

Contrastive Learning and Large Language Models for Depression Detection from Social Media

Dr. Arpana Sinhal
Department of Computer Applications
Manipal University
Jaipur, India
arpana.sinhal@jaipur.manipal.edu

Anay Sinhal
Computer & Information Sci. & Engg.
University of Florida
Gainesville, USA
sinhal.anay@ufl.edu

Dr. Amit Sinhal
Computer Science and Engineering
JK Lakshmipat University
Jaipur, India
amit.sinhal@jklju.edu.in

Abstract— Depression is a global mental health concern, often expressed subtly through social media platforms like Reddit and Twitter. This study explores the application of contrastive learning with large language models (LLMs) for depression detection, utilizing annotated data from the DepressionEmo dataset. Advanced computational techniques, including natural language processing (NLP) and machine learning models like SVM, LightGBM, XGBoost, and deep learning models such as GAN-BERT, BERT, and BART, are employed. By leveraging contrastive learning and data augmentation, the models demonstrate improved precision, recall, and F1-scores. This research addresses gaps in data diversity and contextual understanding, offering scalable, ethical solutions for real-time mental health monitoring and intervention.

Keywords—Contrastive Learning, Large Language Models (LLMs), Natural Language Processing (NLP), SVM, LightGBM, XGBoost, GAN-BERT, BERT, BART, Social Media Analysis, Multimodal Emotion Analysis, Mental Health Monitoring, Sentiment Analysis.

I. INTRODUCTION

Depression is a significant mental health disorder affecting millions globally. Traditional diagnostic methods, such as clinical interviews and self-assessments, are subjective and not scalable for larger populations. With the advent of social media platforms like Reddit and Twitter, a wealth of data reflecting mental states has become available, enabling automated approaches for depression detection [1] [3] [6].

This study investigates the use of contrastive learning with LLMs to identify depressive emotions from social media text. The DepressionEmo dataset, derived from Reddit posts, serves as the primary data source, annotated with multiple depression-related emotions using zero-shot classification methods validated by human and AI annotations. This ensures a robust and diverse dataset for model training [2] [4] [5].

The research employs advanced NLP and ML techniques, integrating models such as BERT, GAN-BERT, and BART alongside traditional ML algorithms like SVM and LightGBM [10]. Contrastive learning, a key focus, enhances model performance by distinguishing subtle linguistic nuances in depressive text.

This work addresses critical challenges in depression detection, including data diversity, context-dependent indicators, and generalizability across real-world scenarios. Additionally, it emphasizes ethical considerations, ensuring data privacy and unbiased model training. The findings highlight the potential of automated, scalable systems for early intervention and mental health monitoring.

II. LITERATURE REVIEW

The review categorizes approaches, highlights critical features, and evaluates their effectiveness compared to existing models.

A. Types of Depression Detection Techniques

1) *Text-Based Analysis*: NLP techniques analyze syntactic and semantic features of text, enabling sentiment analysis and emotion classification. For instance, Deshpande and Rao (2017) demonstrated emotion AI for depression detection, while Kumar and Latha (2019) applied sentiment analysis to Twitter data [11] [14] [22].

2) *Image and Video Analysis*: Visual cues such as facial expressions are analyzed using image processing techniques. Rani and Durgadevi (2017) employed facial emotion recognition, and Victor et al. (2019) combined multimodal data (video, audio) with deep neural networks for comprehensive emotion detection [15] [20] [23].

3) *Hybrid Approaches*: These approaches integrate text, audio, and visual data with advanced machine learning (ML) algorithms. For instance, Khan and Alqahtani (2024) proposed hybrid models that significantly improve detection accuracy by leveraging multimodal inputs [3].

B. Features and Parameters

1) *Linguistic Features*: Word frequency, sentiment scores, and syntactic patterns are critical for identifying depressive tendencies (Aldarwish and Ahmad, 2017). Advanced NLP techniques enhance feature extraction from text data [24].

2) *Visual and Acoustic Features*: Facial expressions and voice characteristics such as tone and pitch provide non-textual indicators of depression [18] [23].

3) *Behavioral Features*: User activity patterns, including posting frequency and engagement on social media, have been explored to detect depressive behaviors (AlSagri and Ykhlef, 2020) [12].

C. Comparison with Existing Models

1) *Traditional Models*: Early approaches like Naive Bayes and Support Vector Machines (SVM) primarily focused on sentiment analysis but lacked the capacity to capture complex patterns [13].

2) *Deep Learning Models*: Advanced models like CNNs and LSTMs offer superior performance by leveraging contextual understanding of text data (Shetty et al., 2020).

3) *Hybrid Approaches*: Integration of multimodal data and techniques has demonstrated notable improvements in accuracy and robustness (Nusrat et al., 2024).

D. Significant Work in the Field

- Nusrat et al. (2024) utilized explainable AI to classify depression types from Twitter data, achieving 96% accuracy with a fine-tuned BERT model [1].
- Wang et al. (2024) automated Beck's Depression Inventory scoring using LLMs, demonstrating superior accuracy and transparency [2].
- Khan et al. (2024) introduced hybrid models combining sentiment analysis with multimodal inputs, achieving 99.4% accuracy [3].
- Tejaswini et al. (2024) leveraged Fasttext embeddings with CNN and LSTM, enhancing model accuracy for Reddit and Twitter datasets [4].
- Iqbal Khan et al. (2024) demonstrated the potential of contrastive learning with self-augmentation to improve F1-scores in disease detection tasks [5].

The reviewed literature underscores significant advancements in automated depression detection through AI and ML. Modern approaches, particularly those integrating deep learning and hybrid methods, consistently outperform traditional models in accuracy and robustness. Additionally, explainable AI holds promise for enhancing model transparency and clinical trust. Future research should focus on addressing challenges like data diversity, subtle linguistic indicators, and ethical considerations to develop scalable, reliable solutions for mental health monitoring.

III. RESEARCH METHODOLOGY

A. Research Design

The design encompasses preliminary analysis, data augmentation, and model development, ensuring a robust and scalable framework for real-time mental health monitoring.

B. Experimental Setup

- *Python*: The primary language for data science and machine learning.
- *TensorFlow and PyTorch*: TensorFlow ensured scalability, while PyTorch facilitated flexible model experimentation.
- *Transformers Library*: Hugging Face's library implemented pre-trained models like BERT and GPT.
- *Jupyter Notebooks*: Enabled interactive development and result visualization.

C. Proposed Methodology

1) *Data Augmentation*: Enhancing dataset diversity to prevent overfitting and improve generalizability:

2) *Model Development and Training*: Contrastive Learning to trains models to distinguish between similar and dissimilar data points using positive and negative text pairs. Embedding Generation with BERT generated contextual embeddings, capturing nuanced linguistic features [5].

D. Data Collection

1) *Data Sources*: Posts were collected from mental health-focused subreddits, including r/depression, r/DepressedPartners, and r/suicide_watch [1].

2) *Selection Criteria*: Posts containing keywords related to depression and a diverse representation of emotions and linguistic variations.

3) *Preprocessing*: Removing irrelevant content and standardizing text with lowercasing and eliminating special characters for uniformity.

The *DepressionEmo* dataset comprises 6,037 annotated Reddit posts with eight depression-related emotions. The annotation method is Zero-shot classification validated by human and AI annotations [25]. The labels are emotions including anger, sadness, worthlessness, hopelessness, and others.

TABLE I. ATTRIBUTES OF THE DATASET

Attribute	Description
Post ID	Unique identifier for each Reddit post
Text	Content of the Reddit post
Anger	Binary label indicating the presence of anger, characterized by expressions of frustration or hostility.
Cognitive Dysfunction	Binary label identifying indications of impaired cognitive functions, such as confusion or difficulty concentrating.
Emptiness	Binary label marking expressions of a pervasive sense of void or lack of fulfillment in the post.
Hopelessness	Binary label capturing sentiments of despair or belief in a negative outlook for the future.
Loneliness	Binary label denoting feelings of social isolation or lack of companionship expressed in the post.
Sadness	Binary label identifying depressive moods characterized by sorrow or melancholy in the post's content.
Suicide Intent	Binary label indicating the presence of expressions related to self-harm or suicidal ideation.
Worthlessness	Binary label marking sentiments of self-deprecation or feelings of being unimportant or undeserving.

IV. IMPLEMENTATION

A. Support Vector Machines (SVM)

SVM models were trained using a linear kernel to classify text data into multiple emotion categories. Hyperparameter tuning focused on optimizing the regularization parameter to balance bias and variance. The model was evaluated using precision, recall, and F1-score metrics to ensure robustness across all emotion categories [13].

B. LightGBM

LightGBM, a gradient boosting framework, was employed to process TF-IDF features for multilabel classification. Its efficient handling of large datasets, enabled by Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), significantly reduced training time. Hyperparameters, including the number of leaves and learning rate, were fine-tuned to maximize performance [8].

C. XGBoost

XGBoost was utilized for its regularization capabilities and ability to handle complex classification tasks. The model was fine-tuned using parameters such as tree depth, learning rate, and boosting iterations. Its performance was evaluated using standard metrics, demonstrating its effectiveness in detecting depressive emotions from textual data [14].

D. GAN-BERT

GAN-BERT integrates BERT's language representation capabilities with a generative adversarial network (GAN). The model alternates between training a generator to create realistic text samples and a discriminator to distinguish between real and synthetic data. This approach enhances the model's robustness by augmenting the dataset with diverse examples. The model was fine-tuned using the AdamW optimizer, and its performance was assessed using precision, recall, and F1-score [6].

E. BERT

BERT's bidirectional transformer architecture enables it to capture the contextual relationships between words in a sentence. Fine-tuning involved adding a classification layer on top of the pre-trained model and optimizing the network using the AdamW optimizer with a learning rate of 2×10^{-5} . The model achieved high accuracy in detecting subtle emotional patterns indicative of depression [2].

F. BART

BART's encoder-decoder structure combines bidirectional and autoregressive transformers, making it effective for text classification and data augmentation. The model was fine-tuned using labeled data, with a focus on optimizing its ability to detect depressive emotions. Evaluation metrics demonstrated BART's capability to classify text data accurately while capturing complex linguistic features [4].

G. Contrastive Learning with LLMs

Contrastive learning enhances model performance by training it to distinguish between similar and dissimilar data points. Applied to BERT, this technique improves the model's ability to understand subtle linguistic cues. The implementation involved creating augmented text pairs through synonym replacement, word deletion, and back-translation. The contrastive loss function optimized the embedding space to maximize similarity between positive pairs and minimize it for negative pairs.

$$\mathcal{L}_{contrastive} = -\log \frac{\exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_k)/\tau)}$$

Where:

- $\mathbf{h}_i$ and $\mathbf{h}_j$ are the embeddings of the augmented pairs.
- $\text{sim}(\mathbf{h}_i, \mathbf{h}_j)$ measures the similarity between the embeddings (e.g., cosine similarity).
- τ is a temperature parameter that controls the sharpness of the distribution.

The training process included pre-training BERT using contrastive learning on a large corpus and subsequently fine-

tuning it on the DepressionEmo dataset. This approach significantly enhanced the model's representation learning, leading to improved classification accuracy.

All models were evaluated on the test set using precision, recall, and F1-score metrics. Cross-validation ensured generalizability and robustness. Results indicated that combining traditional ML models with advanced DL techniques and contrastive learning provides a comprehensive approach to detecting depressive emotions [19]. Each model contributed uniquely to the overall system, with DL models excelling in capturing contextual semantics and ML models offering robust classification capabilities.

V. RESULTS AND DISCUSSION

The performance of each model is evaluated using standard classification metrics, and the comparative analysis highlights their strengths and limitations. The implications of these results and areas for improvement are also discussed.

A. Performance Metrics

1) *Precision*: It measures the proportion of correctly predicted positive instances among all predicted positive instances.

2) *Recall*: It quantifies the proportion of actual positive instances that were correctly identified.

3) *F1 Score*: It balances precision and recall, particularly useful for imbalanced datasets.

B. Preliminary Examination

1) *Distribution of Emotions*: A frequency analysis of the dataset revealed that sadness was the most prevalent emotion, followed by hopelessness and worthlessness. Such distributions provide insights into the emotional characteristics of the data and guide model evaluation priorities.

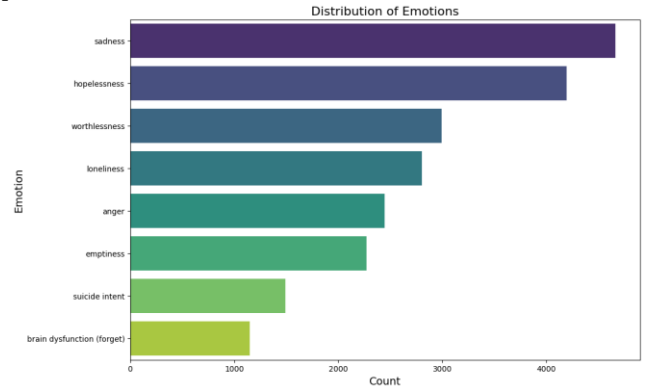


Fig. 1. Distribution of Emotions

2) *Emotion Co-occurrence Matrix*: The co-occurrence matrix demonstrated significant overlap between emotions like sadness and emptiness, reflecting the interrelated nature of depressive states. Identifying such patterns enables better understanding and modeling of depression-related emotions.

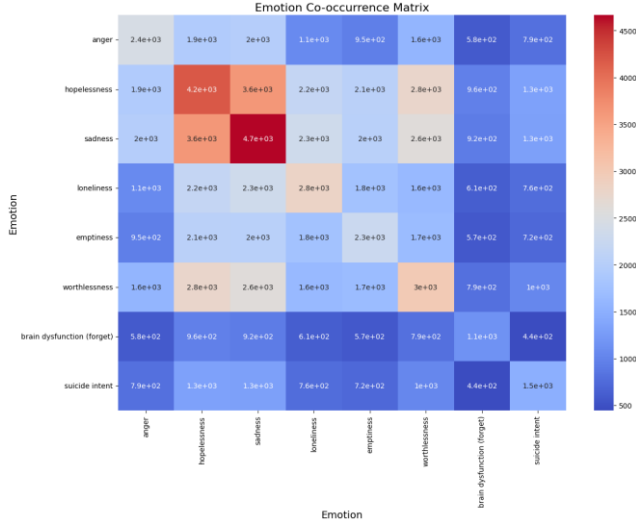


Fig. 2. Emotion Co-occurrence Matrix

3) *Sentiment Analysis*: Sentiment distribution across posts showed that a majority exhibited negative sentiments. These insights help contextualize the dataset's depressive nature.

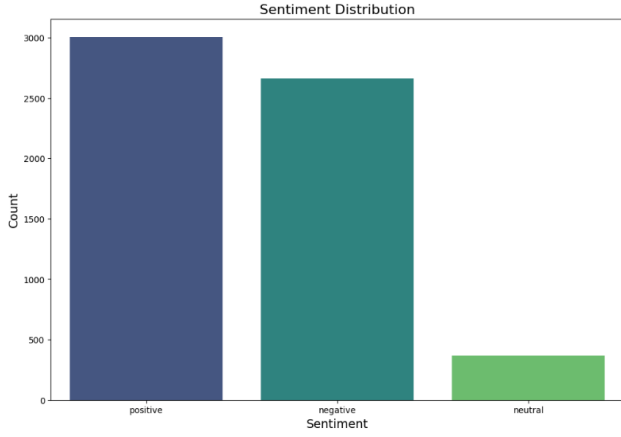


Fig. 3. Sentiment Distribution

C. Results of ML Models

TABLE II. PERFORMANCE OF ML MODELS

Model	F1 Score	Precision	Recall
SVM	0.47	0.72	0.41
LightGBM	0.58	0.48	0.80
XGBoost	0.59	0.63	0.56

1) *SVM*: Demonstrated high precision (0.72) but low recall (0.41), indicating effectiveness in identifying positive instances but a failure to capture many actual positives [9].

2) *LightGBM*: Achieved high recall (0.80), making it suitable for applications prioritizing sensitivity; however, precision (0.48) was suboptimal, reflecting the presence of false positives [7].

3) *XGBoost*: Balanced performance with an F1 score of 0.59, offering reliable predictions without overfitting.

D. Results of DL Models

TABLE III. PERFORMANCE OF DL MODELS

Model	F1 Score	Precision	Recall
GAN-BERT	0.70	0.69	0.72
BERT	0.74	0.72	0.77
BART	0.76	0.70	0.81

1) *GAN-BERT*: The generative capabilities of GANs combined with BERT yielded a balanced performance with strong robustness.

2) *BERT*: Demonstrated superior context understanding, achieving an F1 score of 0.74. Its bidirectional transformer architecture contributed to high recall (0.77).

3) *BART*: Outperformed other DL models with an F1 score of 0.76. Its encoder-decoder architecture effectively captured nuanced patterns, achieving the highest recall (0.81).

E. Results of Contrastive Learning with BERT

Contrastive learning (CL) with BERT yielded the best performance, as shown in Table IV.

TABLE IV. PERFORMANCE OF CL WITH BERT

Metric	Value
F1 Score	0.81
Precision	0.76
Recall	0.85

CL enhanced BERT's representation learning, significantly improving its ability to capture subtle nuances in text data. The high recall (0.85) indicates its effectiveness in identifying actual positive instances, while its precision (0.76) underscores its reliability in making accurate predictions.

TABLE V. EMOTION-WISE PERFORMANCE OF CL WITH BERT

Emotion	F1 Score	Precision	Recall
Anger	0.80	0.78	0.82
Cognitive Dysfunction	0.79	0.80	0.78
Emptiness	0.94	0.93	0.95
Hopelessness	0.84	0.81	0.86
Loneliness	0.92	0.92	0.92
Sadness	0.76	0.78	0.75
Suicide Intent	0.52	0.51	0.53
Worthlessness	0.70	0.74	0.66

“Emptiness and Loneliness” achieved the highest F1 scores, reflecting the model’s exceptional ability to identify these emotions. “Suicide Intent” achieved the lowest performance, highlighting challenges in detecting this complex and sensitive emotion.

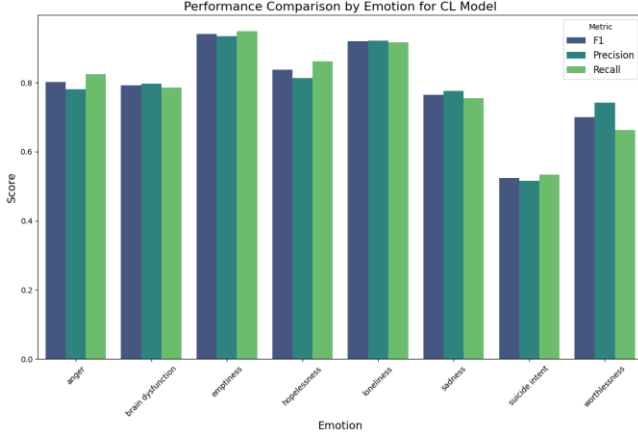


Fig. 4. Performance Comparison by Emotions for CL Model

F. Comparison of All Models

The CL model consistently outperformed traditional ML and DL models across all metrics, demonstrating its potential in detecting depressive emotions with high accuracy and reliability.

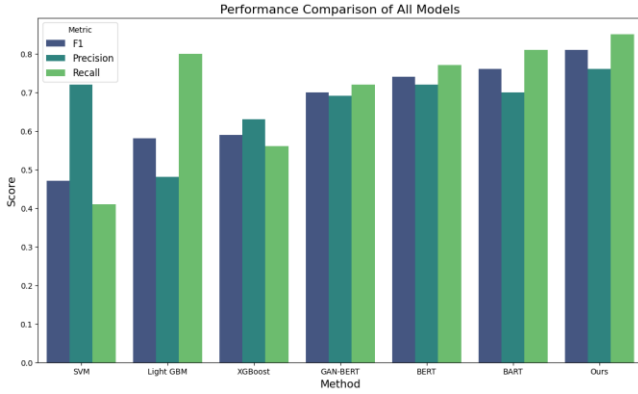


Fig. 5. Performance Comparison of all Models

Radar Chart for Model Performance Comparison

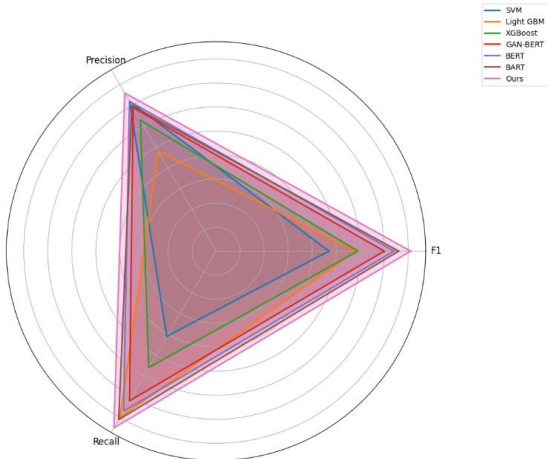


Fig. 6. Performance Comparison of All Models (Radar)

G. Discussion

1) **Strengths:** The superior performance highlights the efficacy of contrastive learning in enhancing contextual representation. DL models demonstrated significant

improvement over ML models due to advanced architectures capable of capturing complex patterns.

2) **Limitations:** Lower performance for certain emotions (e.g., suicide intent) reflects the need for more balanced datasets. Deep learning models require extensive labeled data, which is resource-intensive to obtain [16].

VI. CONCLUSION

This research has comprehensively explored the application of ML and DL techniques to detect depression-related emotions from social media text data. By leveraging the DepressionEmo dataset, advanced augmentation techniques, and state-of-the-art models, the study highlights the potential of automated systems for mental health monitoring.

The DepressionEmo dataset, featuring multilabel classifications of depression-related emotions, served as a robust foundation for model development. Augmentation techniques, including synonym replacement, back translation, noise injection, and random deletion, enhanced data diversity and contributed to improved model generalization. Traditional ML models, such as SVM, LightGBM, and XGBoost, provided baseline performances, with XGBoost demonstrating a balanced trade-off between precision and recall. LightGBM achieved high recall, while SVM was more precise but lacked robustness in recall.

Advanced DL models, particularly BART and BERT, outperformed their ML counterparts. BART achieved the highest F1 score among DL models, reflecting its effectiveness in understanding nuanced depressive emotions. The integration of contrastive learning with BERT further advanced model performance, achieving the highest metrics across precision, recall, and F1 score [2]. This approach demonstrated the efficacy of enhanced representation learning by emphasizing the differentiation between similar and dissimilar data points.

The study also revealed challenges in detecting complex emotions such as "suicide intent," highlighting areas for model improvement. The insights from this research underscore the potential of advanced computational techniques in developing scalable, reliable, and accurate systems for early depression detection and intervention [17].

VII. FUTURE SCOPE

While this research has achieved significant advancements, further development is essential to address current limitations and expand the applicability of depression detection systems. The following areas offer promising directions for future work:

A. Advanced Data Augmentation and Multimodal Integration

Future research should explore generative models, such as GPT-3 or advanced GANs, to synthesize high-quality text data. Integrating text with other modalities, such as images, audio, and video, can provide a more comprehensive representation of depressive states, enhancing detection accuracy [4] [18].

B. Emotion-Specific Model Optimization

Developing emotion-specific architectures tailored to the unique characteristics of individual emotions, such as "suicide intent" or "worthlessness," can improve detection precision. Incorporating domain-specific linguistic, psychological, and contextual features will further refine emotion classification.

C. Transfer Learning and Domain Adaptation

Leveraging pre-trained models fine-tuned on large mental health datasets or similar domains can enhance feature extraction and improve generalizability. Domain adaptation techniques can enable models to be effective across diverse social media platforms and user demographics [21].

D. Model Explainability and User Trust

To build trust in automated systems, efforts should focus on improving model interpretability. Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can provide insights into model decisions, facilitating user and clinician confidence [1] [9].

E. Temporal and Behavioral Analysis

Longitudinal studies examining the temporal progression of depressive emotions can provide deeper insights into user behavior and the evolution of depression. These studies can enable the prediction of critical mental health events and inform preventative measures.

REFERENCES

- [1] Nusrat, M. O., Shahzad, W., & Jamal, S. A. (2024). Multi Class Depression Detection Through Tweets using Artificial Intelligence. arXiv preprint arXiv:2404.13104.
- [2] Wang, Y., Inkpen, D., & Gamaarachchige, P. K. (2024, March). Explainable depression detection using large language models on social media data. In Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024) (pp. 108-126).
- [3] Khan, S., & Alqahtani, S. (2024). Hybrid machine learning models to detect signs of depression. Multimedia Tools and Applications, 83(13), 38819-38837.
- [4] Tejaswini, V., Sathya Babu, K., & Sahoo, B. (2024). Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(1), 1-20.
- [5] Khan, P. I., Dengel, A., & Ahmed, S. (2024). Improving Disease Detection from Social Media Text via Self-Augmentation and Contrastive Learning. arXiv preprint arXiv:2405.01597.
- [6] Khafaga, D. S., Auvdaiappan, M., Deepa, K., Abouhawwash, M., & Karim, F. K. (2023). Deep learning for depression detection using Twitter data. Intelligent Automation & Soft Computing, 36(2), 1301-1313.
- [7] Vasha, Z. N., Sharma, B., Esha, I. J., Al Nahian, J., & Polin, J. A. (2023). Depression detection in social media comments data using machine learning algorithms. Bulletin of Electrical Engineering and Informatics, 12(2), 987-996.
- [8] Saraf, M., & Jadhao, Y. (2023). Survey On Design Efficient Depression Detection Using Machine Learning. International Journal of Ongoing Research in Science and Engineering.
- [9] Joshi, M. L., & Kanoongo, N. (2022). Depression detection using emotional artificial intelligence and machine learning: A closer review. Materials Today: Proceedings, 58, 217-226.
- [10] Nikam, S. S., Patil, A. V., Patil, G. S., Patil, S. P., & Jitkar, B. D. (2020). AI Therapist Using Natural Language Processing. International Journal of Research in Engineering, Science and Management, 3(2), 715.
- [11] Rajaraman, P. V., Nath, A., Akshaya, P. R., & Bhuja, G. C. (2020). Depression detection of tweets and A comparative test. International Journal of Engineering Research, 9(03), 422-425.
- [12] AlSagri, H. S., & Ykhlef, M. (2020). Machine learning-based approach for depression detection in twitter using content and activity features. IEICE Transactions on Information and Systems, 103(8), 1825-1832.
- [13] Shetty, N. P., Muniyal, B., Anand, A., Kumar, S., & Prabhu, S. (2020). Predicting depression using deep learning and ensemble algorithms on raw twitter data. International Journal of Electrical and Computer Engineering, 10(4), 3751.
- [14] Hemanthkumar, M., & Latha, A. (2019). Depression detection with sentiment analysis of tweets. International Research Journal of Engineering and Technology (IRJET), 5.
- [15] Victor, E., Aghajan, Z. M., Sewart, A. R., & Christian, R. (2019). Detecting depression using a framework combining deep multimodal neural networks with a purpose-built automated evaluation. Psychological assessment, 31(8), 1019.
- [16] Ramalingam, D., Sharma, V., & Zar, P. (2019). Study of depression analysis using machine learning techniques. Int. J. Innov. Technol. Explor. Eng. 8(7C2), 187-191.
- [17] Fonseka, T. M., Bhat, V., & Kennedy, S. H. (2019). The utility of artificial intelligence in suicide risk prediction and the management of suicidal behaviors. Australian & New Zealand Journal of Psychiatry, 53(10), 954-964.
- [18] Biradar, A., & Totad, S. G. (2018). Detecting depression in social media posts using machine learning. In International conference on recent trends in image processing and pattern recognition (pp. 716-725). Singapore: Springer Singapore.
- [19] Islam, M. R., Kamal, A. R. M., Sultana, N., Islam, R., & Moni, M. A. (2018). Detecting depression using k-nearest neighbors (knn) classification technique. In 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2) (pp. 1-4). IEEE.
- [20] Thosar, D., Gothe, V., Bhorkade, P., & Sanap, V. (2018). Review on mood detection using image processing and chatbot using artificial intelligence. life, 5(03).
- [21] Oak, S. (2017). Depression detection and analysis. In 2017 AAAI spring symposium series.
- [22] Deshpande, M., & Rao, V. (2017, December). Depression detection using emotion artificial intelligence. In 2017 international conference on intelligent sustainable systems (iciss) (pp. 858-862). IEEE.
- [23] Rani, A. M., & Durgadevi, R. (2017). Image processing techniques to recognize facial emotions. International Journal of Engineering and Advanced Technology (IJEAT), 6(6).
- [24] Aldarwish, M. M., & Ahmad, H. F. (2017, March). Predicting depression levels using social media posts. In 2017 IEEE 13th international Symposium on Autonomous decentralized system (ISADS) (pp. 277-280). IEEE.
- [25] Nadeem, M. (2016). Identifying depression on Twitter. arXiv preprint arXiv:1607.07384.