Detecting Linguistic Traces of Depression in Topic-Restricted Text: Attending to Self-Stigmatized Depression with NLP

JT Wolohan
Department of Information and Library Science
Indiana University - Bloomington
jwolohan@indiana.edu

Misato Hiraga
Department of Linguistics
Indiana University - Bloomington
mhiraga@indiana.edu

Atreyee Mukherjee
Department of Computer Science
Indiana University - Bloomington
atremukh@indiana.edu

Zeeshan Ali Sayyed
Department of Computer Science
Indiana University - Bloomington
zasayyed@indiana.edu

Abstract

Natural language processing researchers have proven the ability of machine learning approaches to detect depression-related cues from language; however, to date, these efforts have primarily assumed it was acceptable to leave depression-related texts in the data. Our concerns with this are twofold: first, that the models may be overfitting on depression-related signals, which may not be present in all depressed users (only those who talk about depression on social media); and second, that these models would under-perform for users who are sensitive to the public stigma of depression. This study demonstrates the validity to those concerns. We construct a novel corpus of texts from 12,106 Reddit users and perform lexical and predictive analyses under two conditions: one where all text produced by the users is included and one where the depression-related posts are withheld. We find significant differences in the language used by depressed users under the two conditions as well as a difference in the ability of machine learning algorithms to correctly detect depression. However, despite the lexical differences and reduced classification performance–each of which suggests that users may be able to fool algorithms by avoiding direct discussion of depression–a still respectable overall performance suggests lexical models are reasonably robust and well suited for a role in a diagnostic or monitoring capacity.

1 Introduction

Major depressive disorder is a serious illness that afflicts more than 1-in-15 Americans and more than 1-in-10 American young adults[1]. Depression is also the number one cause of suicide–the second leading cause of death among adolescents–and a difficult disease to treat, because those suffering from it are often reluctant to report. In part, this is true because depression is a highly stigmatized disease. Not only is stigma a significant contributor to the suffering of both clinically and subclinically depressed individuals, depression stigma is associated with lower rates of help seeking and higher rates of avoidance (Manos et al., 2009). This results in a population that may be motivated to hide or otherwise disguise their depression symptoms.

This paper examines whether a machine learning approach based on linguistic features can be used to detect depression in Reddit users when they are not talking about depression, as would be the case with those wary of depression stigma. We split this effort across two datasets: the first, we allow all the Reddit posts from a sample of 12,106 users, about half of whom are depressed, and in the second, we allow only those posts which were not directly discussing depression. With this second dataset, we intend to approximate the activity of users reluctant to discuss depression online or attempting to hide their depression.

[1] https://www.nimh.nih.gov/health/statistics/major-depression.shtml

Proceedings of the First International Workshop on Language Cognition and Computational Models, pages 11–21
Santa Fe, New Mexico, United States, August 20, 2018.
https://doi.org/10.18653/v1/P17

On each dataset we perform two sets of analysis: a lexical analysis–using LIWC (Pennebaker et al., 2015) and Term-Frequency/Inverse-Document Frequency (TF-IDF) weights–and a classification task–using a number of Support Vector Machine classifiers trained on lexical features. The first analysis reveals differences between the text produced by depressed users when the corpus is allowed to include depression-related text and when depression-related text is withheld. The second analysis reveals that the classification task is more difficult when depression-related text is withheld; however, machine learning classifiers are still able to detect linguistic traces of depression.

Our contributions with this paper are threefold. First we demonstrate the impact and potential importance of removing mental-health topics from a corpus before training natural language processing models; second, we provide attention to the task of detecting stigmatized or otherwise “hidden” depression, which has to date not been looked at by the research community; and third, we find that the linguistic patterns of depressed Reddit users are consistent with popular depression batteries and interventions.

2 Related Work

2.1 Depression detection

Language often reflects how people think, and it has been used in assessing mental health conditions by psychiatrists (Fine, 2006). Recently, computational methods have begun to be employed to study depressed users’ writings and activities on social media. A meta-analysis by Guntuku et al. (2017) summarizes several iterations of the depression detection task, including clinical depression detection (De Choudhury et al., 2013b; Schwartz et al., 2014; Tsugawa et al., 2015; Preoţiuc-Pietro et al., 2015), post-partum depression prediction (De Choudhury et al., 2013a), post-traumatic stress disorder detection (Harman and Dredze, 2014; Preoţiuc-Pietro et al., 2015), and suicidal attempt detection (Coppersmith et al., 2016). For our purposes, it is most important to note how different authors operationalize the depression detection task and what assumptions are included in that approach.

The first such approach, by Coppersmith et al. (2014) (also used by Coppersmith et al. (2015) and Resnik et al. (2015)) , attempts to select a population of users with major depressive disorder by crawling for users’ disclosure of diagnosis. The researchers first scrape a large, broadly relevant assortment of Tweets, before downselecting to only those Tweets which match the regular expression “I was diagnosed with [depression]”. Tweets by the users identified in this way are then scraped to create a gold standard, and a control group of users can be randomly sampled and scraped from the general population.

A second, crowd-sourced-survey approach has also been used effectively (De Choudhury et al., 2013b; Tsugawa et al., 2015). In this approach, the researchers have micro-task workers (e.g., Turkers from Mechanical Turk) take two depression inventories (historically, CES-D (Radloff, 1977) and BDI (Beck et al., 1996)) and provide their social media handle. If the inventory results correlate (both indicating depression or no depression), the authors will scrape the users’ social media data and place them in the depressed group or the control group.

A third, less frequently used, approach is based on community membership or participation. In this approach, users are classified as having a mood disorder–both depression (De Choudhury and De, 2014) and anxiety (Shen and Rudzicz, 2017) have been studied–when they post in a given community (typically a subreddit, as this approach has mostly been used with Reddit-data). This approach has tended more towards descriptive research and past analysis have focused exclusively on content from the identified communities.

Across all three methods, we find a shortcoming: authors largely make no effort to limit the topic of discussion. Given that the gold standards created by the first and third sampling strategies above are constructed by looking for disclosure of diagnosis or at least self-diagnosis, we can assume that these users have a higher probability of discussing depression than a typical, control group user. Algorithms trained upon these samples to predict depression may be cluing in on this topic-proclivity to achieve artificially high results. Further, all three approaches, by not removing explicit discussion of depression from their training data, at the very least can be expected to under perform on an important population: the depressed who are reluctant to speak about their condition. To our knowledge, only three studies have attempted to remedy this and each of those has been computationally (as opposed to psycho-linguistically

	All Subreddits	Depression Withheld	Pct. Change
Users–Depressed	4,947	4,324	-12.6%
Users–Control	7,159	7,153	-0.1%
Users–Total	12,106	11,477	-5.2%
Words–Depressed	55,980,678	48,399,823	-13.5%
Words–Control	93,109,041	92,787,403	-0.3%
Words–Total	149,089,719	141,187,226	-5.3%

Table 1: Dataset Composition by Tasks

oriented) oriented (Yates et al., 2017) or exploratory in nature (Losada and Crestani, 2016; Hiraga, 2017).

2.2 Depression Stigma

One of the reasons we are concerned with previous authors not removing depression-related text from their data is because we are concerned about stigma leading many depressed users to be silent about their depression. Latalova et al. (2014) suggest that stigma-related effects are an important factor preventing depression-related help-seeking among men and that a complex relationship exists between masculinity and depression. Through a narrative review of the research on stigma, they find that masculinity is both a cause of depression and a cause of reduced-help seeking, exemplified by gender norms like "boys don't cry".

Similarly, after having conducted a survey of a random sample (n=5,500+) of college students from 13 American Universities, Eisenberg et al. (2009) suggest that social-norms are a leading cause of perceived public stigma and, in turn, personal stigma. They found that higher self-stigma is associated with lower reported comfort seeking help and that self-stigma was highest among male students, Asian students, young students, poor students and religious students.

In a random sample (n=1,300+) people from the general Australian public, Barney et al. (2006) find this same pattern: higher reported self-stigma scores result in increased hesitation about seeking help for depression. Major sources of this hesitation included personal embarrassment at having depression and the perception that others would respond negatively. This last finding is in contrast to Schomerus et al. (2006), who find that among a sample (n=2,300+) of the German public anticipation of discrimination by others did not prevent help seeking behavior (though again, self-stigma was negatively associated with help seeking).

Our view is that given the consistent findings that self-stigma reduces help-seeking, depression detection efforts using social media and natural language processing have a unique opportunity to reach these individuals. If models can be trained to identify not just the depressed and open about it, but the depressed and hesitant, help could be directed to individuals who would otherwise neglect to seek it. In this study, our aim is to approximate the scenario where the users are hesitant to post about depression.

3 Method

3.1 Data

The data for this analysis are the reddit posts of 12,106 reddit users, totalling 149,089,719 words. The users are divided into two categories: depressed and not-depressed. Of the more than 12,000 users, 4,947 ($\approx$ 40%) are considered depressed and these users account for nearly 56-million words ($\approx$ 38%). The 7,159 ($\approx$ 60%) non-depressed users are responsible for the other 93-million words ($\approx$ 62%).

To gather our depressed users, we used a community participation approach similar to that employed in other Reddit-based research (De Choudhury and De, 2014; Shen and Rudzicz, 2017). We considered a user depressed if they started a thread in Reddit's depression subreddit[2]–which identifies itself as a "a supportive space for anyone struggling with depression."–as a user self-identifying as suffering from depression. On the basis of this heuristic, we scraped the 10,000 most recent post-authors from the

[2] www.reddit.com/r/depression

Depressed		Control	
r/depression_help	r/aww	r/AskReddit	r/news
r/AskReddit	r/Showerthoughts	r/pics	r/gaming
r/depression	r/gaming	r/funny	r/aww
r/pics	r/videos	r/Showerthoughts	r/todayilearned
r/funny	r/todayilearned	r/mildlyinteresting	r/gifs

Table 2: Some of the common subreddits the users participated in

depression subreddit. To construct a control group, we scraped users who had started a thread in Reddit's AskReddit subreddit[3], one of the site's most popular communities with more than 18 million subscribers. We believe AskReddit is a fitting control for the depression community because its question-and-answer format is similar to the information and support seeking of the Depression community, and AskReddit is among the most popular subreddits among depressed users in our sample.

With these two lists of users, we then scraped the entire available post-history of these users. Users from whom we did not collect more than 1,000 words of text were removed from our dataset. By scraping the entirety of our users posts we achieve a diverse range of conversation topics (see Table 3.1), including computer games and internet culture, politics and current events, and more. Most of the discussion sampled ($\approx 96\%$) was unrelated to depression.

Two of the authors validated our heuristic for selecting depressed Reddit users through a systematic, independent review of 150 posts from the front-page of the depression subreddit. The authors agreed on 99% (149/150) of the total classifications and both authors agreed that 147 of the 150 posts indicated at least a self-diagnosis of depression-like symptoms by the authoring user. A 99% confidence interval about this proportion suggests that no less than 92% of users selected by our depressing heuristic are suffering from self-diagnosed depression-like symptoms. We did not attempt to assess the number of depressed users in our control sample; however we would expect the upper-bound on this to be around 1-in-20[4] .

3.2 LIWC Analysis

LIWC, the Linguistic Inquiry and Wordcount Tool, is psychometric analysis software based on the idea that the words a person uses reveal information about their psychological state (Pennebaker et al., 2015). The software has been extensively used in natural language processing tasks for feature-creation, including within the area of mental-illness detection (for more, see Guntuku et al. (2017)). We use LIWC both as a source of features and as part of a stand alone analysis.

For the latter, we estimate the true means of several depression-related indices using 95% T^2 intervals (Hotelling, 1931) for the control and depressed users under our two detection conditions: (1) including all data and (2) withholding depression-related data.

3.3 Classification

With respect to classification, we endeavor to solve two tasks. The first is a benchmark designed to mirror the depression-detection efforts to date. In this task, we use all of the data from the 4,947 depressed users and 7,159 non-depressed users in our dataset. The second task is an expanded version of efforts by Hiraga (2017) which excludes the explicit discussion of depression. We achieve this by witholding posts and comments from 17 subreddits related to depression. We selected subreddits for exclusion by examining subreddits linked from the depression subreddit (e.g., r/SuicideWatch and r/mentalhealth) and snowballing out to other related subreddits. We also examined a list of subreddits frequented by depressed users for those with depression-related names. Limiting our data in this way, our dataset was reduced to only 4,324 depressed users and 7,153 non-depressed users who met our 1,000-word threshold. A comparison of these tasks is shown in Table 1.

[3] www.reddit.com/r/AskReddit

[4] According to the CDC, this is the rate of depression among the general public and AskReddit is a general purpose subreddit.

	All–Dep	Off–Ctrl	Off–Dep
All–Ctrl	950.1*	0.3	460.5*
All–Dep	-	1397.7*	120.4*
Off–Ctrl	-	-	475.7*
*Significant at p<.001			

Table 3: F-values of pairswise two-sample T^2 tests about the LIWC index means

For these tasks, we train two Linear Support Vector Machines (Fan et al., 2008) with TF-IDF weighted combinations of word and character $n$grams and LIWC features. Our **character $n$gram** features include all 2- to 4-grams; our **word $n$gram** features contain unigrams and bigrams; our **LIWC features** contain all the lexical indexes output by LIWC. We use a **smoothed TF-IDF** approach–implemented as $tf(t) \times \log(\frac{N+1}{n_t+1})$–where $tf(t)$ is the number of times the unigram or bigram t occurs, N is the number of documents and n_t is the number of documents containing the unigram or bigram t.

We limit our text prepossessing to sentence segmentation, tokenization, using a simple, social-media aware tokenizer[5], and ignoring case.

4 Results

4.1 LIWC Analysis

The 95% T^2 intervals about the user-level means of select depression-related indices demonstrates a wide-gap between the control users and the depressed users that narrows significantly when depression-related topics are removed from the data. We find significant differences between all group-condition differences, except for the two control groups (control users including depression text and control users with depression text withheld). Table 2 reports the F-values of all pairwise comparisons, with higher numbers indicating a greater difference between the samples.

The intervals about the specific indices reveal that depressed users are less "analytic", with less "clout" and more "authentic" than their control-group counterparts. Further, they use the personal pronoun *I* more, engage in more comparisons, speak with more affect, especially expressing more negative emotion, anxiety and sadness, with a greater emphasis on the present and future. Small to no differences are found between depressed and control users with respect to positive emotion expression (although depressed users may use more), anger, social language, family language, and focus on the past.

Between the depressed users in the all-included condition and the depressed users in the withheld condition, we find that depressed users appear more "analytic" and less "authentic" in the withheld case, with a decreased use of the *I* pronoun, decreased expression of sadness, and a decreased focus on the present. All of these changes make depressed users in the depression withheld condition more similar to control users; however, overall they are still more similar to the depressed users with all data included than to either control group.

4.2 Classification

The results from our two classification tasks in many ways reflect the differences found by the LIWC analysis. Of the four model variants–LIWC scores only, character $n$grams only, word $n$grams only, and the LIWC features plus both sets of $n$gram features–every variant achieved better performance in Task 1, which includes all the data collected, than its counterpart in Task 2. Between the four variants, the LIWC+$n$gram model achieved the best performance (81.8% accuracy in Task 1 and 78.7% accuracy in Task 2).

In the all topic case, as previously noted, we find that the LIWC+$n$gram model performs best. Its accuracy, AUC and F1-score are all better than the second best model, based on word-$n$gram features, that in turn is better than the third best model based on character-$n$gram features. The LIWC-based model performs well, achieving 78.7% accuracy.

[5] We use a modified version of: Christopher Potts' HappierFunTokenizing.

	Task 1: All topics		Task 2: Depression withheld	
	Control	Depression	Control	Depression
Analytic	45.67-48.22	32.79-36.16	45.75-48.30	**36.60-40.10**
Clout	52.07-54.15	43.55-47.04	52.05-54.14	44.64-48.10
Authentic	43.12-46.13	54.76-59.15	43.03-46.04	**49.65-54.15**
I	4.74-5.05	6.31-6.82	4.74-5.04	**5.79-6.29**
Comparisons	2.46-2.55	2.63-2.75	2.46-2.54	2.58-2.71
Affect	6.20-6.50	6.93-7.27	6.20-6.50	6.69-7.05
Pos. Emotions	3.80-4.06	4.05-4.32	3.80-4.06	4.01-4.31
Neg. Emotions	2.30-2.44	2.74-2.94	2.29-2.43	2.54-2.74
Anxiety	0.25-0.28	0.36-0.42	0.25-0.27	0.32-0.37
Anger	0.93-1.03	0.91-1.03	0.93-1.03	0.92-1.05
Sadness	0.37-0.40	0.60-0.68	0.37-0.40	**0.47-0.53**
Social	9.37-9.73	9.42-9.94	9.36-9.72	9.19-9.75
Family	0.34-0.39	0.30-0.36	0.34-0.39	0.29-0.36
Focus:Past	3.60-3.80	3.43-3.67	3.60-3.80	3.50-3.76
Focus:Pres.	11.50-11.82	12.96-13.43	11.49-11.81	**12.31-12.76**
Focus:Fut.	1.17-1.23	1.33-1.43	1.17-1.23	1.25-1.35

Bold text indicates a difference between treatment conditions for depressed users

Table 4: 95% T^2 interval about select LIWC results for groups across treatments

In the depression-topics withheld case, the results are similar. The composite model is the best, with word-$n$grams alone beating character-$n$grams alone and LIWC features performing the worst of all. For this second task, we also tested the best-performing model (the combined-features model) trained on the data from first task. With respect to accuracy, this model out performed all models except its counterpart combined-features model trained on the data from the second task; however, looking more holistically at the measures of performance, underwhelming AUC (73.2%) and an underwhelming F1-score (64.8%) suggest it not be quite as well calibrated as the word-$n$gram feature model.

5 Discussion

We were motivated to do this study by the concern that social media-based approaches to depression detection may be overlooking certain populations of interest, especially those who have high self-stigma. Our analysis reveals that concern to be warranted. Even within the constraints of our study design, which only approximates users who are hiding their depression symptoms, we find that there are significant differences between depressed users when they are talking about depression and depressed users when they are not.

This difference is evident looking at the F-scores presented in Table 2 and the confidence intervals in Table 4. Table 2 indicates large gaps between control and depressed users in both cases: all data permitted and depression-data witheld. Table 4 indicates the specific areas where depressed users modify their language when not discussing their depression. Overall, when not discussing depression, depressed Redditor's become more analytic and less willing to express their personal feelings, especially sadness and their present state.

We find that the depressed Redditor's language use fits within the paradigm one would expect. Beck's depress inventory (Beck et al., 1996) posits a trichotomy of depression: depressed attitude (1) towards the self, (2) towards the world, and (3) towards the future. As reflected by their LIWC scores, it is clear that depressed users more heavily emphasize themselves–seen in *I* usage–and the future–seen in the "Future:Focus" variable–than users who were part of our control group.

Further, these results are also consistent with a mindfulness-linked view of depression (Kabat-Zinn, 2003; Hofmann et al., 2010). Depressed users show an increase in anxious language–especially prevalent when users are talking about depression–decreased analytic language and, as previously mentioned, a

Model	Acc	AUC	F1
Task 1: All topics			
Baseline			
LIWC	.787	.751	.680
Char $n$grams	.810	.771	.707
Word $n$grams	.813	.777	.717
LIWC+$n$gram	**.818**	**.786**	**.729**
Task 2: Depression topic withheld			
Baseline			
Task 1 Best	.780	.732	.648
LIWC	.751	.706	.613
Char $n$grams	.774	.729	.646
Word $n$grams	.778	.738	.660
LIWC+$n$gram	**.787**	**.752**	**.681**

Table 5: Task 1 and Task 2 Results

strong emphasis on the self. This suggests, as the mindfulness research has (Williams, 2008; Michalak et al., 2008), that the wrong 'mode of mind', i.e., ruminating on negative thoughts, may exacerbate depressive mood.

We can further color our understanding of what depressed users are talking about by examining the words with the highest TF-IDF scores. A selection of words from the top-100 highest TF-IDF scores for depressed users is shown in Table 5. We have categorized these words into 5 groups: therapy and medication, people words, dialogic terms, Reddit and games, and porn and masturbation addiction.

Therapy and medication terms Unsurprisingly, the most common class of depression-indicator words are therapy- and medication-related terms. What is interesting, however, is the wide range of treatments about which depressed Redditors talk. They talk about talk-therapy related treatments (e.g., *psychitrist*, *counselor*, *therapist*), standard medications for depression (e.g., *Citalporam*, *Xanax*,*Prozac*, and the general: *antidepressants*), as well as alternative- or self-medications (e.g., *CBD*—THC oil, Kratom—a relatively new psychoactive). This suggests redditors are looking at a wide-range of solutions for their depression, further implying that they have been unsuccessful with previous attempts. It also suggests that Reddit may be a fruitful place to monitor the prevalence un-prescribed treatments.

People words Consistent with our LIWC analysis, in the depressed user all topic results we find personal pronouns like *I'm* and *I've*, which show users talking about themselves. This is also consistent with a notion of depressed individuals emphasizing themselves (Beck et al., 1996).

Dialogic terms Terms that are often used in conversations such as (*you*, *you're*, *yea*, *yeh*, *ur*, *thankyou*) show up with regularity in the top-100. This suggests that depressed users are addressing other redditors with *you* (and *youre*) more than a typical reddit user. This could be because depressed redditors engage more heavily in advice seeking and giving than standard redditors. These narration and response situations would provide ripe opportunity to address others.

Reddit, manga, games Across all user types and conditions we find Reddit-specific terms related to subreddits and gaming, such as *meirl*[6], a meme-sharing sub, *IGN*, a popular gaming website, and various game and manga characters *Nyx, warlock, Goku* and *Vegeta*.

Masturbation and pornography addiction Interestingly, a Reddit community dedicated to male sexual restraint–*nofap*–and one of its core concepts, "porn, masturbation and orgasm avoidance"–*pmo*–appear prominently in the depressed user tf-idf rankings. The stated purpose of the "NoFap" community[7] is to help users "reboot from porn addicition", by abstaining from orgasm for a month or more. This suggests that depressed Redditors, or at least a subset of them, are inclined to side with the research that has linked internet addiction, masturbation and pornography consumption with increases in depression

[6] www.reddit.com/r/meirl

[7] www.reddit.com/r/NoFap

Therapy and medication			People words	Dialogic terms	Reddit, games	Porn addiction
Psychiatrist	mg	Counseling	I'm	Thank you	Nyx	PMO
Xanax	Prozac	NMOM	I've	yea	IGN	nofap
Adderall	Therapist	BDP	ur	yeah	MeIRL	
Anhedonia	Counselor	ug	you			
Lucid	Zoloft	DET	you're			
Psychologist	Citalporam	Kratom	ppl			
Meds	Antidepressants	anhedonia				
ECT	CBD					

Table 6: Assorted words from top-100 most "depressed" words by TF-IDF score

(Chang et al., 2015) and depressive symptoms like loneliness (Yoder et al., 2005), as well as decreases overall health (Brody, 2010). The community appears to be mostly male users, which is perhaps not surprising; however, it is worth noting that depression has also been linked with increased rates of masturbation for women (Cyranowski et al., 2004).

Turning away from the lexical analysis to the predictive modeling, we find that the depression detection tasks mirror the LIWC findings insofar as the first task, which includes all the data, does prove to be more challenging (i.e., the models perform worse in it) than the the second task limited to depression-unrelated data. Across all the models we see a reduction in about 3% points from the all-data condition to the data-withheld condition. The one model trained on the all-data condition and tested on the data-withheld condition suffered more—about 4% points.

Relative to other depression-detection tasks, the models for the first task appear to be above average at depression detection (see Guntuku et al. (2017) for comparisons), and the performance of the LIWC-feature exclusive models suggests that the data here may be noisier than others depression-detection datasets (cf. Preoutic-Pietro et al., 2015). Given that, the 3.4% point reduction in AUC and 3.1% point reduction in accuracy should be taken seriously as a cautionary sign that depression-detection models may be overfitting for situations where social media users are open about their depression.

On a positive note, as Guntuku et al. (2017) note, these AUC scores are still better than the performance of primary-care physicians, which range from 62% to 74% (Mitchell et al., 2011). This suggests that even though social-media trained models may be overtrained, they may still be useful. Further, given that there exists a high-rate of depression-related stigma among primary care goers (Roeloffs et al., 2003), social-media based approaches may be an even more effective diagnostic tool because one can easily imagine patients with depression stigma actively acting to hide their depression from a primary care physician.

6 Conclusion

At the outset of this study, we believed that there was a chance natural language processing depression detection models were at risk of missing depressed individuals who were reluctant to talk about their depressive symptoms publicly, but nevertheless suffer substantially from depression. The results of our analysis, T^2 intervals about LIWC index scores and two classification tasks, are consistent with this belief. There appear to be substantial differences in depressed users language when they are explicitly discussing depression and when depression-related data is withheld.

With respect to the LIWC indexes, we found that depressed users showed differences with our control users as expected by psychological theory: increased anxiety, self-reference, negativity, sadness and affect, paired with decreased analytic language. With respect to the classification tasks, we found that, as expected, the depression data withheld task was more difficult than all topic task. Additionally, we found that the best performing model combined word- and character-$n$grams with LIWC features.

That said, these findings should be considered within the context of this study's limitations. First, the data shows a Reddit-specific bias (exemplified by the presence of porn/masturbation avoidance and a large number of computer, manga and video games terms in the TF-IDF rankings). These findings may not generalize to other social media platforms. Second, while depression diagnosis is temporally bounded, we make no effort to limit our data with respect to time. We may be including data for our depressed users from a time when they were not depressed, adding noise and reducing our accuracy. And third, while we intend to approximate the behavior of users who are both depressed and have high self-stigma, our attempt to do relies on users who presumably are seeking help. Users who have truly high self-stigma may behave differently. These findings and shortcomings naturally lead to future research opportunities. Future research should examine how variations in depression stigma may impact internet language use, how depressed-user language varies across social media platforms, and how language may be used to predict perceptions of public stigma. Lastly, the "NoFap" community appears like it would warrant further study on its own from a sociological perspective.

7 Ethical Considerations

This study aims to add consideration for the needs of high self-stigmatized individuals suffering from depression or depression-like symptoms. With that in mind, there are many valid reasons that people would be reluctant to disclose a mood-disorder or mental-health issue publicly. There is a difference between using computational linguistic technologies to direct targeted help towards these individuals and the use of these same technologies to expose these individuals. As long as the media continues to portray people suffering from mental illness as violent and dangerous (Friedman, 2006) and the public continues to believe that people suffering from mental illness endanger them (Barry et al., 2013), where natural language processing overlaps with health, all applications should strive to meet the classic bioethics principle of non-maleficence: first, do no harm.

Inappropriate uses of depression detection technology—especially on those with high-levels of depression stigma—may alter the way individuals relate to the disease. Individuals who feel targeted by this approach may become less likely to seek support and more likely to perceive the public as judging them for their illness. In those ways, misusing depression detection technology could exacerbate the stigma effects on a stigmatized population that is already at greater risk. Given that the goal of depression-detection for the stigmatized population is to help those individuals above all else, extra care should be paid to how the modeling is perceived by those who are suffering from depression.

References

Lisa J Barney, Kathleen M Griffiths, Anthony F Jorm, and Helen Christensen. 2006. Stigma about depression and its impact on help-seeking intentions. *Australian & New Zealand Journal of Psychiatry*, 40(1):51–54.

Colleen L Barry, Emma E McGinty, Jon S Vernick, and Daniel W Webster. 2013. After newtownpublic opinion on gun policy and mental illness. *New England journal of medicine*, 368(12):1077–1081.

Aaron T Beck, Robert A Steer, and Gregory K Brown. 1996. Beck depression inventory-ii. *San Antonio*, 78(2):490–8.

Stuart Brody. 2010. The relative health benefits of different sexual activities. *The journal of sexual medicine*, 7(4pt1):1336–1361.

Fong-Ching Chang, Chiung-Hui Chiu, Nae-Fang Miao, Ping-Hung Chen, Ching-Mei Lee, Jeng-Tung Chiang, and Ying-Chun Pan. 2015. The relationship between parental mediation and internet addiction among adolescents, and the association with cyberbullying and depression. *Comprehensive psychiatry*, 57:21–28.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60.

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.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 106–117.

Jill M Cyranowski, Joyce Bromberger, Ada Youk, Karen Matthews, Howard M Kravitz, and Lynda H Powell. 2004. Lifetime depression history and sexual function in women at midlife. *Archives of Sexual Behavior*, 33(6):539–548.

Munmun De Choudhury and Sushovan De. 2014. s: Self-disclosure, social support, and anonymity. In *ICWSM*.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013a. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3267–3276. ACM.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013b. Predicting depression via social media. *ICWSM*, 13:1–10.

Daniel Eisenberg, Marilyn F Downs, Ezra Golberstein, and Kara Zivin. 2009. Stigma and help seeking for mental health among college students. *Medical Care Research and Review*, 66(5):522–541.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

Jonathan Fine. 2006. *Language in psychiatry: A handbook of clinical practice*. Equinox London.

Richard A Friedman. 2006. Violence and mental illnesshow strong is the link? *New England Journal of Medicine*, 355(20):2064–2066.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

GACCT Harman and Mark H Dredze. 2014. Measuring post traumatic stress disorder in twitter. *In ICWSM*.

Misato Hiraga. 2017. Predicting depression for japanese blog text. In *Proceedings of ACL 2017, Student Research Workshop*, pages 107–113.

Stefan G Hofmann, Alice T Sawyer, Ashley A Witt, and Diana Oh. 2010. The effect of mindfulness-based therapy on anxiety and depression: A meta-analytic review. *Journal of consulting and clinical psychology*, 78(2):169.

Harold Hotelling. 1931. The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2(3):360–378.

Jon Kabat-Zinn. 2003. Mindfulness-based interventions in context: past, present, and future. *Clinical psychology: Science and practice*, 10(2):144–156.

Klara Latalova, Dana Kamaradova, and Jan Prasko. 2014. Perspectives on perceived stigma and self-stigma in adult male patients with depression. *Neuropsychiatric disease and treatment*, 10:1399.

David E Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. Springer.

Rachel C Manos, Laura C Rusch, Jonathan W Kanter, and Lisa M Clifford. 2009. Depression self-stigma as a mediator of the relationship between depression severity and avoidance. *Journal of Social and Clinical Psychology*, 28(9):1128–1143.

Johannes Michalak, Thomas Heidenreich, Petra Meibert, and Dietmar Schulte. 2008. Mindfulness predicts relapse/recurrence in major depressive disorder after mindfulness-based cognitive therapy. *The Journal of nervous and mental disease*, 196(8):630–633.

Alex J Mitchell, Sanjay Rao, and Amol Vaze. 2011. International comparison of clinicians' ability to identify depression in primary care: meta-analysis and meta-regression of predictors. *Br J Gen Pract*, 61(583):e72–e80.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 21–30.

Lenore Sawyer Radloff. 1977. The ces-d scale: A self-report depression scale for research in the general population. *Applied psychological measurement*, 1(3):385–401.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 99–107.

Carol Roeloffs, Cathy Sherbourne, Jürgen Unützer, Arlene Fink, Lingqi Tang, and Kenneth B Wells. 2003. Stigma and depression among primary care patients. *General hospital psychiatry*, 25(5):311–315.

Georg Schomerus, Herbert Matschinger, and Matthias C Angermeyer. 2009. The stigma of psychiatric treatment and help-seeking intentions for depression. *European archives of psychiatry and clinical neuroscience*, 259(5):298–306.

H Andrew Schwartz, Johannes Eichstaedt, Margaret L Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pages 58–65.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing depression from twitter activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3187–3196. ACM.

J Mark G Williams. 2008. Mindfulness, depression and modes of mind. *Cognitive Therapy and Research*, 32(6):721.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *The Conference on Empirical Methods in Natural Language Processing*, pages 2968–2979.

Vincent Cyrus Yoder, Thomas B Virden III, and Kiran Amin. 2005. Internet pornography and loneliness: An association? *Sexual addiction & compulsivity*, 12(1):19–44.