DOI: 10.21917/ijsc.2025.0555

TRANSFORMING BINARY DEPRESSIVE DATA INTO SEVERITY INSIGHTS THROUGH AUTOMATIC CLUSTERING AND LABELLING

Divya Dewangan[1], Smita Selot[2] and Sreejit Panicker[3]
[1,2]*Department of Computer Application, Shri Shankaracharya Technical Campus, India*
[3]*Techment Technology, India*

Abstract

Depression affects individuals globally, necessitating efficient diagnostic tools. This study introduces an advanced unsupervised hybrid approach, that automatically converts binary-labelled depression datasets into multi-class datasets by integrating a rule-based system with Large Language Models (LLMs). The rule-based system employs the Beck Depression Inventory-II tool that can be used to classify depression levels based on predefined scoring rules, and these rules are segregated into clusters based on score ranges from 0-3. LLMs employ fine-tuned large language Model Meta AI2 (LLaMa2) to generate domain-specific embedding from social media posts. By harnessing LLMs' contextual understanding, both BDI rules and social media posts are embedded, thereafter cosine similarity is applied to calculate semantic similarities. Based on the similarity score, each post is assigned to the most similar BDI cluster, with the highest similarity score, creating a refined multiclass depression cluster. To evaluate clustering effectiveness, the silhouette score was computed, yielding an average score of 0.45, indicating moderate clustering quality. Additionally, 30% of the binary depression dataset was manually labelled by clinical experts. The Normalized Mutual Information (NMI) score of 0.53 further validated the method, showing strong alignment between the generated clusters and expert-labelled data. This approach enhances depression severity classification, providing a scalable, efficient, and accurate tool for researchers and practitioners.

Keywords:

Large Language Model, Multiclass Depression Dataset, Beck Depression Inventory Tool, Semantic Similarity

1. INTRODUCTION

Depression is a common mental health issue faced by almost every person. A lack of communication, overthinking, social stigma, and a lack of proper treatment can increase the severity of this disease [1]. The World Health Organization launched the Mental Health Gap Action Programme (mhGAP) guidelines in 2008 to scale up mental health care services, with proper care and psychological support. Under these circumstances, millions of people in low- and middle-income countries have been able to receive treatment for depression, schizophrenia, and epilepsy; prevent suicide; and start living a normal life [2].

Psychiatrists suggest that sharing feelings can help individuals resolve many issues. However, in today's fast-paced lifestyle, people often struggle to do so, leading to depression and mental distress [3-5]. While many experience this condition, identifying the severity of depression is crucial. People are more comfortable expressing their emotions on social media than with close friends or family [6]. Depression varies widely in severity, symptoms, and personal expression. One person may describe emotional despair, while another highlights physical symptoms like fatigue. Capturing these diverse expressions in a structured format is challenging [7]. Unlike structured data, depressive symptoms are highly variable in wording and meaning. Traditional NLP models struggle with this complexity, as they rely on fixed patterns. Supervised learning approaches also require extensive manual labelling, making psychiatric data annotation costly and time-consuming [8].

Most research on depression detection has focused on identifying its presence, with numerous binary depression datasets available for this purpose and such datasets are used to classify individuals as either depressed or not. However, depression is more complex than mere presence or absence; understanding its severity is equally crucial [9]. There is a fine line between sadness and clinical depression—while sadness is temporary, depression is persistent and tends to intensify over time.

Rather than relying on isolated words like "hopelessness" or "emptiness," modern approaches utilise sentence embeddings, which encode entire sentences into high-dimensional vector spaces [10]. These embeddings capture contextual meaning, addressing the limitations of individual word representations. Similar meanings are placed closer together, enhancing generalisation across different phrasings [12,13]. Pretrained models like BERT, GPT, and LLaMA can be fine-tuned on domain-specific, unlabelled datasets using methods such as LoRA or self-supervised learning, reducing reliance on extensive labelled data [14-16].

By analysing large text corpora, these models generate vector representations of words, phrases, and sentences, positioning semantically similar expressions near each other. For instance, "I feel hopeless" and "There is no hope left for me" would have closely aligned embeddings [17,18]. This technique enables the automatic clustering of depression symptoms, improving detection and classification without requiring exhaustive manual annotation [19].

The scarcity of multi-class depression datasets presents a major challenge in accurately predicting the severity or specific state of an individual's depression. This study introduces a state-of-the-art approach that leverages an unsupervised hybrid methodology, integrating a rule-based system and Large Language Models (LLMs) to automatically convert binary-labelled depression datasets into multi-class datasets.

The rule-based system utilizes the Beck Depression Inventory-II (BDI-II), which categorizes depression levels based on predefined scoring criteria, grouping them into clusters ranging from 0 to 3. Concurrently, the fine-tuned Large Language Model Meta AI2 (LLaMA2) generates domain-specific embeddings from social media posts.

By harnessing the contextual comprehension of LLMs, embeddings are created for both BDI rules and social media posts. The cosine similarity metric is then applied to measure the

semantic resemblance between these embeddings. Each post is assigned to the cluster with the highest similarity score, forming a structured multi-class depression dataset. The silhouette score is used for evaluation, to ensure the effectiveness and cohesiveness of the clustering results. This refined approach enhances the classification of depression severity, providing a more granular and interpretable dataset for mental health analysis.

The rest of the paper is organized as follows. Section 2 presents recent works on depression and depression severity prediction. Section 3 discusses the materials and methods used, including the formulation of the problem, the data preparation, word embeddings, Semantic Similarity Calculation and Label Prediction, and evaluation. Section 4 presents the results and discussion, and Section 5 presents the conclusions of the findings.

2. RELATED WORK

Recharla et al. [7] introduced a method for detecting depression symptoms using Sentence Transformers, leveraging similarity scores between sentences. This approach showed strong performance, particularly when equal weights were applied across options. Their experiments highlighted how varying weight distributions impact sentence relevance. Munoz and Iglesias [10] introduced a feature-based framework that combines linguistic signals, emotional expressions, and cognitive patterns with distributional representations. Using traditional machine learning, it achieved high F1 scores with balanced performance and efficiency. Yang et al. [11] proposed an unsupervised K-means clustering method to classify depressive symptoms, addressing the limitations and outperforming norm-based methods in robustness, model accuracy, AUC, and construct validity.

Babu and Kanaga [12] leverage social media data, including text, emoticons, and emojis, for sentiment analysis, and classifying sentiment into binary, ternary, or multiclass categories. Deep learning-based multiclass classification enhances precision in identifying sentiment polarities, aiding in detecting emotions like apprehension or dejection. Imans et al. [14] focused on a novel explainable multilayer dynamic ensemble framework for depression detection and severity prediction using data from the National Social Life, Health, and Ageing Project. Sampath and Thenmozhi [15] emphasize social media's role in early depression detection, where transformer-based models outperformed traditional methods in the DepSign-LT-EDI@ACL 2022 task, achieving a top macro F1 score of 0.583.

Lau et al. [16] proposed a novel method to tackle data scarcity in depression assessment by utilising a pre-trained large language model with efficient prefix-tuning techniques. Using the DAIC-WOZ dataset, their model predicts PHQ scores while minimizing overfitting and reducing the number of tunable parameters. The prefix-tuned model surpassed existing approaches, offering improved accuracy and flexibility for automatic depression assessment. Perez et al. [17,20] developed a semantic pipeline for assessing depression severity via social media, leveraging BDI-II symptoms through multiclass classification with unsupervised and semi-supervised data selection methods. Additionally, their study on Reddit data highlighted the effectiveness of neural models in estimating depression scores, even with limited training data, demonstrating a scalable alternative to self-report questionnaires.

Miniltsevich et al. [19] applied symptom network analysis and trained a multi-target hierarchical regression model on DIAC-WOZ transcripts to predict individual symptoms. Their model achieved a macro-F1 score of 73.9 in binary classification and effectively assessed depression severity, providing a personalized approach to understanding depression. Mao et al. [21] tackled the challenge of accurate depression diagnosis using an attention-based multimodal model that integrated speech and text features. By combining a Bi-LSTM with a time-distributed CNN, they demonstrated improved severity prediction. Bao et al. [22] explored transformer-based architectures for detecting and explaining depressive symptoms, utilizing both separate models and LLM fine-tuning, enhancing interpretability and achieving strong classification performance.

Agrawal et al. [23] utilized advanced Large Language Models (LLMs) like GPT-4, LLaMA2, and Gemini for enhanced depression detection and treatment. Fine-tuning these models with innovative prompt engineering enables DSM-5-based diagnoses, empathetic conversations, and personalized Cognitive Behavioral Therapy interventions. Shah et al. [24] explored fine-tuned LLMs, including GPT-3 Turbo 1106 and LLaMA2-7B, for analyzing social media posts, achieving a 96% accuracy rate—outperforming existing methods in detecting depressive content. Lan et al. [25] introduced a novel system that leverages large language models (LLMs) to annotate high-risk texts based on medical criteria, summarize emotional patterns, and integrate traditional classifiers for explainable predictions.

3. MATERIALS AND METHODS

The proposed system involves four keys' states, namely problem formulation, data preprocessing, word embedding, similarity calculations and evaluation, as outlined in Fig.1. This figure describes the proposed automatic labelling system.

3.1 PROBLEM FORMULATION

This paper highlights the data scarcity challenge in depression research, particularly for multi-class labelled datasets, which are not publicly available. While binary depression datasets (e.g., depressed vs. non-depressed) exist, the lack of multi-class datasets makes it difficult to train and evaluate models for finer-grained severity classification. Additionally, subjectivity in labelling leads to inconsistencies, as symptom intensity often overlaps across categories [26-28]. Clinician and self-assessment variations further contribute to inconsistent labelling. Data imbalance is another major issue, with mild depression cases being more common, while severe cases are underrepresented, making model training challenging. Moreover, privacy concerns and stigma discourage individuals from sharing their mental health data, limiting dataset availability [29]. The challenges of data scarcity in multi-class depression classification directly justify the need for the unsupervised hybrid approach, since no publicly available labelled dataset exists for multi-class depression, the hybrid approach eliminates reliance on manual labelling. The approach combines a rule-based system (BDI-II) and Large Language Models (LLaMA2) to automatically transform binary-labelled datasets into multi-class datasets. The BDI-II scoring rules provide predefined, objective scoring mechanisms and standardized thresholds, reducing human bias in

labelling. LLMs fine-tuned with domain-specific data can capture linguistic cues and contextualize symptoms in user-generated text, making classification more data-driven than purely subjective.

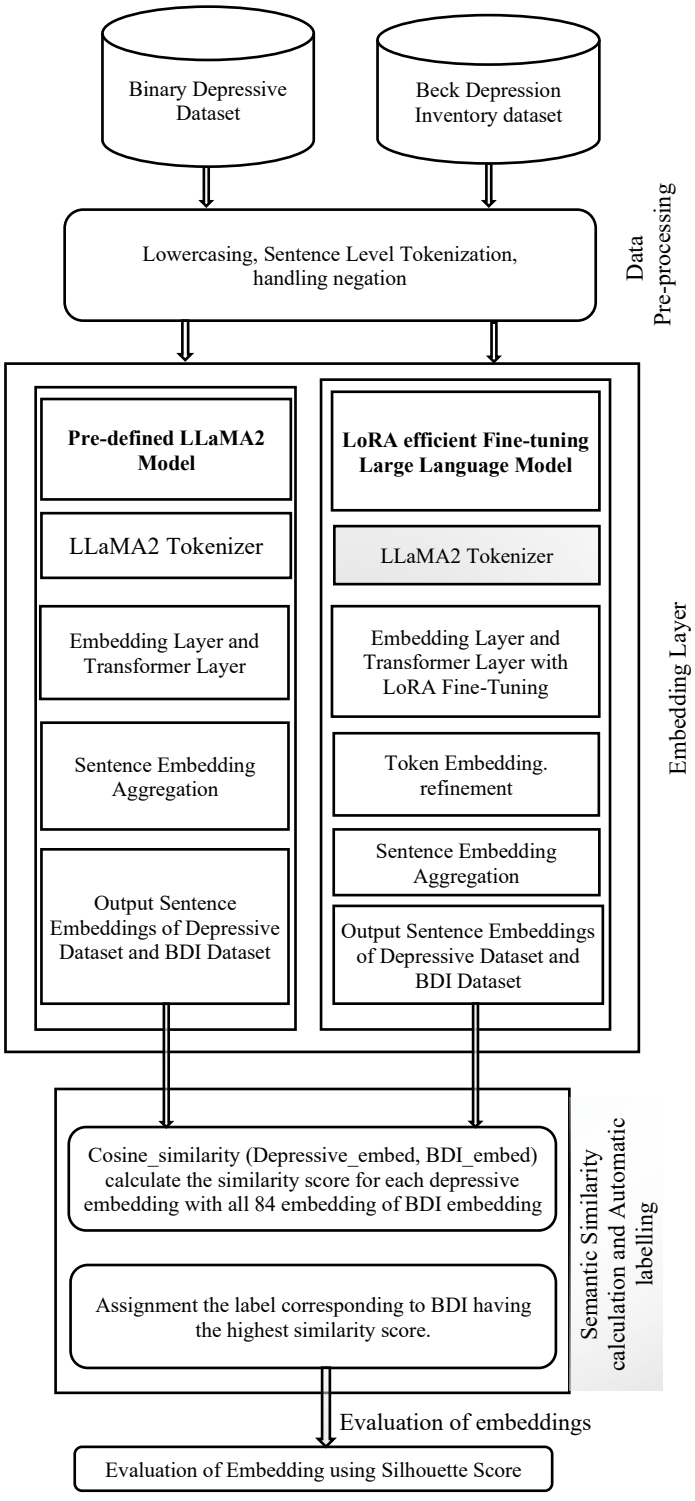

Fig.1. An illustrative description of the proposed Approach to Clustering and Labelling Binary Depressive Data Based on Severity Level

3.2 DATA COLLECTION AND PREPARATION

The dataset used in this study is sourced from online platforms such as the Reddit Depression dataset, which consists of 7,731 instances labelled in a binary format (0 - not depressed, 1 - depressed). However, due to the lack of publicly available multi-class depression datasets, the binary-labelled dataset is mapped to the Beck Depression Inventory-II (BDI-II) tool to transform it into a multi-class dataset [17].

BDI-II is a widely used self-report questionnaire designed to assess the severity of depression. It consists of 21 items; each rated on a four-point scale (0-3). To align this tool with the binary dataset, the BDI-II scores are converted into four distinct labels (0, 1, 2, 3), representing different levels of depression. Each label corresponds to a set of 21 values, resulting in a total of 84 mapped values in the multi-class dataset.

The BDI-II tool is highly beneficial for both clinicians and researchers. Clinicians use it as a standardized tool to assess depression severity, aiding in diagnosis and treatment planning. In our research, it provides an objective framework for automatically classifying depression levels, reducing subjectivity and enhancing model performance [11].

To generate context-aware embeddings, LLaMA2 is employed. Preprocessing steps, including text cleaning, lowercasing, and tokenization, are applied minimally to preserve semantic meaning, ensuring the embeddings effectively capture depression-related linguistic patterns [26].

3.3 WORD EMBEDDINGS

Word embeddings are a representation technique that maps words into a continuous vector space, where similar words have similar vector representations. Traditional word embeddings, like Word2Vec and GloVe, encode individual words as fixed-dimensional vectors [20]. In contrast, sentence embeddings, such as Sentence-BERT and LLMs, represent entire sentences, capturing their semantic meaning [22]. Large Language Models (LLMs) advance word embeddings by generating context-dependent representations using transformer architectures [24]. These models leverage self-attention mechanisms to capture long-range dependencies and contextual nuances within text, enhancing their ability to generate dynamic and meaningful embeddings based on surrounding words [30,31].

3.3.1 General-Purpose Embedding via Large Language Model (LLaMA2):

To efficiently generate word embeddings, we employed the transformer architecture of LLMs [31]. It is a transformer-based language model with layers consisting of:

1. *Attention Mechanism*: For a given input sequence $X \in R^{n \times d}$ (where n is the sequence length, and d is the embedding size), the attention output is computed in Eq.(1) as [31]:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where, $Q = XW_Q, \quad K = XW_K, \quad V = XW_V$

W_Q, W_K, and W_v are learnable weight matrices for the query, key, and value transformations.

2. *Feedforward Layers*: After attention, the output is passed through the feedforward network in Eq.(2) as [32]:

$$\mathrm{FFN}(x) = \sigma(xW_1 + b_1)W_2 + b_2 \quad (2)$$

where W_1, W_2 and b_1, b_2 are weights and biases and σ is a nonlinear activation function.

The pre-trained LLaMa2 model is used directly to generate embeddings without additional fine-tuning. It captures broad semantic relationships between sentences or texts. It tokenizes the input text via the tokenizer provided with the LLaMA2 model. The tokenized input is passed into the model, and the embeddings are extracted from a suitable layer (e.g., the final transformer layer output or a pooled embedding from the CLS token) [31,32].

3.3.2 Domain-Specific Embedding via LoRA Efficient Fine-Tuning Large Language Model (Fine-Tuned LLaMA2):

With many parameters, fine-tuning LLaMA2 is computationally expensive. LoRA provides an efficient solution by injecting low-rank matrices into certain model layers without updating the entire model [32].

• ***Low-Rank Decomposition***:

Let $W \in \mathrm{R}^{d\times k}$ be the original weight matrix (e.g., in the attention layer). Instead of updating W directly, LoRA modifies it in Eq.(3) as:

$$W' = W + \Delta W \quad \text{where} \quad \Delta W = AB \quad (3)$$

where ΔW is the learned low-rank update.

$$A \in \mathrm{R}^{d\times r}, \quad B \in \mathrm{R}^{r\times k}$$

where A and B are smaller matrices trained during fine-tuning. R $\min(d, k)$: ΔW has a much lower rank. This reduces the trainable parameter significantly, as the parameters added by LoRA are defined in Eq.(4):

$$ParamsLoRA = r(d+k) \quad (4)$$

Compared with $d\times k$ for the original weight matrix.

• ***Gradient Update***:

During fine-tuning, only A and B are updated via backpropagation, defined in Eq.(5) [32] as:

$$A \leftarrow A - \eta \frac{\partial L}{\partial A}, \quad B \leftarrow B - \eta \frac{\partial L}{\partial B} \quad (5)$$

where η is the learning rate and L is the loss function.

Once the LLaMa2 model is fine-tuned with LoRA, the texts pass through the fine-tuned LLaMA2 model to generate embeddings tailored to the depression domain. These embeddings capture subtle relationships between the input text and depression-related labels. The model generates the embeddings for both the tweet dataset and the BDI dataset [33].

3.4 SEMANTIC SIMILARITY CALCULATION AND LABEL PREDICTION

Semantic search is an effective method for identifying the most relevant BDI sentences within a paragraph by focusing on meaning rather than keyword matching. It helps to capture the emotional state or theme of the text. By representing sentences as high-dimensional embeddings, semantic similarity is determined by measuring the distance between these embeddings [34].

BDI sentences reflect varying emotional intensities, from normal to severe depression. When analyzing a long paragraph, the goal is to align its emotional tone with the most relevant BDI sentence. Even without identical wording, semantic search captures underlying meanings. For example, a paragraph expressing persistent sadness can be matched with a BDI sentence like, the paragraph: "I feel like no matter what I do, I can't seem to shake off this constant sadness. BDI Sentence: "I am sad all the time, and I can't snap out of it.".

3.4.1 Similarity Calculation of General-Purpose Embeddings:

After generating embeddings, cosine similarity scores are computed between each sentence in the tweet dataset and the BDI sentences for every paragraph in the tweet dataset. Algorithm 1 presents the pseudo-algorithm for calculating similarity using general-purpose embeddings.

Algorithm 1: Pseudo Algorithm for Similarity Calculation of General-Purpose Embeddings.

Input:

- Dt: Tweet_data dataset with N sentences $\{t_1, t_2, ..., t_N\}$.
- D_{BDI}: BDI dataset with M sentences $\{b_1, b_2, ..., b_M\}$ and their corresponding labels.
- Pre-trained LLaMA2 model $\mathrm{Model}_{\mathrm{LLaMA2}}$.

1. *Preprocessing*:

Tokenize all sentences in Dt and D_{BDI} the LLaMA2 tokenizer.

$$T_t = Tokenize(D_t) = \{tok_1, tok_2, \ldots, tok_N\}$$
$$T_{BDI} = Tokenize(D_{BDI}) = \{tok'_1, tok'_2, \ldots, tok'_M\}$$

2. *Embedding Generation*:

Pass each tokenized sentence into LLaMA2 to obtain embeddings Et for tweets and EBDI for BDI sentences

3. *Cosine Similarity Calculation*:

Compute cosine similarity between each tweet embedding (Et) and every BDI embedding (EBDI), identifying the most similar BDI sentence

$$S[i,j] = \frac{E'_t[i] \cdot E'_{BDI}[j]}{\square\, E'_t[i] \square\square\, E'_{BDI}[j] \square}$$
$$\forall i \in \{1,\ldots,N\}, \forall j \in \{1,\ldots,M\}$$

4. *Label Assignment*: Assign the label of the most similar BDI sentence to the corresponding tweet and store the similarity score, matched BDI sentence, and label.
5. *Output*: Return a list containing sentences in tweet_data, its predicted label based on the most similar BDI sentence, and similarity score.

3.4.2 Similarity Calculation of Domain-Specific Embeddings:

Fine-tuned domain-specific embeddings capture the unique semantic relationships in depression-related sentences, enhancing the model's understanding of this specific domain. These refined embeddings are then used for similarity calculations [35]. Cosine similarity measures the similarity between tweet embeddings and BDI sentence embeddings, ensuring the scores reflect domain-specific nuances. The highest similarity score determines the most relevant BDI sentence label, which is then assigned to the corresponding tweet. This process enables the automatic labelling of tweets based on their closest BDI match. Finally, the newly labelled dataset is saved as a pickle or .xlsx file. Algorithm 2 outlines the similarity calculation process.

Similarity calculation of domain-specific embeddings Fine-tuned domain-specific embeddings capture the unique semantic relationships in depression-related sentences, enhancing the model's understanding of this specific domain. These refined embeddings are then used for similarity calculations [35]. Cosine similarity measures the similarity between tweet embeddings and BDI sentence embeddings, ensuring the scores reflect domain-specific nuances. The highest similarity score determines the most relevant BDI sentence label, which is then assigned to the corresponding tweet. This process enables the automatic labelling of tweets based on their closest BDI match. Finally, the newly labelled dataset is saved as a pickle or .xlsx file. Algorithm 2 outlines the similarity calculation process.

3.5 EVALUATION

To evaluate the quality of clustering in the newly labelled dataset, we utilize the silhouette score [36], which assesses how well the embeddings and predicted labels align. This metric quantifies the separation of clusters by measuring how similar a data point is to its assigned cluster compared to other clusters. The silhouette score ranges from -1 to 1, where 1 indicates well-separated clusters, 0 suggests overlapping clusters, and -1 implies incorrect clustering.

Algorithm 2: Pseudo Algorithm for Approach to Similarity Calculation of Domain-Specific Embeddings.

Input:
Dt: Tweet_data dataset with N sentences $\{t_1, t_2, \ldots, t_N\}$.
D_{BDI}: BDI dataset with M sentences $\{b_1, b_2, \ldots, b_M\}$and their corresponding labels.
Pre-trained LLaMA2 model $\text{Model}_{\text{LLaMA2}}$.
LoRA adaptation matrices A and B for domain-specific fine-tuning.

1. *Fine-tuning with LoRA*:
   a. Initialize low-rank adaptation matrices A and B ($A,B \in R^{(d\times r)}$, $r \ll d$)
   b. Fine-tune the LLaMA2 model using the BDI sentence and their corresponding labels.
   c. Compute the weight update $\Delta W = A \cdot B$
   d. Update the original weights: $W' = W + \Delta W$
   e. Save the LoRA-adapted weights.
2. *Preprocessing*:
   a. Tokenize all sentences in D_t and D_{BDI} using the fine-tuned LLaMA2 tokenizer.
$$Tt = Tokenize(Dt),$$
$$TBDI = Tokenize(DBDI).$$
3. *Embedding Generation with fine-tuned model*:
   a. Load LoRA fine-tuned LLaMA2 model.
   b. Generate domain-specific embeddings from the model:
$$E'_{ti} = \text{Model}_{\text{LLaMA2}}(T_{ti}),$$
$$E'_{BDIj} = \text{Model}_{\text{LLaMA2}}(T_{BDIj})$$
4. *Cosine Similarity Calculation*:
   For each sentence embedding in tweet_data E'_t compute the cosine similarity with each sentence embedding in BDI data E'_{BDI} and identify the BDI sentence with the highest similarity score.

$$S[i,j] = \frac{E'_t[i] \cdot E'_{BDI}[j]}{\square\, E'_t[i] \square\square\, E'_{BDI}[j] \square}$$
$$\forall i \in \{1,\ldots,N\}, \forall j \in \{1,\ldots,M\}$$

5. *Label Assignment*:
   a. Assign the label of the most similar BDI sentence to the corresponding sentence in tweet_data
   b. Store the similarity score, BDI sentence, and label.
6. *Output*:
Return a list containing sentences in tweet_data, its predicted label based on the most similar BDI sentence, and similarity score.

For a dataset $X = \{x_1, x_2, \ldots, x_n\}$ consisting of N embedded data points (e.g., tweets or BDI sentences) and their corresponding predicted labels $Y = \{y_1, y_2, \ldots, y_n\}$, the silhouette score for each point x_i is defined in Eq.(6) as.,

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))} \tag{6}$$

where $a(x_i)$ is the average cosine distance to all other points in the same cluster and where $b(x_i)$ is the minimum average cosine distance to points in the nearest different cluster. The mean silhouette score across all the data points is defined in Eq.(7) as (Geronimo, 2024) [36]:

$$S = \frac{1}{N}\sum_{i=1}^{N} s(x_i) \tag{7}$$

To further validate the automatic labelling process, Normalized Mutual Information (NMI) is computed on a manually labelled subset, where 30% of the binary depression dataset was labelled by clinical experts.

NMI is a metric used to compare how well the automatically generated clusters match the ground truth labels. It measures the mutual dependence between the two label sets (predicted vs. actual) by quantifying the shared information between them. Mathematically, NMI is defined as:

$$\text{NMI}(Y, C) = \frac{2 \times I(Y;C)}{H(Y) + H(C)} \tag{8}$$

where $I(Y;C)$ is the mutual information between predicted labels C and ground truth labels Y. $H(Y)$ and $H(C)$ are the entropy of the true and predicted labels, respectively. It ranges from 0 to 1, where 1 indicates perfect alignment between the clusters and the ground truth, while 0 suggests no shared information [37,38].

4. RESULT AND DISCUSSION

This experiment aimed to automatically label tweets using predefined Beck Depression Inventory (BDI) categories based on semantic similarity. Sentence embeddings were generated using either a pre-trained or fine-tuned LoRA LLaMA2 model and compared with BDI embeddings. Cosine similarity determined the most relevant BDI sentence for each tweet, assigning the corresponding label. If multiple labels were possible, the highest similarity score was used. Clustering quality was assessed using the silhouette score. The dataset included 7,731 tweets, with 3,900 non-depressive and 3,831 depressive tweets. Only depressive tweets were labelled into four BDI categories: normal (0), mild

(1), moderate (2), and severe (3). The experimental results of the labelled outputs are presented in Table.1.

Table.1. Label of the most similar sentence is automatically assigned based on a Cosine Similarity Score.

Input sentences	Predicted output
we understand that most people who reply immediately to an op with an invitation to talk privately mean … we have huge admiration and appreciation for the goodwill and good citizenship of so many of you ….	Predicted label: [2] Best Matching BDI Sentence: [["I am very worried about physical problems and it's hard to think of much else."]] Similarity Score: [0.6746545469133172]
i live alone and despite me being prone to loneliness i find myself to be emotionally needy i seem to have … the urge to push people away it feel like i am being consumed by depression again.	Predicted label : [1] Best Matching BDI Sentence: [['i have thoughts of killing myself, but i would not carry them out.']] Similarity Score: : [0.8288883220263281]
idk how to elaborate on it i just started suddenly cry for no real reason and couldn t stop for like 0 minute doe anyone else have this problem i m just wondering.	Predicted label : [2] Best Matching BDI Sentence: [["i am sad all the time and i can't snap out of it."]] Similarity Score: : [0.8181921556265996]
i tried to help his family abandoned him so it wa really hard to change his … i found a necrology of him he killed himself march he texted many time to me about suicide thought im so sad.	Predicted label : [1] Best Matching BDI Sentence: [['i have thoughts of killing myself, but i would not carry them out.']] Similarity Score: [0.7577257431552242]
I don't know how to communicate all of … don't know what to do I feel like a horrible waste of human space stripped down to my bone by my past and my seemingly dim future.	Predicted label : [3] Best Matching BDI Sentence: ['i feel the future is hopeless and that things cannot improve.']] Similarity Score: [0.8000499442387334]

Table 2. Silhouette Score and Distribution of the predicted labels assigned to each label are represented as percentages (%)

	Silhouetted Score	Label 0 (in %)	Label 1 (in %)	Label 2 (in %)	Label 3 (in %)
Pre-trained LLaMA2 model	0.32	5%	43%	34%	18%
Fine-tuned LoRA LLaMA2 model	0.45	3%	32%	42%	23%

This paper presents two main experiments at an embedding level using the pre-trained and fine-tuned LoRA LLaMA2 model. Table.2 presents the silhouette score and distribution of the predicted labels assigned to each label as percentages (%). The silhouette score is computed as 0.32 and 0.45 for the pre-trained LLaMA2 and the fine-tuned LoRA LLaMA2, respectively. This indicates moderately distinct clusters, suggesting that the assigned labels reflect meaningful groupings based on sentence-level embeddings.

The Table.2 shows that the pre-trained LLaMA2 model achieves a relatively low silhouette score of 0.32, suggesting significant overlap among clusters, making it difficult to distinguish between the four depression categories (0-normal, 1-mild, 2-moderate, and 3-severe). In contrast, fine-tuning with LoRA improves clustering, yielding a higher silhouette score of 0.45, which reflects better differentiation among labels. Fig.2. illustrates that pie chart that shows the pre-trained embeddings fail to capture fine-grained depression categories, while LoRA fine-tuning results in a more balanced label distribution. Label 0 remains low at 3% and 5%, as only depressive posts are analyzed. The pie chart provides a comparative view of label distributions (0, 1, 2, and 3) between the pre-trained and fine-tuned models.

Normalized Mutual Information (NMI) measures the similarity between the predicted clusters and the ground truth labels, ranging from 0 to 1, where higher values indicate better alignment. In this study, 30% of the binary depression dataset was manually labelled by clinical experts to serve as a validation benchmark.

For the predefined LLaMA2 model, the NMI score was 0.46, indicating moderate alignment but also suggesting that the model struggled to differentiate between depression severity levels effectively. The embeddings lacked domain-specific knowledge, leading to overlapping clusters.

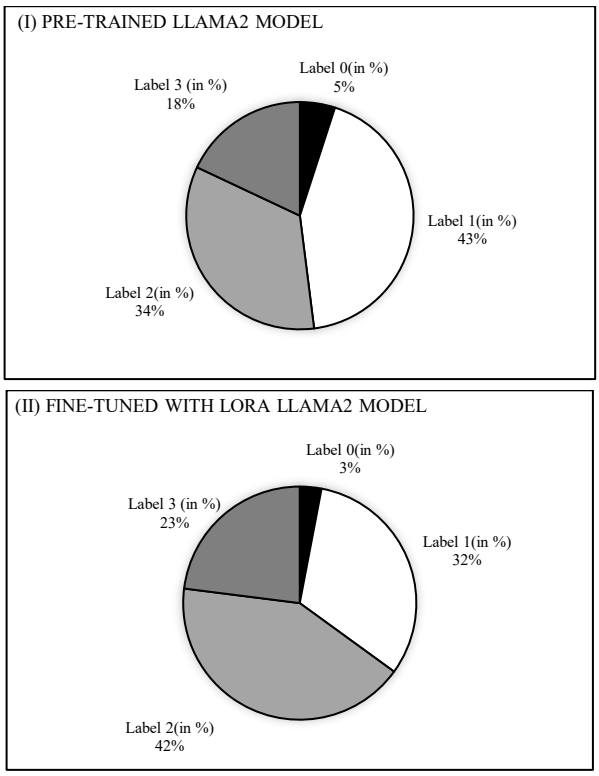

Fig.2. Pie chart represents the comparison of the label distributions with 0,1,2, and 3 between (I) the pre-trained and (II) fine-tuned LoRA LLaMA2 model

After fine-tuning with LoRA, the NMI score improved to 0.53, demonstrating better cluster coherence and alignment with expert labels. This improvement highlights that fine-tuning enhances the model's ability to capture nuanced distinctions between depression categories, making it more effective for automated depression classification based on BDI labels.

5. CONCLUSION

This study investigated the use of Large Language Models (LLMs), specifically the pre-trained LLaMA2 model and its fine-tuned counterpart with LoRA, for the automatic labelling of depression-related datasets. By leveraging the efficacy of large language models, tweets and Beck Depression Inventory (BDI) sentences were embedded to calculate their semantic similarity and predict depression severity categories (normal, mild, moderate, and severe) based on similarity scores. The results highlight the importance of domain-specific fine-tuning, which allowed the model to better align with depression-related language patterns, resulting in improved clustering quality and label distribution.

The unsupervised hybrid approach, which combines a rule-based system and Large Language Models, played a crucial role in enhancing depression classification. The rule-based system utilized BDI scoring rules to define depression severity levels, while LLMs, including fine-tuned LLaMA2, generated embeddings that captured the semantic nuances of social media posts. By embedding both BDI rules and social media text, the hybrid method facilitated the transformation of binary-labelled depression datasets into multi-class labels, improving classification accuracy.

The Normalized Mutual Information (NMI) score further reinforced these findings by quantifying the alignment between predicted clusters and actual depression labels. A higher NMI score for the fine-tuned model demonstrated that LoRA fine-tuning improved label consistency and classification accuracy, reducing semantic overlap among depression categories. Conversely, the low NMI score of the pre-trained model highlighted its limitations in capturing nuanced language variations essential for accurate multi-class depression classification.

Despite these improvements, data scarcity and label ambiguity remain significant challenges, particularly for less-represented depression categories. The low percentage of tweets classified under certain labels affects the model's ability to generalize effectively. Future research should focus on data augmentation, transfer learning, and unsupervised pre-training on domain-specific datasets to address these challenges. Additionally, exploring alternative clustering evaluation metrics and incorporating contextual features may further refine classification accuracy and model performance.

This study underscores the effectiveness of combining rule-based methods with LLMs for depression classification and emphasizes the importance of domain-specific fine-tuning in improving multi-class depression prediction from social media data.

REFERENCES

[1] World Health Organization (WHO), "Depression", Available at https://www.who.int/news-room/fact-sheets/detail/depression, Accessed in 2023.

[2] Mental Health Gap Action Programme (mhGAP), "WHO Mental Health, Brain Health and Substance Use", Available at https://www.who.int/publications/i/item/9789240084278, Accessed in 2023.

[3] K.S. Sowbarnigaa, S. Kogilavani, R.V. Sriraam, S. Malliga, R. Sandhiya and V.E. Sathishkumar, "Leveraging Multiclass Sentiment Analysis on Social Media Text for Detecting Signs of Depression", *Proceedings of 4th International Conference on Computing, Data Science*, pp. 1-8, 2022.

[4] M. Mohit, M. Maruf, H. Ahmed and M. Alam, "Depression and Physical Illnesses: An Update", *Bangladesh Medical Journal*, Vol. 40, No. 1, pp.53-58, 2011.

[5] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt and M. Pantic, "Real-Life Depression, and Affect Recognition Workshop and Challenge", *Proceedings of International Conference on Audio/Visual Emotion Challenge*, pp. 3-9, 2017.

[6] C. Arango, P.D. McGorry, J. Rapoport, I.E. Sommer, J.A. Vorstman, D. McDaid, E. Serrano-Drozdowskyj, R. Freedman and W. Carpenter, "Preventive Strategies for Mental Health", *The Lancet Psychiatry*, Vol. 5, No. 7, pp. 591-604, 2018.

[7] N. Recharla, P. Bolimera, Y. Gupta and A.K. Madasamy, "Exploring Depression Symptoms through Similarity Methods in Social Media Posts", *Proceedings of International Conference on Evaluation Forum*, pp. 1-7, 2023.

[8] Y. Zang, O. Zang, Y. Wu, H.J. Lee, J. Xu, H. Xu and K. Roberts, "Psychiatric Symptoms Recognition without Labelled Data using Distributional Representations of Phrases and On-Line Knowledge", *Journal of Biomedical Informatics*, Vol. 75, pp. 129-137, 2017.

[9] N. Li, L. Feng, J. Hu, L. Jiang, J. Wang, J. Han, L. Gan, Z. He and G. Wang, "Using Deeply Time-Series Semantics to Assess Depressive Symptoms based on Clinical Interview Speech", *Frontiers in Psychiatry*, Vol. 14, pp. 1-16, 2023.

[10] S. Munoz and C.A. Iglesias, "Detection of the Severity Level of Depression Signs in Text Combining a Feature-Based Framework with Distributional Representations", *Applied Sciences*, Vol. 13, No. 11695, pp. 1-9, 2023.

[11] Z. Yang, C. Chen, H.M. Li and X. Zhao, "Unsupervised Classifications of Depression Levels Based on Machine Learning Algorithms Perform Well as Compared to Traditional Norm-Based Classifications", *Frontiers in Psychiatry*, Vol. 11, No. 45, pp. 23-38, 2020.

[12] N.V. Babu and E.G.M. Kanaga, "Sentiment Analysis in Social Media Data for Depression Detection using Artificial Intelligence: A Review", *SN Computer Science*, Vol. 3, No. 74, pp. 1-8, 2022.

[13] H. Faris, M. Faris, M. Habib and A. Alomari, "Automatic Symptoms Identification from a Massive Volume of Unstructured Medical Consultations using Deep Neural and BERT Models", *Heliyon*, Vol. 8, pp. 1-24, 2022.

[14] D. Imans, T. Abuhmed, M. Alharbi and S. El-Sappagh, "Explainable Multi-Layer Dynamic Ensemble Framework

Optimized for Depression Detection and Severity Assessment", *Diagnostics*, Vol. 14, No. 2385, pp. 1-13, 2024.

[15] K. Sampath and D. Thenmozhi, "Finding of the Shared Task on Detecting Signs of Depression from Social Media", *Proceedings of International Conference on Language Technology for Equality, Diversity, and Inclusion*, pp. 1-5, 2022.

[16] C. Lau, Z. Zhu and W.Y. Chan, "Automatic Depression Severity Assessment with Deep Learning using Parameter-Efficient Tuning", *Frontiers in Psychiatry*, Vol. 14, pp. 132-149, 2023.

[17] A. Perez, N. Warikoo, K. Wang, J. Parapar and I. Gurevych, "Semantic Similarity Models for Depression Severity Estimation", *Proceedings of International Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, pp. 1-6, 2023.

[18] M. Bader, M. Abdelwanis, M. Maalouf and H.F. Jelinek, "Detecting Depression Severity using Weighted Random Forest and Oxidative Stress Biomarkers", *Scientific Reports*, Vol. 14, No. 2, pp. 16328-16335, 2024.

[19] K. Miniltsevich, K. Sirts and G. Dias, "Towards Automatic Text-Based Estimation of Depression through Symptom Prediction", *Brain Informatics*, Vol. 10, No. 4, pp. 1-12, 2023.

[20] A. Perez, J. Parapar and A. Barreiro, "Automatic Depression Score Estimation with Word Embeddings Models", *Artificial Intelligence in Medicine*, Vol. 132, No. 1, pp. 1-13, 2022.

[21] K. Mao, W. Zhang, B.D. Wang, R. Li. Y. Jiao, Y. Zhu, T. Zheng, L. Qian, W. Lyu, M. Ye and J. Chen, "Prediction of Depression Severity Based on the Prosodic and Semantic Features with Bi-directional LSTM and Time Distributed CNN", *IEEE Transactions of Affective Computing*, Vol. 45, No. 2, pp. 1-13, 2022.

[22] E. Bao, A. Perez and J. Parapar, "Explainable Depression Symptom Detection in Social Media", *Health Information Science and Systems*, Vol. 12, No. 47, pp. 1-14, 2023.

[23] A. Agrawal, "Illuminate a Novel Approach for Depression Detection with Explainable Analysis and Proactive Therapy using Prompt Engineering", *International Journal of Psychiatry*, Vol. 9, No. 2, pp. 118-132, 2024.

[24] S.M. Shah, S.A. Gillani, M.S.A. Baig, M.A. Saleem and M.H. Siddiqui, "Advancing Depression Detection on Social Media Platforms Through Fine-Tuned Large Language Models", *Proceedings of International Conference on Computing, Data Science*, pp. 1-4, 2024.

[25] X. Lan, Y. Cheng, L. Sheng, C. Gao and Y. Li, "Depression Detection on Social Media with Large Language Models", *Proceedings of International Conference on Artificial Intelligence and Natural Language Processing*, pp. 896-904, 2024.

[26] M. Betianu, A. Malan, M. Aldinucci, R. Brike and L. Chen, "DALLMi: Domain Adaption for LLM-based Multi-label Classifier", *Advances in Knowledge Discovery and Data Mining*, Vol. 14647, pp. 1-23, 2024.

[27] F. Sufi, "Addressing Data Scarcity in the Medical Domain: A GPT-Based Approach for Synthetic Data Generation and Feature Extraction", *Information*, Vol. 15 No. 264, pp. 23-36, 2024.

[28] B. Thapa, I. Torres, S.F. Koya, G. Robbins, S.M. Abdalla, O.A. Arah, W.B. Weeks, L. Zang, S. Asma, J.V. Morales, S. Galea, K. Rhee and H.J. Larson, "Use of Data to Understand the Social Determinants of Depression in Two Middle-Income Countries: the 3-D Commission", *Journal of Urban Health*, Vol. 98, pp. 41-50, 2021.

[29] J. Bogatinovski, L. Todorovski, S. Dzeroski and D. Kocev, "Comprehensive Comparative Study of Multi-Label Classification Methods", *Expert Systems with Applications*, Vol. 87, No. 1, pp. 1-18, 2022.

[30] K.M.A. Rai, R. Goyal and M. Choudhary, "Depression Detection Model using Word and Sentence Embedding with Different Classifiers", *International Journal of Engineering Applied Sciences and Technology*, Vol. 23, No. 2, pp. 16-26, 2023.

[31] H. Naveed, A.U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes and A. Mian, "A Comprehensive Overview of Large Language Models", *Expert Systems with Applications*, Vol. 89, No. 1, pp. 1-14, 2022.

[32] M.A.K. Raiaan, M.S.H. Mukta, K. Fatema, N.M Fahad, S. Sakib, M.M.J. Mim, J. Ahmad, M.E. Ali and S. Azam, "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges", *IEEE Access*, Vol. 12, pp. 26839-26874, 2024.

[33] C. Jeong, "Fine-tuning and Utilization Methods of Domain-specific LLMs", *Journal of Intelligence and Information Systems*, Vol. 18, No. 2, pp. 1-12, 2024.

[34] F. Wei, R. Keeling, N. Huber-Fliflet, J. Zhang, A. Dabrowski and J. Yang, "Empirical Study of LLM Fine-Tuning for Text Classification in Legal Document Review", *Proceedings of International Conference on Big Data*, pp. 2786-2792, 2023.

[35] H. Kang and J. Jung, "Similarity Evaluation and Fine-Tuning of Embedding Models from Various Linguistic Perspectives", *KSII Transactions on Internet and Information Systems*, Vol. 19, No. 9, pp. 1-16, 2025.

[36] Geronimo, "SemScore: Evaluating LLMs with Semantic Similarity", Available at https://huggingface.co/blog/g-ronimo/semscore, Accessed in 2024.

[37] R. Koopman and S. Wang, "Mutual Information based Labelling and Comparing Clusters", *Scientometrics*, Vol. 111, No. 2, pp. 78-86, 2017.

[38] A. Mahmoudi and D. Jemielniak, "Proof of Biased Behaviour of Normalized Mutual Information", *Scientific Reports*, Vol. 14, No. 9021, pp. 1-12, 2024.