Qualitative Analysis of Depression Models by Demographics

Carlos Aguirre
Johns Hopkins University
caguirr4@jhu.edu

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

Abstract

Models for identifying depression using social media text exhibit biases towards different gender and racial/ethnic groups. Factors like representation and balance of groups within the dataset are contributory factors, but difference in content and social media use may further explain these biases. We present an analysis of the content of social media posts from different demographic groups. Our analysis shows that there are content differences between depression and control subgroups across demographic groups, and that temporal topics and demographic-specific topics are correlated with downstream depression model error. We discuss the implications of our work on creating future datasets, as well as designing and training models for mental health.

1 Introduction

Models of mental health trained on social media data exhibit biases in downstream performance on different gender and racial/ethnic demographic groups (Aguirre et al., 2021). An important factor is that minority groups (People of Color in general) are underrepresented in datasets and thus models under perform compared to majority groups. While size and balance of datasets contribute to the gap in performance, there may be differences in the manner in which depressive behavior is exhibited across demographic groups, creating problems in generalization.

Difference in depression prevalence across demographics have long been known (Brody et al., 2018), although there is no clear explanation for why this is the case (Hasin et al., 2018). On social media, demographic-based mental health analyses have used matched control samples (Dos Reis and Culotta, 2015), which allow for comparison of behaviors across groups (Coppersmith et al., 2014; Amir et al., 2019). These types of analyses have focused on downstream performance of trained models (Aguirre et al., 2021) and how they show differences in depression rates, but there have been no qualitative studies investigating these demographic differences (Chancellor and De Choudhury, 2020; Harrigian et al., 2020b).

Others have used qualitative studies to analyze behaviors and performance of machine learning models in general (Chen et al., 2018). Previous work has analyzed representative sentences (Ettinger, 2020), hashtags (Sykora et al., 2020), performed a thematic analysis by using the Linguistic Inquiry and Word Count dictionary (Wolohan et al., 2018) or trained topic models (Harrigian et al., 2020a; Yazdavar et al., 2017; Mitchell et al., 2015).

We propose a qualitative language analysis to reveal what differences occur, and how these differences can contribute to downstream performance. What language trends characterize depression and how do these vary across demographic groups? We use an analysis method similar to Mueller et al. (2021) but instead of training an Latent Dirichelt Allocation (LDA) (Blei et al., 2003) topic model and performing Point-Wise Mutual inference to obtain topics related to demographics, we train a Partially-Labeled LDA model (Ramage et al., 2011) which allows us to assign labels to demographic groups as well as depression and control groups to obtain label-specific topics to our user groups.

We base our analysis on datasets from previous work using Twitter. We train simple text-based models based on previous work on these datasets (Harrigian et al., 2020a; Aguirre et al., 2021). We use a labeled topic model to characterize what content indicates depression and how this content varies by demographic group.

Our analysis shows variations in content between depression and control subgroups across demographic groups, however, most of these differences are due to non-clinical phenomena e.g. vi-

Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology, pages 169–180
June 11, 2021.

ral content trends such as *TV shows awards*. Further, model error analysis corroborates that temporal trends and nongeneralizable topics of demographic groups are correlated with downstream model error. Our qualitative analysis approach can be utilized to analyze language differences across demographics on other datasets and mental health tasks. We discuss the implications of our work on creating new datasets, as well as designing and training language models for mental health.

2 Ethical Considerations

Given the sensitive nature of mental health topics and demographics of individuals, additional precautions (based on *depression diagnoses* (Benton et al., 2017a); *gender identity* (Larson, 2017); *race/ethnicity identity* (Wood-Doughty et al., 2020)) were taken during this study. Data sourced from external research groups was retrieved according to each datasets respective data use policy. For gender labels, due to current limitations on datasets and methods, we consider the *folk perception* of gender, as described in Larson (2017), and for race/ethnicity labels we use the mutually-exclusive non-Hispanic White, non-Hispanic Black,non-Hispanic Asian and Hispanic/Latinx, following Wood-Doughty et al. (2020). We acknowledge that both our gender and racial/ethnic categories do not fully capture many individuals' gender and/or race/ethnicity. Additionally, we acknowledge the limitations of the demographic inference methods employed to obtain the demographic labels that have been raised in multiple previous studies (Mueller et al., 2021; Aguirre et al., 2021). While we carefully consider these issues, we believe the urgency of understanding mental health models (Aguirre et al., 2021) warrants our work and hope that our results provide sufficient evidence to justify further study in this area. This research was deemed exempt from review by our Institutional Review Board (IRB) under 45 CFR ğ 46.104.

3 Data

We use two datasets for depression identification on Twitter from previous studies: the *CLPsych 2015 Shared Task* (Coppersmith et al., 2015b), and the multi-disorder *multitask* learning for mental health dataset (Benton et al., 2017b).

CLPsych. The dataset contains publicly available tweets of individuals where the diagnosed group was collected by self-report through regular expression matching, e.g. `"I was diagnosed with <disorder>"`. Control individuals were approximated by matching inferred age and gender using tools from the World Well-Being Project (Sap et al., 2014) from a pool of random accounts. While the original dataset collected four conditions, we select the depression users (475) and their matched control users resulting in 950 individuals.

Multitask. This dataset combines subsets of several datasets (Coppersmith et al., 2015a,b,c). All methods used the same collection process: self-report through regular expression matching, and control individuals by matching inferred age and gender with the same tool. Additionally, the complete public history of tweets is collected for each individual as opposed to the latest 3000 tweets on *CLPsych* resulting in a bigger dataset. We select the depression users (1400) and their matched control users resulting in 2800 individuals.

label	topic	tokens	Δ	E
		Female White		
Depression	mental health	men mental ppl trans #mentalhealth sex		0.936
Female White	Body Neg.	fat weight eating line cross die cut body		0.906
Depression	UK language	lovely mum favourite uk mate cos london		0.876
Depression	Game/Media	anime luigi art games draw mario character		0.919
Depression	One Direction	harry louis zayn direction niall liam		0.903
Female White	School	class college weekend homework break		0.294
Control	Sports	team football fine state season congrats		0.093
Latent	AAVE	nigga gotta yo niggas bout bitches tho		0.345
Control	Arabic	libra يا و ما ، لا على الله في من beer		0.111
Latent	Rap/Music	yall lmfao smh nah tho drake kanye album		0.433
		Female PoC		
Depression	Game/Media	anime luigi art games draw mario character		0.944
Female White	Body Neg.	fat weight eating line cross die cut body		0.971
depression	mental health	men mental ppl trans #mentalhealth sex		0.922
Depression	UK language	lovely mum favourite uk mate cos london		0.925
Depression	5 SOS	#vote5sos luke #kca michael calum ashton		0.992
Female White	School	class college weekend homework break		0.29
Control	Sports	team football fine state season congrats		0.14
Control	Arabic	libra يا و ما ، لا على الله في من beer		0.111
		Male White		
Depression	UK language	lovely mum favourite uk mate cos london		0.851
Depression	mental health	men mental ppl trans #mentalhealth sex		0.827
latent	Politics	police trump president state america		0.614
Latent	Media	book movie star film story books episode		0.602
Depression	Game/Media	anime luigi art games draw mario character		0.729
Female White	School	class college weekend homework break		0.205
Latent	Relationship	text boyfriend care relationship not_want		0.347
Control	Sports	team football fine state season congrats		0.116
Latent	Spanish	que la el en es te un mi se lo por los		0.135
Control	Arabic	libra يا و ما ، لا على الله في من beer		0.111
		Male PoC		
Depression	Game/Media	anime luigi art games draw mario character		0.845
depression	One Direction	harry louis zayn direction niall liam		0.997
Depression	5 SOS	#vote5sos luke #kca michael calum ashton		0.996
Female White	Body Neg.	fat weight eating line cross die cut body		0.902
Female PoC	Pop Culture	jacob jack vine dm fans #fifthharmony		0.999
Female White	School	class college weekend homework break		0.207
Control	Sports	team football fine state season congrats		0.051
Control	Arabic	libra يا و ما ، لا على الله في من beer		0.111

Table 1: Top and bottom 5 topics, as measured by the change of prevalence between depression and control group Δ, per demographic group on *Multitask* dataset. Only showing topics were Δ is statistically significant with bootstrapping (iterations $= 1000$, CI $= 0.95$).

While both dataset collection methods are nearly identical, the time period in which the tweets were collected, and the number of tweets and individuals are different for each dataset, likely leading to different types of depression indicators. Note that while there is an overlap between **Multitask** and **CLPsych** of 110 individuals, it is a small percentage of both datasets ($\sim 4\%$ and $\sim 10\%$ respectively).

4 Methodology

Demographic Labels. While both datasets utilized gender and age inferences to match control and disorder groups at collection time, these models are now out-dated and labels for race/ethnicity were not made available. We obtain new race/ethnicity and gender labels from the work of Aguirre et al. (2021). Demographic statistics for both datasets are available in Appendix A. Since the race/ethnicity minority groups are extremely underrepresented in the datasets, we combine them to create a Person of Color (PoC) group.

Mental Health Models. We create mental health models for these datasets based on recent work (Harrigian et al., 2020a; Aguirre et al., 2021). Following standard pre-processing procedures, we filter numeric values, username mentions, retweets and urls from raw tweet text. For model features, we considered TF-IDF vector representations, mean-pooled 200 dimensional Twitter GloVe embeddings (Pennington et al., 2014), Linguistic Inquiry Word Count (LIWC) representations (Pennebaker et al., 2007), and features based on topic distributions learned via LDA. We train ℓ_2-regularized logistic regression models on both datasets and follow hyper-parameter tuning procedures from Harrigian et al. (2020a); Aguirre et al. (2021).

4.1 Topic Model Analysis

Model. We use a topic model analysis to identify topic distribution differences between demographic groups. We train on each dataset (separately) a Partially Labeled LDA model (Ramage et al., 2011), which incorporates per-label latent topics into an LDA model. We assign both *depression* and *demographic* labels to individuals, with $K = 5$ topics per label and 20 latent topics not associated with any labels for a total of 50 topics, following the number of topics from previous work. Intuitively, this has the effect of *credit attribution* – associating words to either *depression*, *demographic* groups or latent to dataset topics for each individual.

Metrics. To measure topic prevalence between groups, we use the enrichment (E) metric from Marlin et al. (2012); Ghassemi et al. (2014):

$$E(c') = \frac{\sum_i^d \mathbb{1}(c_i = c')y_i \cdot q_{ic}}{\sum_i^d \mathbb{1}(c_i = c')q_{ic}}$$

The metric E has the effect of highlighting topics regardless of topic importance within the group. In order to preserve topic importance, we take the non-normalized average difference in E between control and depression groups (Δ). For each document i, and corresponding label y_i, topic c_i, and topic probability q_{ic}:

$$\Delta(c') = \frac{1}{n}\sum_i^n \mathbb{1}(y_i = 1)\mathbb{1}(c_i = c')q_{ic} - \mathbb{1}(y_i = 0)\mathbb{1}(c_i = c')q_{ic}$$

Where negative values are topics most aligned with control group and positive values are aligned with depression.

Finally, To measure error rate attributions to topics, we use the topic error rate $\hat{E}$ metric from Chen et al. (2018):

$$\hat{E}(c') = \frac{\sum_i^d \mathbb{1}(y_i \neq \hat{y})\mathbb{1}(c_i = c')q_{ic}}{\sum_i^d \mathbb{1}(c_i = c')q_{ic}}$$

Data Processing. In addition to removing numeric values, username mentions, retweets and urls, we also remove English stopwords, pronouns[1] and emojis in order to create more coherent topics for our annotators. Removing stopwords and pronouns has the potential to erase depression signals as previous studies have found signals on pronoun usage, and also suppress voices and languages that do not fit certain norms. A full list of stopwords and pronouns is provided in Appendix B. We excluded topics from our results that did not have any coherent semantic groupings as annotated by one of the authors and 2 volunteers by looking at the top 15 most probable words per topic, obtaining a fair multi-annotator agreement Fleiss' Kappa $\kappa = 0.332$. After, topics were the majority of annotators selected as coherent where labeled by one of the authors.

[1]English stopwords and pronouns were obtained from *NLTK* tool (Bird et al., 2009)

	label	topic	tokens	Female		Male	
				White	PoC	White	PoC
Multitask	Female-PoC	Pop Culture	jacob jack vine dm fans ugly #theyretheone meet	0.172	0.002	0.137	**0.999**
	depression	One Direction	harry louis zayn direction niall liam	0.087	0.024	0.155	**0.951**
	Male-White	Video Games	tap games gta stream gg pc xbox glitch #gamergate	0.156	0.001	0.341	**0.988**
	Female-White	Justin Bieber	justin retweet bieber ily tour babe meet #mtvstars	0.143	0.039	0.164	**0.766**
	depression	Game/Media	anime luigi art games draw mario character	0.140	0.136	0.307	**0.776**
CLPsych	Female-White	Justin Bieber	bieber #emazing #mtvstars beliebers	0.141	0.709	**0.717**	0.017
	Female-White	One Direction	direction niall liam louis zayn leo #mtvhottest fandom	0.315	**0.837**	0.106	0.563
	Female-White	People's Choice	demi miley austin lovato #peopleschoice vote album	0.231	0.154	**0.698**	0.001
	control	Beauty	wedding #love #fashion #nails #beauty #hair #beautiful	**0.592**	0.051	0.004	0.008
	Female-PoC	AAVE	n**gas smh yo gone somebody everybody mad ima	0.247	**0.577**	0.151	0.097

Table 2: Top 5 topics as measured by topic error rate $\hat{E}$. Higher value represents higher prevalence on individuals that mental health models misclassified.

5 Analysis

The topic model identified label-specific topics for depression and control. Appendix C shows the topics for both datasets as well as the top 10 most probable words per topic. Some *depression* topics are reasonable e.g. `mental-health` (in both datasets) and `social media stats` (may be related to internet statistics and popularity). Similarly, *control* topics like `sports` and `beauty` are active, positive and self-caring topics that are reasonable for being representative of our control group. However, some topics in both *depression* and *control* groups are not clearly tied to the groups e.g. for depression group, topics like `One Direction` and `5 Seconds of Summer`. These might be topics introduced by temporal phenomena impeding model generalization (Harrigian et al., 2020a), rather than representative topics for those labels.

5.1 Content Differences

We characterize the difference in content between depression and control groups for each demographic. Table 1 shows the top and bottom 5 most prevalent topics with respect to the depression subgroups per demographic category as measured by Δ on the *Multitask* dataset where only the topics with statistically significant Δ are shown, as computed by bootstrapping with 1000 iterations with a CI of 95%.

Some *depression* topics (`One Direction` and `5 Seconds of Summer`) are not representative across demographics, while reasonable topics e.g. `mental-health` are representative of depression across demographic groups. Additionally, the topics most prevalent in the control subgroups (`School` and `Sports`) are the same across all demographics and represent qualities that are not related to depression, showing the robustness of these indicators and the well-formed nature of the control group in the dataset.

The `Body Negative` topic, attributed to the *Female-White* label, is very prevalent on depression subgroups for both female groups but is not prevalent on male subgroups, suggesting that there are differences on depression language online between gender groups.

For Male PoC individuals, the `mental-health` topic for depression is not prevalent in the depression subgroup while `One Direction` is prevalent. Given that the Male PoC group has the fewest users in the dataset, this suggests that its depression subgroup is not a representative group of individuals for depression yielding spurious topics, confirming prior work on dataset size being a factor on difference in performance across demographics (Aguirre et al., 2021).

Further, topics representing non-English language (`Arabic` and `Spanish`) or minority accent (AAVE) are more prevalent in control subgroups of demographics where those are not expected e.g. `Arabic` and `AAVE` on *Female-White* group. Perhaps this is evidence of demographic label noise, further exacerbating the need of obtaining self-reported demographic labels on mental health datasets for more concrete analysis.

5.2 Depression Model Errors

We analyze the predictions of our depression models to identify content differences between demographics that are correlated to models errors. Table 2 shows the top 5 topics that are most prevalent on individuals that were wrongly classified by the models on each dataset. Expanding results from Section 5.1, we find that topics that are not representative across demographics e.g. `One Direction`, are correlated with downstream errors in classification of mental health models. This suggests that topics that are prevalent of depression subgroups and are not related to depression are misleading the model.

Additionally, the majority of topics most prevalent on model errors (e.g. `Justin Bieber`, `One Direction` and `People's Choice`) apart from not being related to mental health and not representative across demographics, are influenced by temporal phenomena e.g. short term events such People's Choice Awards, that stem from the time period in which the dataset was collected. Such topics are not generalizable, corroborating evidence from prior work on challenges in model generalizations of temporal themes (Harrigian et al., 2020a).

Further, some topics prevalent on model errors are the effect of dataset balance. For example, the topic `AAVE` is a *Female-PoC* labeled topic, but it also is very prevalent on model errors, suggesting that there are very few examples of AAVE in the dataset and the mental health model is oversensitive to this language. On the other hand, the topic `beauty`, labeled as *control*, is over-represented in the dataset. This suggest that datasets should be balanced based on demographics, following prior work (Aguirre et al., 2021).

6 Conclusion

We performed a qualitative analysis to find content differences related to mental health across demographic groups. We showed that there are content differences between depression and control subgroups, while most of these differences are due to non-clinical phenomena e.g. temporal topics. Additionally, we find that dataset size might be a factor in these content differences. Furthermore, model error analysis corroborates that temporal topics and demographic-specific topics are correlated with downstream model error.

Our findings support prior work on the importance of methods that seek to generalize temporal topics (Harrigian et al., 2020a). We also find supporting evidence of the importance of dataset size as well as dataset balance in order to generalize to minority groups (Aguirre et al., 2021). Though in our analysis we only consider one mental health disorder (*depression*), our methodology was able to generalize across two datasets. This suggests that it is a valid method for qualitative analysis on finding content differences in other mental health datasets. Additionally, while we were limited in our demographic labels by current demographic models and dataset sizes, we showed that our approach is valid across two demographic axes and could be expanded to include other demographic axes (such as age and economic status), and include genders and racial/ethnic groups outside of the ones considered in this work. We hope our work warrants further studies of mental health language differences across more diverse demographic groups yielding more inclusive datasets and research.

Acknowledgements

We thank the anonymous reviewers for their helpful feedback. Also, we thank Keith Harrigian and Rachel Wicks for volunteering time and effort for topic annotations, and Elizabeth Salesky for helpful feedback.

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)*.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

S Bird, E Loper, and E Klein. 2009. Natural language processing with python oreilly media inc.

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

Debra Brody, Laura Pratt, and Jeffery Hughes. 2018. Prevalence of depression among adults aged 20 and over: United states, 2013-2016. *NCHS data brief*, pages 1–8.

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 Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002*.

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, Denver, Colorado. Association for Computational Linguistics.

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.

Glen Coppersmith, Ryan Leary, Eric Whyne, and Tony Wood. 2015c. Quantifying suicidal ideation via language usage on social media. In *Joint Statistics Meetings Proceedings, Statistical Computing Section, JSM*, volume 110.

Virgile Landeiro Dos Reis and Aron Culotta. 2015. Using matched samples to estimate the effects of exercise on mental health via twitter. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. 2014. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020a. Do models of mental health based on social media data generalize? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3774–3788, Online. Association for Computational Linguistics.

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

Deborah S. Hasin, Aaron L. Sarvet, Jacquelyn L. Meyers, Tulshi D. Saha, W. June Ruan, Malka Stohl, and Bridget F. Grant. 2018. Epidemiology of Adult DSM-5 Major Depressive Disorder and Its Specifiers in the United States. *JAMA Psychiatry*, 75(4):336–346.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Benjamin M Marlin, David C Kale, Robinder G Khemani, and Randall C Wetzel. 2012. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*, pages 389–398.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 11–20.

Aaron Mueller, Zach Wood-Doughty, Silvio Amir, Mark Dredze, and Alicia L Nobles. 2021. Demographic representation and collective storytelling in the me too twitter hashtag activism movement. *Proceedings of the ACM on Human-Computer Interaction, CSCW*.

James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Operators manual: Linguistic inquiry and word count: LIWC2007. *Austin, Texas: LIWC. net http://homepage. psy. utexas. edu/HomePage/Faculty/Pennebaker/Reprints/LIWC2007_OperatorManual. pdf (accessed 1 October 2013)*.

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 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.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and H Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151.

Martin Sykora, Suzanne Elayan, and Thomas W Jackson. 2020. A qualitative analysis of sarcasm, irony and related #hashtags on twitter. *Big Data & Society*, 7(2):2053951720972735.

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, Paiheng Xu, Xiao Liu, and Mark Dredze. 2020. Using noisy self-reports to predict twitter user demographics. *arXiv preprint arXiv:2005.00635*.

Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 1191–1198.

A Dataset Demographics

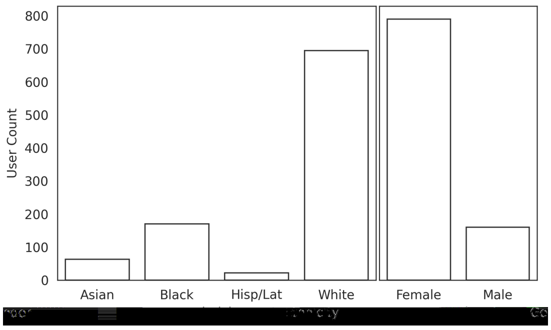

Figure 1: User count by gender and racial/ethnic demographic groups on **CLPsych** dataset.

Figure 2: User count by gender and racial/ethnic demographic groups on **Multitask** dataset.

B Partially Labeled LDA

We need a procedure to identify topic distribution differences between demographic groups. Prior work have accomplished this by training an LDA topic model and either using pointwise mutual inference (PMI) (Mueller et al., 2021) or an enrichment metric E (Marlin et al., 2012; Ghassemi et al., 2014) to measure how distinctive a given topic is of a given demographic group.

Instead, we train[2], on both datasets separately, a Partially Labeled LDA model (Ramage et al., 2011), which incorporates per-label latent topics to an LDA model. Unlike LDA, each document d can only use the topics associated with the set of labels L_d assigned to d, where each label $l \in L_d$ is assigned some number of topics K. The model computes the joint likelihood of observed words w, observed labels l and topic assignments z, given available labels Λ, and document-topic α, topic-word η priors from a Dirichtlet distribution $P(w, l, z|\Lambda, \alpha, \eta)$. We assign both *depression* and *demographic* labels to individuals, with $K = 5$ topics per label and 20 latent topics not associated with any labels for a total of 50 topics. Intuitively, this has the effect of *credit attribution* – associating words to either *depression*, *demographic* groups or latent to dataset topics for each individual.

To measure topic difference between groups (**RQ1**) we use the enrichment (E) metric from Marlin et al. (2012); Ghassemi et al. (2014):

$$E(c') = \frac{\sum_i^d \mathbb{1}(c_i = c')y_i \cdot q_{ic}}{\sum_i^d \mathbb{1}(c_i = c')q_{ic}}$$

To measure error rate (**RQ2**) we use the topic error rate $\hat{E}$ metric from Chen et al. (2018):

$$\hat{E}(c') = \frac{\sum_i^d \mathbb{1}(y_i \neq \hat{y})\mathbb{1}(c_i = c')q_{ic}}{\sum_i^d \mathbb{1}(c_i = c')q_{ic}}$$

Additionally, in order to preserve topic importance, we take the non-normalized average difference in E between control and depression groups (Δ). For each document i, and corresponding label y_i, topic c_i, and topic probability q_{ic}:

$$\begin{aligned}\Delta(c') = \frac{1}{n}\sum_i^n & \mathbb{1}(y_i = 1)\mathbb{1}(c_i = c')q_{ic} \\ & - \mathbb{1}(y_i = 0)\mathbb{1}(c_i = c')q_{ic}\end{aligned}$$

Where negative values are topics most aligned with control group and positive values are aligned with depression. In addition to filtering numeric values, username mentions, retweets and urls, we also filter stopwords, pronouns and emojis to obtain more coherent topics. We excluded topics from our results that did not have any coherent semantic groupings as annotated by one of the authors by looking at top 10 most probable words per topic.

[2]Model implementation based on Tomotopy python library `https://bab2min.github.io/tomotopy/v0.10.2/en/` which provides Gibbs-sampling based implementations of multiple *LDA models.

i	me	my	myself	we	our	ours	ourselves	you
you’ve	you’ll	you’d	your	yours	yourself	yourselves	he	him
himself	she	she’s	her	hers	herself	it	it’s	its
they	them	their	theirs	themselves	what	which	who	whom
that	that’ll	these	those	am	is	are	was	were
been	being	have	has	had	having	do	does	did
a	an	the	and	but	if	or	because	as
while	of	at	by	for	with	about	against	between
through	during	before	after	above	below	to	from	up
in	out	on	off	over	under	again	further	then
here	there	when	where	why	how	all	any	both
few	more	most	other	some	such	no	nor	not
own	same	so	than	too	very	s	t	can
just	don	don’t	should	should’ve	now	d	ll	m
re	ve	y	ain	aren	aren’t	couldn	couldn’t	didn
doesn	doesn’t	hadn	hadn’t	hasn	hasn’t	haven	haven’t	isn
ma	mightn	mightn’t	mustn	mustn’t	needn	needn’t	shan	shan’t
shouldn’t	wasn	wasn’t	weren	weren’t	won	won’t	wouldn	wouldn’t

Table 3: English stopwords from NLTK (Bird et al., 2009)

he	she	they	i	him	her	we	me	it
us	them	myself	ourselves	yourself	yourselves	himself	itself	herself
my	our	ours	your	yours	their	its	mine	theirs

Table 4: English pronouns from NLTK (Bird et al., 2009)

C PLDA Topics

label	Topic	words	weight
depression	Mental Health	men mental ppl trans #mentalhealth sex health woman racist depression illness racism rape gender	0.0166
	UK Language	xx lovely bit xxx mum favourite uk mate cos london ffs australia brilliant bloody	0.0140
	One Direction	harry louis zayn direction niall liam в stats unfollowers followers #emabiggestfans1d и на не fandom	0.0107
	5 Seconds of Summer	#vote5sos luke #kca michael calum ashton seconds clifford hood summer hemmings #mtvstars #5sosfam	0.0058
control	Arabic	libra الله في من enter #nyc لا على aries و ما ، check يا beer	0.0069
	Portuguese	#android que e é discovered location não j eu lyn pra london um com streets	0.0021
	Sports	team football fine state season nails posted touch college congrats basketball proud coach	0.0225
Female White	Body Negative	fat weight eating line cross die cut body anymore skinny loves kill pain	0.0154
	School	class summer college weekend car homework dad break friday semester dog netflix room hour	0.0563
	Justin Beiber	justin retweet bieber dm #callmecam ily tour gain babe meet #mtvstars jacob pls proud	0.0136
	TV Shows	proud #thewalkingdead season episode #supernatural strong dead #love :d saved sam	0.0061
Female PoC	Dating	se dating singles z je surveys polls za yahoo politics health si po pro	0.0003
	Spanish	la en el que con por un los es gracias para del las se una	0.0019
	Pop culture	jacob jack vine dm fans ugly #fifthharmony #theyretheone meet sebastian af indirect #shawnformmva	0.0032
Male White	German	ich die und daily der das eyes hazel #supergirl ist nicht es top stories zu	0.0006
	Cities	#albuquerque #tpp israel vote u.s. obama #tcot #newmexico support war #faceofmlb transport	0.0021
	Canada/Music	#nowplaying team season #winnipeg load #spotify #canada football ask band final #music #indie player	0.0068
	Video Games	full added tap games liked beer menu gta stream gg pc xbox glitch #gamergate xd	0.0025
Male PoC	Social Media	followers #retweet goodmorning fast #teamfollowback retweets mentions #follow2befollowed followed	0.0010
	Video Games	#gamergate #notyourshield http anti- games gg sjws htt h sjw gamers anti harassment ht wu	0.0003
	AAVE	bro smh yall gotta bruh tho vine im team lebron season fam nba its dont	0.0015
latent	Politics	police trump president state america obama law country killed news vote gun government american rights	0.0226
	Zodiac	cancer others although current seems leo seem capricorn energy mind surgery gemini	0.0193
	AAVE	nigga gotta yo niggas bout bitches tho lil af bruh cuz n hoes bro dude	0.0419
	Media	book movie star film story books episode series reading writing art write post blog	0.0232
	Social Events	check party friday album top tickets weekend adam posted tour meet fans congrats vote	0.0382
	Music	yall lmfao smh nah tho drake kanye mad men bae gotta saying album boo wtf	0.0304
	People	kids child woman mother lady sister season married brother movie dad daughter sex	0.0348
	Spanish	que la el en es te un mi se lo por los con las para	0.0087
	Relationship	tired text anymore boyfriend care relationship honestly mood kinda forever babe leave sick	0.1121

Table 5: Label, topic title, top words and topic importance of *Multitask* dataset.

label	Topic	words	weight
depression	Mental Health	#mentalhealth mental depression link submitted comment health asked anxiety disorder illness	0.0063
	Social Stats	stats loves unfollowers photoset followed happen daily follower unfollower	0.0049
	Pop culture	pls luke michael ilysm hemmings babe ashton penguin zayn pizza clifford calum niall	0.0045
control	Peoples Choice	katy perry #peopleschoice others skinny share roar fiber sticker unlocked glee darren #musicvideo	0.0031
	Beauty	wedding #love #fashion #cute #nails #me #beauty #hair #beautiful #taurus #instagood	0.0035
	Life	movies virgo college favorite win shopping seeing classes study puppies studying loved	0.0062
Female White	Peoples Choice	demi miley austin lovato #peopleschoice vote album tour sign cyrus selena xoxo miley's	0.0080
	Justin Bieber	bieber #emazing de que #mtvstars el la beliebers #mtvhottest en justin's reason #kca foto te	0.0070
	One Direction	direction niall liam louis zayn leo #mtvhottest fandom luke fans album xx story babe styles	0.0136
	Internet Abv.	ily omfg crying aw dm meet retweet bye idk picture quote literally ill cry bby	0.0228
Female PoC	AAVE	n**gas smh yo gone somebody everybody mad ima hoes swear nobody af lil yall	0.0190
	Spanish	mi :p que de la te tu ke en r el se luv b h	0.0042
Male White	Nascar	bob fans mate race #bbq #nascar palace bbq win top racing league grilling fame	0.0037
	Work	č dundee salary hiking camping scotland engineer manager angus trail jobs sales	0.0016
Male PoC	AAVE	niggas bout wit yo smh aint bro gone yu everybody yea lil hoes bitches tryna	0.0043
	Music	outta line #soundcloud #new #retweet feat #rt download mixtape prod essay ft	0.0037
latent	Relationships	care anymore hurt smile relationship alone reason enough fall not_want change thinking feelings	0.1058
	AAVE	lmfao fuckin thats ill yo bitches dude lil high bout smoke wtf cuz hit black	0.0422
	Life	awesome needs r friday vote lady w pic chris retweet la halloween h movie lead	0.0229
	Politics	welcome west america state history star obama bill v national second john question david government	0.0255
	Sports	team season win games football goal fans vs congrats playing run beat final ball fan	0.0287
	Social Stats	followed stats unfollowers bro follower moment question yea hahahaha bus awkward teacher thats la unfollower	0.0287
	Relationships	babe idk ugly rn bae boyfriend wtf mad bored honestly annoying w k tbh kiss	0.0467

Table 6: Label, topic title, top words and topic importance of *CLPsych* dataset.