Diverse Perspectives, Divergent Models: Cross-Cultural Evaluation of Depression Detection on Twitter

Nuredin Ali[1], **Charles Chuankai Zhang**[1], **Ned Mayo**[2] , **Stevie Chancellor**[1]

[1]University of Minnesota
{ali00530,zhan6914,steviec}@umn.edu
[2]Macalester College
emayo@macalester.edu

Abstract

Social media data has been used for detecting users with mental disorders, such as depression. Despite the global significance of cross-cultural representation and its potential impact on model performance, publicly available datasets often lack crucial metadata related to this aspect. In this work, we evaluate the generalization of benchmark datasets to build AI models on cross-cultural Twitter data. We gather a custom geo-located Twitter dataset of depressed users from seven countries as a test dataset[1]. Our results show that depression detection models do not generalize globally. The models perform worse on Global South users compared to Global North. Pre-trained language models achieve the best generalization compared to Logistic Regression, though still show significant gaps in performance on depressed and non-Western users. We quantify our findings and provide several actionable suggestions to mitigate this issue.

1 Introduction

According to the data from World Health Organization, depression is a global issue affecting 240 million people worldwide[2]. In response to these trends, in the last decade, there has been a surge in studying the mental health status of users from social media based on their content and interaction (Ji et al., 2018). Research has focused on various disorders, including depression, anxiety, and eating disorders, and has used many methods (Wongkoblap et al., 2017). Specifically - depression is among the most widely studied disorders (and the most commonly diagnosed), and Twitter is a common source of data in these studies (Chancellor and De Choudhury, 2020). This work tries to predict if someone may have depression based on data from social media (Chancellor and De Choudhury, 2020; Harrigian et al., 2021).

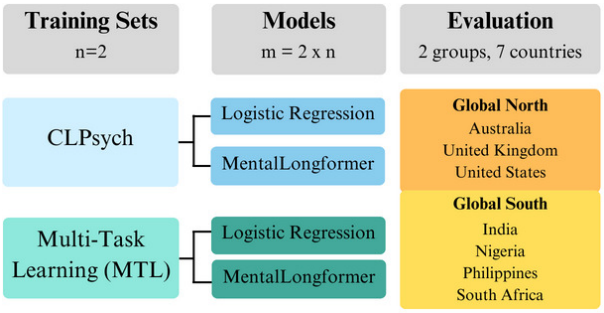

Figure 1: Flow chart of the overall design of the work. This shows the training and evaluation process. n=datasets, m=models.

Given this area's popularity and potential reach to clinical settings, NLP has also called for careful evaluations of bias, performance gaps, and generalizability of claims from small datasets (Aguirre et al., 2021; Harrigian et al., 2020; Hovy and Spruit, 2016). One source of underexplored bias in these datasets and models is the impact of a person's geographic location (and consequently, their culture) on their communication style. The importance of cultural consideration in social media studies about mental health is critical (Lee et al., 2014). In prior work, De Choudhury et al. showed cross-cultural differences in mental health communication styles in cultures such as the US, India, and the Philippines. Cross-cultural users also have different identity dimensions, language use, and support behavior (Pendse et al., 2019; Mittal et al., 2023); sentiment detection can vary across cultures (Pruksachatkun et al., 2019). However, interaction in international forums does not affect their clinical mental health language use (Pruksachatkun et al., 2019). Recent literature reviews (Chancellor and De Choudhury, 2020) and persuasive calls (Garg, 2023) point to the need to study the generalizability of models to distinctive user populations for mental health research. Similar audits have been

[1]Details of the cross-cultural evaluation dataset used in this work: https://grouplens.org/datasets/twitter-depression-dataset-2024/

[2]https://vizhub.healthdata.org/gbd-results/

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 672–680

instrumental in identifying gaps in performance across clinical/non-clinical populations (Ernala et al., 2019) and in gender and racial groups (Harrigian et al., 2020; Aguirre et al., 2021; Aguirre and Dredze, 2021).

Building on this prior research, in this paper, we analyze the generalization of depression detection models on cross-cultural data trained on existing benchmark datasets. Inspired by (Harrigian et al., 2020; Aguirre et al., 2021), we ask: do models built on popular social media benchmark datasets to predict depression generalize to people who live in different countries, yet speak English? If the prior work is correct about geographic and cultural biases impacting predictions, what countries may be most affected? How stark are the performance differences between countries?

We audit depression detection models generalization on cross-cultural data gathered from Twitter (now X). We collected data from seven countries using a strict location verification technique and the prevalence of English content in users' feeds, using keyword matching and manual annotation to identify genuine depression disclosures. We trained two models, Logistic Regression and MentalLongformer, on benchmark depression datasets (CLPsych and MentalLongformer). We assessed their generalization by both country and socioeconomic development classification (Global North vs. Global South).

We show that models on broad Twitter benchmarks do not generalize well to the cross-cultural data. Models generalize much better to evaluation data from users in the Global North (US, UK, Australia) than to users in the Global South that use English as a national language (India, Nigeria, Philippines, South Africa). Distinct gaps emerge between countries, with models generalizing very poorly to posters from Nigeria and India. Our findings demonstrate that existing benchmark datasets are not representative of training generalized models that could detect depressed users from various cultures. We provide suggestions for building better datasets and models.

2 Datasets

We carefully selected two popular benchmark datasets for constructing depression models on Twitter data and then created a geolocated dataset of depression posts. Table 1 summarizes our datasets.

Dataset	Classes	Train	Val
CLPsych	Depression	327	150
(Coppersmith et al., 2015)	Control	570	301
Multi-Task Learning	Depression	1520	320
(Shen et al., 2017)	Control	1520	320
Ours (Evaluation)	Depression	-	267
	Control	-	264

Table 1: Datasets used in our experiments.

CLPsych: This dataset comes from the CLPsych 2015 Shared Task (Coppersmith et al., 2015). The shared task contains two mental disorder identifications, identified with keywords and manual annotation: depression and PTSD, of which we use the depression treatment data and control. The data comprises the users' most recent posts around the date of depression disclosure, up to a maximum of 3000 posts per user.

Multi Task Learning (MTL): This dataset is from (Shen et al., 2017), which contains Twitter user profile information and their posts within one month. This dataset also identified people who may be depressed in Twitter. Both CLpsych and MTL are gathered based on a strict set of keywords/keyphrases such as "(I'm/I was/ I am/ I've been) diagnosed depression," etc., to identify the candidate depressed users. We leverage the text data only (as this dataset does contain images).

Our global dataset: At the time of writing, there are no public benchmark datasets of global expressions of depression in Twitter data. Therefore, we collect a corpus from public posts from Twitter using the Twitter Research API (now defunct)[3]. We used the search terms/phrases from De Choudhury et al.'s cross-cultural depression study on Twitter to identify people discussing depression or suicidality (a common co-morbid symptom of depression). These include phrases such as "I am/I'm depressed" and "I want to hurt myself", and were verified by psychologists by the collaborators of De Choudhury et al. We searched the sample of the Twitter data made available between January 2015 and December 2022[4].

Given our focus on cross-cultural content, we leveraged geotagged tweets. We specifically gathered users from seven countries: Australia, South Africa, Nigeria, the Philippines, India, the United Kingdom, and the United States. We selected these

[3]https://developer.twitter.com/en/use-cases/do-research/academic-research

[4]Note that the Twitter API gave a sample of data, but not all of it.

countries for geographic diversity, their large volume of geotagged disclosures, and the fact that English is a first language or is a business/government-listed language in those countries. To verify that a user was in the country, we looked at their 3200 geotagged posts before disclosure and took the country where the user posted the most.

We manually verified each user's veracity of depression disclosure with human raters, similar to the process in (Coppersmith et al., 2015). We developed and applied a codebook to identify users who had genuine disclosures of depression (see Appendix A.1 for details). We then gathered control users with similar demographics whose posts did not include the search terms but with the same geotagged Tweet rules as discussed. Therefore, we made a matched, "control" sample of users from the same country who disclosed having depression and those who did not.

At the outset, our dataset comprised 16,112 potentially depressed users from the seven countries. Of these, 1,556 were manually reviewed, leading to the identification of 267 authentic disclosures. The annotation encompassed all original sample users from the Global South countries but did not cover all users from the Global North due to the substantial volume of data. Cohen's Kappa (McHugh, 2012) between the raters resulted in 0.65, showing a substantial agreement. The disagreements were resolved through two rounds of discussion. Appendix A.3 shows genuine and non-genuine posts obtained through human annotation. These examples are paraphrased and lightly edited to protect the identity of the posters (Ayers et al., 2018).

Our geo-located dataset encompasses a total of 531 users, with 267 users identified as having depression through manual verification. We define two groups of countries based on the United Nations categorization of countries[5] - the Global North and the Global South. 140 users were in the Global North (64 United States, 45 United Kingdom, 31 Australia), while 127 users hail from the Global South (58 Philippines, 35 South Africa, 19 India, 15 Nigeria).

2.1 Preprocessing

We applied the same preprocessing pipeline across the datasets (including the benchmarks) for consistency, following recommendations from (Harrigian et al., 2020). Specific retweet tokens, username mentions, URLs, and numeric values were removed. English contractions were expanded. We removed the disclosure words from the training and evaluation sets. Users with fewer than 20 posts were excluded, and only those with a minimum of 20 English posts were considered for inclusion.

[5]https://unctad.org/system/files/official-document/tdstat47_en.pdf

3 Baseline Models

For mental health prediction tasks, Logistic Regression is a popular and performant statistical baseline due to its quick training time, success in prediction, and highly interpretable feature relevance (Benton et al., 2017b; Jiang et al., 2018; Harrigian et al., 2020). In this experiment, we extracted the features using the term frequency-inverse document frequency (TF-IDF). Using scikit-learn, we applied grid search on 5-fold cross-validation. The best hyperparameters for the logistic regression are *'penalty': 'l2', 'solver': 'lbfgs', 'max_iter': 10000*, and 7000 TF-IDF features.

Model	CLPsych		MTL	
	Recall	F1	Recall	F1
Logistic Regression	0.83	0.80	0.89	**0.89**
MentalLongformer	0.76	0.73	0.93	**0.92**

Table 2: The F1 score of the baseline model on both datasets.

Pretrained language models such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) have significantly improved text classification on many general (Mularka et al., 2020) and domain-specific tasks (Ji et al., 2021, 2023). We finetune the MentalLongformer language model for our baseline, which outperforms the other pre-trained models on this specific task and has an extended sequence modeling capacity (Ji et al., 2023). We use it to investigate its generalization capabilities to our task.

For this experiment, the pre-trained head of the MentalLongformer is replaced with a randomly initialized classification head. We set the learning rate to 5e-5. Adam is used as an optimizer (Kingma and Ba, 2014). We trained for 'num_train_epochs=50' and applied an 'early_stopping_patience=10'. The remaining parameters were set to the default hyperparameters of MentalLongformer on Huggingface. Table 2 presents the results of the baseline models on the test of both datasets. To evaluate the model's performance, we report F1 and recall, selected for their effectiveness in handling unbalanced datasets.

Training Data	Australia		Nigeria		South Africa		Philippines		India		UK		US	
	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1
CLPsych	0.61	0.63	0.13	0.23	0.45	0.53	0.39	0.46	0.10	0.19	0.53	0.61	0.53	**0.66**
Multi Task Learning	0.53	0.60	0.13	0.23	0.28	0.36	0.08	0.15	0.26	0.35	0.84	**0.69**	0.75	0.61

Table 3: F1 scores of *Logistic Regression* trained on CLPsych and Multi-Task Learning datasets.

Training Data	Australia		Nigeria		South Africa		Philippines		India		UK		US	
	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1
CLPsych	0.64	**0.72**	0.06	0.12	0.2	0.32	0.18	0.3	0.15	0.27	0.37	0.5	0.42	0.56
Multi Task Learning	0.93	0.69	0.33	0.45	0.71	0.67	0.31	0.43	0.68	0.7	0.95	**0.72**	0.84	0.64

Table 4: F1 scores of *MentalLongformer* trained on CLPsych and Multi-Task Learning datasets.

Model	Training set	Global North		Global South	
		Recall	F1	Recall	F1
Logistic Regression	CLPsych	0.47	**0.58**	0.17	0.28
	Multi-Task Learning	0.76	**0.63**	0.17	0.26
MentalLongformer	CLPsych	0.45	**0.58**	0.17	0.28
	Multi-Task Learning	0.9	**0.68**	0.48	0.56

Table 5: F1 scores of both models trained on CLPsych and Multi-Task Learning datasets evaluated on Global North and Global South eval sets.

4 Results

In Table 2, we present the results of our ML models on two benchmark datasets (CLPsych and MTL). Our baseline models closely replicate prior research of benchmark datasets (logistic regression (Aguirre et al., 2022), and MentalLongformer (Ji et al., 2023)). We evaluate model performance on our custom dataset, split into two groups - Global North vs. Global South and then country-level.

4.1 Global North vs. Global South

Our baseline models trained on benchmark datasets perform much worse on data from the Global South than the Global North. Table 5 shows the results of the two groups and our model's performance. There is an expected drop in performance between the baseline model and the Global North and the Global South evaluation datasets (due to them being out-of-domain). However, all four models have a superior F1 and recall in identifying the Global North evaluation users. This finding aligns with prior research that there is a gap in performance between these categories (Pruksachatkun et al., 2019), though it confirms it at a larger cultural scale.

Table 4 shows that the MentalLongformer model trained on the MTL data has much better recall (or sensitivity) for detecting the presence of depression in the Global North countries compared to the Global South. This is particularly useful in these settings where identifying depression users is essential or where models are used for downstream interventions.

4.2 Country Level Analysis

To investigate the performance gap in Global North vs. Global South, we analyze country-level outcomes, presented in Tables 3 for the Logistic Regression and 4 for the MentalLongformer model. We separate each country into groups for this analysis, noting that the size of each country's dataset is imbalanced (see Datasets 2).

There is a significant difference (p-value: 0.001) in accurately identifying depressed users among various countries. Further analysis within two groups, (Australia, US, UK) and (India, Nigeria, South Africa, and the Philippines), revealed no statistical differences (p-values: 0.47 and 0.39, respectively). Notably, all models struggled to correctly identify users from Nigeria and India. This disparity indicates the need for more generalizable training benchmarks and models.

4.3 Qualitative Error Analysis

To understand the disparities in detection, we explore the model with the highest variance in F1 score between the Global North and Global South (the Logistic Regression model trained on the MTL dataset, with a 0.37 F1 score gap between these two regions). We conducted a qualitative error analysis focusing on users from Nigeria and the Philippines. We initially look at the word distributions between these two regions to discern potential similarities/differences in the most frequent words. We present a few qualitative observations of trends.

First, users in the Global South, particularly those from Nigeria and India, express common words such as 'god,' 'life,' 'love,' and 'people.' Within Global North countries, words like 'work,' 'day,' 'time,' and 'people' rank prominently among common words. There are shared linguistic fea-

tures between both regional sets of countries, such as 'love,' 'life,' 'like,' and 'one.' Still, we note that many of the users from Nigeria and India have discrepancies in how they communicate in general (not just about mental health). This aligns with prior research that highlights variations in linguistic patterns (Pendse et al., 2019; De Choudhury et al., 2017). Such differences in language usage might account for the subpar performance observed across these countries.

Second, we also note that some users in non-Western countries rarely engage in code-mixing, where they use two or more languages in speech at a given time. Take this example user, who contains both English and other languages that the model misclassifies.- (e.g. *'here it goes no one wants me i am worthless even though i am alive feeling dead inside gusto ko magbakasyon ng mahabang mahaba...'*. Similar trends happen in Nigerian users, e.g. *'wani abu ma sai dan shaye shaye sai ma lokacin iftar zata ga abubuwa amin lemme just pretend i did not see that'*. However, there are no such examples of code-mixing in the Global North countries where English is the primary official language. Recall that we picked countries where English is an official language or would be used in business settings and identified Twitter users who primarily Tweeted in English. However, identifying code-mixed tweets is challenging, and Twitter's language detector has limitations.

5 Recommendation and Conclusion

In this work, we quantified the generalization capability of depression detection models in cross-cultural data. We specifically quantified that models have higher discrepancies in identifying users from different cultures. We provide the following suggestions for improving the identified gaps.

Construct datasets with more geographical examples. Similar to (Harrigian et al., 2020; Aguirre et al., 2021), we hypothesize that mental health detection from social media suffers from small datasets. Existing benchmark datasets lack the location meta-data of users (Garg, 2023) and lack different demographic representations (Aguirre et al., 2021), meaning that fairness audits are challenging to execute post-hoc.

We propose a few solutions to this problem. First, researchers could adapt techniques to infer geo-location if larger datasets were available (Mittal et al., 2023; Shaikh et al., 2022) to conduct audits. Larger datasets could be composited from comparable sources, pointing to evidence from (Harrigian et al., 2020) that more data helps alleviate racial disparities in predictions. Balancing the datasets effectively makes the algorithms fair in different groups (Pessach and Shmueli, 2023). Ultimately, the field needs to find paths forward to identify and supplement datasets for this task. As an initial stride in this direction, we provide the details for our dataset and how other users may replicate our findings[6]

Investigating the cross-cultural detection capabilities of proposed models. Current work has a considerable gap in ethical consideration and transparent reporting (Ajmani et al., 2023). Fine-grained subgroup analysis reporting leads towards building more inclusive and transparent models (Buolamwini and Gebru, 2018). We call for critical consideration when reporting these metrics when introducing algorithms.

6 Ethical Considerations

Predicting mental health via social media data is ripe with ethical challenges (Benton et al., 2017a; Chancellor and De Choudhury, 2020). Yet, this area also holds promise in identifying early indications of mental disorders, potentially averting risky behaviors, and getting people access to treatment. This requires careful consideration and application in ways that benefit society while mitigating risks.

Our study follows standard procedures for deanonymizing participants in our data (Chancellor and De Choudhury, 2020; Benton et al., 2017a). The IRB at the University of Minnesota (study ID: STUDY00018665) ruled that our work was not human subjects research because our data was publicly available and we did not interact with users. The CLPsych data is accessed through IRB approval, and the Multi-Task Learning data is publicly available. Before computational modeling, we still took procedures to protect participants' identities, such as removing URLs, usernames, and personal identifiers from data. We do not report any data about individuals in certain countries nor provide examples of data to protect people in these situations. It is imperative to underscore that these datasets should exclusively be used for research purposes.

[6]The details of the cross-cultural evaluation dataset used in this work is provided https://grouplens.org/datasets/twitter-depression-dataset-2024/

One risk we highlight is cross-cultural factors such as differences in stigma and the consequences of disclosure. Countries have major differences in the stigma and social consequences that the prediction of mental illness may have in those spaces. Individuals can be shamed for disclosing mental illness, prevented from opportunities (employer use in screening processes), or denied dignity. In some cultures, mental illness can be trivialized or ignored. These same factors may lead to different strategies for disclosure in public forums like Twitter. Nonetheless, this data should not be used to draw conclusions about which countries might have higher depression rates or who is "better" at caring for people with mental illness. Nor should this data be used to profile people based on inferences from social media data.

7 Limitations

The dataset used for evaluating the six countries might not be representative for three reasons.

1. Individuals from different countries might convey their mental health status in unique ways, involving using different sets of key phrases compared to those in our study (Pendse et al., 2019). To comprehensively understand these potential variations, additional research is required to pinpoint and incorporate these specific keywords and research culture-specific means of disclosure. Moreover, there is also the critical challenge of self-disclosure bias that affects the underlying user sample and modeling output of depressed users (Chancellor et al., 2023).

2. During the qualitative error analysis, we found that users from countries like Nigeria, India, and the Philippines use code-mixing in their posts. Although we filtered for English-only content using Twitter language detection, it missed some posts, resulting in code-mixed content for some users. This could potentially be the source of some of the disparities identified. Therefore, future research could investigate methods to effectively handle code-mixing, enhancing technical capabilities in NLP and cross-cultural mental health detection.

3. The geo-tagged tweets play a vital role in our research. This constitutes approximately 1% of Twitter's daily content on Twitter (X) (Lamsal et al., 2022). However, our reliance on this specific subset of data also limits the volume of data in our study.

4. We focused solely on two models, two widely used benchmark datasets, and Twitter (X) as our platform. While this provides valuable insights into disparities, conducting further studies on additional models and platforms could offer a more comprehensive understanding.

References

Carlos Aguirre and Mark Dredze. 2021. Qualitative analysis of depression models by demographics. In *Proceedings of the seventh workshop on computational linguistics and clinical psychology: improving access*, pages 169–180.

Carlos Aguirre, Keith Harrigian, and Mark Dredze. 2021. Gender and racial fairness in depression research using social media. *arXiv preprint arXiv:2103.10550*.

Maia Aguirre, Laura García-Sardiña, Manex Serras, Ariane Méndez, and Jacobo López. 2022. BaSCo: An annotated Basque-Spanish code-switching corpus for natural language understanding. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3158–3163, Marseille, France. European Language Resources Association.

Leah Hope Ajmani, Stevie Chancellor, Bijal Mehta, Casey Fiesler, Michael Zimmer, and Munmun De Choudhury. 2023. A systematic review of ethics disclosures in predictive mental health research. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1311–1323.

John W Ayers, Theodore L Caputi, Camille Nebeker, and Mark Dredze. 2018. Don't quote me: reverse identification of research participants in social media studies. *NPJ digital medicine*, 1(1):30.

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 *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

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):43.

Stevie Chancellor, Jessica L Feuston, and Jayhyun Chang. 2023. Contextual gaps in machine learning for mental illness prediction: The case of diagnostic disclosures. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–27.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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*, CHI '19, page 1–16, New York, NY, USA. Association for Computing Machinery.

Muskan Garg. 2023. Mental health analysis in social media posts: a survey. *Archives of Computational Methods in Engineering*, 30(3):1819–1842.

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

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 15–24.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Shaoxiong Ji, Celina Ping Yu, Sai-fu Fung, Shirui Pan, and Guodong Long. 2018. Supervised learning for suicidal ideation detection in online user content. *Complexity*, 2018.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*.

Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, Erik Cambria, and Jörg Tiedemann. 2023. Domain-specific continued pretraining of language models for capturing long context in mental health. *arXiv preprint arXiv:2304.10447*.

Haihua Jiang, Bin Hu, Zhenyu Liu, Gang Wang, Lan Zhang, Xiaoyu Li, and Huanyu Kang. 2018. Detecting depression using an ensemble logistic regression model based on multiple speech features. *Computational and mathematical methods in medicine*, 2018.

Shivani Kapania, Alex S Taylor, and Ding Wang. 2023. A hunt for the snark: Annotator diversity in data practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Rabindra Lamsal, Aaron Harwood, and Maria Rodriguez Read. 2022. Where did you tweet from? inferring the origin locations of tweets based on contextual information. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3935–3944. IEEE.

Hye-Ryeon Lee, Hye Eun Lee, Jounghwa Choi, Jang Hyun Kim, and Hae Lin Han. 2014. Social media use, body image, and psychological well-being: A cross-cultural comparison of korea and the united states. *Journal of health communication*, 19(12):1343–1358.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Juhi Mittal, Abha Belorkar, Vinit Jakhetiya, Venu Pokuri, and Sharath Chandra Guntuku. 2023. Language on reddit reveals differential mental health markers for individuals posting in immigration communities. In *Proceedings of the 15th ACM Web Science Conference 2023*, pages 153–162.

Ankit Murarka, Balaji Radhakrishnan, and Sushma Ravichandran. 2020. Detection and classification of mental illnesses on social media using roberta. *arXiv preprint arXiv:2011.11226*.

Sachin R Pendse, Kate Niederhoffer, and Amit Sharma. 2019. Cross-cultural differences in the use of online mental health support forums. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–29.

Dana Pessach and Erez Shmueli. 2023. Algorithmic fairness. In *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*, pages 867–886. Springer.

Yada Pruksachatkun, Sachin R Pendse, and Amit Sharma. 2019. Moments of change: Analyzing peer-based cognitive support in online mental health forums. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13.

Samira Shaikh, Thiago Ferreira, and Amanda Stent, editors. 2022. *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*. Association for Computational Linguistics, Waterville, Maine, USA and virtual meeting.

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, Wenwu Zhu, et al. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, pages 3838–3844.

Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. 2017. Researching mental health disorders in the era of social media: systematic review. *Journal of medical Internet research*, 19(6):e228.

A Appendix

A.1 Human Verification of Authentic Mental Health Disclosures

The keyphrases used to search the candidate depression disclosure include words such as *'i [*] diagnosed [*] depression', 'i attempted suicide', 'i am depressed', ' i [have/had] depression', 'i want to die', etc*. However, the candidate depression disclosure data is prone to noise. Often, users use these candidate keyphrases in their posts while they are not depressed. For instance, *"I haven't been to the gym in about a week and a half and I'm depressed."* is not a genuine disclosure according to the annotation rules but would match our keywords.

We constructed a codebook to manually verify genuine disclosures, building on prior work (Coppersmith et al., 2015). To classify a post as genuinely about depression, the post must demonstrated that the user states they are sincere about being depressed; a dark joke or sarcasm directly disclosing that they are depressed, suicidal, or thinking about self injury; or the links associated with a post (i.e. images, texts, etc) are related to genuine depression expressions.

Two annotators were involved during the annotation process of the dataset (the first two authors). This includes two PhD students with non-Western backgrounds, and they were supervised by the final author with a Western background and experience in the research area. The two annotators took three rounds of annotation to discuss disagreements on identifying genuine disclosures and refine the process. These discussions were critical to reducing random disagreements (Kapania et al., 2023). The final author consulted on the codebook creation and served as a third deliberation point when needed.

A post is a non-genuine disclosure if the post talks about feelings about a transient situation that uses "depressed" as a stand-in for being sad and the state of mental disorder is unclear, e.g. 'Manchester United lost the game, I'm depressed.' or being depressed because you have to go to work when you don't want to. The majority of posts with language about "being depressed" were ambiguous in these less serious uses of the term depression.

To apply the codebook, we followed the following approach. First, we consulted the post directly to see if it aligned with the codebook. If the post does not provide a full context or was borderline, we looked at the history of the users' posts before the disclosure. If the prior posts do not indicate that the user is depressed, we consider the disclosure as inauthentic.

A.2 Distribution of Tokens

The token distribution differs among the three datasets, with CLPsych containing more tokens than Multi-Task Learning and our dataset, which share a similar proportion see Figure 2. However, this variation doesn't significantly impact the models. The MentalLongformer is specifically designed to handle 4096 tokens (Ji et al., 2023). For logistic regression, we opt for a reduced number of features.

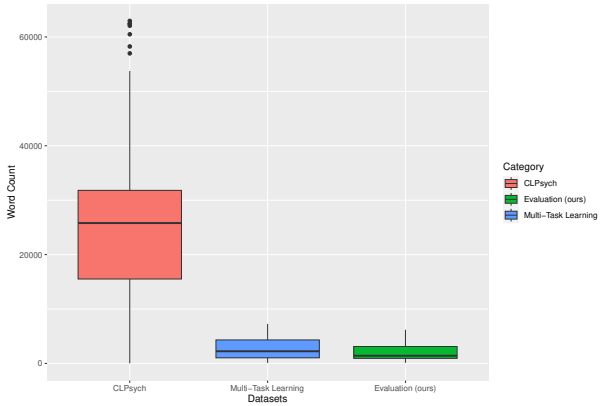

Figure 2: The box plot illustrates the distribution of tokens across the datasets.

A.3 Example of Genuine and Non-Genuine Disclosures

- **Genuine Disclosure:** *"His song means more to me now because like I told you I have depression and anxiety. It got so much worse in last few months. Listening that saved me from having a severe mental breakdown and wanting to jump out of my window or do even worse"*, *"I can't pretend to be happy anymore. I cut because I am depressed. I have tried*

killing myself because I get bullied. I'm not happy.", *"It's my favorite holiday, and I'm depressed I'm fighting it, but that's exhausting, and so is everything else"*

- **Non-Genuine Disclosure:** *"Haven't driven my toyota in so longggg. I'm depressed now haha"*, *"These next few months will be dedicated to finally dropping some fucking merch. I've been killing myself over it."*, *"I'm killing myself I'm killing myself I'm killing myself...LoL LoL :-D"*