Your Model Is Not Predicting Depression Well And That Is Why: A Case Study of PRIMATE Dataset

Kirill Milintsevich[1,2] and **Kairit Sirts**[2] and **Gaël Dias**[1]
[1]Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, France
[2]Institute of Computer Science, University of Tartu, Estonia
{first_name}.{last_name}@{unicaen.fr[1]|ut.ee[2]}

Abstract

This paper addresses the quality of annotations in mental health datasets used for NLP-based depression level estimation from social media texts. While previous research relies on social media-based datasets annotated with binary categories, i.e. depressed or non-depressed, recent datasets such as D2S and PRIMATE aim for nuanced annotations using PHQ-9 symptoms. However, most of these datasets rely on crowd workers without the domain knowledge for annotation. Focusing on the PRIMATE dataset, our study reveals concerns regarding annotation validity, particularly for the lack of interest or pleasure symptom. Through reannotation by a mental health professional, we introduce finer labels and textual spans as evidence, identifying a notable number of false positives. Our refined annotations, to be released under a Data Use Agreement, offer a higher-quality test set for anhedonia detection. This study underscores the necessity of addressing annotation quality issues in mental health datasets, advocating for improved methodologies to enhance NLP model reliability in mental health assessments.

1 Introduction

Applying various NLP techniques to automatically estimate the depression level from social media texts has been a widely researched topic in the field of NLP applied for mental health. Most of these datasets consist of online posts gathered from popular social media platforms, such as Twitter or Reddit. These posts are usually annotated by crowd workers who had only a brief training with a mental health professional (MHP) or sometimes only had access to the annotation instructions.

While there exist multiple depression-related datasets based on social media texts, most of them only present binary annotation, i.e. whether the user is depressed or not. The most common sources of data are Reddit (Losada and Crestani, 2016; Yates et al., 2017; Pirina and Çöltekin, 2018) and X (former Twitter) (Coppersmith et al., 2014; Syarif et al., 2019). Most of the studies use automatic methods of annotations, such as regular expression matching of self-reported terms, like "I have been diagnosed with depression". Some of them perform manual verification and annotation either via layman crowd workers (Yates et al., 2017) or by the authors themselves (Coppersmith et al., 2014; Losada and Crestani, 2016).

Recently, the interest in more fine-grained depression annotation has emerged. In particular, the two recent datasets D2S (Yadav et al., 2020) and PRIMATE (Gupta et al., 2022), identify depressed social media posts from X and Reddit, respectively and annotate them with PHQ-9 symptoms (Kroenke and Spitzer, 2002). Both datasets have been annotated with the help of crowd workers and later verified by MHPs. However, the verification process was different. For D2S, conflicting annotations were resolved with the majority voting, and the psychiatrist resolved the ties. After that, 100 random samples were selected for quality control and verified by a psychiatrist. Additionally, Zirikly and Dredze (2022) annotated a random sample of D2S with the explanations for each symptom with the help of two MHPs[1], increasing the validity of the data. In the case of PRIMATE, no information is given on the quality control procedure. This raises concerns about the validity of the annotations; thus, we selected PRIMATE for our case study.

In this study, on the example of the PRIMATE dataset, we show that the validity of the annotations for the mental health data is a concern when performed by layman crowd workers. Our MHP reannotated 170 posts from the PRIMATE dataset for the lack of interest or pleasure (anhedonia) symp-

[1]Zirikly and Dredze (2022) did not report any conflicts between their annotation and the labels provided with D2S.

tom. The MHP is the second author of the paper, who is also a practising clinical psychology intern. Our annotations include more fine-grained labels ("mentioned" vs "answerable", as well as an additional "writer's symptom" label) as well as spans of texts that serve as evidence of the labels. We observe a high number of false positives in the PRIMATE labels, which can be related to the high difficulty of conceptualizing anhedonia (Rizvi et al., 2016). The annotations are to be released under a Data Use Agreement (DUA), and we believe that it can serve as a higher-quality test set for anhedonia detection.

2 Dataset

PRIMATE (Gupta et al., 2022) is a dataset based on the Reddit posts from the r/depression_help subreddit. Each post is annotated with binary labels for each PHQ-9 question, where "yes" means that a post contains the answer to a PHQ-9 question and "no" otherwise. The nine symptoms are shortly described as follows: lack of interest or pleasure in doing things (LOI), feeling down or depressed (DEP), sleeping disorder (SLE), lack of energy (ENE), eating disorder (EAT), low self-esteem (LSE), problems with concentrating (CON), hyper or lower activity (MOV), suicidal thoughts (SUI).

The annotation was performed by five crowd workers with additional quality control by an MHP. The information about the annotation procedure or crowd worker training, as well as how exactly the MHPs were involved in the quality control, are not provided in the paper. The only metric on the annotation process is an annotator agreement using Fleiss' kappa, which is reported to be 67% for initial annotation and 85% after involvement of the MHPs.

In total, the dataset consists of 2003 posts. Table 1 shows the distribution of the labels[2]. Note that the exact numbers of labels are slightly different from the ones presented by Gupta et al. (2022). The dataset is not pre-split into train, validation and test sets; thus, we randomly sample 200 posts for validation and another 200 posts for testing.

Figure 1 shows the label co-occurrence matrix of the training set. Two symptoms, DEP and LSE, co-occur the most with all the other symptoms, which can be explained by their general prevalence in the dataset. The connection between the lack of interest or pleasure (LOI) and lack of energy (ENE) is also seen in the dataset, which reflects high comorbidity of these symptoms (van Borkulo et al., 2015; Park and Kim, 2020).

PHQ-9 Symptom	Number of Posts	
	Present	Absent
LOI	949	1054
DEP	1664	339
SLE	374	1629
ENE	688	1315
EAT	194	1809
LSE	1680	323
CON	195	1808
MOV	527	1476
SUI	743	1260

Table 1: Label distribution in PRIMATE.

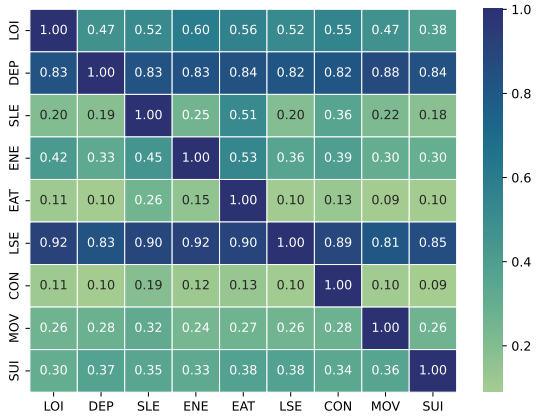

Figure 1: Symptom label co-occurrence matrix of the PRIMATE training set. Each value is normalized column-wise by dividing it by the highest value in the column.

3 Experimental Setup

In our experiments, we aimed to test how well current pre-trained language models can model the depression symptom detection problem using the PRIMATE dataset. We first chose DistilBERT (Sanh et al., 2019) as a baseline and BERT-Base (Devlin et al., 2018), RoBERTa-Base, RoBERTa-Large (Liu et al., 2019), DeBERTa-Base, and DeBERTa-Large (He et al., 2020) as higher-performing models. In particular, DeBERTa has shown constant improvements in various NLP tasks and replaced BERT and RoBERTa as the state-of-the-art model for many of them[3].

For fine-tuning, we used the implementation from Transformers library (Wolf et al., 2020). Each

[2]The order of the symptoms in the original work by Gupta et al. (2022) is different from the one of PHQ-9. In our work, we reordered the symptoms to match PHQ-9.

[3]https://gluebenchmark.com/leaderboard

Model	LOI	DEP	SLE	ENE	EAT	LSE	CON	MOV	SUI
DistilBERT	.64	.88	.67	.58	.60	.90	.50	.67	.81
BERT-Base	.55	.88	.66	.55	.63	.90	.46	.66	.79
RoBERTa-Base	.54	.88	.70	.57	.57	.90	.51	.69	.85
RoBERTa-Large	.57	.86	.75	.63	.65	.91	.52	.71	.85
DeBERTa-Base	.58	.91	.69	.52	.42	.90	.36	.61	.81
DeBERTa-Large	.60	.90	.68	.64	.47	.91	.50	.73	.83

Table 2: Symptom-wise F1-scores on the validation set.

Mentioned:

I simply want everything to finish. **I have no drive to do anything.** I am very irritable. Nothing is going as I want to and even if it was I probably wouldn't appreciate it.

Answerable:

I feel like I'm spending my life for nothing. **I used to escape my problems by browsing Youtube and Reddit for hours, but now I don't even find that enjoyable anymore.**

Not author's symptoms:

I've tried to talk about looking for other options or just ways to deal with the stress, but he's not really interested now.

Figure 2: Examples of reannotated posts. Evidences are highlighted in **bold**.

Predictions	Against PRIMATE				Against "mentioned"				Against "answerable"			
	A	P	R	F1	A	P	R	F1	A	P	R	F1
DistilBERT	.58	.56	.62	.58	.56	.30	.71	.42	.51	.10	.75	.18
PRIMATE Labels	-	-	-	-	.56	.27	.58	.37	.54	.09	.58	.15

Table 3: Results on the reannotated part of the validation set. Here, **A** stands for Accuracy, **P** for Precision, **R** for Recall, and **F1** for F1-score for the positive class.

model consists of a pre-trained encoder with a classification head on the top of the [CLS] token. The classification head is represented by a linear layer; in the case of DeBERTa, another linear layer followed by GELU (Hendrycks and Gimpel, 2016) is added before the classification head. We trained each model for 20 epochs using AdamW optimizer with the learning rate of $2e^{-5}$, ϵ of $1e^{-6}$, β_1, β_2 of $(0.9, 0.999)$, and weight decay λ of 0.01. Additionally, a linear learning rate scheduler is applied with a warmup ratio of 0.1. Finally, the training batch size was set to 16.

4 Results and Discussion

Table 2 shows that larger models, such as RoBERTa-Large and DeBERTa-Large, perform better for ENE, LSE, MOV, and SUI. Additionally, DEP shows slight improvement with DeBERTa models, however, decreased performance for EAT. RoBERTa models perform better for SLE and SUI prediction. Nevertheless, DistilBERT sets a strong baseline and performs on par with larger models overall. Finally, LOI shows a decrease in performance for all the models compared to the DistilBERT.

We investigate the diminished performance of the LOI symptom since it is a core symptom of a major depressive disorder (Association, 2013) and shows unstable results for our models. Furthermore, LOI is one of the symptoms of schizophrenia (Association, 2013) and is associated with both anxiety and depression (Winer et al., 2017). Thus, we selected a subset of 170 posts from the validation set based on the DistilBERT predictions: if at least one symptom was predicted incorrectly, the post was selected. Next, an MHP read all the posts in the subset and labelled them for the presence of loss of interest or pleasure (LOI). The MHP as-

signed three labels to each post: a) "mentioned" if the symptom is talked about in the text, but it is not possible to infer its duration or intensity; b) "answerable" if there is clear evidence of anhedonia; c) "writer's symptoms" which shows whether the author of the post discusses themselves or a third person. Additionally, the MHP selected the part of the text that supports the positive label.

Figure 2 shows examples for the reannotated posts[4]. The first example is labelled as "mentioned" since it contains evidence of a symptom but does not contain information about the *loss* of interest. The second example is labelled as "answerable" because it is possible to infer that the person used to have interest in what they were doing before but lost it at some point in time. Finally, the last example shows the post without signs of `LOI` that describes the condition of another person.

Table 3 shows accuracy, precision, recall and F1-score for positive class against different sets of labels on our manually reannotated subset. DistilBERT, when measured against "mentioned" and "answerable" labels, performs considerably worse than against original labels from PRIMATE. It is unsurprising given the extremely low agreement between these sets of labels with Cohen's kappa of 9% and 3%, respectively. Furthermore, the most common error type is a false positive, i.e., a symptom marked as present in PRIMATE when our MHP found no evidence of it in the text. Additionally, using PRIMATE labels as predictions and comparing their performance against our labels shows lower performance than the DistilBERT model.

Considering the "writer's symptom" label, in 18 out of 170 selected posts, the author describes a symptom of another person rather than themselves. This raises the question of how these posts should be annotated and whether they should be included in the dataset at all. We suspect that the language of describing one's condition or feelings in the first person is different from the third person. We leave this question for future debate and assign "mentioned" and "answerable" labels to the posts describing a third person in the same manner as to the personal posts.

Our findings are consistent with the original results presented by Gupta et al. (2022). Similar to our experiment, they also trained a classifier based on the BERT-Base model and reported low MCC for `LOI`. However, we provided the evidence that this might be caused by annotation errors. Additionally, we noticed that many posts that were mistakenly labelled with `LOI` are more closely related to the "inner tension" symptom from the Montgomery-Åsberg Depression Rating Scale (MADRS) (Montgomery and Åsberg, 1979).

While we agree that our reannotated test set is also, to some extent, susceptible to errors, we believe that it serves as a more reliable benchmark for the anhedonia symptom. A more fine-grained, evidence-based labelling scheme reduces the risk of mislabelling and is more transparent for further verification. Finally, it lays the foundation for future collaboration to produce a higher-quality Reddit-based dataset for depression symptom estimation.

5 Conclusion

In conclusion, this study highlights the importance of evaluating and enhancing the quality of annotations in mental health datasets, particularly within the context of automated depression level estimation from social media texts. While recent datasets such as PRIMATE introduce commendable efforts toward nuanced annotations using PHQ-9 symptoms, our examination of the PRIMATE dataset reveals concerns about annotation validity, specifically regarding the lack of interest or pleasure symptom. Through careful reannotation by a mental health professional, we discerned a considerable number of false positives among the original labels indicative of challenges in conceptualizing anhedonia.

The findings presented here advocate for a more rigorous and standardized approach to mental health dataset annotation, emphasizing the need for greater involvement of domain experts in the annotation process. The release of our refined annotations under a Data Use Agreement (DUA) contributes a valuable resource for future research, offering a higher quality test set for anhedonia detection. Moving forward, a concerted effort toward refining annotation methodologies and promoting collaboration between domain experts and NLP practitioners is imperative to foster advancements in this crucial intersection of technology and mental health research.

6 Availability of Data

The instructions for accessing the annotations presented in this paper can be found here: `https://github.com/501Good/primate-anhedonia`.

[4]All example posts are paraphrased for privacy.

7 Ethical Considerations

According to Benton et al. (2017), studies involving user-generated content are exempt from Institutional Review Board (IRB) requirements if the data source is public and user identities are not identifiable. We access and use the data according to the Data Use Agreement provided with the PRIMATE dataset. Finally, we are going to release our annotations under another Data Use Agreement and separate them from the original PRIMATE data. We also acknowledge that no automatic system can replace a real mental health professional and cannot be used as a sole instrument of diagnostics.

8 Limitations

We acknowledge the limitations inherent in our work and findings. First, the manually annotated explanations serve as a proxy for what clinicians might find informative in assessing Reddit posts flagged as depressive. While evaluating the informativeness of explanations in a true clinical setting would provide more insight, it falls beyond the scope of this paper. Furthermore, our reannotation was carried out by only one mental health professional, which does not allow for performing an inter-annotator agreement analysis. However, we believe that our evidence-based labelling scheme partially mitigates this problem. Finally, anhedonia is extremely challenging to conceptualize and binary labels may not be the best choice in situations when the difference between the presence or absence of the symptom is marginal. In this case, labels based on the Likert scale, as in PHQ-9, would be more appropriate and allow us to capture the intensity of the symptom more accurately. Furthermore, different demographics, for example, adolescents and adults, express signs of anhedonia differently (Watson et al., 2020).

Acknowledgements

This research was supported by the Estonian Research Council Grant PSG721 and the FHU A^2M^2P project funded by the G4 University Hospitals of Amiens, Caen, Lille and Rouen (France). The calculations for model's training and inference were carried out in the High Performance Computing Center of the University of Tartu (University of Tartu, 2018).

References

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5™ (5th ed.)*. American Psychiatric Publishing, Inc.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. 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, Baltimore, Maryland, USA. Association for Computational Linguistics.

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*.

Shrey Gupta, Anmol Agarwal, Manas Gaur, Kaushik Roy, Vignesh Narayanan, Ponnurangam Kumaraguru, and Amit Sheth. 2022. Learning to automate follow-up question generation using process knowledge for depression triage on Reddit posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 137–147, Seattle, USA. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Kurt Kroenke and Robert L Spitzer. 2002. The PHQ-9: a new depression diagnostic and severity measure.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

David E Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *International conference of the cross-language evaluation forum for European languages*, pages 28–39. Springer.

Stuart A Montgomery and MARIE Åsberg. 1979. A new depression scale designed to be sensitive to change. *The British journal of psychiatry*, 134(4):382–389.

Seon-Cheol Park and Daeho Kim. 2020. The centrality of depression and anxiety symptoms in major depressive disorder determined using a network analysis. *Journal of affective disorders*, 271:19–26.

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

Sakina J Rizvi, Diego A Pizzagalli, Beth A Sproule, and Sidney H Kennedy. 2016. Assessing anhedonia in depression: Potentials and pitfalls. *Neuroscience & Biobehavioral Reviews*, 65:21–35.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Iwan Syarif, Nadia Ningtias, and Tessy Badriyah. 2019. Study on mental disorder detection via social media mining. In *2019 4th International conference on computing, communications and security (ICCCS)*, pages 1–6. IEEE.

University of Tartu. 2018. UT rocket.

Claudia van Borkulo, Lynn Boschloo, Denny Borsboom, Brenda WJH Penninx, Lourens J Waldorp, and Robert A Schoevers. 2015. Association of symptom network structure with the course of depression. *JAMA psychiatry*, 72(12):1219–1226.

Rebecca Watson, Kate Harvey, Ciara McCabe, and Shirley Reynolds. 2020. Understanding anhedonia: A qualitative study exploring loss of interest and pleasure in adolescent depression. *European Child & Adolescent Psychiatry*, 29:489–499.

E Samuel Winer, Jessica Bryant, Gregory Bartoszek, Enrique Rojas, Michael R Nadorff, and Jenna Kilgore. 2017. Mapping the relationship between anxiety, anhedonia, and depression. *Journal of affective disorders*, 221:289–296.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. 2020. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 696–709, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Ayah Zirikly and Mark Dredze. 2022. Explaining models of mental health via clinically grounded auxiliary tasks. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 30–39, Seattle, USA. Association for Computational Linguistics.