Do Models of Mental Health Based on Social Media Data Generalize?

Keith Harrigian, Carlos Aguirre, Mark Dredze
Johns Hopkins University
kharrigian@jhu.edu, caguirr4@jhu.edu, mdredze@cs.jhu.edu

Abstract

Proxy-based methods for annotating mental health status in social media have grown popular in computational research due to their ability to gather large training samples. However, an emerging body of literature has raised new concerns regarding the validity of these types of methods for use in clinical applications. To further understand the robustness of distantly supervised mental health models, we explore the generalization ability of machine learning classifiers trained to detect depression in individuals across multiple social media platforms. Our experiments not only reveal that substantial loss occurs when transferring between platforms, but also that there exist several unreliable confounding factors that may enable researchers to overestimate classification performance. Based on these results, we enumerate recommendations for future mental health dataset construction.

1 Introduction

In the last decade, there has been substantial growth in the area of digital psychiatry. Automated methods using natural language processing have been able to detect mental health disorders based on a person's language in a variety of data types, such as social media (Mowery et al., 2016; Morales et al., 2017), speech (Iter et al., 2018) and other writings (Kayi et al., 2017; Just et al., 2019). As in-person clinical visits are made increasingly difficult by socioeconomic barriers and public-health crises, such as COVID-19, tools for measuring mental wellness using implicit signal become more important than ever (Abdel-Rahman, 2019; Bojdani et al., 2020).

Early work in this area leveraged traditional human subject studies in which individuals with clinically validated psychiatric diagnoses volunteered their language data to train classifiers and perform quantitative analyses (Rude et al., 2004; Jarrold et al., 2010). In an effort to model larger, more diverse populations with less overhead, a substantial portion of research in the last decade has instead explored data annotated via automated mechanisms (Coppersmith et al., 2015a; Winata et al., 2018).

Studies leveraging proxy-based annotations have supported their design by demonstrating alignment with existing psychological theory regarding language usage by individuals living with a mental health disorder (Cavazos-Rehg et al., 2016; Vedula and Parthasarathy, 2017). For example, feature analyses have highlighted higher amounts of negative affect and increased personal pronoun prevalence amongst depressed individuals (Park et al., 2012; De Choudhury et al., 2013). Given these consistencies, the field has largely turned its attention toward optimizing predictive power via state of the art models (Orabi et al., 2018; Song et al., 2018).

The ultimate goal of these efforts has been threefold—to better personalize psychiatric care, to enable early intervention, and to monitor population-level health outcomes in real time. Nonetheless, research has largely trudged forward without stopping to ask one critical question: do models of mental health conditions trained on automatically annotated social media data actually generalize to new data platforms and populations?

Typically, the answer is no—or at least not without modification. Performance loss is to be expected in a variety of scenarios due to underlying distributional shifts, e.g. domain transfer (Shimodaira, 2000; Subbaswamy and Saria, 2020). Accordingly, substantial effort has been devoted to developing computational methods for domain adaptation (Imran et al., 2016; Chu and Wang, 2018). Outcomes from this work often provide a solid foundation for use across multiple natural language processing tasks (Daume III and Marcu, 2006). However, it is unclear to what extent factors specific to mental health require tailored intervention.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3774–3788
November 16 - 20, 2020.

In this study, we demonstrate that at a baseline, proxy-based models of mental health status *do not* transfer well to other datasets annotated via automated mechanisms. Supported by five widely used datasets for predicting depression in social media users from both Reddit and Twitter, we present a combination of qualitative and quantitative experiments to identify troublesome confounds that lead to poor predictive generalization in the mental health research space. We then enumerate evidence-based recommendations for future mental health dataset construction.

Ethical Considerations. Given the sensitive nature of data containing mental health status of individuals, additional precautions based on guidance from Benton et al. (2017a) were taken during all data collection and analysis procedures. Data sourced from external research groups was retrieved according to each dataset's respective data usage policy. The research was deemed exempt from review by our Institutional Review Board (IRB) under 45 CFR § 46.104.

2 Domain Adaptation in Mental Health

Domain adaptation (or "transfer") of statistical classifiers is a well-studied computational problem with high relevance across several areas of natural language processing (Jiang, 2008; Peng and Dredze, 2017). It is particularly useful in situations where acquiring ample training data for a target application is intractable (e.g. monetary, time constraints) or impossible (e.g. privacy constraints) (Rieman et al., 2017). For example, in the sub-field of machine translation, significant effort is devoted to finding ways to effectively use large corpora of formal parallel text to train models for application in domains with informal and dynamic language, such as social media and conversational speech (Wang et al., 2017; Murakami et al., 2019).

Traditional challenges encountered when transferring models between domains include variance in source and target class distributions (Japkowicz and Stephen, 2002), semantic misalignment (Wu and Huang, 2016), and sparse vocabulary overlap (Stojanov et al., 2019). Fortunately, once these issues are identified, it is typically possible to decrease the transfer performance gap via methods such as structural correspondence learning, feature subspace mapping, and adversarial training (Blitzer et al., 2006; Bach et al., 2016; Tzeng et al., 2017).

Domain adaptation is of particular interest in the mental health space, where there exist numerous complexities in obtaining a sufficient sample of training data. For instance, the sensitive nature of mental health data necessitates extra care when creating and supporting new datasets (Benton et al., 2017a). Additionally, behavioral disorders are known to display variable clinical presentations amongst different populations, which can make identification of ground truth difficult (De Choudhury et al., 2017; Arseniev-Koehler et al., 2018).

The latter point highlights the presence of label noise inherent in mental health data (Mitchell et al., 2009; Shing et al., 2018). This facet serves as one of two primary issues unique to this research space that may hinder attempts at domain transfer. Indeed, prior work found that diverse and sometimes conflicting views humans have regarding suicidal ideation can make obtaining reliable gold-standard labels fundamentally challenging and lead to degradation in model performance (Liu et al., 2017).

Sampling-related biases present the other main area of concern for successful domain transfer by mental health classifiers. Attributes such as personality, gender, age, and disorder co-morbidity have been found to significantly affect the presentation of mental health disorders in language data (Cummins et al., 2015; Preoţiuc-Pietro et al., 2015). Moreover, the proxy-based annotation mechanisms used to label large social media data sets with mental health status invite the introduction of self-disclosure bias into the modeling task (Amir et al., 2019). Specifically, labels sourced from populations of individuals who self-disclose certain attributes may contain activity-level and thematic biases that cause poor generalization in larger populations (Lippincott and Carrell, 2018).

Research leveraging text data for mental health status classification has primarily only considered a constrained form of domain transfer. In a within-subject analysis, Ireland and Iserman (2018) examined differences in language usage by Reddit users who had posted in an anxiety support forum within and outside mental health forums. Similarly, Wolohan et al. (2018) explored the predictive power of models trained to detect depression within Reddit users as a function of access to text from explicit mental health related subreddits. Both studies highlighted a mitigation of overt mental health discussion outside of the support forums, but still detected linguistic nuances in individuals with an affiliation to the mental health forums.

Dataset	Platform	Years	Size (Individuals)	Annotation Mechanism
CLPsych	Twitter	2011-2014	Control: 477 Depression: 477	Regular expressions; Manual verification; Age- & gender-matched controls
Multi-Task Learning	Twitter	2013-2016	Control: 1,400 Depression: 1,400	Regular expressions; Manual verification; Age- & gender-matched controls
RSDD	Reddit	2008-2017	Control: 107,274 Depression: 9,210	Regular expressions; Manual verification; Subreddit-based controls
SMHD	Reddit	2010-2018	Control: 127,251 Depression: 14,139	Regular expressions; Subreddit-based controls
Topic-Restricted Text	Reddit	2014-2020	Control: 7,016 Depression: 6,853	Community participation

Table 1: Summary statistics for each dataset. All datasets leverage proxy-based annotations. Distribution over time and sample size varies significantly between datasets.

Shen et al. (2018) attempted to use transfer learning with large amounts of English Twitter data annotated with individual-level depression labels to improve predictive performance of depression classifiers in Chinese Weibo data. Using the English and Chinese versions of the Linguistic Inquiry and Word Count tool (LIWC) (Pennebaker et al., 2001; Huang et al., 2012) in conjunction with other modalities of social data (e.g. profile metadata, images), the authors showed that signal from Twitter was useful for classification on Weibo.

Recent work from Ernala et al. (2019) was the first to explore some of aforementioned difficulties with domain transfer in the mental health space. Multiple different annotation mechanisms were used to train Twitter-based models for identifying schizophrenia and then applied to Facebook data from an independent population of clinically diagnosed schizophrenia patients. Three different types of proxy signals with varying degrees of manual supervision were each found to generalize poorly to the clinical population. While the authors' analysis suggested the domains were similar enough to justify transfer attempts, only limited post-hoc analysis of the data platform effect was carried out. Thus, it remains unclear to what extent the annotation methodologies as opposed to platform effects (or other confounds) caused the degradation.

3 Data

We select depression classification as our task because it is perhaps the most widely studied, has multiple datasets from different platforms, and is of critical importance to society. Estimated to affect 4.4% of the global population, depression presents a significant economic burden and remains the most common psychiatric disorder associated with deaths by suicide (Hawton et al., 2013; Organization et al., 2017). Occupying a lion's share of the computational literature, depression classification is a critical first target for evaluating generalization of mental health models in social media (Chancellor and De Choudhury, 2020).

To quantify the nature of domain transfer loss, we consider five datasets. Datasets were selected based on their common adoption in the literature (Preoţiuc-Pietro et al., 2015; Gamaarachchige and Inkpen, 2019) and their use of proxy-based annotations (Coppersmith et al., 2014). We use two Twitter—*CLPsych 2015 Shared Task* (Coppersmith et al., 2015b), *Multi-Task Learning* (Benton et al., 2017b)—and three Reddit datasets—*RSDD* (Yates et al., 2017), *SMHD* (Cohan et al., 2018), and *Topic-Restricted Text* (Wolohan et al., 2018). Table 1 presents summary statistics. Construction details are in Appendix A as a courtesy to the reader.

3.1 Mitigating Bias

Each dataset was curated in part by a system of simple rules (e.g. matches to "I was diagnosed with depression," participation in a depression support forum). While these heuristics are useful for identifying candidates to include within each dataset, they also risk introducing bias that may render the modeling task trivial. For example, individuals who disclose a depression diagnosis are likely to also share their experience with other psychiatric conditions (Benton et al., 2017b), while language used in dedicated mental-health subreddits systematically differs from the rest of Reddit (De Choudhury and De, 2014; Ireland and Iserman, 2018).

To encourage our mental health classifiers to learn subtle linguistic nuances that cannot be easily captured using straightforward logic, we make efforts to exclude unambiguous mental health content from all training and evaluation procedures. In line

with prior work, we discard posts that include mentions of clinically-defined psychiatric conditions, adopting the list of mental health terms enumerated by Cohan et al. (2018) as a reference. This list (*N*=458) extends work from Yates et al. (2017) by including disorders tangential to depression, common misspellings, and colloquial references.

As is standard for mental health modeling, we also discard posts made in subreddits dedicated to providing mental health support (Yates et al., 2017; Cohan et al., 2018; Wolohan et al., 2018). Since new subreddits are created daily and our version of the Topic-Restricted Text dataset contains posts made after collection of RSDD and SMHD, we create an updated list of mental health support subreddits. To do so, we examine the empirical distribution of posts amongst subreddits within the Topic-Restricted Text dataset and rank each subreddit S based on pointwise mutual information (PMI) for the depression group D, $\log(p(S|D)/p(S))$. We manually examined the top 1000 subreddits based on PMI and identified all subreddits whose description affirmed an association to mental health.

Our list (*N*=242) expands existing resources from Yates et al. (2017) and Cohan et al. (2018) by providing 162 additional mental health subreddits, many of which were actually created before the collection of RSDD and SMHD.[1] While this step diminishes the risk of mental health content saturating the Topic-Restricted Text dataset, the list's expansion beyond that of the RSDD and SMHD lists suggests that the former two Reddit datasets may indeed still have overt mental health content. We explore how different degrees of subreddit-based filtering may affect generalization in §6.4.

4 Models

We begin by training classification models for predicting depression on each dataset. All classification experiments leverage the same training procedure and features (see Appendix D for details). As a classifier, we use ℓ_2-regularized logistic regression. Despite our model's relative simplicity we are able to achieve respectable within-domain classification performance while maintaining an ability to interpret learned parameters. Logistic regression has served as a difficult benchmark to beat given access to appropriate engineered features for prior mental health studies (Benton et al., 2017b).

4.1 Model Validation

To validate our modeling framework against prior work, we first establish *within-domain* predictive baselines. This step also allows us to contextualize performance by estimating the intrinsic difficulty of modeling each dataset (DeMasi et al., 2017).

Methods. We use train/development/test splits if they have been established by the dataset distributor; otherwise, we sample 20% from the available data to be used as a held-out test set and then create an additional 80/20 train/dev split using the remaining data. For each dataset, we use an independent grid search to select regularization strength C that maximizes F1 in the dataset's development split (see Appendix E). We use a binarization threshold of 0.5 (noninclusive) for all datasets.

Results. We report test set F1 for each dataset in the bottom row of Table 2. Our models perform on par with prior research for the two Twitter datasets and the Topic-Restricted Text dataset. Results for RSDD and SMHD improve upon their respective baseline models, but are inferior to neural methods.

5 Transfer Experiments

We conduct a series of experiments to measure the generalization of models between depression datasets and explain sources of model degradation.

5.1 Cross-domain Transfer

Task formulation and dataset design remain a significant source of nuance across prior studies for mental health status prediction (Morales et al., 2017; Chancellor and De Choudhury, 2020). As such, we hypothesize that standardizing training settings (e.g. class balance, sample size) will account for discrepancies in cross-domain performance.

Methods. We consider two experimental designs. In the first experiment (†), we downsample all datasets to have the same training/development size of the smallest class in the smallest dataset (i.e. CLPsych). In the second experiment (††), we balance class distributions independently for each dataset based on the dataset's smaller class, but allow sample size to vary between datasets. The former experiment allows us to establish equitable baselines between datasets, while the latter experiment enables us to explore whether access to additional training data ameliorates transfer loss.

[1] Subreddits and code are made available to other researchers: https://github.com/kharrigian/emnlp-2020-mental-health-generalization

	Test Data									
	CLPsych		Multi-Task		RSDD		SMHD		Topic-Restricted Text	
Train Data	†	††	†	††	†	††	†	††	†	††
CLPsych	**.774 ± .009**	**.774 ± .009**	.635 ± .054	.635 ± .054	.169 ± .011	.169 ± .011	.064 ± .006	.064 ± .006	.638 ± .034	.638 ± .034
Multi-Task	.533 ± .111	.739 ± .004	**.802 ± .018**	**.830 ± .005**	.149 ± .001	.149 ± .001	.054 ± .000	.054 ± .001	.648 ± .007	.655 ± .011
RSDD	.247 ± .034	.284 ± .046	.338 ± .041	.407 ± .051	**.338 ± .010**	**.405 ± .003**	–	–	.487 ± .046	.434 ± .003
SMHD	.335 ± .048	.355 ± .028	.543 ± .040	.464 ± .028	–	–	**.186 ± .007**	**.212 ± .006**	.626 ± .011	.631 ± .007
Topic-Restricted Text	.624 ± .018	.668 ± .008	.516 ± .060	.648 ± .026	.173 ± .017	.218 ± .004	.105 ± .014	.106 ± .008	**.686 ± .007**	**.735 ± .002**
Baseline	.77		.82		.59		.38		.75	

Table 2: F1 score ($\mu \pm \sigma$) for the *Balanced & Downsampled* (†) and *Balanced* (††) cross-domain transfer experiments. Baselines described in §4.1, which preserve class imbalance during training, are presented in the bottom row. Increasing dataset size (10x in some cases) does *not* unanimously improve transfer.

For both experiments, we start by combining training and development splits. Then, for each dataset, we sample from the combined splits based on the parameters of the experiment and split the resulting sample into 5 class-stratified folds. We train 5 classifiers per dataset, using 4 folds for training each time, and apply the classifiers to each dataset's test set. Since a substantial portion of individuals in SMHD are part of RSDD, we refrain from conducting experiments between the two datasets.

Results. We report F1 score ($\mu \pm \sigma$) for both experiments in Table 2. In line with existing research, within-domain training outperforms cross-domain training in each of our datasets for both sampling settings. While additional samples available for training in the second experiment moderately improve within-domain performance, they are not uniformly helpful for mitigating transfer loss to other datasets. Models generally outperform a random classifier at ranking depression risk in cross-domain transfer scenarios. However, some models are poorly calibrated for new domains and consequently obtain low F1 scores (e.g. CLPsych → SMHD). Addressing miscalibration in domain transfer scenarios remains an open research question (Pampari and Ermon, 2020; Park et al., 2020).

We find that models trained on Twitter data transfer to Reddit data better than models in the reverse direction. Not surprisingly, given their overlap in training samples, models trained on the SMHD and RSDD datasets transfer to other domains in an equitable manner, trading improvements with each other across transfer settings. These results indicate that sample size and class balance are not solely responsible for generalization loss.

5.2 Temporal Transfer

Typical sources of transfer loss concern differences in features between domains (Blitzer et al., 2007; Ben-David et al., 2010). However, other factors may govern model degradation for depression classification. One such cause of loss is temporal misalignment between the datasets (Table 1). Prior work has shown that language dynamics may hinder models upon deployment (Dredze et al., 2016; Huang and Paul, 2018). In social media, where users adopt new linguistic norms rapidly, performance may be more volatile (Brigadir et al., 2015).

5.2.1 Class Misalignment

As an exercise to understand whether temporal artifacts are present in the datasets, we first consider training and evaluating single-domain models with a temporal misalignment between the control and depression groups. By training on mutually-exclusive time periods for each class, we hypothesize the classifier will not only able to learn how to distinguish between groups, but also to distinguish between time periods. If this hypothesis holds true, we expect performance metrics to be artificially inflated when a temporal exclusivity per class exists.

Methods. We split each dataset into one year periods based on the calendar year. For each year, we identify individuals in the Twitter datasets with at least 200 posts and individuals in the Reddit datasets with at least 100 posts.[2] We balance the number of individuals across time periods and groups within each dataset, but allow this sample size to vary across datasets. To account for growth in post frequency over time (which increases the number of documents that generate individual feature vectors), we perform additional post-level sampling. We randomly select 200 posts per year in the Twitter datasets and 100 posts per year in the Reddit datasets. Samples of individuals within each time period are additionally separated into 5 stratified folds. Folds are established so that individuals in the training data of one time period are never present in the test data of another time period.

[2]We use 2x more posts in the Twitter data to account for posts in the Reddit datasets having roughly twice as many words as Tweets do on average.

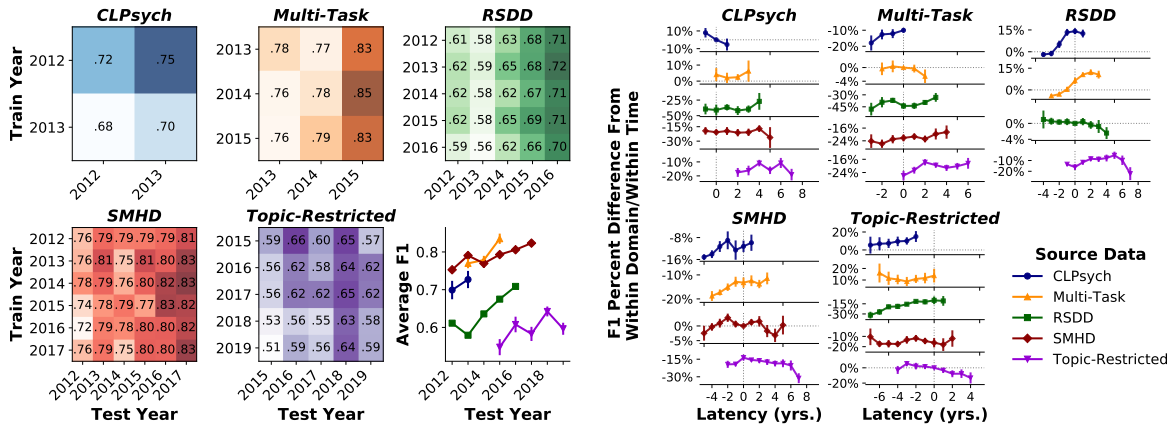

Figure 1: Temporal-transfer results. **(Left)** Average within-domain F1 score as a function of training and evaluation periods. Predictive performance tends to be better for more recent temporal splits regardless of training period. **(Right)** Percent difference in F1 score relative to within-domain, no-latency model. Models trained on Twitter data benefit most from temporal alignment. Performance suffers when applying new models to old data.

To evaluate the degree to which temporal effects are present, we sample groups from all possible combinations of time periods. For example, in one setting, both the control and depression groups are sampled from 2013; in another setting, the control group is sampled from 2013, while the depression group is sampled from 2015. For each combination, we use 4 of the stratified folds for training and use the remaining fold for evaluation, and then repeat the process for all folds. We compare performance when classes are sampled from the same time period against performance when classes are sampled from mutually exclusive time periods.

Results. We achieve a 3-22% increase in F1 across all datasets when classes are sampled from mutually exclusive time periods instead of being temporally-aligned. The improvement suggests that temporal artifacts exist, as the classifier is able to not only identify signal relevant to classifying depression, but also to classifying data from different periods of time. This result highlights the importance of sampling classes evenly over time.

5.2.2 Latency

We now measure the effect temporal artifacts have on cross-domain performance. We hypothesize model degradation scales with deployment latency.

Methods. We use the same data sampling mechanism described in §5.2.1. However, we now only consider the case in which control and depression groups are sampled from the same time period. As before, we train a classifier on 4 of the 5 stratified folds for a time period in one dataset. We then evaluate *within-domain* performance using the remaining fold and *cross-domain* performance using one fold from each time period in the other datasets. We assume ground truth is consistent over multiple time periods; given the episodic nature of depression, we recognize this may promote pessimistic results for some periods (Tsakalidis et al., 2018).

Results. Examining *within-domain* results in Figure 1 (left), predictive performance tends to be better for more recent temporal splits regardless of training period. Classifiers trained on old data (relative to the evaluation period) tend to perform on par with aligned regimens, while classifiers trained on new data show linear losses over time. Losses are significant after 2-3 years depending on the dataset.

Though some trends do emerge, *cross-domain* performance as a function of temporal latency is relatively variable. Visualized in Figure 1 (right), models trained on the Twitter datasets benefit most from temporal alignment in cross-domain settings. Models trained on Topic-Restricted Text show significant drop offs in predictive performance when applied to older samples within all Reddit datasets. While models trained on RSDD perform better on Topic-Restricted Text as latency is reduced, models trained on SMHD do not exhibit the same trend.

6 Post-hoc Analysis

In the previous section, we identified the degree to which loss occurs under a variety of domain transfer settings. However, these settings do not account for all performance disparities. In this section, we measure differences between the datasets to understand the source of loss.

6.1 Vocabulary Overlap

Traditionally, different feature vocabularies account for domain transfer loss (Serra et al., 2017; Chen and Gomes, 2019; Stojanov et al., 2019). Therefore, we hypothesize that limited feature overlap and poor vocabulary alignment across datasets could hinder cross-domain generalization.

Methods. We explore this phenomenon by computing the Jaccard Similarity (JS) of vocabularies between each dataset. We examine correlations between JS and F1 scores from the cross-domain transfer experiments discussed in §5.1.

Results. We find the minimum similarity occurs between the CLPsych and RSDD datasets ($JS = 0.10$) while the maximum occurs between the Topic-Restricted Text and SMHD datasets ($JS = 0.65$).[3] Only a weak correlation between similarity and performance exists (Pearson $\rho < 0.18$), suggesting poor generalization is not solely due to differences in vocabulary.

6.2 Topical Alignment

Our classification models leverage reduced feature representations in the form of LDA topic-distributions (Blei et al., 2003) and mean-pooled pre-trained GloVe embeddings (Pennington et al., 2014). Designed to capture and reflect semantics, we hypothesized these low-dimensional features would mitigate transfer loss due to poor vocabulary alignment. Lacking support from our cross-domain transfer results, we look closer at the themes present within each dataset.

Methods. We identify the unigrams that are most unique to each dataset and group. For each dataset, we use scores assigned by our KL-divergence-based feature selection method (see Appendix D) to rank the most informative features per class (Chang et al., 2012). We jointly examine the top-500 most informative unigrams per class, noting high-level themes common across the datasets.

Results. With respect to similarities, we note that words used in discussion about gender and sexuality are strongly associated with each of the depression groups (e.g. 'cis', 'homophobia', 'masculine'), likely a reflection of marginalized groups being at higher risk of depression (Budge et al., 2013). Also ubiquitous amongst each of the datasets are references to self-injurious behavior (e.g 'wrists', 'self-harm', 'hotline'). Increased emoji usage and references to athletics ('nbafinals', 'scorer') are strong indicators of the control group in each dataset, as well as terms reflecting current events.

With respect to differences, associations between word usage and depression are subjectively easier to interpret within the Reddit datasets. For example, discussion of mental-health treatment (e.g. 'counselor', 'therapy', 'wellbutrin') and familial and intimate relationships ('brother-in-law', 'soulmate') are prominent within the Reddit datasets. In contrast, language associated with depression within the Twitter datasets tends to reflect slightly more nuanced elements of the condition—e.g. social inequity ('sexism', '#yesallwomen') and fantasy ('fanfics', 'cosplay', 'villians'). These themes align with empirical findings that women are at a higher risk of depression (Kessler, 2003) and depressed individuals often find solace in niche subcultures (Blanco and Barnett, 2014; Bowes et al., 2015).

Additionally, we find several temporally-isolated references within the Twitter datasets (e.g. '#RIPRobinWilliams', '#SDCC'). In the Multi-task Learning dataset, we also see several terms using non-American English (e.g. 'colour', 'favourite') which may represent a geographic imbalance amongst the sampled individuals.

6.3 Stability of LIWC

The Linguistic Inquiry and Word Count (LIWC) dictionary has been an effective tool for measuring linguistic-nuances of mental health disorders regardless of textual formality (Mowery et al., 2016; Turcan and McKeown, 2019). Our version of the dictionary (2007) maps approximately 12k words to 64 dimensions (e.g. negative emotion, leisure) that have been empirically validated to capture an individual's social and psychological states (Tausczik and Pennebaker, 2010).[4] A single LIWC feature value represents the proportion of words used across an individual's post history that match the given LIWC dimension. In the same way that we expect semantic distributions (§6.2) to ameliorate transfer loss, we hypothesize that models trained on this representation will be more robust when vocabulary overlap is sparse.

Methods. We explore this hypothesis from three angles: 1) We perform cross-domain transfer experiments using LIWC as the only feature set provided

[3] JS is moderately deflated in RSDD due to the dataset's large vocabulary, causing SMHD and Topic-Restricted Text to have the highest similarity instead of SMHD and RSDD.

[4] The 2007 version of LIWC has a high similarity with the 2015 version amongst dimensions most strongly associated with depression (Pennebaker et al., 2015).

for training and evaluation; 2) We fit LIWC-based classifiers 100 times per dataset using random 70% samples and examine correlations of the learned coefficients; 3) We compute the average feature value of each LIWC dimension per class and measure the difference between classes.

Results. We note that domain-transfer experiments using LIWC as the only feature set maintain high degrees of transfer loss while sacrificing within-domain performance. Moreover, correlations between coefficients of models between datasets are relatively low across all comparisons, maxing out at a Spearman R value of 0.338 for the comparison between RSDD and SMHD datasets, which happen to have significant user overlap as is. In general, LIWC coefficients tend to be more correlated within platforms than between them.

Examination of the underlying class differences provides insight into linguistic differences between each dataset's depression group. In line with prior work, function word use, first-person pronoun use, and cognitive mechanisms are more common within the depression group of each dataset, though their relative prevalence varies. Conversation regarding relativity (i.e. space, motion, time) is strongly associated with the control groups in the Twitter data, but is more associated with the depression groups in the Reddit data. Anger and perceptual topics are more prevalent within the depression groups for Twitter than Reddit.

6.4 Self-disclosure Bias

In the aforementioned analysis, posts from mental health subreddits and those including mental health terms were excluded. Nonetheless, individuals within each of the depression groups for the Reddit datasets displayed language that was unambiguously associated with seeking support or sharing personal experience with mental health issues. Accordingly, we hypothesize that existing filters are unable to remove confounds in individuals who disclose a depression diagnosis on Reddit.

Methods. To measure this effect, we examine differences in the distribution of subreddits that individuals in the depression group of the Topic-Restricted Text data post in relative to individuals in the control group. Specifically, we fit a logistic regression model mapping the subreddit distribution of individuals' posts to their mental health status after applying each subreddit filter list (e.g. RSDD, SMHD, Ours). We compare predictive performance of these models and the learned coefficient weights to understand the effect of filtering. As a baseline, we maintain posts from the r/depression subreddit in the feature set. Then, in sequence of coverage from least to most, we apply subreddit filters from RSDD, SMHD, and our study, and measure classification performance. For each filter, we examine the learned coefficient weights to develop a sense for the personality and interests of individuals in the depression group.

Results. The baseline F1 score in the development set maxes out at 0.83, representing the fact that several individuals in the control group had posted in the r/depression subreddit at some point in their history, but were not labeled as having depression due to the sole use of recent original posts by the automatic annotation procedure. Performance degrades with the expansion of excluded subreddits from each filter, settling at an F1 of 0.72. Coefficients from the model highlight subreddits related to themes of sexuality (r/bisexual, r/actuallesbians), gender (r/ftm), personality (r/introvert, r/INFP), drugs (r/Trees, r/LSD), and relationships (r/MakeNewFriendsHere, r/BreakUps) as being predictive of depression.

The strong classification performance achieved after our filtering measures is evidence that distributional differences in online interaction remain in the "cleaned" Topic-Restricted Text dataset. As our subreddit list is more robust than both the RSDD and SMHD lists, there is reason to believe similar confounds exist in these datasets. The coefficient analysis provides a window into the types of themes that could incorrectly confuse a classification model during generalization attempts.

7 Recommendations

We have demonstrated that issues of transfer loss persist in the mental health space, at least for the proxy-based social media datasets considered in our study. Importantly, we identified confounds that emerge as a result of each dataset's respective design. Critically, existing datasets have flaws that make them difficult to use for constructing models for new data types and populations.

Topical Alignment. *Researchers must account for self-disclosure bias and confounds of personality when curating new datasets.* First discussed in §6.2, models trained on the Reddit datasets learn dependencies between support-driven topics, such as medication usage and relationship advice, and

depression. In contrast, models trained on the Twitter datasets identify the same correlations between sexuality, gender, and depression that Reddit-based models detect, but also learn about the recreational outlets (i.e. fantasy) and social concerns (i.e. racism, sexism) common in depressed individuals.

We hypothesize that semantic divergences reflect self-disclosure bias and differences in platform interaction patterns (Malik et al., 2015; Shelton et al., 2015). Twitter's status- and reply-based structure serves as a place for individuals to share personal thoughts and experiences in reaction to their daily life. Meanwhile, Reddit's community-based forums require active engagement with specific topics and may silo individuals who wish to discuss their mental health beyond defined areas. The latter gains support from our analysis of subreddit distributions in the Topic-Restricted Text data (§6.4).

Topical nuances in language may appropriately reflect elements of identity associated with mental health disorders (i.e. traumatic experiences, coping mechanisms). However, if not contextualized during model training, this type of signal has the potential to raise several false alarms upon application to new populations. Accordingly, we urge researchers to minimize the presence of overt topical disparities between classes in their datasets.

Mitigating Temporal Artifacts. *Researchers must take steps to remove temporal artifacts in new datasets.* Experiments conducted in §5.2 reveal that group-based temporal alignment and latency between model training and deployment can have a significant effect on predictive performance. Variability of performance over time is surprising, as there is no clinical evidence to suggest that the underlying symptoms of depression (on a population level) change over time (APA, 2013).

We hypothesize two reasons for this observation. First, since depression presents in an episodic manner, we may expect data closest to the date of annotation to be the most predictive of an individual's labeled mental status (Melartin et al., 2004). If most posts used for annotation occurred in recent time windows, then it is possible that content in older posts is less relevant to the depressive state of individuals in our data sets. Second, and more problematic, is the possibility that signal used by our classifiers is only a spurious correlation.

At a bare minimum, our results highlight the importance of sampling classification groups so that post volume is equal over time. Discrepancies may wrongly suggest that temporal artifacts are useful for detecting mental health disorders. Going further, researchers should remove temporally-specific references and minimize highly-dynamic language in their datasets. Avenues for accomplishing the latter include using NER to redact n-grams that serve as spurious correlations (Ritter et al., 2011) and leveraging adversarial training to evaluate the degree to which mental health signal may be learned without a notion for time (Tzeng et al., 2017).

8 Limitations and Future Work

Though our study provides a robust perspective toward understanding generalization capabilities of mental health classifiers for social media, we recognize that more learning opportunities exist. Our study only considers a handful of datasets, two platforms, a single mental health disorder, and homogeneous annotation mechanisms. Still unexplored, in large part due to the precautions necessary for securing sensitive mental health data, is how well models trained on data from actual clinical populations generalize to proxy-based datasets and other clinical populations. While high co-morbidity rates between depression and other mental health disorders may allow us to infer model behavior for alternative conditions, we also recognize that presentations of different psychiatric disorders can be quite variable and warrant their own research (Benton et al., 2017b; Arseniev-Koehler et al., 2018).

Another limitation in our work is the lack of depression to control group matches from original reference material. Preoţiuc-Pietro et al. (2015) and De Choudhury et al. (2017) demonstrate that mental health disorders such as depression can have variable presentations based on demographic attributes. The attributes used to construct our Twitter datasets originally were inferred via now-outdated text-based models. Accordingly, demographic inference errors may be propagated to and correlated with depression classification errors. Moreover, these attributes were not considered within the construction of any of the Reddit datasets we explored. The effect of demographics on generalization remains a valuable insight for future exploration.

Finally, our attempts at domain transfer are constrained. Namely, we do not invoke explicit domain adaptation methods (Peng and Dredze, 2017; Li et al., 2018; Huang and Paul, 2019). Moving forward, we plan to explore algorithmic strategies to mitigate the biases discovered in this study.

References

Omar Abdel-Rahman. 2019. Socioeconomic predictors of suicide risk among cancer patients in the united states: A population-based study. *Cancer epidemiology*, 63:101601.

Silvio Amir, Mark Dredze, and John W Ayers. 2019. Mental health surveillance over social media with digital cohorts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 114–120.

American Psychiatric Association APA. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Alina Arseniev-Koehler, Sharon Mozgai, and Stefan Scherer. 2018. What type of happiness are you looking for?-a closer look at detecting mental health from language. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 1–12.

Ngo Xuan Bach, Vu Thanh Hai, and Tu Minh Phuong. 2016. Cross-domain sentiment classification with word embeddings and canonical correlation analysis. In *Proceedings of the Seventh Symposium on Information and Communication Technology*, pages 159–166.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017a. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017b. Multitask learning for mental health conditions with limited social media data. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics.

Joel A Blanco and Lynn A Barnett. 2014. The effects of depression on leisure: Varying relationships between enjoyment, sociability, participation, and desired outcomes in college students. *Leisure Sciences*, 36(5):458–478.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128.

Ermal Bojdani, Aishwarya Rajagopalan, Anderson Chen, Priya Gearin, William Olcott, Vikram Shankar, Alesia Cloutier, Haley Solomon, Nida Z Naqvir, Nicolas Batty, et al. 2020. Covid-19 pandemic: Impact on psychiatric care in the united states, a review. *Psychiatry Research*, page 113069.

Lucy Bowes, Rebecca Carnegie, Rebecca Pearson, Becky Mars, Lucy Biddle, Barbara Maughan, Glyn Lewis, Charles Fernyhough, and Jon Heron. 2015. Risk of depression and self-harm in teenagers identifying with goth subculture: a longitudinal cohort study. *The Lancet Psychiatry*, 2(9):793–800.

Igor Brigadir, Derek Greene, and Pádraig Cunningham. 2015. Analyzing discourse communities with distributional semantic models. In *Proceedings of the ACM Web Science Conference*, pages 1–10.

Stephanie L Budge, Jill L Adelson, and Kimberly AS Howard. 2013. Anxiety and depression in transgender individuals: the roles of transition status, loss, social support, and coping. *Journal of consulting and clinical psychology*, 81(3):545.

Patricia A Cavazos-Rehg, Melissa J Krauss, Shaina Sowles, Sarah Connolly, Carlos Rosas, Meghana Bharadwaj, and Laura J Bierut. 2016. A content analysis of depression-related tweets. *Computers in human behavior*, 54:351–357.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):1–11.

Hau-wen Chang, Dongwon Lee, Mohammed Eltaher, and Jeongkyu Lee. 2012. @ phillies tweeting from philly? predicting twitter user locations with spatial word usage. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 111–118. IEEE.

Di Chen and Carla P Gomes. 2019. Bias reduction via end-to-end shift learning: Application to citizen science. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 493–500.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319.

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. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497.

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.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

Hal Daume III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

Munmun De Choudhury, Sanket S Sharma, Tomaz Logar, Wouter Eekhout, and René Clausen Nielsen. 2017. Gender and cross-cultural differences in social media disclosures of mental illness. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 353–369.

Orianna DeMasi, Konrad Kording, and Benjamin Recht. 2017. Meaningless comparisons lead to false optimism in medical machine learning. *PloS one*, 12(9):e0184604.

Mark Dredze, Miles Osborne, and Prabhanjan Kambadur. 2016. Geolocation for twitter: Timing matters. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1064–1069.

Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Prasadith Kirinde Gamaarachchige and Diana Inkpen. 2019. Multi-task, multi-channel, multi-input learning for mental illness detection using social media text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 54–64.

Keith Hawton, Carolina Casañas i Comabella, Camilla Haw, and Kate Saunders. 2013. Risk factors for suicide in individuals with depression: a systematic review. *Journal of affective disorders*, 147(1-3):17–28.

Chin-Lan Huang, Cindy K Chung, Natalie Hui, Yi-Cheng Lin, Yi-Tai Seih, Ben CP Lam, Wei-Chuan Chen, Michael H Bond, and James W Pennebaker. 2012. The development of the chinese linguistic inquiry and word count dictionary. *Chinese Journal of Psychology*.

Xiaolei Huang and Michael Paul. 2018. Examining temporality in document classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 694–699.

Xiaolei Huang and Michael Paul. 2019. Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4113–4123.

Muhammad Imran, Prasenjit Mitra, and Jaideep Srivastava. 2016. Cross-language domain adaptation for classifying crisis-related short messages. In *13th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2016*. Information Systems for Crisis Response and Management, ISCRAM.

Molly Ireland and Micah Iserman. 2018. Within and between-person differences in language used across anxiety support and neutral reddit communities. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 182–193.

Dan Iter, Jong Yoon, and Dan Jurafsky. 2018. Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 136–146.

Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

William L Jarrold, Bart Peintner, Eric Yeh, Ruth Krasnow, Harold S Javitz, and Gary E Swan. 2010. Language analytics for assessing brain health: Cognitive impairment, depression and pre-symptomatic

alzheimer's disease. In *International Conference on Brain Informatics*, pages 299–307. Springer.

Jing Jiang. 2008. A literature survey on domain adaptation of statistical classifiers. *URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey*, 3:1–12.

Sandra Just, Erik Haegert, Nora Kořánová, Anna-Lena Bröcker, Ivan Nenchev, Jakob Funcke, Christiane Montag, and Manfred Stede. 2019. Coherence models in schizophrenia. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 126–136.

Efsun Sarioglu Kayi, Mona Diab, Luca Pauselli, Michael Compton, and Glen Coppersmith. 2017. Predictive linguistic features of schizophrenia. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 241–250.

Ronald C Kessler. 2003. Epidemiology of women and depression. *Journal of affective disorders*, 74(1):5–13.

Hongmin Li, Doina Caragea, Cornelia Caragea, and Nic Herndon. 2018. Disaster response aided by tweet classification with a domain adaptation approach. *Journal of Contingencies and Crisis Management*, 26(1):16–27.

Tom Lippincott and Annabelle Carrell. 2018. Observational comparison of geo-tagged and randomly-drawn tweets. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 50–55.

Tong Liu, Qijin Cheng, Christopher M Homan, and Vincent Silenzio. 2017. Learning from various labeling strategies for suicide-related messages on social media: An experimental study. *arXiv preprint arXiv:1701.08796*.

Momin M Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. 2015. Population bias in geotagged tweets. In *Ninth international AAAI conference on web and social media*.

Tarja K Melartin, Heikki J Rytsala, Ulla S Leskela, Paula S Lestela-Mielonen, T Petteri Sokero, and Erkki T Isometsa. 2004. Severity and comorbidity predict episode duration and recurrence of dsm-iv major depressive disorder. *Journal of Clinical Psychiatry*, 65(6):810–819.

Alex J Mitchell, Amol Vaze, and Sanjay Rao. 2009. Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet*, 374(9690):609–619.

Michelle Morales, Stefan Scherer, and Rivka Levitan. 2017. A cross-modal review of indicators for depression detection systems. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pages 1–12.

Danielle L Mowery, Y Albert Park, Craig Bryan, and Mike Conway. 2016. Towards automatically classifying depressive symptoms from twitter data for population health. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 182–191.

Soichiro Murakami, Makoto Morishita, Tsutomu Hirao, and Masaaki Nagata. 2019. Ntt's machine translation systems for wmt19 robustness task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 544–551.

Brendan O'Connor, Michel Krieger, and David Ahn. 2010. Tweetmotif: Exploratory search and topic summarization for twitter. In *Fourth International AAAI Conference on Weblogs and Social Media*.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97.

World Health Organization et al. 2017. Depression and other common mental disorders: global health estimates. Technical report, World Health Organization.

Anusri Pampari and Stefano Ermon. 2020. Unsupervised calibration under covariate shift. *arXiv preprint arXiv:2006.16405*.

Minsu Park, Chiyoung Cha, and Meeyoung Cha. 2012. Depressive moods of users portrayed in twitter. In *Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)*, volume 2012, pages 1–8.

Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. 2020. Calibrated prediction with covariate shift via unsupervised domain adaptation. *Statistics (AISTATS)*.

Nanyun Peng and Mark Dredze. 2017. Multi-task domain adaptation for sequence tagging. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 91–100.

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

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

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.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142. Piscataway, NJ.

Daniel Rieman, Kokil Jaidka, H Andrew Schwartz, and Lyle Ungar. 2017. Domain adaptation from user-level facebook models to county-level twitter predictions. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 764–773.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1524–1534. Association for Computational Linguistics.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

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.

Joan Serra, Ilias Leontiadis, Dimitris Spathis, Gianluca Stringhini, Jeremy Blackburn, and Athena Vakali. 2017. Class-based prediction errors to detect hate speech with out-of-vocabulary words. In *Proceedings of the First Workshop on Abusive Language Online*, pages 36–40.

Martin Shelton, Katherine Lo, and Bonnie Nardi. 2015. Online media forums as separate social lives: A qualitative study of disclosure within and beyond reddit. *iConference 2015 Proceedings*.

Tiancheng Shen, Jia Jia, Guangyao Shen, Fuli Feng, Xiangnan He, Huanbo Luan, Jie Tang, Thanassis Tiropanis, Tat Seng Chua, and Wendy Hall. 2018. Cross-domain depression detection via harvesting social media. International Joint Conferences on Artificial Intelligence.

Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36.

Hoyun Song, Jinseon You, and Jin-Woo Chung Jong C Park. 2018. Feature attention network: Interpretable depression detection from social media. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*.

Petar Stojanov, Ahmed Hassan Awadallah, Paul Bennett, and Saghar Hosseini. 2019. On domain transfer when predicting intent in text.

Adarsh Subbaswamy and Suchi Saria. 2020. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Adam Tsakalidis, Maria Liakata, Theo Damoulas, and Alexandra I Cristea. 2018. Can we assess mental health through social media and smart devices? addressing bias in methodology and evaluation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 407–423. Springer.

Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A reddit dataset for stress analysis in social media. *EMNLP-IJCNLP 2019*, page 97.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176.

Nikhita Vedula and Srinivasan Parthasarathy. 2017. Emotional and linguistic cues of depression from social media. In *Proceedings of the 2017 International Conference on Digital Health*, pages 127–136.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488.

Genta Indra Winata, Onno Pepijn Kampman, and Pascale Fung. 2018. Attention-based lstm for psychological stress detection from spoken language using distant supervision. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6204–6208. IEEE.

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.

Fangzhao Wu and Yongfeng Huang. 2016. Sentiment domain adaptation with multiple sources. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 301–310.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978.

A Data

CLPsych 2015 Shared Task. This Twitter dataset was constructed using regular expressions matching phrases similar to "I was diagnosed with depression" by Coppersmith et al. (2015b). The authors manually verified the authenticity of each candidate self-disclosure and then sampled an age- and gender-matched "control" population using tweet-based inferences (Schwartz et al., 2013). To approximate the depression-control pairs in the original dataset, which has since been anonymized, we sampled from the full set of available control group candidates based on their inferred demographics.

Multi-Task Learning. Compiled by Benton et al. in 2017, this Twitter dataset for multiple mental health disorders was constructed in the same manner as the CLPsych 2015 Shared Task.[5] Although depression-control linkages remain in our version of the dataset, we only use them to isolate an appropriate control group for the depression group. Individuals who were annotated as part of both the Multi-Task Learning and the CLPsych 2015 Shared Task data were removed from the CLPsych data (55 depression, 0 control).

RSDD. The Reddit Self-disclosed Depression Diagnosis (RSDD) dataset is a Reddit-based data asset in which individuals who self-disclosed they were living with depression were identified via regular expressions and manually verified much like the two aforementioned Twitter datasets (Yates et al., 2017). Individuals selected for the control group were required not to have posted in a list of 24 mental health related subreddits or to have used any of 19 mental health terms.

To align the theme of language generated by individuals across classification groups, each individual in the depression group was greedily matched with 12 individuals from the candidate control pool based on Hellinger distance between each individual's post distribution over subreddits. To preserve privacy of individuals within the dataset, usernames were anonymized and post metadata was redacted. Accordingly, linkages between each individual within the depression group and their respective control group pairs could not be recreated.

SMHD. The Self-Reported Mental Health Diagnoses (SMHD) dataset was constructed in a similar manner as RSDD, albeit being expanded to support 9 conditions, leverage more precise regular expressions, and abide by a more conservative term/subreddit filter set (Cohan et al., 2018). As with RSDD, linkages between individuals in the depression group and their controls were not preserved in our version of the dataset nor could they be readily reproduced. A substantial portion of individuals in SMHD are also part of RSDD; for this reason, we refrain from conducting domain transfer experiments between the two datasets.

Topic-Restricted Text. To expand the scope of our analysis, we follow methods described in Wolohan et al. (2018) to curate an additional Reddit dataset in which annotations are assigned based on community participation and explicit mental health signal is removed (hence "topic-restricted text"). Per the original paper, individuals who initiated one of 10k recent posts in r/depression were considered members of the depression group, while individuals who initiated one of 10k recent posts in r/AskReddit (but not in the recent r/depression query) were considered to be members of the control group. Due to the anonymous nature of the RSDD and SMHD datasets, we were unable to determine if any individuals found within the Topic-Restricted Text dataset were also in RSDD or SMHD.

B Temporal Filtering

To limit the introduction of temporal artifacts into the classification process, all datasets were truncated in time so that at least 100 unique data points (e.g. Tweets, Reddit comments) were present in the first and final month across individuals in both

[5] While we were able to reproduce the class distributions of the dataset described in Benton et al. (2017b), we identified discrepancies between the dates that tweets in this version of the dataset were posted relative to the dates that the original component datasets were published.

classes. Date ranges selected based on this criteria are presented in Table 1.

C Tokenization

To maintain our ability to interpret results consistently, the same preprocessing pipeline was applied across all datasets. Text within both Tweets and Reddit comments was tokenized using a modified version of the Twokenizer (O'Connor et al., 2010). English contractions were expanded, while specific retweet tokens, username mentions, URLs, and numeric values were replaced by generic tokens. As pronoun usage tends to differ in individuals living with depression (Vedula and Parthasarathy, 2017), we removed any English pronouns from our stop word set.[6] Case was standardized across all tokens, with a single flag included if an entire post was made in uppercase letters.

D Features

Text from all documents for an individual are concatenated together and tokenized using the procedure described in Appendix C. The vocabulary of each training procedure is fixed to a maximum of 100-thousand unigrams selected based on KL-divergence of the class-unigram distribution with the class-distribution of stop words (Chang et al., 2012). This reduced bag-of-words representation is then used to generate the following additional feature dimensions: a 50-dimensional LDA topic distribution (Blei et al., 2003), a 64-dimensional LIWC category distribution (Tausczik and Pennebaker, 2010), and a 200-dimensional mean-pooled vector of pretrained GloVe embeddings (Pennington et al., 2014). The reduced bag-of-words representation is transformed using TF-IDF weighting (Ramos et al., 2003).[7]

E Hyperparameter Selection

Each model is trained using a hyperparameter grid search over the regularization strengths {1e-3, 1e-2, 1-e1, 1, 10, 100, 1e3, 1e4, 1e5}. Hyperparameters were selected to maximize F1 score within the development splits of each dataset.

[6] English Stop Words (nltk.org)

[7] All data-specific feature transformations (e.g. LDA, TF-IDF) are learned without access to development or test data. We use Scikit-learn implementations of LDA and TF-IDF.