Robust User Representations for Cross Domain Depression Classification using Partially-labeled Latent Dirichlet Allocation

Anonymous NAACL-HLT 2021 submission

Abstract

Recent studies have suggested that existing methods for detecting psychiatric conditions based on language usage in social media fail to generalize to new platforms and populations. In this study, we attempt to improve generalization by learning domain-invariant feature representations for depression classification using Partially-labeled Latent Dirichlet Allocation (PLDA). Experiments using four benchmarks from Twitter and Reddit provide an understanding of PLDA's effectiveness at representing users compared to vanilla Latent Dirichlet Allocation (LDA).

1 Introduction

An estimated 16.2% of individuals will experience at least one major depressive episode during their lifetime (Kessler et al., 2003; Brody et al., 2018). Unfortunately, stigma towards discussing mental health and overwhelming socioeconomic barriers continue to drive a substantial portion of this population away from receiving adequate care (Gary, 2005). Accordingly, there has been significant interest by clinicians and social workers alike to develop non-invasive tools that increase equity in mental health care, improve the efficiency of clinical treatment, and effectively monitor mental well-being at scale (Chen et al., 2019; Galea et al., 2020).

In response to this demand, researchers have devoted significant effort towards developing systems that measure mental health using non-clinical data sources (De Choudhury et al., 2013; Jaques et al., 2015). Offering large amounts of user-generated language, social media platforms have served as the most prolific resource for researchers to translate theory regarding language usage and mental health into the digital world (Ramirez-Esparza et al., 2008; Guntuku et al., 2017; Chancellor and De Choudhury, 2020). Unfortunately, several recent studies have provided evidence that models trained on existing social media datasets fail to generalize to new data platforms and populations due to underlying sampling biases (Ernala et al., 2019; Harrigian et al., 2020a; Aguirre et al., 2021).

In this study, we evaluate the use of Partially-labeled Latent Dirichlet Allocation (PLDA) (Ramage et al., 2011) for performing unsupervised domain adaptation in the context of a depression inference task. We hypothesize PLDA's use of label-specific and label-invariant topics will encourage learning of user representations that are robust across domains. To explore this hypothesis, we systematically compare domain transfer abilities of vanilla Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and PLDA using four social media datasets from Twitter and Reddit.

Ethics Statement. This research was deemed exempt from review by our Institutional Review Board (IRB) under 45 CFR § 46.104. All data was analyzed in accordance to protocols for mental health research enumerated in Benton et al. (2017a). Due to the sensitive nature of mental health data, we are unable to share any raw and/or processed forms of our datasets. However, we provide code for replicating our experiments and instructions for accessing datasets from their respective providers.[1]

2 Background

The process of training statistical models that generalize from one distribution to another is referred to as *domain adaptation* (Ben-David et al., 2006; Kouw and Loog, 2019). Historically, research regarding generalization in models of mental health based on social media has been quite narrowly focused. Several researchers have explored language usage amongst individuals with depression as a function of the subreddits they posted comments within (De Choudhury and De, 2014; Ireland and Iserman, 2018; Wolohan et al., 2018). Others have evaluated the feasibility of detecting depression as

[1] https://link-deanonymized-after-review

a function of the manner in which control groups are defined (Pirina and Çöltekin, 2018).

Shen et al. (2018) was the first study to explicitly attempt transfer across social media platforms, ultimately suggesting that data from Twitter could be used to inform the prior for models deployed on Sina Weibo. However, later work from Ernala et al. (2019) and Harrigian et al. (2020a) contradicted the optimism of these findings by providing evidence that models of schizophrenia and depression, respectively, transferred poorly between social media platforms as a result of underlying dataset biases. To date, no study has explicitly attempted to reduce error when transferring mental health classifiers from one platform or population to another.

Fortunately, the natural language processing (NLP) community has a rich foundation of domain adaptation research for non-mental-health applications (Blitzer et al., 2007; Daumé III, 2007; Dredze and Crammer, 2008; Daumé III, 2009). Broadly, these adaptation techniques can be categorized as being either feature-based — e.g. subspace mapping (Fernando et al., 2013; Sun and Saenko, 2015), invariant feature selection (Zhao et al., 2019) — or model-based — e.g. instance weighting (Jiang and Zhai, 2007; Wang et al., 2017), adversarial learning (Tzeng et al., 2017; Meng et al., 2018). Unfortunately, many existing methods are poorly suited for application in the mental health domain, either due to significant hyperparameter sensitivity (Plank et al., 2014; Xia et al., 2018) or non-interpretable transformations (Saria and Subbaswamy, 2019).

Bayesian adaptation methods provide a natural framework for sharing covariance between domains and have been used successfully in various modeling tasks (Finkel and Manning, 2009; Daumé III, 2009). Topic models have had a particularly strong adoption rate with adaptation-focused NLP research, due both to their performance and their generation of interpretable representations (Blei and Lafferty, 2009). For example, Chen et al. (2012) used topics learned using Google search results to contextualize latent topics in microblog messages (i.e. Twitter, Sina Weibo), while Yang et al. (2019) presented a topic model that jointly aligns concepts across multiple language domains to improve document classification. Serving as the foundation for our work, Bao et al. (2013) and Jing et al. (2018) constructed partially-supervised topic models that learn to distinguish between domain-specific and domain-invariant semantics.

Importantly, topic models are familiar to those working on mental-health related applications of NLP and thus present an increased chance of clinical adoption. For instance, Resnik et al. (2013, 2015b,a) demonstrated that LDA and its variants could be used to discover themes discussed by individuals with heightened levels of depression, neuroticism, and post traumatic stress disorder. Topic models have also been used to identify suicidal ideation (Huang et al., 2015, 2017) and monitor population-level well-being (Biester et al., 2020).

3 Methods

We begin by introducing the topic models that will be used for generating user representations for mental health inference. Thereafter, we describe our downstream classifier and its associated training procedure.

3.1 Topic Modeling

Topic models can be seen as a Bayesian generative extension to the matrix factorization procedure Latent Semantic Analysis, offering a mechanism for representing the semantics of a document in a low-dimensional hyperspace (Blei et al., 2003; Blei and Lafferty, 2009). In this study, we compare two formulations — "vanilla" Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Partially-labeled Latent Dirichlet Allocation (PLDA) (Ramage et al., 2010, 2011).[2]

LDA. LDA was first presented by Blei et al. (2003) and has since been cited in over 30-thousand studies across a wide array of disciplines. We present a plate diagram for the sampling model in Figure 1 (left). We are interested in learning the document-topic distributions θ to use as user representations in the downstream inference task.

PLDA. While LDA has proven to be a powerful tool for representing documents in a low-dimensional hyperspace (Henderson and Eliassi-Rad, 2009), its unsupervised nature limits the degree to which researchers can take advantage of rich metadata that is often available with language data. Partially-labeled Latent Dirichlet Allocation (PLDA) was proposed by Ramage et al. (2010, 2011) as an extension of LDA that could use annotations to inform, but not fully restrict, the learning of topic distributions in a corpus. We present PLDA's plate diagram in Figure 1 (right).

[2] We use versions of LDA and PLDA implemented in the `tomotopy` Python package.

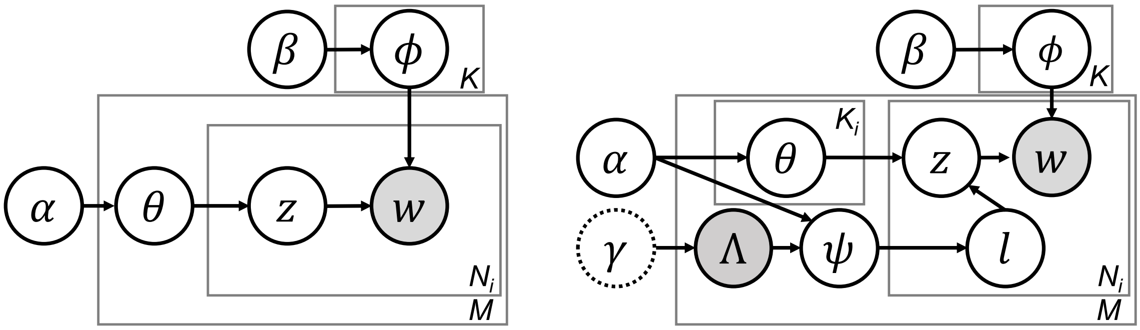

Figure 1: Comparison of LDA (left) and PLDA (right) sampling models. Shaded variables are observed. If we assume only latent topics (i.e. no label-specific topics) are present, the PLDA model reduces to LDA.

PLDA is structured similarly to LDA, with the primary difference being that documents are generated from both latent topics general to the entire training corpora and a subset of latent topics specific to the labels associated with each document. For the purposes of facilitating adaptation, we assign labels indicating whether a document comes from the source or target domain. Importantly, we do not require any ground truth mental health status in the target domain at training time, which can be difficult or even impossible to acquire for some populations. Since PLDA assumes topics are each associated with a single label, we are restricted to representing documents in our downstream task using the general latent topics only. For consistency between LDA and PLDA, we re-normalize each document's general latent topic distribution using an l_2 norm before feeding them into the downstream classifier.

3.2 Mental Health Inference

We feed user representations inferred from each topic model into an l_2-regularized logistic regression classifier whose parameters are learned using the Scikit-learn implementation of LBFGS (Liu and Nocedal, 1989; Pedregosa et al., 2011). Indeed, logistic regression has served as an adequate baseline in several prior studies regarding inference of mental health status (Benton et al., 2017b; Cohan et al., 2018). Furthermore, we argue that the model's relative simplicity will allow us to more acutely estimate differences in user representation quality that arise between LDA and PLDA.

4 Data

We choose to focus on the task of depression inference for this study, which remains one of the most popular tasks in this application domain (Chancellor and De Choudhury, 2020; Harrigian et al., 2020b). To evaluate our hypothesis that PLDA can be used effectively for performing unsupervised domain adaptation, we consider four social media datasets containing ground-truth depression status. Specifically, we consider two Twitter datasets — *2015 CLPsych Shared Task* (Coppersmith et al., 2015) and *Multi-task Learning* (Benton et al., 2017b) — and two Reddit datasets — *SMHD* (Cohan et al., 2018) and *Topic-Restricted Text* (Wolohan et al., 2018). Full descriptions of each dataset are provided in the appendix as a courtesy to the reader.

To form "documents" for modeling, all posts made by an individual are concatenated together and tokenized into unigrams using a modified version of Twokenizer (O'Connor et al., 2010). Per the recommendations of Schofield et al. (2017), we remove the 250 most frequent tokens from each corpus. We also ignore tokens used by less than 10 users throughout the training corpus. We preserve existing train/development/test splits for our analysis, but downsample each subset so that positive and negative instances of depression are balanced. We recognize this sampling decision does not reflect the true prevalence of depression amongst the population, but it facilitates comparisons to results from Harrigian et al. (2020a) and is arguably sufficient for comparing performance between PLDA and LDA.

5 Experiments

In line with prior work from Harrigian et al. (2020a), we consider a standard domain transfer experimental design in which we train a mental health classifier using one dataset and evaluate on another. At training time, we use data both from the source and target domains (training subsets only) to fit a

Source Domain	Target Domain: CLPsych Shared Task	Target Domain: Multi-Task Learning	Target Domain: SMHD	Target Domain: Topic-Restricted Text
CLPsych Shared Task		- 0.002 ± 0.014	**- 0.011 ± 0.005**	**0.016 ± 0.006**
Multi-Task Learning	- 0.013 ± 0.035		**- 0.010 ± 0.006**	**0.018 ± 0.011**
SMHD	**0.044 ± 0.026**	- 0.002 ± 0.021		- 0.003 ± 0.004
Topic-Restricted Text	0.029 ± 0.051	**0.013 ± 0.006**	0.001 ± 0.001	

Table 1: Difference in test set AUC between LDA and PLDA ($\mu \pm \sigma$). Positive deltas indicate PLDA has superior performance. Bolded values indicate a significant difference based on a pairwise t-test.

topic model. After a burn-in period of 1000 MCMC iterations, we generate 1000 additional samples from the topic model's posterior predictive distribution for users in the source domain's training data and the target domain's evaluation data. Based on guidance from Nguyen et al. (2014), final user representations are computed by averaging over the 1000 post burn-in samples and re-normalizing using an l_2 norm.

To most fairly compare user representations learned by LDA and PLDA, we first conduct a multi-stage hyperparameter search for each topic model and their respective downstream classifiers. The first stage consists of a joint search over the prior hyperparameters α and β, and the classifier's regularization strength C. The second stage consists of a search over the number of latent topics K, the number of per-domain topics K_d (PLDA only), and regularization strength C. At each stage, we select parameters that maximize area under the curve (AUC) in the target domain's validation data. Our hyperparameter search space and resulting parameter selections are enumerated in the appendix. Using these hyperparameters, we train 5 independent models per source-target combination using a stratified sample (80%) of the combined training and development subsets. We choose AUC as our primary performance metric since it is robust to model calibration errors that may arise under covariate shift (Pampari and Ermon, 2020; Park et al., 2020). We use a paired t-test to evaluate the significance of differences in performance that arise under the two types of user representations.

Results. From our hyperparameter optimization procedure, we observe that inference performance is relatively sensitive to topic model parameterization for both LDA and PLDA models. The number of general latent topics K and domain-specific topics K_d has the strongest impact on performance, though these relationships are non-linear and not necessarily correlated between topic models. This sensitivity may suggest that nonparametric topic models are more appropriate for an unsupervised adaptation task (Teh et al., 2006), though we leave this hypothesis for future work.

We present the average difference in test set AUC ($\mu \pm \sigma$) that arises under our two "optimal" topic model user representations in Table 1. We note that user representations derived using PLDA are *not* uniformly superior to user representations using LDA. In general, differences in performance are relatively minor for most source-target comparisons. That said, we note that all significant differences arise in cross-platform transfer scenarios (as opposed to cross-dataset, same platform).

6 Discussion

In this paper, we evaluated the effectiveness of using Partially-labeled Latent Dirichlet Allocation (PLDA) to construct user representations for cross-domain mental health status inference. Specifically, we showed that PLDA is able to separate domain-specific topics from domain-invariant topics and thus promote generalization under certain data regimes. However, PLDA-based user representations are not uniformly superior for all adaptation scenarios and, furthermore, are non-trivially sensitive to hyperparameter choices.

We would be negligent not to recognize limitations of our study. We have only explored transfer between datasets with a prevalence of depressed individuals that is non-representative of the general population (e.g. 50/50 group balance). Moving forward, researchers should consider fully random samples of data from the desired target domain to most accurately reflect a deployment scenario. It is also worth more thoroughly evaluating the impact that factors such as target domain data availability, demographic representation, and the degree of covariate shift have on performance.

References

Carlos Aguirre, Keith Harrigian, and Mark Dredze. 2021. Gender and racial fairness in depression research using social media. In *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Yang Bao, Nigel Collier, and Anindya Datta. 2013. A partially supervised cross-collection topic model for cross-domain text classification. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 239–248.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137–144.

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. Multi-task learning for mental health using social media text. *arXiv preprint arXiv:1712.03538*.

Laura Biester, Katie Matton, Janarthanan Rajendran, Emily Mower Provost, and Rada Mihalcea. 2020. Quantifying the effects of covid-19 on mental health support forums. *arXiv preprint arXiv:2009.04008*.

David M Blei and John D Lafferty. 2009. Topic models. *Text mining: classification, clustering, and applications*, 10(71):34.

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.

Debra J Brody, Laura A Pratt, and Jeffery P Hughes. 2018. *Prevalence of depression among adults aged 20 and over: United States, 2013-2016*. US Department of Health and Human Services, Centers for Disease Control and Prevention.

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.

Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. 2019. Can ai help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21(2):167–179.

Yan Chen, Zhoujun Li, Liqiang Nie, Xia Hu, Xiangyu Wang, Tat-seng Chua, and Xiaoming Zhang. 2012. A semi-supervised bayesian network model for microblog topic classification. In *Proceedings of COLING 2012*, pages 561–576.

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, 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.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Hal Daumé III. 2009. Bayesian multitask learning with latent hierarchies. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 135–142.

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. *Icwsm*, 13:1–10.

Mark Dredze and Koby Crammer. 2008. Online methods for multi-domain learning and adaptation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 689–697.

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.

Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. 2013. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pages 2960–2967.

Jenny Rose Finkel and Christopher D Manning. 2009. Hierarchical bayesian domain adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 602–610.

Sandro Galea, Raina M Merchant, and Nicole Lurie. 2020. The mental health consequences of covid-19 and physical distancing: The need for prevention and early intervention. *JAMA internal medicine*, 180(6):817–818.

Faye A Gary. 2005. Stigma: Barrier to mental health care among ethnic minorities. *Issues in mental health nursing*, 26(10):979–999.

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.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020a. 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.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020b. On the state of social media data for mental health research. *arXiv preprint arXiv:2011.05233*.

Keith Henderson and Tina Eliassi-Rad. 2009. Applying latent dirichlet allocation to group discovery in large graphs. In *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 1456–1461.

Xiaolei Huang, Xin Li, Tianli Liu, David Chiu, Tingshao Zhu, and Lei Zhang. 2015. Topic model for identifying suicidal ideation in chinese microblog. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 553–562.

Xiaolei Huang, Linzi Xing, Jed R Brubaker, and Michael J Paul. 2017. Exploring timelines of confirmed suicide incidents through social media. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 470–477. IEEE.

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.

Natasha Jaques, Sara Taylor, Asaph Azaria, Asma Ghandeharioun, Akane Sano, and Rosalind Picard. 2015. Predicting students' happiness from physiology, phone, mobility, and behavioral data. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 222–228. IEEE.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 264–271.

Baoyu Jing, Chenwei Lu, Deqing Wang, Fuzhen Zhuang, and Cheng Niu. 2018. Cross-domain labeled lda for cross-domain text classification. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 187–196. IEEE.

Ronald C Kessler, Patricia Berglund, Olga Demler, Robert Jin, Doreen Koretz, Kathleen R Merikangas, A John Rush, Ellen E Walters, and Philip S Wang. 2003. The epidemiology of major depressive disorder: results from the national comorbidity survey replication (ncs-r). *Jama*, 289(23):3095–3105.

Wouter M Kouw and Marco Loog. 2019. A review of single-source unsupervised domain adaptation. *arXiv preprint arXiv:1901.05335*.

Dong C Liu and Jorge Nocedal. 1989. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528.

Zhong Meng, Jinyu Li, Zhuo Chen, Yang Zhao, Vadim Mazalov, Yifan Gang, and Biing-Hwang Juang. 2018. Speaker-invariant training via adversarial learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5969–5973. IEEE.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2014. Sometimes average is best: The importance of averaging for prediction using mcmc inference in topic modeling. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1752–1757.

Brendan O'Connor, Michel Krieger, and David Ahn. 2010. Tweetmotif: exploratory search and topic summarization for twitter. In *ICWSM*, pages 384–385.

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

Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. 2020. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pages 3219–3229. PMLR.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Inna Pirina and Çağrı Çöltekin. 2018. Identifying depression on reddit: The effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 9–12.

Barbara Plank, Anders Johannsen, and Anders Søgaard. 2014. Importance weighting and unsupervised domain adaptation of pos taggers: a negative result. In

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 968–973.

Daniel Ramage, Susan T Dumais, and Daniel J Liebling. 2010. Characterizing microblogs with topic models. *icwsm*, 10(1):16.

Daniel Ramage, Christopher D Manning, and Susan Dumais. 2011. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–465.

Nairan Ramirez-Esparza, Cindy K Chung, Ewa Kacewicz, and James W Pennebaker. 2008. The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches. In *ICWSM*.

Philip Resnik, William Armstrong, Leonardo Claudino, and Thang Nguyen. 2015a. The university of maryland clpsych 2015 shared task system. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 54–60.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015b. 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.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1348–1353.

Suchi Saria and Adarsh Subbaswamy. 2019. Tutorial: safe and reliable machine learning. *arXiv preprint arXiv:1904.07204*.

Alexandra Schofield, Måns Magnusson, and David Mimno. 2017. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 432–436.

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.

Baochen Sun and Kate Saenko. 2015. Subspace distribution alignment for unsupervised domain adaptation. In *BMVC*, volume 4, pages 24–1.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581.

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.

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.

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.

Rui Xia, Zhenchun Pan, and Feng Xu. 2018. Instance weighting for domain adaptation via trading off sample selection bias and variance. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4489–4495. AAAI Press.

Weiwei Yang, Jordan Boyd-Graber, and Philip Resnik. 2019. A multilingual topic model for learning weighted topic links across corpora with low comparability. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1243–1248.

Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J Gordon. 2019. On learning invariant representation for domain adaptation. *arXiv preprint arXiv:1901.09453*.