Grenze International Journal of Engineering and Technology, June Issue

Real Time AI Mental Health Analysis: A Comprehensive Review

Mehran Arsh Hussain[1], Hemant Kumar[2], Shashank Shekhar[3] and Deepti Sahu[4]
[1-4]Department of Computer Science Engineering, Sharda School of Computing Science & Engineering, Greater Noida, India
Email: mehranarsh786@gmail.com, hemant27134@gmail.com, sshkhar21@gmail.com, deepti.sahu@sharda.ac.in

***Abstract*— People are increasingly using blogs, journals, and social media to share their personal and mental health struggles as the world grows more digitally connected. This trend offers a rare chance to detect psychological distress early with artificial intelligence. Models like BERT have revolutionized the field of Natural Language Processing, whereas earlier text analysis techniques were restricted to basic positive or negative sentiment and frequently overlooked important context. Like a human, this sophisticated model comprehends the meaning of words in context. Going beyond basic sentiment analysis, this paper offers a precise, step-by-step framework for using the BERT model to specifically identify subtle and complex emotional cues in online writing. The ultimate goal is to create a practical tool that can assist clinicians, enable timely intervention, and fundamentally improve the approach to mental well-being in the digital age.**

***Index Terms*— Mental Health, Artificial Intelligence (AI), Natural Language Processing (NLP), BERT Model, Psychological Distress, Early Detection, Sentiment Analysis.**

I. INTRODUCTION

People post pretty much everything online these days. Whether it’s on a blog, a random forum like Reddit, or just all-over social media. Because of this, there is now a huge, always growing pile of public writing. It basically shows how people are feeling in real time.
This whole opportunity. It’s especially important when considering what the World Health Organizations says: depression is a top cause of disability around the world [1]. A major challenge is that many people don't get help in time [2]. Technology could be used to spot the early warning signs of distress in the things people write online, creating a way to offer support much earlier than currently possible.
However, teaching a computer to understand psychological distress is incredibly hard. Human emotions are complex, and the way people write about them is filled with subtlety. For years, the computer programs designed to understand language, a field known as Natural Language Processing (NLP), just weren't up to the task.
Historically, progress in NLP was slow. Experts would have to manually create rules and features to teach a computer what to look for in a piece of text, a process known as feature engineering [3]. This approach was very limited and couldn't understand context [4]. Despite these advances, many existing approaches focuses primarily on binary sentiment. Therefore, there is a need for a multi-label, context-aware solution capable of identifying multiple emotions. This paper proposes a BERT-based architecture for fine-grained emotion detection and evaluates its performance against a classic machine learning baseline using metrics.

Grenze ID: 01.GIJET.12.2.286_20

Recently, the field has been transformed by a new method known as the "pre-train, fine-tune" model [5,6]. This powerful two-step process has completely changed what's possible, shattering performance records on nearly every task. This paper will explore how this technology can be applied to the critical challenge of mental health monitoring.

A. Problem Statement

Despite advancements in Natural Language Processing, existing approaches to mental health text analysis often rely on binary sentiment classification. Such methods are insufficient for identifying the complex and multi-dimensional nature of human emotional expression. Mental health indicators are rarely expressed as purely positive or purely negative sentiment.

Furthermore, traditional feature-engineered models struggle to capture contextual dependencies within text, limiting their ability to recognize subtle emotional cues. This creates a gamp between emotional expression and the representation capability of classical machine learning approaches.

Therefore, there is a need for a context-aware, multi-label emotion classification framework capable of identifying multiple emotional categories within textual data. This paper addresses this challenge by evaluating a transformer-based model against a classical baseline to determine whether contextual representation learning provides measurable advantages in mental health-oriented emotion detection.

II. ORGANISATION

This paper provides a comprehensive review of modern AI techniques for mental health analysis and proposes a detailed framework for a real-time distress detection system.

- The Introduction discusses the shift of people sharing mental health struggles online and the opportunity this presents for AI intervention.
- The Literature Review traces the history of Natural Language Processing (NLP), from early sentiment analysis with NLTK to the revolutionary, context-aware BERT model.
- The Proposed Method outlines a 5-step plan to train a BERT-based model using specialized datasets to identify both general distress and specific emotions.
- The Discussion section explores the advantages (like deep contextual understanding), limitations (like data bias), and ethical considerations of the proposed model.
- The Future Scope section explores potential enhancements such as multilingual support and multimodal analysis.

Finally, the Conclusion summarizes the paper's key arguments, discusses the significant implications of this technology, and underscores the critical ethical considerations that must guide its development.

III. LITERATURE REVIEW

Stage 1: The Foundations of Text Analysis

The Natural Language Toolkit (NLTK) [7], from Steven Bird et al, led to the first widespread use of techniques like Sentiment Analysis. This early method used a pre-defined dictionary of "positive" and "negative" words to create a score [8]. This was fine for simple tasks, but it lacked contextual understanding and couldn't differentiate between the human emotions.

Stage 2: A More Systematic, but Still Manual, Approach

The next stage was defined by the work of Fabian Pedregosa et al., who in 2011 released Scikit-learn [15]. These older methods relied on a process called "feature engineering"[9]. Manual feature engineering limited scalability and generalization [4].

Stage 3: Deep Learning and Sequential Understanding

The third stage brought a fundamental shift with the rise of Deep Learning and, specifically, models called Long Short-Term Memory networks (LSTMs) [11]. Unlike the previous models that looked at text as a "bag of words," LSTMs were designed to process text in sequence, one word at a time [10].

Stage 4: The Unsolved Problem of One-Way Context

While LSTMs were a major improvement, they suffered from a critical flaw: they were unidirectional. ELMo, introduced by Peters et al., used a "shallow" bidirectional approach but lacked fully integrated contextual modelling [12, 16].

Stage 5: The BERT Revolution and True Two-Way Understanding

In 2018, Devlin et al. introduced the BERT model [7]. It is built on the Transformer architecture, which uses a mechanism called "self-attention." This lets the model look at the whole sentence at once. Devlin et al. came up with a new method called the Masked Language Model (MLM) where approximately 15% of input tokens are randomly masked and predicted using bidirectional context [20]. This training strategy enables deep contextual representation learning and established the pre-train and fine-tune paradigm as a dominant approach in NLP [6].

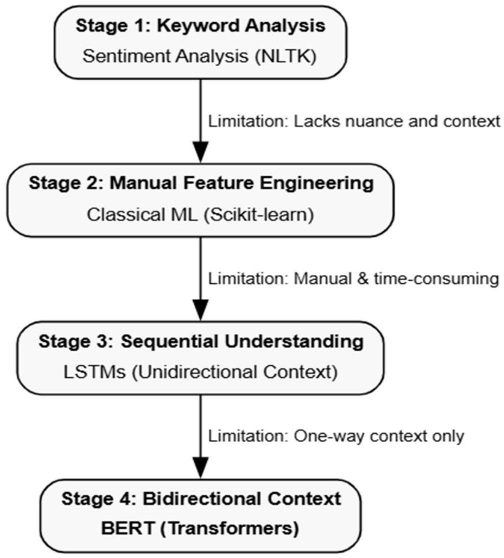


Figure 1. Stages leading to BERT

The development of modern transformer models is supported by open-source ecosystems including NumPy [14], TensorFlow [15], PyTorch [14], and the HuggingFace Transformers library [21], which provide computational infrastructure and access to pre-trained architectures.

TABLE I. SUMMARY OF KEY STUDIES

Study / Tool	Key Idea	Impact on the Field
Scikit-learn (Pedregosa et al., 2011) [15]	Created a simple, consistent interface for common machine learning tasks.	Made machine learning much more accessible and became the standard for many projects.
TensorFlow (Abadi et al., 2016) [15]	Built a powerful and scalable system for large-scale deep learning.	Provided the industrial-strength "engine" needed to train massive models.
PyTorch (Paszke et al., 2019) [14]	Offered a more flexible and user-friendly "define-by-run" deep learning framework.	Became the favourite tool in the research community, speeding up experimentation.
BERT (Devlin et al., 2019) [7]	Introduced the "Masked Language Model" for deep bidirectional training.	Revolutionized NLP by solving the context problem and setting a new standard for performance.
Hugging Face (Wolf et al., 2020) [21]	Created a unified library and a central hub for sharing pre-trained models.	Made advanced models like BERT easy for anyone to download and use.

IV. PROPOSED METHOD

The goal is to build a system that gets progressively smarter, starting with a basic understanding of distress and gradually learning to recognize specific, nuanced emotions in the context of personal journal entries.

A. Dataset Description

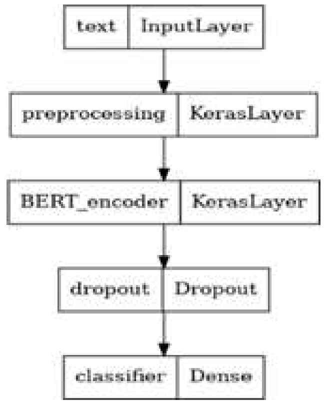
The GoEmotions dataset [6] is used for supervised training. It consists of Reddit comments annotated with one or more of 27 emotion categories, making it a multi-label classification task. The dataset is split into 80% training and 20% validation sets using a fixed random seed (42) to ensure reproducibility.

B. Data Preprocessing

The text data was first normalized by converting all characters to lowercase and removing URLs, non-alphabetic characters and extra whitespace. The processed text was then tokenized using the pre-trained BERT-base-uncased tokenizer. Each input sentence padded to a maximum length of 128. The 27-dimensional binary vector tokenized inputs and label vectors were converted into PyTorch tensors for model training.

C. Model Architecture

The proposed architecture is built upon a pre-trained BERT-base-uncased-model. The contextualized representation corresponding to the pooled CLS token is extracted from the BERT encoder and passed through a dropout layer (p=0.3) to reduce overfitting. A fully connected linear layer maps the 768-dimensional hidden representation to 27 output neurons corresponding to the emotion categories. Since the task is multi-label classification, sigmoid activation is applied during inference and the model is optimized using Binary Cross-Entropy with Logits (BCEWithLogitsLoss).



D. Training Configuration

A classical TF-IDF + Logistic Regression model was first implemented as a baseline for comparison. The proposed BERT-based model was then fine-tuned on the GoEmotions dataset.

The model was trained using BCEWithLogitsLoss to adapt to the multi-label classification setup. The AdamW optimizer was employed with a learning rate of 2e-5 and weight decay of 0.01. Training was conducted for 3 epochs with a batch size of 16. A liner rate scheduler with 10% warmup steps was applied to stabilize optimization.

E. Evaluation Strategy

During evaluation, sigmoid activation was applied to output logits. Threshold tuning was performed in the range of 0.1 to 0.5 to determine the optimal classification threshold. The best performance was achieved at a threshold of 0.2.

Model performance was evaluated using Micro F1, Macro F1, and Weighted F1 scores. Micro F1 aggregates contributions from all labels and reflects overall classification performance. Macro F1 computes the unweighted average across emotion categories. Weighted F1 accounts for label frequence and provides a balanced overall measure.

V. Results and Discussion

TABLE II. PERFORMANCE COMPARISON

Model	Micro F1	Macro F1	Weighted F1
TF-IDF + Logistic Regression	0.409	0.297	0.367
BERT	0.450	0.373	0.440

As shown in Table 2, the proposed BERT-based model demonstrates clear improvements over the classical TF-IDF + Logistic Regression baseline across all evaluation metrics. The Micro F1 score increases from 0.409 to 0.450, indicating stronger overall performance in multi-label emotion classification. This suggests that the transformer-based model is better at jointly predicting multiple emotions within the same text instance.

The most significant improvement observed in the Macro F1 score, which rises from 0.297 to 0.373. Since Macro F1 assigns equal importance to each emotion class regardless of frequence, this result shows that BERT improves performance on less frequent emotions too. In the context of mental health analysis, accurate recognition of subtle or rare emotional states is particularly important, as these signals may reflect critical psychological conditions.

The Weighted F1 score also improves from 0.367 to 0.440, confirming that the performance gains are consistent across the overall class distribution. Unlike TF-IDF, which represents text as independent token frequencies, BERT captures contextual relationships between words through bidirectional self-attention. This enables deeper understanding and allows the model to interpret emotional expressions that depend on context and word interactions.

Threshold optimization further contributed to improved classification performance. A decision threshold of 0.2 provided the best balance between precision and recall in the multilabel setting, highlighting the importance of calibrated probability selection rather than relying on a fixed default threshold.

Overall, the results demonstrate that contextual transformer models provide measurable and meaningful advantages over traditional feature-based machine learning approached for fine-grained emotion detection in contextual data.

VI. Future Scope

- Expansion of Emotional Understanding: Future work could expand the emotional spectrum to include more nuanced states like loneliness, guilt, or hopelessness, which would require the curation of new, more detailed datasets.
- Multilingual Support: A major area for development is adding multilingual capabilities by fine-tuning models pre-trained in other languages (e.g., mBERT) to make this tool globally accessible.
- Multimodal Analysis: Future systems could integrate text analysis with other data sources, such as correlating journal entries with biometric data from wearable devices for a more holistic view of well-being.
- Adaptive Learning Mechanisms: Research could explore an adaptive learning mechanism where the system improves its accuracy over time through continuous user feedback, creating a personalized model.
- Model Efficiency and Accessibility: The quest for greater efficiency through techniques like model distillation and quantization is a critical frontier to make such tools more cost-effective and environmentally friendly to deploy.

VII. Conclusion

This paper presented a transformer-based approach for fine-grained multi-label emotion classification using the GoEmotions dataset. A comparative evaluation was conducted against a classical TF-IDF + Logistic Regression baseline to assess the effectiveness of contextual modelling for emotion detection.

Experimental results demonstrated that the proposed BERT-based model consistently outperformed the baseline across Micro, Macro, and Weighted F1 metrics. The most significant improvement was observed in the Macro F1 score, indicating enhanced performance on minority and less frequent emotion categories. This highlights the importance of bidirectional contextual representation when analysing emotionally nuanced textual data.

The findings confirm that transformer architectures provide measurable advantages over traditional feature-engineered approaches in multi-label emotion recognition tasks. By leveraging self-attention mechanisms and contextual embeddings, the proposed system is better equipped to interpret subtle emotional expressions within text.

The rapid growth of online self-expression has created a profound opportunity for AI-driven mental health analysis, and as this paper has argued, recent breakthroughs in Natural Language Processing, particularly with context-aware models like BERT, have provided the key to unlock this potential. The goal is not an automated therapist, but a powerful support system that helps society listen more effectively to the millions of voices in the digital world, making this a critically important area for continued and careful research.

References

[1] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016). TensorFlow: A System for Large-Scale Machine Learning. In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16).

[2] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc.

[3] Chollet, F. (2018). Deep Learning with Python. Manning Publications.

[4] Cohan, A., Young, S., Hoon, A., & O'Connor, K. (2020). A large-scale dataset of manually annotated Reddit posts for suicide risk assessment. In Proceedings of the 6th Workshop on Computational Linguistics and Clinical Psychology.

[5] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting Depression via Social Media. In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.

[6] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
[8] Gaur, M., Al-Haleem, H., Kursuncu, U., Alambo, A., Sheth, A., & Ruan, Y. (2021). "Let me tell you about your mental health": A deep learning system for analysing mental health from online text. In Proceedings of the 13th International Conference on Web Search and Data Mining.
[9] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362.
[10] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
[11] Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed.). Prentice Hall.
[12] Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In Proceedings of the 8th International Joint Conference on Natural Language Processing.
[13] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.
[14] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems.
[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.
[16] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics.
[17] Rude, S. (2019). An Overview of Multi-Task Learning in Deep Neural Networks. arXiv. https://arxiv.org/abs/1706.05098
[18] Saeidi, S., & Dvorkin, Y. (2018). The role of open-source software in the energy system modeling community. Energy Strategy Reviews, 20, 1-7.
[19] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems.
[21] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv preprint arXiv:1910.03771.
[22] World Health Organization. (2023). Depression. WHO. https://www.who.int/health-topics/depression