Towards Understanding the Role of Gender in Deploying Social Media-Based Mental Health Surveillance Models

Eli Sherman
Johns Hopkins University
esherman@jhu.edu

Keith Harrigian
Johns Hopkins University
kharrigian@jhu.edu

Carlos Aguirre
Johns Hopkins University
caguirr4@jhu.edu

Mark Dredze
Johns Hopkins University
mdredze@cs.jhu.edu

Abstract

Spurred by advances in machine learning and natural language processing, developing social media-based mental health surveillance models has received substantial recent attention. For these models to be maximally useful, it is necessary to understand how they perform on various subgroups, especially those defined in terms of protected characteristics. In this paper we study the relationship between user demographics – focusing on gender – and depression. Considering a population of Reddit users with known genders and depression statuses, we analyze the degree to which depression predictions are subject to biases along gender lines using domain-informed classifiers. We then study our models' parameters to gain qualitative insight into the differences in posting behavior across genders.

1 Introduction

The United States Centers for Disease Control and Prevention estimates that 8% of American adults suffer from major depression at a given time (Brody et al., 2018). This represents a critical public health threat, as depression is associated with downstream physical health complications (Rush, 2007; Alboni et al., 2008) and an increased risk of suicide (Richards and O'Hara, 2014). Among the many efforts to address this crisis is a line of research at the intersection of language modeling, social media analysis, and mental health. Seminal papers by De Choudhury et al. (2013) and Coppersmith et al. (2014) demonstrated the general feasibility of predicting mental health status from social media data.

A major obstacle to the practical use of mental health surveillance models is differential performance for different subgroups of the population. This behavior can arise either because the training data is not sufficiently representative of the population, or because some groups are simply harder to predict given the same data. The former case is well-studied in the machine learning literature and can be addressed by careful data collection and training regimes. The latter case, however, is often more subtle and harder to address. *Not* identifying and addressing these differences in performance degrades the utility of the models. In particular, if the performance is worse for historically marginalized populations it can reinforce existing inequities such as under-diagnosis of depression (Elazar and Goldberg, 2018).

In this work we aim to assess the scope of the differential performance problem by studying the relationship between gender and predictions of depression. The most useful insight we could gain would be determining whether or not gender is a confounder for depression predictions; that is, whether gender both causally affects the way in which users post on Reddit and causally affects our predictions of the user's depression status. Unfortunately, testing whether this causal dynamic is true is very difficult with the purely observational data available to us. Towards testing this phenomena, we will instead test the slightly weaker hypotheses i) that depression predictions exhibit gender bias (i.e., there are differences in performance across genders) and ii) that these differences are due, at least in part, to differing uses of language between men and women in talking about their mental state. Together these hypotheses serve as a sort of associational version of the causal phenomenon we'd like to study. They can tell us whether depression predictions are correlated with gender and whether certain terms are likely to have different meanings based on the gender of the author.

We test hypothesis (i) quantitatively by fitting depression prediction models to a novel data set collected from Reddit with *ground truth* genders, derived from self-disclosures, and comparing the performances across genders. We test hypothesis (ii) qualitatively by looking at features strongly predictive of depression for each gender. We identify themes that are concordant across genders and consistent with the literature (De Choudhury et al., 2016) as well as themes that are discordant across genders and support our hypothesis that men and women use many terms differently to talk about (non-) depression. We follow these analyses with a discussion of open questions that follow from this work. In particular, we discuss the use of causal methodologies to assess our stronger hypothesis that gender confounds depression prediction. We highlight the types of methods that could be used and the data that is necessary to test the causal hypothesis. We conclude with a discussion of limitations and the ethical implications of this work.

Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology, pages 217–223

2 Related Work

Several existing papers have considered the role of demographics in mental health prediction. Elazar and Goldberg (2018) demonstrated that demographics are implicitly encoded in text data. Wood-Doughty et al. (2017) and Loveys et al. (2018) both studied differing language use across cultures. The former used a Twitter data set with *inferred* demographic labels, while the latter used a carefully-curated proprietary data set from 7 Cups of Tea. Amir et al. (2019) explored the role of cohort selection in assessing mental health disorder prevalence. Aguirre et al. (2021) is the closest to the present work. The authors characterized the biases present in depression prediction models by showing there are differences in performance for different demographic subgroups. This work studied biases that arise due to the specific data set used for training,focusing on the popular, publicly available data sets CLPsych (Coppersmith et al., 2015) and MULTITASK (Benton et al., 2017).

The present work differs from those cited in that we seek to quantify demographic bias in depression prediction using self-disclosures in a publicly available data set. This approach improves scalability and reproducibility compared to hand-labeled and proprietary data sets. Additionally, while self-disclosures are not perfect, they are not subject to the same degree of noise and error that is induced when using genders inferred by using a pre-trained model, trained on an auxiliary data source. Our estimates of the depression prediction performance across genders are therefore likely to be of a higher quality. Moreover, our analyses of features that are predictive of depression for each gender are also likely to be less noisy than they would be if we were also inferring genders from those same features.

3 Data Collection

To obtain a dataset with ground truth gender, we mined all posts and comments from the r/AskMen and r/AskWomen subreddits between January 1, 2019 and December 31, 2019 using the Pushshift API (Baumgartner et al., 2020). In total, we collected 251,487 original submissions and 4,481,354 comments.

For each post, we consider the flair – an optional tag users can apply to their posts to reveal information about themselves or the content of their post – to determine the ground truth gender of the post author. We considered the author of a post to be true-male if they used one of 'Male', 'male', 'Dude', or ♂ for their flair, and true-female if they used one of 'Female', 'female', ♀, or ♀♡. Of the mined posts, 1,002,079 had some sort of flair, while 660,684 had one of the male or female indicator flairs. This process yielded a data set of 15,140 unique male and 11,241 unique female users, as well as 59 users whose gender-related flair use was inconsistent (i.e. at least one post each with a male- and female-indicating flair). While people who identify as non-binary are known to have higher rates of depression (Budge et al., 2013; Wolohan et al., 2018) and thus could benefit from the studies like this one, we did not have a reliable method for identifying non-binary users beyond the list of inconsistent users and the sub-population in our cohort was too small to yield meaningful analysis. For the remainder of the paper we restrict attention to binary genders under the folk conception of gender (Larson, 2017).

For each of the 26,381 gender-binary users, we collected the user's entire Reddit posting and commenting history from January 1, 2019 to December 31, 2019, totaling 1,035,782 original submissions and 19,029,981 comments across 64,162 subreddits. Following the literature on social media-driven mental health surveillance (De Choudhury et al., 2013; Yates et al., 2017), we defined a user as true-depressed if they authored an original submission or comment in r/depression during the study period and true-control otherwise. The breakdown of gender and depression classes is 721 and 713 depressed males and females respectively, and 14,416 and 10,526 control males and females respectively. Replication data for this study can be found at `https://github.com/esherma/CLPsych2021_Gender_and_Depression` and is available under a data usage agreement.

4 Methods

We fit user-level models to predict depression status from our harvested Reddit data. To enable analysis of the impact of gender as a confounder, we fit separate models on two separate data sets: a random sample of the true-men users in our data set, and a random sample of the true-women users. To reduce noise induced by 'throwaway' or 'lurker' accounts, we excluded users who made fewer than 5 posts (submissions + comments) during the study period. This decision could reduce our results' generalizability since throwaway accounts may be owned by users with separate primary accounts and post with the throwaway differently (e.g. posting more personal information).

Because depression is a rare outcome in our data, our initial train and test sets had very few depressed individuals (109 train, 26 test). This proved too few to draw meaningful conclusions about the role of gender in depression prediction. We therefore report the performance of our models trained on data sets constructed by performing *balanced sampling* from the full data. The resulting class breakdowns are: 721 and 613 depressed males and females respectively, and 820 and 712 control males and females respectively.

We split each of these sampled data sets 80-20 into train and test sets, stratifying by user. We then constructed a Bag-of-Words (BoW) vocabulary from the submissions and comments for each user in the training sets. We included 1-, 2-, and 3-grams, as well as LIWC (Pennebaker et al., 2007) and TF-IDF (Jones, 1972) features. We imposed that features must be used by a

minimum of 25 users to be included in the vocabulary. We also removed posts from the r/depression subreddit from each user's BoW vector and filtered out terms and subreddits commonly associated with self-disclosure of mental health disorders using the SMHD dataset (Cohan et al., 2018). To model depression, we used the scikit-learn implementation of regularized logistic regression (Pedregosa et al., 2011). At the end of training, we discarded all but the top 100,000 features using the pairwise mutual information criterion as an additional regularization step.

5 Results

5.1 Model Performance

The performance of each model on each test set is shown in Figure 1. The most striking result is that the performance of both models is considerably higher on the men-only test set than on the women-only test set (.770 vs. .702 and .758 vs. .707 respectively). This difference indicates that predicting depression among men is easier than among women. Looking at the distribution graphs, it appears that women are *over diagnosed* as depressed. Mechanically, this difference in predictions likely arises due to the existence of a few key features that indicate depression for one gender but not the other. We identify candidate features in the analysis below.

5.2 Feature Analysis

We extracted the regression coefficients from each of our models and generated a scatter plot in Figure 2 of the 50,967 features the two models had in common. Towards identifying strongly predictive features, we scored each feature using the sum of the absolute value of the coefficient from each model for that feature. In the figure, we labeled the 50 highest-scoring features in each plot quadrant.

Concordant Depression Features (top right) Even though we filtered out self-disclosure tokens (e.g. 'depression' and 'depressed'), we see that many of the most predictive features are consistent with themes discussed in the mental health surveillance literature (De Choudhury et al., 2016): emotion ('feel', LIWC affect, LIWC negemo), physical symptoms of depression ('sleep'), and indicators of social isolation ('alone', 'porn', and personal pronouns 'me', 'my', and 'I'). One notable feature is the token 'jews'. This feature could indicate that many depressed Jewish people of both genders frequently discuss their religious identity on Reddit, possibly in the context of their peoples' historically marginalized status (McCullough and Larson, 1999). Also plausible is that the token is indicative of anti-semitic tendencies which are correlated with depression (e.g. blaming one's personal struggles on a scapegoat minority group). This phenomenon has been documented in the largely-male 'incel' community (Hoffman et al., 2020) but we could not find a clear connection between anti-semitism and depression among women in the psychology or sociology literature.

Concordant Control Features (lower left) These feature themes are also consistent with findings in the literature. Features indicative of social interactions are quite common ('church', 'wedding', 'couple') as well as features that suggest positive affect regarding life activities ('fun', 'cool', LIWC leisure).

Discordant Features (top left, lower right) These features are of primary interest for identifying potential gender-based confounding. Here we find features that are predictive of depression in women but control in men or vice-versa. We observe that there are several terms that likely have different meanings for men and women users. Many of these pertain to social interactions.

For instance, 'gay', 'gay men' and 'my husband' are all strongly predictive of control for men. This suggests that men who are comfortable discussing non-straight sexualities online are also in a relatively healthy mental state. In contrast, these terms (along with 'my wife') indicate increased mental health struggles for (possibly gay) women. We suspect 'my husband' is neutral for women because there are roughly equal numbers of users praising and condemning their husbands.

Beyond sexuality, we see that some familial terms have differing predictive interpretations across genders. 'my mum' is predictive of depression for men and control for women, while the reverse is true for 'my son'. This suggests a substantial difference in parent-child relationships depending on the gender of each: each gender appears to have an affinity for family members of the same gender.

We also highlight a few features with broader societal interpretations. 'trump' is strongly predictive of depression among women but neutral for men. This is consistent with the well-known 'gender gap' phenomenon and could also indicate that mental health is in part a function of political climate. The LIWC category 'money' is slightly depression predictive for women and control-predictive for men. Similar to the above, this could be an artifact of the wage gap: money topics may be more stressful for women because they tend to earn less money for the same amount or more work.

6 Discussion

In this paper we showed that depression predictions do indeed exhibit gender bias. This was evidenced by a substantially better performance when predicting depression among males than when predicting among females. We also identified terms that are used differently between men and women, providing insight into the manifestations of depression beyond modeling dynamics.

6.1 Open Questions and Future Work

As hinted in the introduction, the key open question is **does gender confound depression predictions?** In

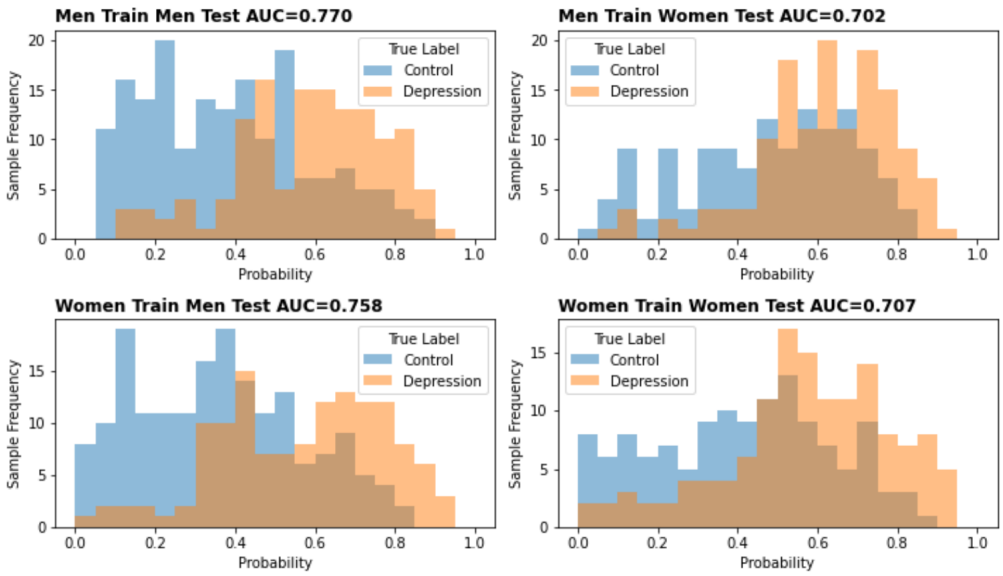

Figure 1: Performance of each model, trained to predict depression on either male users or female users only, when evaluated on each test set

other words, does gender *both* affect depression predictions *and* the features we use to predict it? There are numerous plausible explanations for why both of these causal relationships may hold or not hold, but without a rigorous causal analysis, it is not possible to rule any one explanation out in favor of another.

To properly evaluate whether a associational relationship is in fact causal, the causal framework requires 'intervening' on an independent variable while holding other variables in the system constant to see whether there are changes in the dependent variable. Here, that means intervening on gender, which is infeasible to carry out directly.

There may however, be some viable proxy approaches for simulating the intervention on gender. One such approach would entail fitting a model to predict the ground truth gender and then using a clustering algorithm to find male and female centroids based on the most predictive features in the gender prediction model. The analyst could then simulate an intervention on gender for the purposes of analyzing changes to depression prediction by replacing the user's feature vector in the depression inference model with each gender centroid vector. This approach will not permit a true causal interpretation but it could provide insights into the relationship between gender and depression prediction beyond those gained from the simple models studied in this work. Unfortunately this approach cannot be applied to analyzing the relationship between gender and the text features since it entails changing those text features.

Outside of the explicit question of confounding, we can ask **how do we correct for the performance differentials across demographic groups when predicting depression?**. As hinted earlier, an obvious approach with support in the literature (Amir et al., 2019) is to simply collect 'better' data. This is an unsatisfying answer, however, since good data is often hard to come by or expensive to collect. Instead, we can again turn to causal inference ideas to try to address data quality issues. We can potentially use methods from the causal fairness literature to impose constraints on depression models to ensure negligible differences in prediction performance. For instance, following (Nabi et al., 2019), we could impose a constraint that requires that the total effect of gender on depression predictions is zero, or, plainly, that there is no difference in model performance when we do or don't condition on gender.

6.2 Limitations

Aside from the limitations described above, i) all users in our cohort posted in r/AskMen or r/AskWomen (which we used to derive ground truth) and ii) we rebalanced our data sets due to insufficient numbers of depressed users in the 'representative' population. These decisions could reduce the generalizability of our results. One way to address this would be to collect data on more users by expanding the study period and by consulting other subreddits with gender self-disclosure such as r/relationships (Wang and Jurgens, 2018).

Additionally, while our use of self-disclosed genders increases scalability, this could induce bias in two ways. Users could be dishonest in their disclosure and, even if they aren't, users who choose to self-disclose could be fundamentally different from the general population. It's likely that the only solution is to collect data external to Reddit about Reddit users' genders as a more reliable supplement to our data.

Finally, our depression labels were not obtained via self-disclosures. Rather, they were defined based on

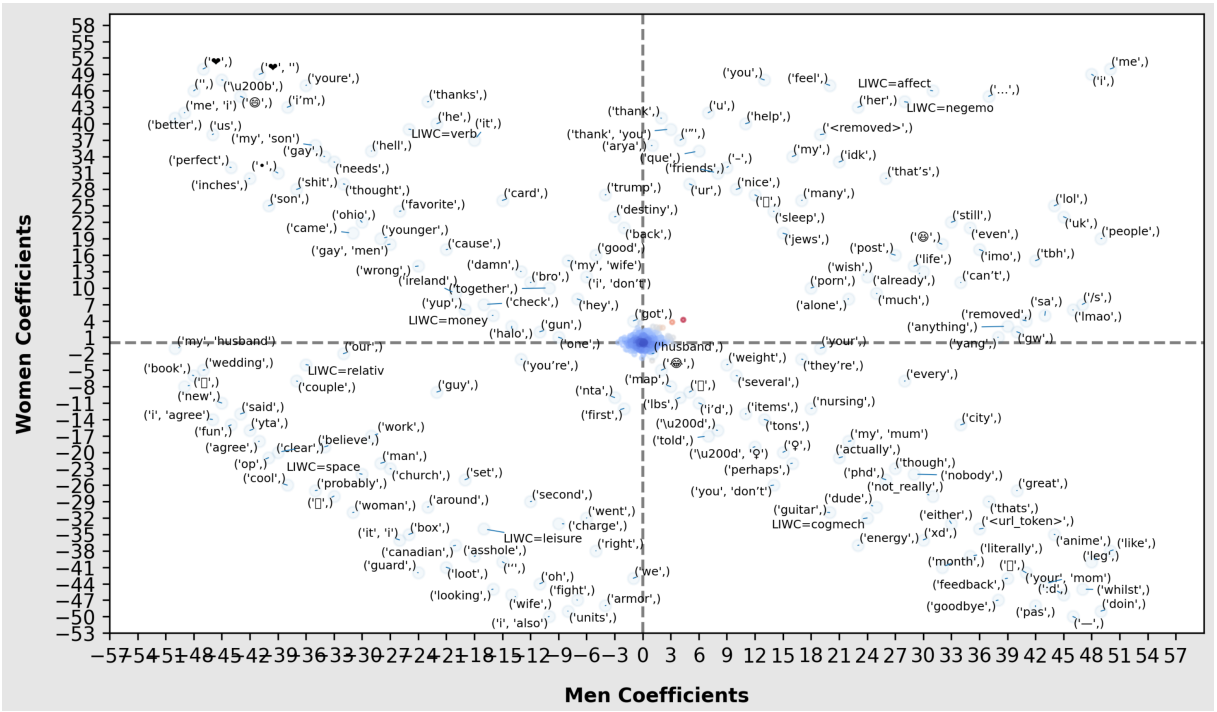

Figure 2: Features in common between the male- and female- trained models with the 50 highest scoring features in each quadrant labeled

whether the user posted in the r/depression subreddit. While this approach is consistent with data collection approaches from the literature (De Choudhury et al., 2013), it is likely to induce some noise. For instance, a user could post in the subreddit to seek support for a friend or relative, rather than for themself and would therefore be incorrectly labeled as depressed. One way to address this would be to take a more nuanced approach to labeling. For instance, we could use regular expressions matched on the text of r/depression posts to develop a more exclusive labeling policy that filters out users who are not seeking personal support.

6.3 Ethics

As in any applied setting it is necessary to weigh the potential advantages and harms of carrying out our research agenda. This work has the potential to cause harm in a couple key ways.

First, as previously mentioned, we restrict attention to users satisfying a narrow and dated 'folk' definition of gender in line with much of the existing research in the space of computational psychology. This is done at the cost of excluding non-binary individuals, who potentially stand to benefit the most from this work due to the increased prevalence of depression in gender nonconforming populations. Furthermore, excluding any marginalized population from a study of this type has the potential to reinforce existing biases. For instance, if our model had demonstrated improved prediction performance for the binary genders, that could lead to an incorrect assumption that the model will perform well on the general population, which includes non-binary genders. This could lead to *worse* performance for the unstudied groups.

Second, while we infer depression status from Reddit users with the goal of alleviating harms, these approaches could be harnessed with malice to identify and target already vulnerable individuals whose screen names and posting behavior are public.

On the other hand, there is great potential in this study and the work that will follow it. Identifying obstacles to model deployment for a restricted population will likely aid in correcting those obstacles for the entire population. This would substantially improve the performance and, more importantly, the clinical utility of mental health surveillance models. Given the potential benefits of this study we feel it is better to proceed, with care and transparency, rather than sit idle for lack of perfect answers to address the issues the work poses.

Acknowledgements

The first author was sponsored by a Google PhD Fellowship.

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)*.

Paolo Alboni, Elisa Favaron, Nelly Paparella, Massimo Sciammarella, and Mario Pedaci. 2008. Is there an association between depression and cardiovascular mortality or sudden death? *Journal of Cardiovascular Medicine*, 9(4):356–362.

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.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

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

Debra J Brody, Laura A Pratt, and Jeffery P Hughes. 2018. Prevalence of depression among adults aged 20 and over: United states, 2013-2016.

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.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640*.

Bruce Hoffman, Jacob Ware, and Ezra Shapiro. 2020. Assessing the threat of incel violence. *Studies in Conflict & Terrorism*, 43(7):565–587.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Brian N Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 78–87.

Michael E McCullough and David B Larson. 1999. Religion and depression: A review of the literature. *Twin Research and Human Genetics*, 2(2):126–136.

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. 2019. Learning optimal fair policies. In *International Conference on Machine Learning*, pages 4674–4682. 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.

James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Liwc2007: Linguistic inquiry and word count. *Austin, Texas: liwc. net*.

C Steven Richards and Michael W O'Hara. 2014. *The Oxford handbook of depression and comorbidity*. Oxford University Press.

A John Rush. 2007. The varied clinical presentations of major depression disorder. *The Journal of clinical psychiatry*.

Zijian Wang and David Jurgens. 2018. It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 33–45, Brussels, Belgium. Association for Computational Linguistics.

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.

Zach Wood-Doughty, Michael Smith, David Broniatowski, and Mark Dredze. 2017. How does twitter user behavior vary across demographic groups? In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 83–89.

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, Copenhagen, Denmark. Association for Computational Linguistics.