Enhanced Labeling Technique for Reddit Text and Fine-Tuned Longformer Models for Classifying Depression Severity in English and Luganda

Richard Kimera
Department of Advanced Convergence
Handong Global University
Pohang, South Korea
kimrichies@handong.ac.kr

Daniela N. Rim
School of Computer Science and Electrical Engineering
Handong Global University
Pohang, South Korea
danielarim@handong.ac.kr

Joseph Kirabira
Department of Psychiatry
Busitema University (Faculty of Health Sciences)
Mbale, Uganda
jkirabira.fhs@busitema.ac.ug

Ubong Godwin Udomah
Department of Psychiatry
University of Uyo teaching hospital
Uyo, Akwa Ibom, Nigeria
bloominasant@yahoo.com

Heeyoul Choi
School of Computer Science and Electrical Engineering
Handong Global University
Pohang, South Korea
hchoi@handong.edu

***Abstract*—Depression is a global burden and one of the most challenging mental health conditions to control. Using the BDI questionnaire, experts can detect its severity early, administer appropriate medication to patients, and impede its progression. Owing to the fear of potential stigmatization, many patients turn to social media platforms such as Reddit for advice and assistance at various stages of their journey. This research extracts text from Reddit to facilitate the diagnostic process, employs a proposed labelling approach to categorize the text, and subsequently fine-tunes the Longformer model. The model's performance is compared against the baseline models, including Naive Bayes, Random Forest, Support Vector Machines, and Gradient Boosting. Our findings reveal that the Longformer model outperforms the baseline models in both English (48%) and Luganda (45%) languages on a custom-made dataset.**

***Index Terms*—Depression Severity, Longformer, BART, fine-tuning, Luganda, Reddit**

I. Introduction

Mental health disorders are prevalent worldwide and are predicted to be the leading cause of disease burden by 2030 [1]. Amongst them, *major depressive episodes* (depression) is a common psychiatric condition that can be challenging to manage due to its various presentations, unpredictable course and prognosis, and variable response to treatment [2]. Access to professional mental health assessment, care, and resources is often limited to a general demographic. First, detecting a patient's severity level is important to provide proper medication and prevent its advancement to stages that can lead to suicide tendencies and death. To this end, when diagnosing depression, doctors use the Beck Depression Inventory (BDI) questionnaire. This 21-question multiple-choice has a set of four (4) possible choices for each question, ranging in severity. It can be self-administered, and the obtained score will determine a depression severity categorized into either of the six; *normal*, *mild*, *moderate*, *borderline*, *severe*, and *extreme* [3]. This resource-intensive task is hard to administer, especially in resource-constrained settings and societies where mental health patients are stigmatized. With this fear, many patients resort to using social media to anonymously share and receive global feedback from people with similar conditions, recovering patients, and experts. One example is Reddit[1], a platform allowing users to post and exchange ideas freely. This social media platform has over 57 million daily active users, with over 100,000 active communities. It also has over 50,000 daily active moderators [4]. The modulation can be done by moderators, administrators and a modulation tool (AutoMod), and if the content violates the platform terms of use, it is removed. In 2022 3.7% of the total content created

This research was supported by the Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No. 2018-0-00749, Development of virtual network management technology based on artificial intelligence), Korea International Cooperation Agency (KOICA), and by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2022R1A2C1012633).

[1]https://www.reddit.com/

was removed [5].

Text collected from Reddit for types of depression classification purposes is usually labelled by experts, trained personnel [6], or according to the corresponding category of the mental illness it is associated with [7], [8]. This classified text is then used to train machine learning or deep learning models, an approach that is not new. [9]. Razavi et al. [10] used the BDI-II to measure the severity of depression on a wide array of machine learning classification algorithms. Similarly, BDI-II was used in a study exploring the diagnostic ability of three machine learning methods for evaluating the depression status of Chinese recruits [11].

Transformer based models such as BERT [12], [13] have been fine-tuned on Reddit text for classification tasks [14], [15]. [16] modified BERT for Multiple Choice Question Answering (MCQA) to predict users' answers to the BDI-II questionnaire. Variants such as RoBERTa [17] have been used to classify mental health disorders such as depression, anxiety, bipolar disorder, Attention Deficit Hyperactivity Disorder, and Post Traumatic Stress Disorder [8]. Using pre-trained models can maximize data efficiency, allowing for effective fine-tuning on smaller task-specific datasets. One of the setbacks of models that use the original attention mechanism is that they are limited to handling a maximum of 512 tokens. This is because the self-attention mechanism scales quadratically. The Longformer model was invented to solve this limitation. It uses the local window attention to scale linearly and global attention to attend to the entire sequence [18]. It can handle eight times longer tokens than BERT, an attribute necessary to process longer texts (e.g., from Reddit) to detect severity levels more precisely. The Longformer model has been used in clinical text and outperforms clinical-BERT and clinical-Big Bird models [19]. It has been used to detect depression in users from web-based forums [20], and predicts differential responses to antidepressant classes using electronic health records [21].

Most of these models have been trained in English, leaving low-resource languages like Luganda unattended. Luganda is one of the morphologically rich Bantu languages spoken in Uganda by over half of the population and neighboring East African countries. Funding for mental health services in Uganda remains low by international standards, with only 1% of GDP allocated for mental health services [22]. This lack of resources has limited access to mental health services, particularly in rural areas [23]. Furthermore, language and cultural norms influence communication about mental health topics experienced by patients receiving mental health treatment [24]. Therefore, in such a multilingual society, it is important to provide various avenues to handle mental health problems, particularly depression.

In this research, our objectives encompass two main aspects;

1) introducing a method for labelling social media text through a combination of keyword matching, a context-aware BART model, and an expert, and
2) refining the Longformer model's performance in classifying depression severity using fine-tuning, while focusing on both English and Luganda languages.

II. Methods

A. Data collection

Using the PRAW API, we collected a total of 1807 sentences from the r/depression subreddit[2]. The community ranks 805 with 972,203 subscribers [25]. It was assumed that the majority of the people who post under this subreddit have had a long-standing period of depression, as seen in the sample text below [sic];

"*I feel like a complete failure. I can't hold down a job and am getting a medical withdrawal from my semester in university. I feel like a waste of time and so much wasted effort for this school semester plus all the money. I'm having the hardest time doing basic self care activities like showering and changing my clothes. I feel so incredibly isolated and yet I keep ignoring peoples texts. I do not know what to do and I am sleeping around 12 hrs a night. I feel very very hopeless. Was just diagnosed with major depressive disorder and c-ptsd. Don't know where to go from here but trying to hold on.*"

The collected text was pre-processed to input for the models. First, every paragraph was converted to lowercase words. Then, stop words, unnecessary punctuations, spaces, and hyperlinks were removed.

B. Data labelling

1) **Keyword extraction and matching.**
   Using NLTK [26], a list of keywords was extracted from the 21 questions of the BDI questionnaire, while retaining the corresponding scores ranging from 0 to 3. A score of 0 indicates the absence of a symptom, while a score of 3 indicates the most severe manifestation of a symptom.
   Pattern matching was then conducted on each of the extracted sentences. A score value was assigned for every word match and aggregated to calculate a total score. The label was assigned based on the score range, as depicted in Figure 1, following the original BDI scoring standard.

2) **Classification with BART.** In this approach (Figure 2), we supplied the extracted text and the BDI severity labels to a pre-trained transformer-based BART model [27]. This model includes scientific text and medical content in its training stage.

3) **Domain expert.**
   This labelling was performed by a Psychiatrist who assigned the labels considering the BDI questionnaire and their expertise (Figure 3).

[2]https://www.reddit.com/r/depression/

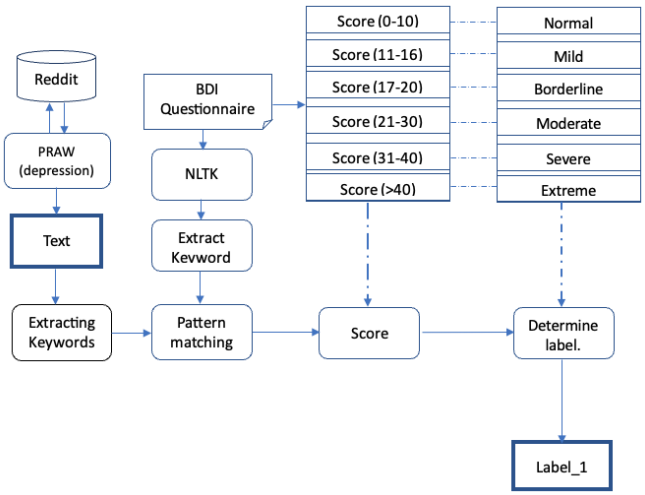

Fig. 1. Keyword extraction and matching.

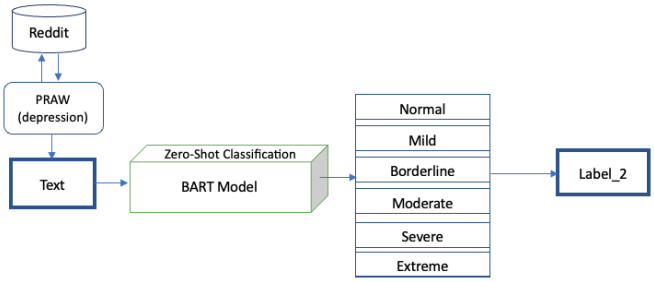

Fig. 2. BART model for labelling.

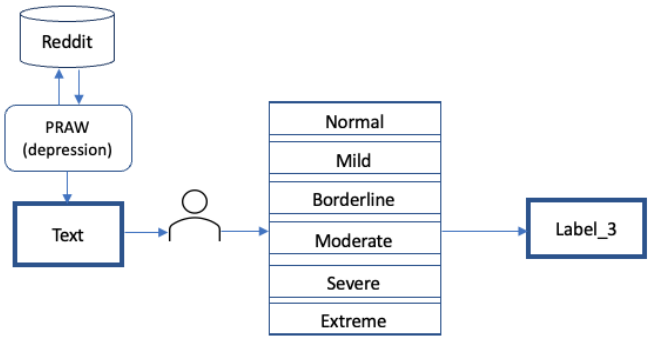

Fig. 3. Expert manual labelling.

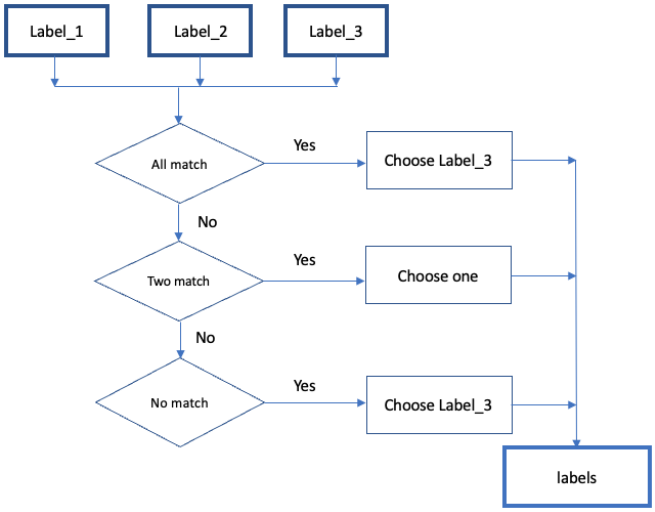

Fig. 4. Final label selection using weighted majority voting.

4) **Selecting the final label.**
The final label for the input text was assigned by weighted majority voting. The labels from **1)**, **2)** and **3)** were aggregated; the label was immediately assigned if the three agreed. If only two agreed, then that was the resulting label, and if none of them agreed, the final label was the expert's (Figure 4).

Finally, we obtained a labelled dataset with six classes: normal, mild, borderline, moderate, severe, and extreme. However, the "borderline" class was merged with the "mild" class labels, and the "extreme" class was merged with the "severe" class due to a low sample count of 17 and 23, respectively. Table I shows the final class distribution.

TABLE I
DATA DISTRIBUTION FOR THE FINAL SYNTHETIC LABELLED DATASET, FOR VALIDATION (V) AND TESTING (T).

Label	Total	English(V/T)	Luganda(V/T)
Normal	301	12/14	12/11
Mild	255	17/16	12/12
Moderate	372	15/14	21/21
Severe	215	13/14	12/14

C. Classification models

We used Google[3] to translate the input text from English to Luganda. The dataset was shuffled for each language and divided into training, testing, and validation. The training set was oversampled with the synthetic minority over-sampling technique (SMOTE) algorithm to address class imbalances [28].

For the final classification, we fine-tuned the Longformer model. This model has the ability to handle longer sequences of more than 512 tokens. It adjusts its attention mechanism for long sequences by combining strategies to lessen the computational constraints of processing such sequences. It attends to key parts of the input sequence using a combination of global and local attention methods, allowing it to capture long-range dependencies of the input sequence. It also employs an inter-attention mechanism that computes a distinct representation of the input for each output step, allowing the decoder to effectively "look at" the input's relevant part(s) for each output step. As a result, the encoder is relieved of the burden of encoding all information about the input sequence into a fixed-

[3]https://translate.google.com/

size rich representation vector. It has performed well in various tasks such as machine reading comprehension, summarization, and question answering [29].

We used the Longformer[4] model provided by the HuggingFace library [30]. The model consists of a default configuration of a hidden size of 768, 12 attention heads, an attention window size of 512, and 12 layers. The average number of characters per input varied between both languages, with English comprising 893 characters and Luganda 1776.

To optimize the model, we performed a hyperparameter search involving 8, 16, and 32 batch sizes, and learning rates ranging from 1e-1 to 5e-1. Additionally, early stopping was implemented. Model performance evaluation on validation and testing sets was based on precision, recall, F1-score and accuracy metrics.

After converting our data using TF-IDF, the following machine learning models were employed as baseline models: Naïve Bayes (NB) classifier, Random Forest (RF), Support Vector Machines (SVM), and Gradient Boosting (GB).

III. Results

The models were trained on Luganda and English, separately, and validated. Subsequently, they were evaluated using the test dataset (Table I). Precision, recall, and F-1 score were calculated for each severity level. The overall accuracy was calculated for each of the models employed (Table II and Table III).

Regarding the English dataset in Table II, the Random Forest model achieved the highest accuracy among the baseline models on the testing set (43%). The gradient boosting model failed as it could not predict any of the samples in the severe class. SVM had the lowest performance (41%). On the other hand, the Longformer model outperformed the baseline models with an accuracy of 48%. The hyperparameters used included a batch size of 16, a learning rate of 5e-5, and a dropout rate of 0.1.

For the Luganda dataset, similar to English, the random forest achieved the highest accuracy (40%), while SVM and GB performed poorly with failure to predict any of the samples in Mild and Severe classes respectively. However, unlike English, Naive Bayes achieved the lowest value (34%). The fine-tuned Longformer model achieved the highest accuracy of 45%, as shown in Table III. The hyperparameters used were a batch size of 16, and a learning rate of 4e-4.The performance did not require a dropout layer, as its inclusion negatively impacted the experiments.

TABLE II
Performance Metrics for Different Models on English

		Model				
Class	Metric	NB	RF	SVM	GB	Longformer
	Precision	1.00	0.69	0.10	0.58	1.00
	Recall	0.50	0.79	0.50	0.50	0.57
Normal	F-1	0.67	0.73	0.67	0.54	0.73
	Precision	0.23	0.45	0.50	0.28	0.37
	Recall	0.19	0.31	0.12	0.44	0.69
Mild	F-1	0.21	0.37	0.20	0.34	0.48
	Precision	0.15	0.27	0.28	0.24	0.34
	Recall	0.14	0.57	0.79	0.36	0.43
Moderate	F-1	0.15	0.36	0.21	0.29	0.39
	Precision	0.44	0.10	0.57	0.00	1.00
	Recall	0.79	0.07	0.29	0.00	0.21
Severe	F-1	0.56	0.13	0.38	0.00	0.35
Accuracy		0.40	0.43	0.41	0.33	0.48

Performance Analysis by Severity Level

TABLE III
Performance Metrics for Different Models on Luganda

		Model				
Class	Metric	NB	RF	SVM	GB	Longformer
	Precision	1.00	0.43	1.00	0.22	1.00
	Recall	0.45	0.83	0.67	0.92	0.55
Normal	F-1	0.62	0.57	0.80	0.35	0.71
	Precision	0.07	0.27	0.00	0.50	0.30
	Recall	0.08	0.25	0.00	0.17	0.50
Mild	F-1	0.08	0.26	0.00	0.25	0.37
	Precision	0.38	0.41	0.47	0.33	0.45
	Recall	0.29	0.43	0.90	0.05	0.43
Moderate	F-1	0.32	0.42	0.62	0.08	0.44
	Precision	0.35	1.00	0.86	0.00	0.42
	Recall	0.57	0.08	0.50	0.00	0.36
Severe	F-1	0.43	0.15	0.63	0.00	0.38
Accuracy		0.34	0.38	0.40	0.24	0.45

Performance Analysis by Severity Level

IV. conclusion and future work

With our success in fine-tuning the Longformer model on Luganda and English text to detect the severity of depression using a custom-made dataset from Reddit, we have shown that the longformer model can be finetuned on both languages outperforming machine learning models, that acted as baseline models.

We have also demonstrated that despite Luganda not being one of the languages originally trained with the Longformer model, it was successfully fine-tuned. However, we believe that training the Longformer model with the Luganda language and subsequently fine-tuning it could potentially yield better results.

The dataset used was small, and hence this is presumably the reason why the baseline machine learning models failed. A large dataset could improve the overall performance of all models and provide more samples for the severe and moderate classes.

The use of Google Translate as a machine translation model also affected the performance of the Luganda experiments. The service of a linguistic expert could be employed for a better translation output.

[4]https://huggingface.co/allenai/longformer-base-4096

References

[1] F. Fogarty, G. McCombe, K. Brown, T. Van Amelsvoort, M. Clarke, and W. Cullen, "Physical health among patients with common mental health disorders in primary care in europe: a scoping review," *Irish journal of psychological medicine*, vol. 38, no. 1, pp. 76–92, 2021.

[2] R. McAllister-Williams, C. Arango, P. Blier, K. Demyttenaere, P. Falkai, P. Gorwood, M. Hopwood, A. Javed, S. Kasper, G. Malhi, *et al.*, "The identification, assessment and management of difficult-to-treat

depression: an international consensus statement," *Journal of Affective Disorders*, vol. 267, pp. 264–282, 2020.
[3] A. T. Beck, R. A. Steer, G. K. Brown, *et al.*, *Beck depression inventory*. Harcourt Brace Jovanovich New York:, 1987.
[4] Reddit Inc., "Reddit Blog: Apifacts." https://www.redditinc.com/blog/apifacts. Retrieved on August 14, 2023.
[5] "Reddit Transparency Report 2022." https://www.redditinc.com/policies/2022-transparency-report. Retrieved on August 14, 2023.
[6] N. H. Kim, J. M. Kim, D. M. Park, S. R. Ji, and J. W. Kim, "Analysis of depression in social media texts through the patient health questionnaire-9 and natural language processing," *Digital Health*, vol. 8, p. 20552076221114204, 2022.
[7] J. Kim, J. Lee, E. Park, and J. Han, "A deep learning model for detecting mental illness from user content on social media," *Scientific reports*, vol. 10, no. 1, p. 11846, 2020.
[8] A. Murarka, B. Radhakrishnan, and S. Ravichandran, "Classification of mental illnesses on social media using roberta," in *Proceedings of the 12th international workshop on health text mining and information analysis*, pp. 59–68, 2021.
[9] U. Naseem, A. G. Dunn, J. Kim, and M. Khushi, "Early identification of depression severity levels on reddit using ordinal classification," in *Proceedings of the ACM Web Conference 2022*, pp. 2563–2572, 2022.
[10] R. Razavi, A. Gharipour, and M. Gharipour, "Depression screening using mobile phone usage metadata: a machine learning approach," *Journal of the American Medical Informatics Association*, vol. 27, no. 4, pp. 522–530, 2020.
[11] R. Francese and P. Attanasio, "Emotion detection for supporting depression screening," *Multimedia Tools and Applications*, vol. 82, no. 9, pp. 12771–12795, 2023.
[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[14] R. Shounak, S. Roy, V. Kumar, and V. Tiwari, "Reddit comment toxicity score prediction through bert via transformer based architecture," in *2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 0353–0358, IEEE, 2022.
[15] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, "Hatebert: Retraining bert for abusive language detection in english," *arXiv preprint arXiv:2010.12472*, 2020.
[16] J. Gabín, A. Pérez, and J. Parapar, "Multiple-choice question answering models for automatic depression severity estimation," *Engineering Proceedings*, vol. 7, no. 1, p. 23, 2021.
[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
[18] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.
[19] Y. Li, R. M. Wehbe, F. S. Ahmad, H. Wang, and Y. Luo, "A comparative study of pretrained language models for long clinical text," *Journal of the American Medical Informatics Association*, vol. 30, no. 2, pp. 340–347, 2023.
[20] D. Owen, D. Antypas, A. Hassoulas, A. F. Pardiñas, L. Espinosa-Anke, J. C. Collados, *et al.*, "Enabling early health care intervention by detecting depression in users of web-based forums using language models: Longitudinal analysis and evaluation," *JMIR AI*, vol. 2, no. 1, p. e41205, 2023.
[21] Y.-h. Sheu, C. Magdamo, M. Miller, S. Das, D. Blacker, and J. W. Smoller, "Ai-assisted prediction of differential response to antidepressant classes using electronic health records," *NPJ Digital Medicine*, vol. 6, no. 1, p. 73, 2023.
[22] S. A. Iversen, J. Nalugya, J. N. Babirye, I. M. S. Engebretsen, and N. Skokauskas, "Child and adolescent mental health services in uganda," *International journal of mental health systems*, vol. 15, pp. 1–12, 2021.
[23] M. W. Newman, M. Hawrilenko, M. Jakupcak, S. Chen, and J. C. Fortney, "Access and attitudinal barriers to engagement in integrated primary care mental health treatment for rural populations," *The Journal of Rural Health*, vol. 38, no. 4, pp. 721–727, 2022.
[24] I. Dagsvold, S. Møllersen, and V. Stordahl, "What can we talk about, in which language, in what way and with whom? sami patients' experiences of language choice and cultural norms in mental health treatment," *International journal of circumpolar health*, vol. 74, no. 1, p. 26952, 2015.
[25] "r/depression statistics." https://subredditstats.com/r/depression. Retrieved on August 15, 2023.
[26] E. Loper and S. Bird, "Nltk: The natural language toolkit," *arXiv preprint cs/0205028*, 2002.
[27] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.
[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
[29] H. Basafa, S. Movahedi, A. Ebrahimi, A. Shakery, and H. Faili, "Nlp-iis@ ut at semeval-2021 task 4: Machine reading comprehension using the long document transformer," *arXiv preprint arXiv:2105.03775*, 2021.
[30] S. M. Jain, "Hugging face," in *Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve Problems*, pp. 51–67, Springer, 2022.