interaction variable to analyse the change, if any, in linguistic markers' association with depression. The interaction coefficient provides the additional change in slope of any given linguistic marker over PHQ-9 x race. We plot interaction plots for each linguistic marker having significant interaction with race to examine the usage trend of language markers with varying severity of depression namely *none=(PHQ score of 0-4), mild=(5-9), moderate=(10-14), moderate-severe=(15-19)* and *severe=(20-27)*.

2.2 Evaluating differential predictive performance for PHQ-9 across racial groups: After identifying language markers differentially associated with depression between Black and White individuals, we then study the difference in the predictive utility of language as a whole at estimating PHQ9 scores between the two racial groups. To identify a matched sample of Black and White individuals, we first performed coarsened exact matching (CEM) using Matchit package[3] in R to remove the effect of confounders. For every sample in the minority class in our dataset (that is, Black individuals), CEM identifies a statistical twin having similar age and gender distribution. We obtained a set of 438 Black individuals paired with demographically similar 438 White individuals. We randomly sample 10% of samples from Black and White individuals from the matched set separately which serve as the unseen test set for performance evaluation. The remaining samples were used to train ridge

[1] AAVE - https://en.wikipedia.org/wiki/African-American_Vernacular_English
[2] DLATK: https://dlatk.wwbp.org/tutorials.html
[3] MatchIt: https://cran.r-project.org/web/packages/MatchIt/vignettes/MatchIt.html

regression models (alpha=10000) to predict PHQ-9 scores in three different training setups: (a) Training on White samples only (M_{white}) (b) Training on Black samples only (M_{black}) and c) Training on White and Black samples ($M_{white+black}$). All three trained models were tested on White Test set (T_{white}), Black Test set (T_{black}) and combined test set ($T_{black+white}$) formed by merging individuals from the White and the Black test sets.

3. Results

3.1 Linguistic markers associated with depression and race: The top-3 LIWC features positively associated with PHQ9 are negative tone, mental and cognition whereas the top-3 LIWC features associated with not being depressed are leisure, positive emotions and time (see Table 1). On considering interaction with race, there were only three LIWC categories namely I, Pronouns, and Personal Pronouns demonstrating significant interaction with race. The regression coefficients for I and Personal Pronouns increased by 23% due to interaction whereas Pronouns increased by over 16%. These features grow linearly with increasing severity of depression for White individuals but not for Black individuals (see Figure 1). The top-3 open vocabulary topics are on *negative feelings, confusion* and *worry* whereas the top-3 negatively associated significant topics are related to *get-togethers*, *fun* and *weekends*. A total of 26 topics having significant correlation with PHQ-9 were found to have significant interaction with race (See Table 2). The top-3 open-vocabulary topics with highest interaction discussed *physical actions* and *guilt* (See Figure 2). We manually identify five themes for a few FB topics with their usage trend for both races as shown in Figure 3. The identified themes such as *feelings, belongingness* and *event_descriptions* have either an opposite or no effect in the usage pattern of depressed Black individuals compared to depressed White individuals. The use of *exclamations* drastically drops for Black individuals when depressed however, has a growing trend for White individuals. Likewise, the use of *physical descriptives*-(hungry, grumpy, thirsty, lonely) is more prominent amongst white individuals whereas (walk, stare, watch, jump) is more frequently used by Black individuals.

3.2 Predictive performance for PHQ-9 across racial groups: The predictions by M_{white} on T_{white} has the strongest correlation and M_{black} predictions' on T_{black} is strongest, as might be expected (See Table 3). However, while M_{black} performed well on T_{white} and $T_{white+black}$' M_{white} and $M_{white+black}$ had no signal on T_{black}. The poor performance on T_{black} is likely due to the contrasting linguistic usage patterns between both races when depressed.

4. Findings

Significant interaction effects between language usage and race as it relates to depression suggests the presence of racial heterogeneity in language markers associated with mental health. We also discover linguistic expressions and topics that are widely associated with depression, are prominently used by White individuals and have an opposite or no effect for Black individuals. While inventories such as PHQ-9 are tuned to obtain signals from individuals across racial categories, similar efforts are warranted in the Natural Language Processing efforts associated with Computational Psychology.

References

Aguirre, Carlos, and Mark Dredze. 2021. "Qualitative Analysis of Depression Models by Demographics." In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, 169–80. Online: Association for Computational Linguistics.

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.

Tables & Figures

1. LIWC Features

Table 1. LIWC features derived from Facebook Messages having significant ($p<0.05$) interaction with race. p values are corrected using Benjamini-Hochberg correction for multiple comparisons. Under heading “without interaction”, we provide ‘regression coefficient’ without race as interaction variable for comparison.

feature	With race as interaction				without interaction	
	Regression Coefficient	CI (l,u)	interaction coefficient	CI (l,u)	Regression coefficient	CI (l,u)
I	0.174	[0.132, 0.215]	0.065	[0.022,0.107]	0.141	[0.099,0.183]
PRONOUN	0.171	[0.130, 0.212]	0.069	[0.027,0.112]	0.147	[0.105,0.188]
PPRON	0.153	[0.111, 0.195]	0.063	[0.020,0.105]	0.125	[0.083,0.166]

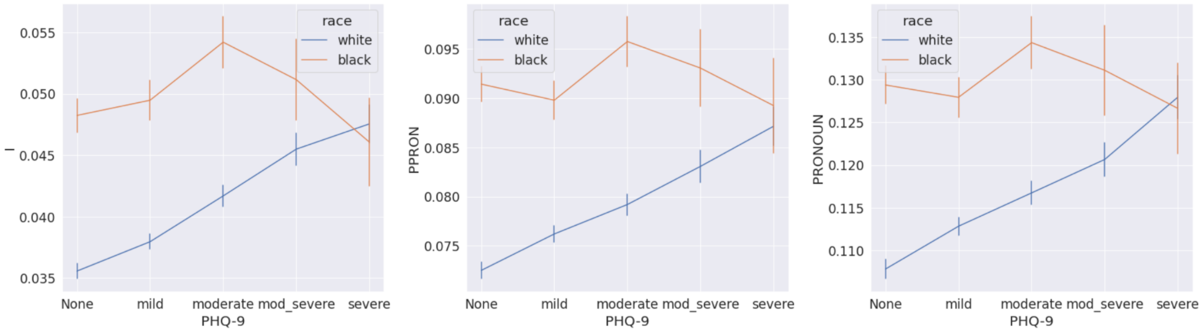

Figure 1. Interaction Plot for LIWC features - [I, PPRON, PRONOUN] found to have significant interaction with race. The use of these LIWC features grows linearly with the severity of depression for white individuals whereas the usage of these features has a downward trend for Black individuals.

2. Open Vocabulary Topics

Table 2. Fb Topics derived using LDA from facebook messages having significant interaction ($p<0.05$) with race. p values are corrected using Benjamini-Hochberg correction for multiple comparisons. Under heading "without interaction", we provide 'regression coefficient' without race as interaction variable for comparison.

feature	term	With race as interaction				without interaction	
		Regression coefficient	CI (l,u)	Interaction coefficient	CI (l,u)	Regression coefficient	CI (l,u)
1848	i'ma, mess, wreck, messing, cursed, confess, that's, whoa, don't, i've	0.179	[0 .138, 0.220]	0.074	[0.032, 0.116]	0.153	[0.111, 0.194]
1945	dreamer, terrified, shy, scared, misunderstood, afraid, pretending, drowning, i'm, aware	0.161	[0 .119, 0.202]	0.069	[0.027, 0.112]	0.148	[0.106, 0.189]
1340	weirdo, creep, liking, curious, ignoring, belong, concerned, i'm, i've, don't	0.153	[0.111, 0.194]	0.075	[0.033, 0.118]	0.137	[0.095, 0.179]
1136	argue, sinner, hopeful, joking, joker, drag, idiot, tempted, lover, practically	0.151	[0.109, 0.192]	0.075	[0.033, 0.117]	0.138	[0.096, 0.179]
1548	hearing, sick, tired, assuming, fed, tire, surface, i'm, feeling, numb	0.146	[0.104, 0.187]	0.070	[0.027, 0.112]	0.136	[0.094, 0.177]
1251	balcony, tearing, stopping, i'ma, doubt, sooooooooooooooooooooooo, i'm, gonna, set, causing	0.145	[0.103, 0.186]	0.077	[0.034, 0.119]	0.128	[0.086, 0.169]
1935	punch, sucker, slap, smiley, punching, face, punched, pinch, fist, stuffing	0.142	[0.100, 0.184]	0.061	[0.019, 0.103]	0.078	[0.035, 0.120]
1768	guilty, accomplished, conscience, pleasure, betrayed, pleasures, roosevelt, feel, guilt, superior	0.139	[0.097, 0.180]	0.080	[0.038, 0.122]	0.133	[0.091, 0.175]
958	-_-, urgh, effin, friggin, gah, >.<, gawd, grr, >_<, dammit	0.138	[0.096, 0.180]	0.068	[0.026, 0.111]	0.056	[0.014, 0.099]
1708	gonna, struck, blow, i'm, figure, twelve, fond, it's, assuming, prince	0.137	[0.095, 0.178]	0.081	[0.039, 0.123]	0.117	[0.075, 0.159]
250	thrilled, faded, i'm, doomed, psyched, i've, yeah, wanting, folks, promising	0.136	[0.094, 0.178]	0.072	[0.029, 0.114]	0.123	[0.080, 0.164]
850	wanna, undo, don't, consuming, brag, touch, surrender, blew, explode, stay	0.136	[0.094, 0.178]	0.066	[0.023, 0.108]	0.097	[0.055, 0.139]
1247	gonna, shotgun, shook, slapped, ain't, i'm, cigarette, load, doll, he's	0.136	[0.094, 0.177]	0.081	[0.039, 0.123]	0.109	[0.066, 0.151]
489	hungry, sooooooooo, tired, grumpy, pretending, exhausted, i'm, thirsty, irritated, stressed	0.133	[0.091, 0.175]	0.074	[0.031, 0.116]	0.117	[0.075, 0.159]
1881	jealous, sayin, insanely, gonna, dread, i'm, yeah, screw, practically, complaining	0.128	[0.086, 0.170]	0.071	[0.029, 0.113]	0.114	[0.072, 0.156]
429	regret, regretting, dwell, undo, importantly, things, risks, accept, regrets, inform	0.126	[0.084, 0.168]	0.064	[0.021, 0.106]	0.095	[0.052, 0.137]

562	gonna, nothings, i'm, nobody's, miss, hadn't, ='(, craziness, generally, gawd	0.122	[0.080, 0.164]	0.076	[0.034, 0.118]	0.101	[0.059, 0.143]
568	nothing's, gonna, puke, how's, dunno, happen, i'm, someday, it'll, goooood	0.121	[0.079, 0.163]	0.077	[0.035, 0.119]	0.100	[0.058, 0.142]
524	welp, dang, guess, messed, sigh, wel, smh, idk, geez, hmm	0.102	[0.060, 0.144]	0.058	[0.016, 0.100]	0.042	-0.001, 0.084]
1625	alright, it'll, it's, i'm, what's, convince, guaranteed, pretending, nostalgia, unbelievable	0.101	[0.059, 0.143]	0.070	[0.028, 0.112]	0.095	[0.053, 0.137]
415	wasn't, wasnt, expected, expecting, bad, thankfully, aware, disappointing, he'd, suppose	0.082	[0.040, 0.124]	0.067	[0.025, 0.110]	0.078	[0.036, 0.120]
1199	walks, replies, asks, pulls, sits, jumps, picks, runs, watches, tells	0.081	[0.039, 0.123]	-0.175	[-0.216, -0.134]	0.005	-0.038, 0.048]
1856	happen, happening, accidents, things, reason, occur, somethings, unexpected, happened, outcome	0.078	[0.036, 0.120]	0.065	[0.023, 0.108]	0.067	[0.025, 0.109]
1236	swine, flu, aids, symptoms, jab, bug, shot, infected, fl, recovered	0.066	[0.024, 0.109]	-0.101	[-0.143, -0.058]	0.044	[0.002, 0.087]
1759	phew, missed, finally, survey, glad, back, =d, figured, sweetie, soooo	-0.084	[-0.127, -0.042]	-0.065	[-0.107, -0.022]	-0.049	[-0.092, -0.007]
1461	finished, assignment, phew, relieved, exams, handed, assignments, whew, completed, submitted	-0.114	[-0.156, -0.072]	-0.084	[-0.126, -0.041]	-0.026	[-0.069, 0.016]

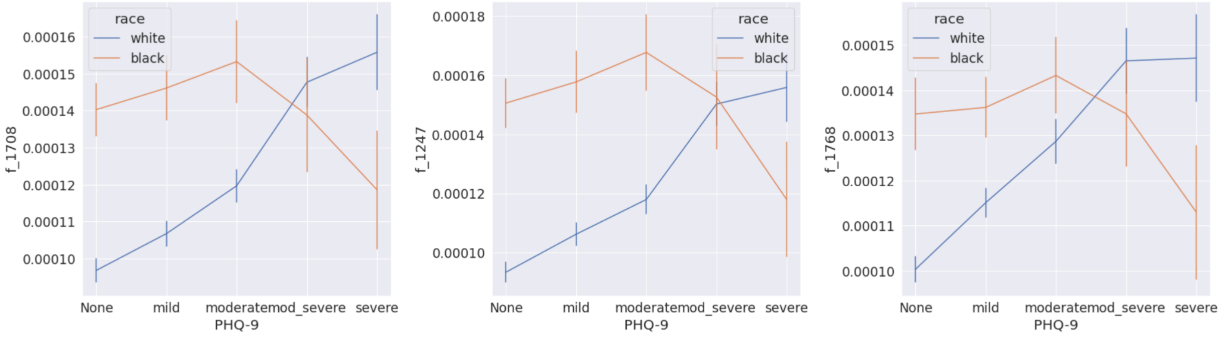

Figure 2. Interaction Plot for Top-3 Facebook Topics with the highest interaction coefficient with race (see Table 2). These topics are positively correlated with depression for White individuals however, the usage of these topics has a downward trend for Black individuals with growing depression. Here, topic words for Topic-*1708*:{gonna, struck, blow, i'm, figure, twelve, fond, it's, assuming, prince}, Topic-*1247*: {gonna, shotgun, shook, slapped, ain't, i'm, cigarette, load, doll, he's}, and Topic-*1768*: {guilty, accomplished, conscience, pleasure, betrayed, pleasures, roosevelt, feel, guilt, superior}.

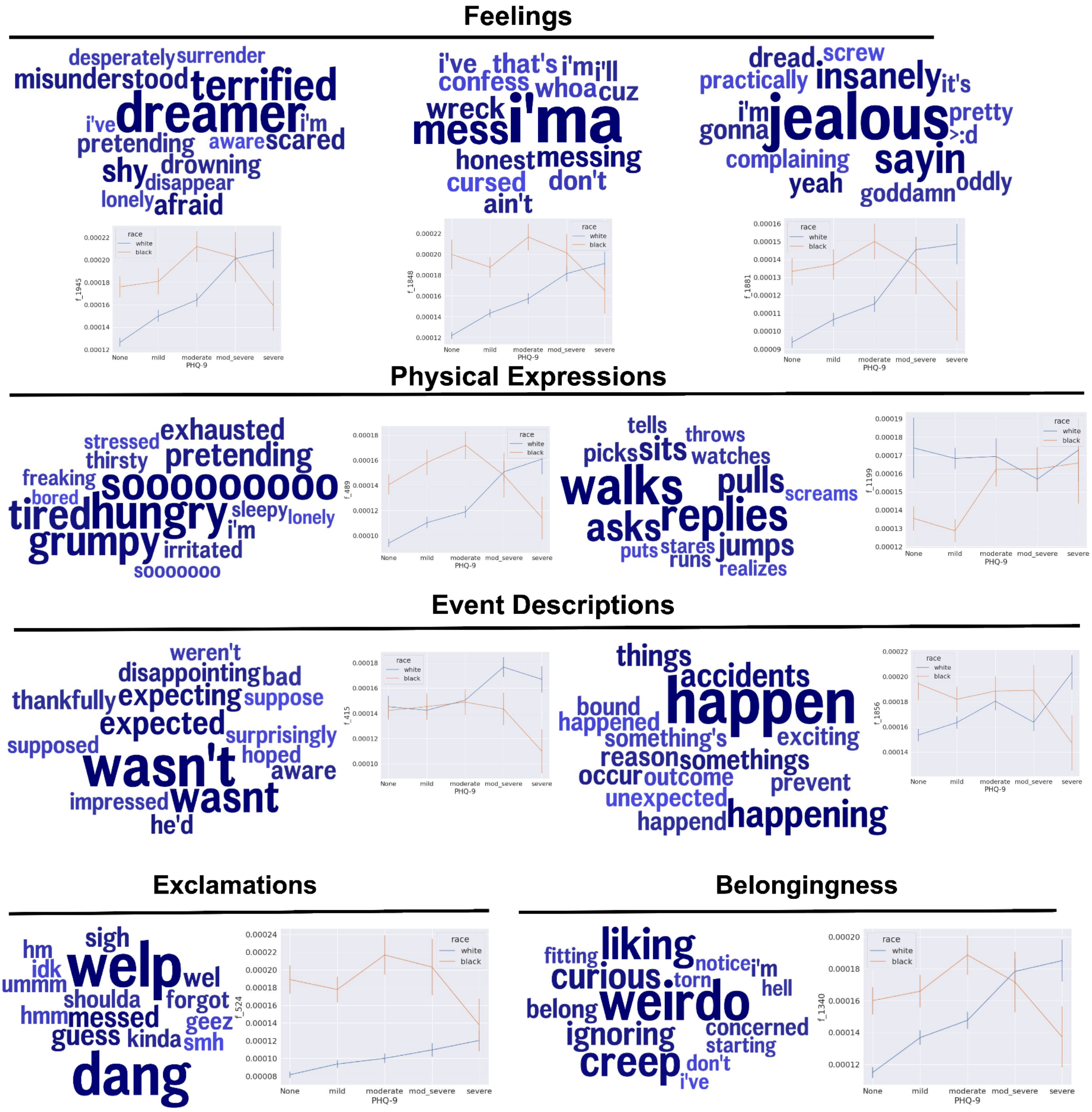

Figure 3. Themes for selected Facebook Topics having significant ($p<0.05$) interaction with race . The interaction plot indicates the trend for topic usage for both races with growing severity of depression (Blue color indicates the trend for White race and the orange color is for Black race.).

3. Prediction Model

Table 3. Model Performance. $M_{\$train-set\$}$ indicates the model trained on the \$train-set\$. Each model is tested on three test sets $T_{\$test-set\$}$ as described in Section 2.2. We report Pearson r value. Higher the Pearson r, the stronger is the correlation between predictions and original values. Pearson r with p value >0.05 is considered insignificant and is highlighted in red. The best performance on each test set is marked in bold. The model M_{white} performs the best on T_{white} whereas M_{black} gives the best performance on T_{black} and $T_{white+black}$.

	Pearson r (p value)		
$M_{\$train-set\$}$ / $T_{\$test-set\$}$	T_{white}	T_{black}	$T_{white+black}$
M_{white}	**0.467 (0.0013)**	0.0193 (0.90)	0.2514 (0.018)
M_{black}	0.334 (0.02)	**0.514 (0.0003)**	**0.425 (0.00004)**
$M_{white+black}$	0.44 (0.002)	-0.0775 (0.617)	0.21 (0.045)