A Data

We leverage four social media datasets containing ground truth mental health status to facilitate our analysis. Our datasets are a subset of those used by Harrigian et al. (2020); we choose to exclude the Reddit Self-disclosed Depression Diagnosis (RSDD) dataset (Yates et al., 2017) from our analysis because it contains substantial user overlap with the Self-reported Mental Health Diagnoses (SMHD) dataset (Cohan et al., 2018). All datasets were acquired, stored, and analyzed under the terms of their respective data usage agreements. We describe each dataset below and provide summary statistics in Table 1. Note that our datasets vary across platform, size, and time period dimensions.

CLPsych 2015 Shared Task. Introduced in Coppersmith et al. (2014) and refined in Coppersmith et al. (2015b) for the 2015 Computational Linguistics and Clinical Psychology Workshop (CLPysch), this Twitter dataset contains tweet histories of up to 3,000 posts from 844 individuals (422 depressed). Individuals in the depression cohort were identified using regular expressions that matched phrases similar to "I was diagnosed with depression" and manually vetted by the authors. The control group was sampled randomly from the Twitter Gardenhose so that inferred age and gender attributes were closely aligned with the distribution of the depression group (Schwartz et al., 2013). Several of the individuals in the original version of the dataset are also part of the Multi-Task Learning dataset (Benton et al., 2017) and were accordingly removed from our version to facilitate an unbiased domain transfer analysis.

Multi-Task Learning. Constructed using the same procedure as the 2015 CLPsych Shared Task dataset, this Twitter dataset was introduced by Benton et al. (2017) as an amalgamation of data from Coppersmith et al. (2015b,a) to facilitate an analysis of multi-task learning methods for mental health status inference. The dataset contains tweet histories for 2,800 individuals (1,400 depressed). Our version includes data from the same users as the original article, but has been updated to include tweets posted after the original articles were published.

Topic-Restricted Text. This Reddit dataset was introduced in Wolohan et al. (2018) and reproduced by Harrigian et al. (2020) as a resource to facilitate analysis of language by depressed individuals outside explicit mental health subreddits.[1]. Annotations for this dataset are derived as a function of subreddit participation. The depression group was sampled by identifying unique users who posted one of the 10,000 most recent top-level threads in the r/depression subreddit (recent as of October 25, 2019); the control group was sampled by identifying users who posted one of the 10,000 most recent top-level threads in the r/AskReddit subreddit during the same time period *and* had not posted a top-level thread in the r/depression subreddit. Users with fewer than 1,000 words in their entire post history were excluded from the final dataset.

Dataset	Platform	Time Period	# Individuals Per Class
2015 CLPsych Shared Task	Twitter	Jan. 2011 - Dec. 2013	422
Multi-Task Learning	Twitter	Jan. 2013 - Jan. 2016	1,400
Topic-Restricted Text	Reddit	Jan. 2014 - Jan. 2020	6,859
SMHD	Reddit	Jan. 2010 - Jan. 2018	7,847

Table 1: Summary statistics for annotated datasets. Note that we use balanced datasets based on the smallest available class size.

SMHD. Our final dataset (SMHD) was introduced by Cohan et al. (2018) as a Reddit-centric alternative to the Twitter dataset introduced by Coppersmith et al. (2015a). The official version available for distribution contains ground truth annotations for 9 mental health conditions and a large control group. SMHD was constructed using a similar procedure as our Twitter datasets — candidate users for the condition groups were identified using regular expressions that matched self-disclosures of a mental health diagnosis and were then manually vetted by domain experts. Individuals in this dataset have been completely anonymized and the dataset's usage agreement prohibits deanonymization. For this reason, we were unable to tell whether any individuals in SMHD are also part of the Topic-Restricted Text dataset.

B Preprocessing

Unlike work from e.g. Wolohan et al. (2018); Harrigian et al. (2020), we do not perform any additional dataset filtering in this analysis (e.g. removing posts with overt mental health content, excluding

[1]We make use of the reproduced dataset using code made available by the authors of Harrigian et al. (2020)

Source	Target	LDA		PLDA		
		C	K	C	K	K_d
CLPsych Shared Task	Multi-task Learning	1	150	5	75	25
	Topic-Restricted Text	0.3	50	0.001	25	50
	SMHD	1	50	10	75	50
Multi-task Learning	CLPsych Shared Task	10	200	100	50	75
	Topic-Restricted Text	50	100	50	75	75
	SMHD	5	50	10	200	25
Topic-Restricted Text	CLPsych Shared Task	100	75	0.1	100	50
	Multi-task Learning	0.1	200	5	100	50
	SMHD	0.001	200	0.001	150	75
SMHD	CLPsych Shared Task	100	75	50	75	100
	Multi-task Learning	0.3	25	10	200	100
	Topic-Restricted Text	50	150	100	75	25

Table 2: Parameters that maximize AUC in the target domain's development set for both LDA and PLDA topic models.

mental health support subreddits). As our objective is to directly compare user representations learned via LDA and PLDA, we do not expect this decision to introduce bias into our conclusions. Indeed, since we hypothesize PLDA is well-suited for mitigating the effect of spurious correlations introduced during the data annotation process, we argue that our decision not to make attempts to exclude sampling related artifacts is beneficial for evaluating the strength of our hypothesis.

C Hyperparameters

We consider a two-stage hyperparameter optimization process to most fairly compare user representations from LDA and PLDA in the downstream depression inference task. The parameters part of the grid search are enumerated below:

- α :{0.001, 0.01, 0.1, 1}
- β :{0.001, 0.01, 0.1, 1}
- K :{25, 50, 75, 100, 150, 200}
- K_d :{25, 50, 75, 100}
- C :{0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 5, 10, 50, 100}

In the first stage, we fix the number of general latent topics $K = 100$ and the number of per-domain topics $K_d = 25$, and then perform a grid search over all combinations of the document-topic prior α, topic-word prior β, and regularization strength C. Priors behave similarly for both LDA and PLDA models and consistently across regularization strengths. We do not observe any significant difference in AUC for models trained with α, β <1, though models trained with α =1 and β =1 consistently underperform in comparison to other priors across all datasets. Ultimately, we decide to fix $\alpha = 0.01$ and $\beta = 0.01$ for subsequent optimization stages as a compromise between 0.001 and 0.1.

In the second stage, we look to maximize predictive performance as a function of the number of topics, K and K_d, and regularization strength C. Limited by computational expense, we make the (strong) assumption that optimal priors discovered in the first stage generalize across all values of K and K_d. We execute a grid search over all combinations of K, K_d, and C. Parameters that maximize AUC in the target domain's development set for each source-target dataset combination are provided in Table 2.

References

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multi-task learning for mental health using social media text. *arXiv preprint arXiv:1712.03538*.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. *arXiv preprint arXiv:1806.05258*.

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, and Kristy Hollingshead. 2015a. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 1–10.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015b. 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.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3774–3788.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

JT Wolohan, Misato Hiraga, Atreyee Mukherjee, Zeeshan Ali Sayyed, and Matthew Millard. 2018. Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with nlp. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 11–21.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.