Explainable Detection of Depression in Social Media Contents Using Natural Language Processing

[1] KODIPYAKA SAIRAM
Computer Science and Engineering
Guru Nanak Institutions Technical Campus Hyderabad, India.
kodipyakasairam4@gmail.com

[2] KONDA SHREEYA
Computer Science and Engineering
Guru Nanak Institutions Technical Campus, Hyderabad,India.
shreeyakonda@gmail.com

[3] GANAPURAM SWAPNA
Computer Science and Engineering
Guru Nanak Institutions Technical Campus, Hyderabad,India.
Swapna.gangapuram@gmail.com

***Abstract*— This paper an explainable deep learning approach using Long Short-Term Memory (LSTM) networks to detect depression from social media posts. The model classifies text into depression or control categories by capturing linguistic patterns and sequential dependencies. An attention mechanism is integrated to enhance interpretability, highlighting key features influencing predictions. Evaluated on a public mental health dataset, the model shows high accuracy and transparency, offering a scalable solution for early depression detection and supporting timely mental health interventions.**

I. INTRODUCTION

Depression is a common and serious mental health disorder that can significantly affect an individual's quality of life. Early detection is essential for timely intervention, yet traditional machine learning models struggle with limited labeled data and lack interpretability. To address these challenges, this study explores an explainable deep learning approach using Natural Language Processing (NLP) techniques, specifically a hybrid model combining Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks. This model is designed to classify social media posts as either depression-related or not, while also providing transparency through attention mechanisms.

II. EASE OF USE

The proposed system is simple to use and designed for easy operation. It requires only basic input in the form of social media text data and automatically processes it through the model. The system runs on commonly available hardware and uses a simple interface through the Spyder IDE in Python. No advanced technical knowledge is needed to use the model, making it accessible for researchers, mental health professionals, and developers. The interpretability feature allows users to see which factors influenced the model's decisions.

III. EXISTING SYSTEM

Convolutional Neural Networks (CNNs) are a specialized type of deep neural network designed primarily for processing structured grid-like data, such as images, audio, and time-series data. CNNs have proven to be highly effective in tasks like image recognition, object detection, and natural language processing (NLP) due to their ability to automatically learn spatial hierarchies of features from data.

A CNN works by applying various convolutional layers that use filters (also known as kernels) to perform convolutions on input data. This allows the network to detect local patterns, such as edges, textures, and shapes, which are then combined to form higher-level features as the data progresses through deeper layers.

DISADVANTAGES

CNNs require significant computational resources for training, especially with large datasets.

CNNs typically perform well only when trained on large amounts of labled data.

Although CNNs are powerful and effective, they can be considered black box models.

IV. TECHNIQUES/ALGORITHMS USED

Long Short-Term Memory (LSTM):

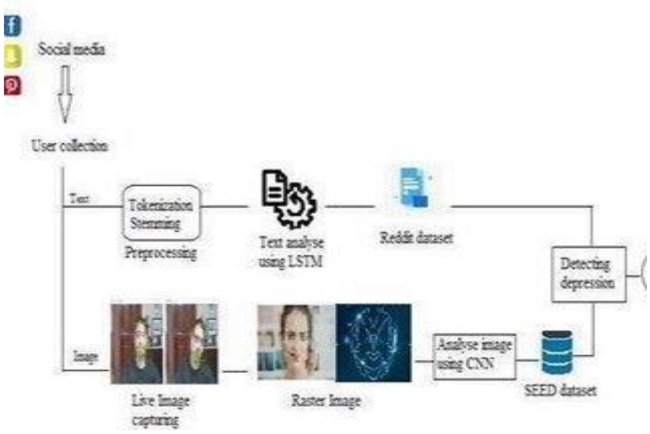
LSTM is a type of Recurrent Neural Network (RNN) designed to handle long-term dependencies in sequential data. It uses memory cells and gates (input, forget, and output) to retain important information over time, making it suitable for analyzing text data like social media posts..

Gated Recurrent Unit (GRU):

GRU simplifies the LSTM structure by merging the input and forget gates into one update gate, making it more efficient. It is faster and more efficient while still effectively learning from sequential data..

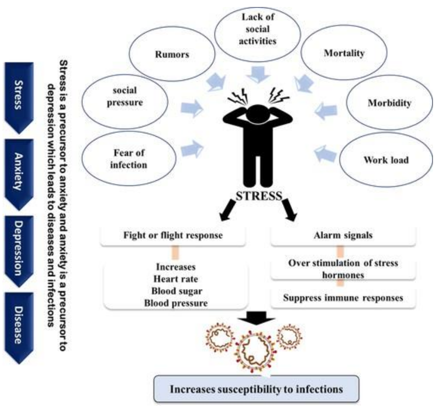
Attention Mechanism:

This technique helps the model focus on the most relevant parts of the input text when making predictions. It improves interpretability by highlighting which words or phrases contributed most to the decision.



B. ADVANTAGES

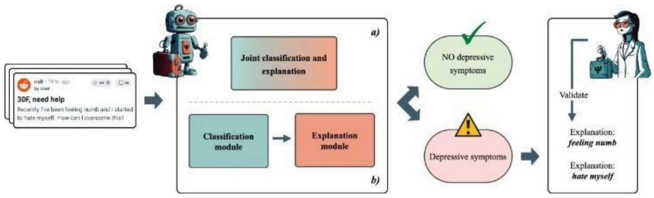
- Improved Accuracy.
- Efficient and Scalable
- Handles Long-Term Dependencies
- Better Integration of Attention Mechanisms



C. POSITIONING MODEL OF FOREIGN OBJECTS

Numerous studies employ machine learning algorithms for data processing, and similarly, this study utilizes machine learning regression algorithms for handling data with continuous distributions. Initially, the constructed locating dataset is divided into a training set and a test set in an 8:2 ratio. Subsequently, the coefficient of determination R^2 serves as the performance evaluation metric during the regression model training stage, indicating the extent of agreement between the predicted and actual values. The calculation procedure for R^2 is illustrated in equation, where R^2 values range between zero and one. A higher R^2 value suggests enhanced interpretability of the corresponding variable by the independent variable as it approaches one.

$$R2=1-\sum mi=0(yi-yi\hat{})2/\sum mi=0(yi-\overline{y}i)2$$



V. DATA PROCESSING

UAVs captured high-resolution images and 4K videos during low-traffic hours for safe and clear data collection. A dataset of 7,625 FOD images (1280×1280) was created under various conditions. To fit the model's 320×320 input, bounding boxes were extracted from key regions to retain objects and minimize background loss.

VI. CONCLUSION

This project presents an effective and explainable approach for detecting depression in social media content using a hybrid LSTM-GRU model. By capturing sequential patterns and integrating attention mechanisms, the system not only achieves high accuracy but also provides transparency in its predictions. The model's ease of use and efficiency make it a practical tool for supporting early mental health intervention through automated text analysis.

VII. REFERENCES

[1] (2021). National Institute of Mental Health. Accessed: Apr. 23, 2024. [Online]. Available: https://www.nimh.nih.gov/health/statistics/mentalillness

[2] (2019). World Health Organization. Accessed: May 5, 2024. [Online]. Available: https://www.who.int/teams/mental-health-and-substanceuse/promotion-prevention/mental-health-in-the-workplace VOLUME 12, 2024 161211 C. Xin, L. Q. Zakaria: Integrating Bert With CNN and BiLSTM for Explainable Detection of Depression

[3] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, ''Detection of depressionrelated posts in Reddit social media forum,'' IEEE Access, vol. 7, pp. 44883– 44893, 2019.

[4] A. Husseini Orabi, P. Buddhitha, M. Husseini Orabi, and D. Inkpen, ''Deep learning for depression detection of Twitter users,'' in Proc. 5th Workshop Comput. Linguistics Clin. Psychol., From Keyboard to Clinic, 2018, pp. 88–97.

[5] S. G. Burdisso, M. Errecalde, and M. Montes-y-Gómez, ''A text classification framework for simple and effective early depression detection over social media streams,'' Exp. Syst. Appl., vol. 133, pp. 182–197, Nov. 2019.

[6] C. Sun, X. Qiu, Y. Xu, and X. Huang, ''How to fine- tune BERT for text classification?''

2019, arXiv:1905.05583.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep
bidirectional transformers for language understanding,'' 2018, arXiv:1810.04805.

[8] M. Choras, M. Pawlicki, D. Puchalski, and R. Kozik, ''Machine learning—The results are
not the only thing that matters! What about security, explainability and fairness?'' in
Computational Science—ICCS, V. V. Krzhizhanovskaya, G. Závodszky, M. H. Lees, J. J.

Dongarra, P. M. A. Sloot, S. Brissos, and J. Teixeira, Eds., Cham, Switzerland: Springer, 2020,
pp. 615–628.

[9] A. Das and P. Rad, ''Opportunities and challenges in explainable artificial intelligence
(XAI): A survey,'' 2020, arXiv:2006.11371.

[10] F. Cacheda, D. Fernandez, F. J. Novoa, and V. Carneiro, ''Early detection of depression:
Social network analysis and random forest techniques,'' J. Med. Internet Res., vol. 21, no. 6, Jun.
2019, Art. no. e12554.

[11] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia,
S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, ''Explainable artificial
intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,''
Inf. Fusion, vol. 58, pp. 82–115, Jun. 2020.

[12] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, ''MentalBERT: Publicly
available pretrained language models for mental healthcare,'' 2021, arXiv:2110.15621.

[13] H. Chefer, S. Gur, and L. Wolf, ''Transformer interpretability beyond attention
visualization,'' in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Nashville,
TN, USA, Jun. 2021, pp. 782–791.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I.
Polosukhin, ''Attention is all you need,'' 2017, arXiv:1706.03762.

[15] M. R. Islam, A. R. M. Kamal, N. Sultana, R. Islam, M. A. Moni, and A. Ulhaq, ''Detecting
depression using K-nearest neighbors (KNN) classification technique,'' in Proc. Int. Conf.
Comput., Commun., Chem., Mater. Electron. Eng. (ICME), Feb. 2018, pp. 1–4.

[16] K. A. Govindasamy and N. Palanichamy, ''Depression detection using machine learning
techniques on Twitter data,'' in Proc. 5th Int. Conf. Intell. Comput. Control Syst. (ICICCS),
Madurai, India, May 2021, pp. 960–966.

[17] Z. Peng, Q. Hu, and J. Dang, ''Multi-kernel SVM based depression recognition using social
media data,'' Int. J. Mach. Learn. Cybern., vol. 10, no. 1,
pp. 43–57, Jan. 2019.

[18] H. Dinkel, M. Wu, and K. Yu, ''Text-based depression detection on sparse data,'' 2019,
arXiv:1904.05154.

[19] A. Amanat, M. Rizwan, A. R. Javed, M. Abdelhaq, R. Alsaqour, S. Pandya, and M. Uddin,
''Deep learning for depression detection from textual data,'' Electronics, vol. 11, no. 5, p. 676,
Feb. 2022.

[20] S. M. Lundberg and S.-I. Lee, ''A unified approach to interpreting model predictions,''
2017, arXiv:1705.07874.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, '''Why should I trust you?' Explaining the
predictions of any classifier,'' in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data
Mining, San Francisco, CA, USA, 2016, pp. 1135–1144.

[22] (2022). Depression: Reddit Dataset (Cleaned).
Accessed: May 3, 2024. [Online]. Available:
https://www.kaggle.com/datasets/infamouscoder/ depression-reddit-cleaned

[23] (2021). Sentimental Analysis for Tweets. Accessed: May 3, 2024. [Online]. Available:
https://www.kaggle.com/datasets/gargmanas/sentimentalan alysis-for-tweets
[24] (2023). Mental
Health Corpus. Accessed: May 3, 2024. [Online]. Available:
https://www.kaggle.com/datasets/reihanenamdari/mentalhe alth-corpus

[25] A. A. Falaki and R. Gras, ''Attention visualizer package: Revealing word importance for
deeper insight into encoder-only transformer models,'' 2023, arXiv:2308.14850.

[26] R. H. Chassab, L. Q. Zakaria, and S. Tiun, ''An adjusted BERT architecture for the
automatic essay scoring task,'' in Proc. 5th Int. Multi-Conf. Artif. Intell. Technol., 2021, pp.
40–41.

[27] R. H. Chassab, L. Q. Zakaria, and S. Tiun, ''An optimized LSTMbased augmented language
model (FLSTM-ALM) using fox algorithm for automatic essay scoring prediction,'' IEEE

Access, vol. 12, pp. 48713–48724, 2024, doi:
10.1109/ACCESS.2024.3381619.

[28] Y. Y. Chang and N. Omar, ''Data annotation architecture for automatic depression
detection,'' Asia–Pacific J. Inf. Technol. Multimedia, vol. 12, no. 1, pp. 39–56, 2023.

[29] S. K. Hamed, M. J. A. Aziz, and M. R. Yaakub, ''Enhanced feature representation for
multimodal fake news detection using localized finetuning of improved BERT and VGG-19
models,'' Arabian J. Sci. Eng., pp. 1–17, Aug. 2024, doi: 10.1007/s13369-024-09354-2.

[30] N. N. W. N. Hashim, N. A. Basri, M. A.-E. A. Ezzi,
and N. M. H. N. Hashim, ''Comparison
of classifiers using robust features for depression detection on bahasa Malaysia speech,'' IAES
Int. J. Artif. Intell. (IJ-AI), vol. 11, no. 1, p. 238, Mar. 2022, doi: 10.11591/ijai.v11.i1.pp238-
253.