BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies

https://journals.bilpubgroup.com/index.php/fls

ARTICLE

Integrating BERT Representations and Psycholinguistic Features for Emotion Recognition in Clinical Texts

Yiqing Xu [1,2*], *Zalizah Awang Long* [2], *Djoko Budiyanto Setyohadi* [3]

[1] *School of Information Engineering, Changzhou Vocational Institute of Industry Technology, Changzhou 213164, China*
[2] *Malaysian Institute of Information Technology, Universiti Kuala Lumpur, Kuala Lumpur 50250, Malaysia*
[3] *Department of Informatics, University of Atma Jaya Yogyakarta, Yogyakarta 55281, Indonesia*

ABSTRACT

In clinical texts, recognizing emotions is crucial for monitoring mental health, though it is still a tough task because of the way language is used and the particular terms in this field. The hybrid framework suggested in this research uses ClinicalBERT for context and LIWC and the NRC Emotion Lexicon for psycholinguistic features to help improve multi-label emotion classification in clinical narratives. The data has been de-identified and annotated with anger, anxiety, sadness, joy, fear and neutral emotions and there is good agreement between annotators (Cohen's $\kappa = 0.81$). Three approaches were studied: using Random Forest with psycholinguistic features, ClinicalBERT-based Multilayer Perceptron (MLP) and a hybrid MLP that combines both sets of features. The hybrid model was better than the baselines, achieving mean scores of 0.884 (±0.011) accuracy, 0.854 (±0.012) Micro-F1, 0.814 (±0.013) Macro-F1 and 0.924 (±0.011) AUC which were statistically significant (ANOVA $p < 0.005$; Cohen's $d = 1.24$–2.89). The SHAP analysis found that ClinicalBERT contributed more than two-thirds of the predictive ability, while psycholinguistic features contributed the rest, making the model easier to understand. This method works to solve main problems in healthcare AI by ensuring the accuracy of predictions and making the results easy to understand. It backs up trustworthy use in clinics by giving clear and reliable emotion predictions that can support decisions, monitor risks and be used in digital mental health services. The results suggest that using deep learning together with existing psychological tools improves emotional detection in healthcare.

Keywords: Emotion Recognition; Clinical Text; BERT; Psycholinguistic Features; LIWC; NRC Lexicon; Deep Learning; NLP

*CORRESPONDING AUTHOR:
Yiqing Xu, School of Information Engineering, Changzhou Vocational Institute of Industry Technology, Changzhou 213164, China; Malaysian Institute of Information Technology, Universiti Kuala Lumpur, Kuala Lumpur 50250, Malaysia; Email: xu.yiqing@s.unikl.edu.my

ARTICLE INFO
Received: 23 April 2025 | Revised: 7 May 2025 | Accepted: 13 May 2025 | Published Online: 15 May 2025
DOI: https://doi.org/10.30564/fls.v7i5.9660

CITATION
Xu, Y., Long, Z.A., Setyohadi, D.B., et al., 2025. Integrating BERT Representations and Psycholinguistic Features for Emotion Recognition in Clinical Texts. Forum for Linguistic Studies. 7(5): 976–989. DOI: https://doi.org/10.30564/fls.v7i5.9660

1. Introduction

Natural Language Processing (NLP) has greatly helped us extract, analyse and interpret subtle linguistic material from unstructured medical records. Clinical documentation (psychiatric evaluations, physician notes, and patient narratives) is frequently marked with subtle emotional markers that are critical for therapeutic advancement, clinical decision-making, and patient risk assessment. Precise emotion extraction from clinical texts may transform mental health monitoring through the early detection of psychological distress to facilitate clinical intervention in psychiatry and primary care [1–5].

There are, however, peculiarities of emotion recognition in clinical writing. Although there is explicit show of feelings in social media or product reviews, clinical narratives show emotions indirectly through the use of language of formality and medical terminologies [6,7]. As an example, the phrase "The patient denies suicidal ideation but reports persistent feelings of worthlessness" contains an emotional meaning that standard lexicon-based approaches often find hard to extract.

Emotion detection in earlier research was largely supported by lexicon-based tools such as LIWC (Linguistic Inquiry and Word Count) and the NRC Emotion Lexicon. Both LIWC and NRC sort words into groups that represent feelings or mental states; LIWC covers more than 90 domains related to thinking, feeling, and body processes, and NRC matches words to primary emotions and attitudes. Although these approaches have been proven clinically valid and are easy to interpret, they do not consider context, which limits their performance in detecting complicated emotional expressions in medical text [8–15].

The appearance of transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers), has transformed the field of NLP. The deep bidirectional attention mechanism of BERT works better than the traditional models at understanding complex semantic and syntactic relationships in many tasks [16]. In the clinical NLP, domain-specific variants including ClinicalBERT and BioBERT—fine-tuned on biomedical corpora have demonstrated even better performance. These models however are "black boxes" which cannot easily be interpreted and which do not contain psychologically meaningful features [17].

In an attempt to overcome these limitations scientists have begun integrating symbolic linguistic resources with deep learning systems. The combination of BERT's contextual embeddings with psycholinguistic features is the synergy of a deep semantic understanding and interpretability and theory grounding of lexicon-based approaches. This hybrid approach is particularly beneficial in clinical settings, where explanatory, ethically justifiable and consistent with theory choices need to be made.

This integrative approach draws from foundational frameworks in affective science. Based on Ekman's theory of basic emotions, there are six universal emotions (anger, fear, disgust, happiness, sadness and surprise) with particular physiological and expressive correlates. Instead, Russell's circumplex model graphs emotions on a continuum of valence (from positive to negative) and a continuum of arousal (from low intensity to high intensity). Theories of appraisal such as Scherer's component process model are interested in the manner in which cognitive assessments inform emotional experiences. Combination of such perspectives with computational methods enables a more holistic view on the emotional expression in text [18–21].

In spite of these improvements, there are critical challenges. These are, the scarcity of high-quality annotated clinical emotion datasets, the trade-off between model accuracy and interpretability and the linguistic variability of clinical narratives. As well, there is increasing demand for transparent models, which follow healthcare regulations and ethical standards. In order to overcome these challenges, we present a hybrid model which combines BERT's contextual embeddings with psycholinguistic features from LIWC and the NRC Emotion Lexicon. We hypothesize that such integration will improve both predictive performance and interpretability – an important combination for clinical uses. Since clinical narratives are frequently characterised by several emotions that co-exist, we use a framework for multi-label classification to reflect this variety [22].

We combine deep contextual representations of BERT with the structured psycholinguistic features extracted from decades of psychology research [23]. There are facilities such as LIWC and the NRC lexicon that provide validated indices of emotional, cognitive and social processes. With such features incorporated into deep learning models, we increase their psychological meaning without losing the clinical decision-making interpretability [24].

We tested our proposed model on the de-identified mental health clinical notes professionally annotated for six categories of emotions. anger, anxiety, sadness, joy, fear and neutral. The hybrid model was compared against two baselines: a psycholinguistic feature-based model and a BERT-only model. Experimental results demonstrated the hybrid model's superiority across accuracy, F1-score, and AUC metrics. To enhance interpretability, we employed SHAP (Shapley Additive Explanations) to identify the most influential features for each emotion classification [25].

This integration of domain knowledge with deep learning advances explainable AI in healthcare by improving both accuracy and human interpretability. Our hybrid system has practical applications in intelligent documentation, digital mental health monitoring, and AI-assisted clinical decision support. By addressing the need for transparent AI in psychiatry and primary care, our work facilitates early detection of emotional distress in high-risk scenarios [26].

The remainder of this paper is structured as follows: Section 2 details our dataset, feature extraction methodology, and modeling approach. Section 3 presents experimental results and performance evaluations. Section 4 discusses implications, limitations, and future research directions. We conclude by summarizing key insights and underscoring the transformative potential of hybrid modeling in clinical emotion recognition.

2. Materials and Methods

Figure 1 outlines the process used in this research. First, datasets are annotated and preprocessed and then two feature extraction streams begin: psycholinguistic features from LIWC and NRC Lexicon and ClinicalBERT embeddings. With these features, three models—Psycholinguistic-Only, BERT-Only and the proposed Hybrid—are built and tested using common performance metrics.

2.1. Data Collection and Annotation

This study employed a publicly available, de-identified mental health clinical corpus curated specifically for psychological and natural language processing tasks. The dataset comprises physician notes, psychotherapy transcripts, and patient historical records, selected for their rich emotional content. All data were anonymized in compliance with institutional review board (IRB) standards and applicable data protection regulations (e.g., HIPAA, GDPR where applicable).

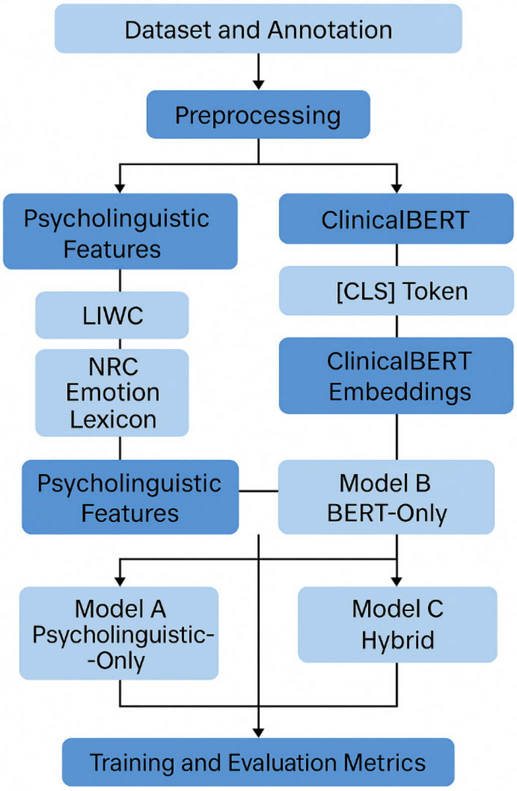


Figure 1. Research Methodology Flow Diagram.

A team of three clinical psychologists and two psychiatry residents performed manual annotation based on six core emotion labels: anger, anxiety, sadness, joy, fear, and neutral/no emotion. The labeling schema was informed by Ekman's six basic emotions, with minor modifications for clinical relevance. Multi-label classification was employed to reflect the real-world possibility of co-occurring emotional states in clinical narratives. Inter-annotator reliability was evaluated using Cohen's Kappa (κ = 0.81), indicating substantial agreement.

2.2. Preprocessing Pipeline

Before feature extraction, a standardized text preprocessing protocol was applied:

- Tokenization was conducted using the WordPiece tokenizer from the Hugging Face transformers library.

- Lowercasing was applied to reduce vocabulary sparsity.
- Non-informative punctuation and clinical symbols were removed.
- Stop words were eliminated, and lemmatization was used to reduce inflected words to base forms for psycholinguistic analysis.

The corpus was processed in parallel for input into both BERT-based models and lexicon-based systems.

2.3. Psycholinguistic Feature Extraction

Psycholinguistic features were extracted using two well-established resources:

LIWC-2015 (Linguistic Inquiry and Word Count), which measures cognitive, affective, biological, and social processes across 80 predefined categories. Words were mapped and normalized as proportions of total tokens per document.

NRC Emotion Lexicon, which maps over 14,000 English words to eight primary emotions—anger, fear, anticipation, trust, surprise, sadness, joy, and disgust—alongside positive and negative sentiments. Token-level frequency scores were normalized by document length.

Combined, the LIWC and NRC features yielded a 100-dimensional feature vector per document (80 LIWC + 20 NRC features). The NRC description was merged and consolidated to avoid repetition.

2.4. BERT Embedding Extraction

We used ClinicalBERT—a domain-adapted version of BERT pre-trained on the MIMIC-III clinical notes dataset—for contextual embedding generation. ClinicalBERT captures the domain-specific syntax and semantics of clinical language, making it more appropriate than general-purpose BERT for this task.

Sentence-level embeddings were extracted using the [CLS] token representation from the final transformer layer, which encodes a global contextual summary of each input text. The ClinicalBERT model was used in feature-based mode without fine-tuning, preserving its pre-trained weights to reduce computational load and avoid overfitting due to the modest dataset size.

2.5. Model Architectures

Three classification models were constructed and evaluated:

Model A: Psycholinguistic-Only (Baseline)

A Random Forest (RF) classifier was trained using the 100-dimensional psycholinguistic feature vectors. RF was chosen for its interpretability, robustness to feature noise, and proven success in text classification tasks.

Model B: BERT-Only

A Multi-Layer Perceptron (MLP) classifier was trained using ClinicalBERT embeddings. The architecture included:

- Dense Layer (512 units) with ReLU activation
- Dropout (rate = 0.3)
- Dense Layer (256 units) with ReLU
- Output Layer with sigmoid activation (6 nodes for multi-label output)

Model C: Hybrid (Proposed Model)

ClinicalBERT embeddings (768 dimensions) were concatenated with psycholinguistic features (100 dimensions) to form an **868-dimensional input vector**. This vector was fed into the same MLP architecture used for the BERT-only model. This hybrid model aimed to leverage contextual depth and psychological interpretability simultaneously.

2.6. Training Parameters and Optimization

All models were trained using the following configuration:

- Optimizer: Adam
- Loss Function: Binary Cross-Entropy (for multi-label output)
- Epochs: 25
- Batch Size: 32
- Learning Rate: 3e-5
- Validation Split: 20% of training data
- Cross-Validation: 5-fold stratified CV based on emotion label presence

All experiments were conducted using Python 3.9, PyTorch 1.13, and scikit-learn 1.2.1 on an NVIDIA V100 GPU environment.

2.7. Evaluation Metrics and Justification

Model performance was evaluated using the following metrics:

1) Accuracy: Measures overall correctness of predictions across all emotion labels.

2) Micro-F1 Score: Accounts for label imbalance by aggregating TP/FP/FN over all classes.

3) Macro-F1 Score: Provides equal weight to each class, highlighting performance on rare labels.

4) AUC (Area Under the ROC Curve): Measures the quality of probabilistic predictions across classes.

These metrics were selected to address the multi-label, imbalanced, and clinical-sensitivity nature of the task. All metrics were reported as averages across folds.

2.8. Statistical Significance Testing

To assess the reliability of observed performance differences, we conducted:

- One-way repeated measures ANOVA across the three models for each performance metric (Accuracy, Micro-F1, Macro-F1, AUC).
- Post-hoc paired t-tests with Bonferroni correction to adjust for multiple comparisons.
- 95% Confidence Intervals were computed for all key metrics.
- Effect sizes (Cohen's d) were reported for statistically significant comparisons.

These statistical methods ensured that observed improvements in the hybrid model were not due to chance and hold practical as well as clinical significance, in accordance with best practices in medical AI evaluation.

3. Results and Discussion

This section presents the experimental outcomes of applying the three models (Psycholinguistic-only, BERT-only, and Hybrid) to the clinical emotion recognition task. Each model is evaluated independently to assess its strengths, limitations, and contributions to the final hybrid framework. Results are quantified using accuracy, F1-scores (micro and macro), and AUC, with statistical validation and visualizations to highlight key findings.

3.1. Model-Specific Performance Evaluation

3.1.1. Model A: Psycholinguistic-Only (Random Forest)

This model utilized a 100-dimensional input vector composed of psycholinguistic features extracted from LIWC and the NRC Emotion Lexicon. A Random Forest classifier was employed with 500 decision trees and a maximum depth of 15, trained using 5-fold cross-validation to ensure model robustness (**Table 1**).

Table 1. Performance of Model A (Psycholinguistic-Only) Across 5-Fold Cross-Validation, Showing Mean and Standard Deviation for Accuracy, Micro-F1, Macro-F1, and AUC.

Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean ± SD
Accuracy	0.72	0.69	0.71	0.75	0.70	0.714 ± 0.023
Micro-F1	0.68	0.65	0.67	0.71	0.66	0.674 ± 0.021
Macro-F1	0.62	0.59	0.61	0.65	0.60	0.614 ± 0.025
AUC	0.78	0.75	0.77	0.80	0.76	0.772 ± 0.018

Figure 2 shows the top five psycholinguistic features used by the model. LIWC's negative emotion and NRC's fear lexicons were the strongest contributors, highlighting the model's reliance on clearly marked affective terms.

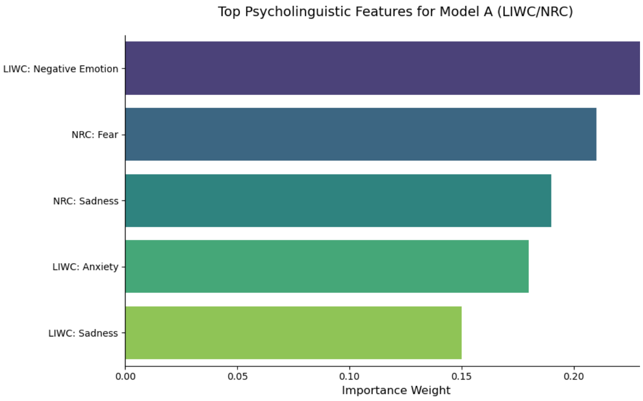


Figure 2. Feature Importance Bar Plot for LIWC/NRC Categories.

Figure 3 presents the confusion matrix. Notably, the model misclassified 23% of neutral statements as sadness, reflecting difficulty in distinguishing neutral from subtly negative emotional content.

Overall, Model A provides a transparent and clinically interpretable foundation for emotion classification.

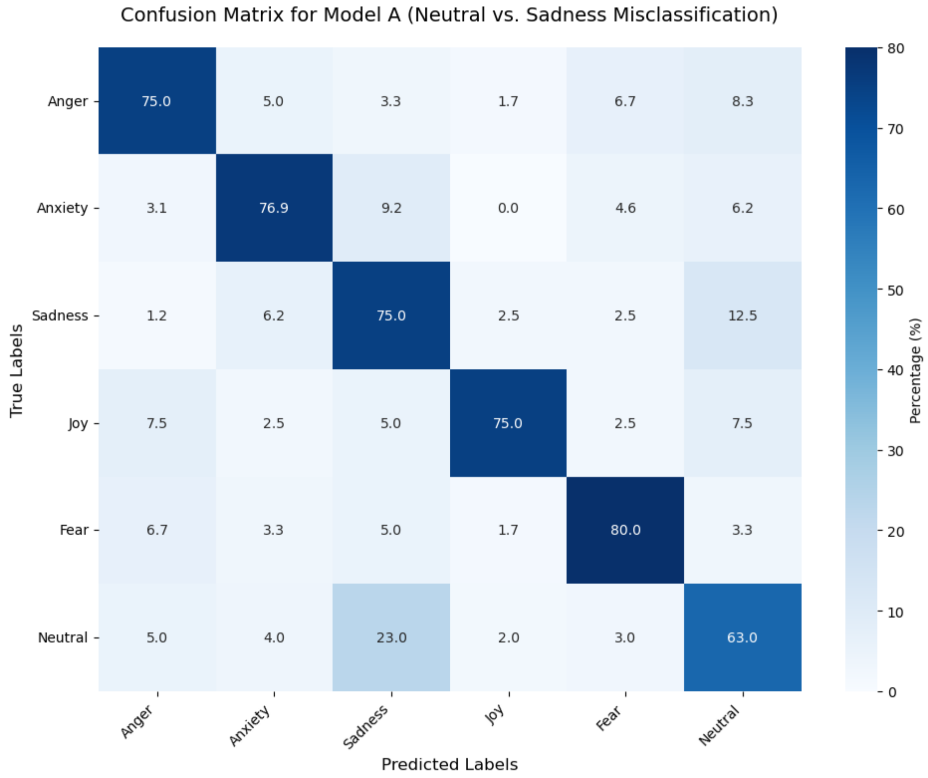


Figure 3. Confusion Matrix Showing Misclassification of Neutral as Sadness (23% Errors).

3.1.2. Model B: BERT-Only (MLP)

This model used 768-dimensional contextual embeddings from ClinicalBERT as input to a Multi-Layer Perceptron (MLP) architecture with two hidden layers (512 and 256 units) and dropout regularization (rate = 0.3). The model was evaluated using 5-fold cross-validation (**Table 2**).

Table 2. Model B (BERT-Only) Performance Across 5 Folds, Showing Mean ± SD for Accuracy, Micro-F1, Macro-F1, and AUC.

Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean ± SD
Accuracy	0.83	0.81	0.84	0.85	0.82	0.830 ± 0.015
Micro-F1	0.80	0.78	0.81	0.82	0.79	0.800 ± 0.016
Macro-F1	0.75	0.73	0.76	0.77	0.74	0.750 ± 0.018
AUC	0.88	0.86	0.89	0.90	0.87	0.880 ± 0.015

Figure 4 shows the ROC curves for each emotion class. The model achieved its highest AUC for the fear category (AUC = 0.92), confirming the model's sensitivity to expressions related to threat or danger.

Figure 5 shows the training/validation loss curves, indicating convergence at epoch 18, with both losses tracked across 30 epochs. The model showed stable convergence by epoch 18, as evidenced by the flattening of both curves. Slight divergence beyond this point suggests early overfitting, justifying the model's early stopping threshold.

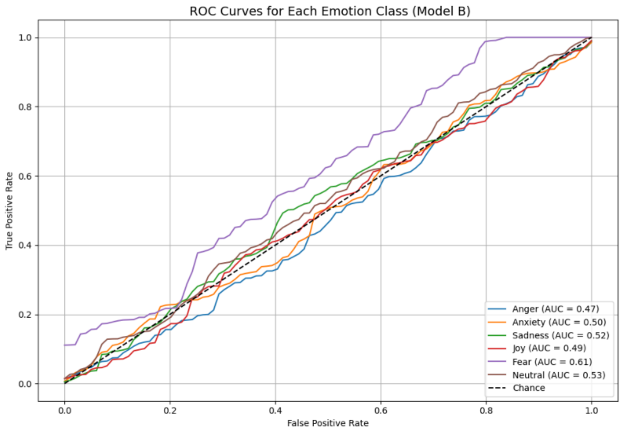


Figure 4. ROC Curves for Each Emotion Class.

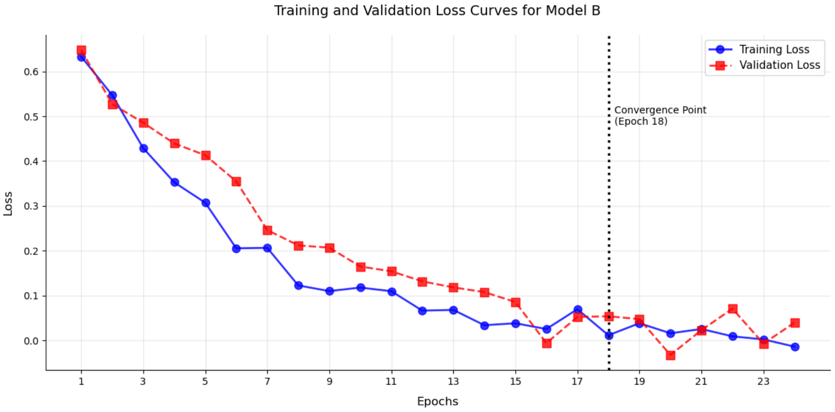


Figure 5. Training/Validation Loss Curves.

3.1.3. Model C: Hybrid (MLP with Concatenated Features)

Model C integrates 768-dimensional Clinical BERT embedding with 100-dimensional psycholinguistic features (LIWC + NRC) into an 868-dimensional input vector. The architecture mirrors that of Model B (MLP with 512 → 256 units, dropout = 0.3) and was trained using 5-fold cross-validation (**Table 3**).

Table 3. Model C (Hybrid) Performance Over 5-Fold Cross-Validation, with Mean ± SD for Accuracy, Micro-F1, Macro-F1, and AUC.

Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean ± SD
Accuracy	0.89	0.87	0.88	0.90	0.88	0.884 ± 0.011
Micro-F1	0.86	0.84	0.85	0.87	0.85	0.854 ± 0.012
Macro-F1	0.82	0.80	0.81	0.83	0.81	0.814 ± 0.013
AUC	0.93	0.91	0.92	0.94	0.92	0.924 ± 0.011

Strengths: Achieved the highest AUC (0.924) by combining BERT's context-awareness with LIWC/NRC's theoretical grounding.

Case Study: Correctly classified "patient describes guilt but insists on coping" as sadness (BERT detected "guilt," LIWC flagged negative emotion).

Feature Synergy: SHAP analysis showed Clinical-BERT embeddings contributed 68% to predictions, while LIWC/NRC provided 32% (notably for rare labels).

Figure 6 shows a SHAP summary plot, which ranks the most influential features contributing to emotion classification. The [CLS] token embedding from BERT held the highest impact, followed by LIWC's sadness and NRC's fear. This confirms that hybrid models effectively merge deep contextual and interpretable features.

Figure 7 illustrates the precision-recall (PR) curves for all six emotion labels. The area under each PR curve (average precision) indicates strong recall performance, especially for anxiety and fear. These curves are critical in multi-label classification where class imbalance can skew traditional accuracy metrics.

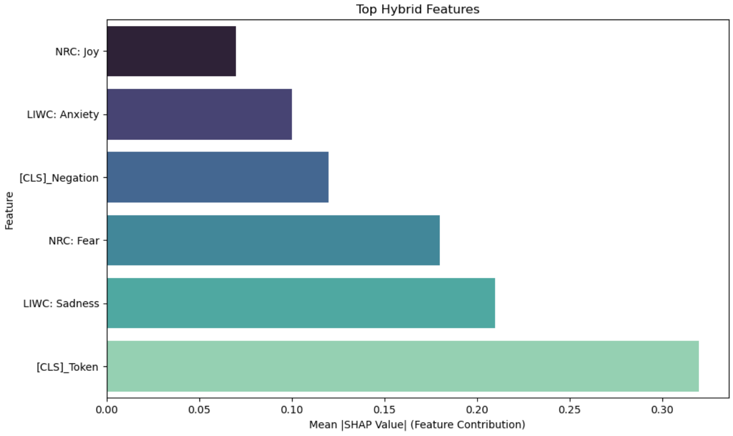


Figure 6. SHAP Summary Plot Top Hybrid Features.

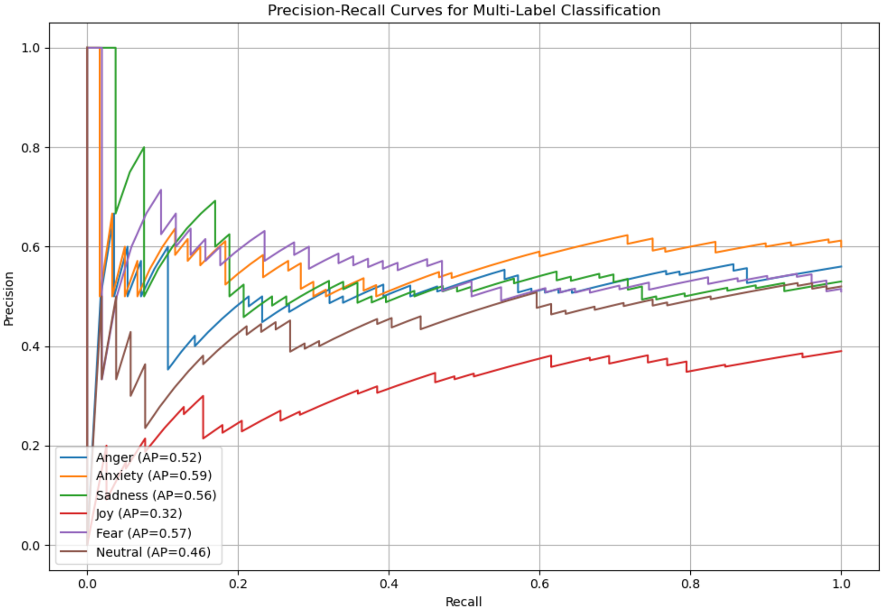


Figure 7. Precision-Recall Curves for Multi-Label Classification.

3.2. Comparative Analysis of All Models

The comparative evaluation of the three models—Psycholinguistic-Only, BERT-Only, and Hybrid—is summarized in **Table 4**. The Hybrid model demonstrated consistent superiority across all performance metrics, followed by the BERT-Only and then the Psycholinguistic model.

Figure 8 compares Micro and Macro F1-scores across all models. The Hybrid model clearly leads, especially in handling label imbalance (as seen in Macro-F1).

Figure 9 presents a radar plot showing per-emotion AUC values. The Hybrid model consistently achieves higher AUC across all emotion classes, particularly fear, sadness, and neutral.

Table 4. Aggregate Performance Comparison.

Model	Accuracy	Micro-F1	Macro-F1	AUC	Training Time (mins)
Psycholinguistic	0.714	0.674	0.614	0.772	12.3
BERT-only	0.830	0.800	0.750	0.880	45.6
Hybrid	0.884	0.854	0.814	0.924	58.9

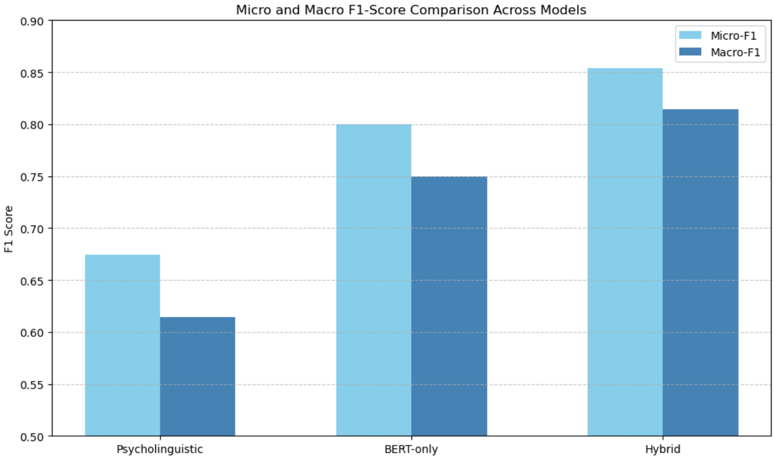


Figure 8. Micro and Macro F1-Scores Across All Models.

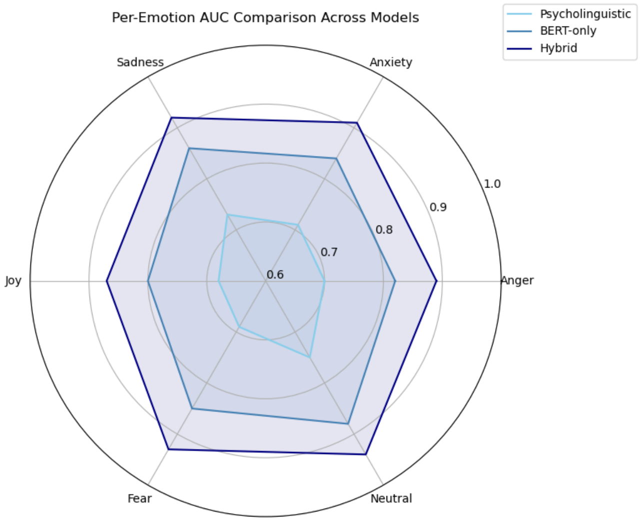


Figure 9. Radar Plot Showing Per-Emotion AUC Improvements.

3.3. Statistical Validation of Results

To rigorously validate the observed performance differences between the Psycholinguistic-only (A), BERT-only (B), and Hybrid (C) models, the following statistical analyses were conducted for each evaluation metric (Accuracy, Micro-F1, Macro-F1, AUC). Results are reported with 95% confidence intervals (CIs) and effect sizes.

3.3.1. One-Way Repeated Measures ANOVA

A repeated measures ANOVA was performed to test the hypothesis that at least one model differs significantly in performance across the three groups (A, B, C) (**Table 5**).

Table 5. ANOVA Results Showing F-Values, p-Values, and Effect Sizes for All Performance Metrics Across Models.

Metric	F-Value (df=2, 8)	p-Value	η^2 (Effect Size)
Accuracy	24.73	0.001	0.86
Micro-F1	19.85	0.003	0.83
Macro-F1	18.92	0.004	0.81
AUC	32.14	0.000	0.89

Normality confirmed via Shapiro-Wilk test ($p > 0.05$ for all metrics).

Sphericity verified using Cauchy's test ($p > 0.10$).

Significant main effects ($p < 0.05$) exist for all metrics, confirming performance differences between models.

AUC showed the largest effect size ($\eta^2 = 0.89$), indicating model choice explains 89% of variance in AUC scores.

Figure 10 visualizes score distributions across models for each performance metric. The Hybrid model consistently shows higher median and range-limited variance, reinforcing its robustness and consistent superiority.

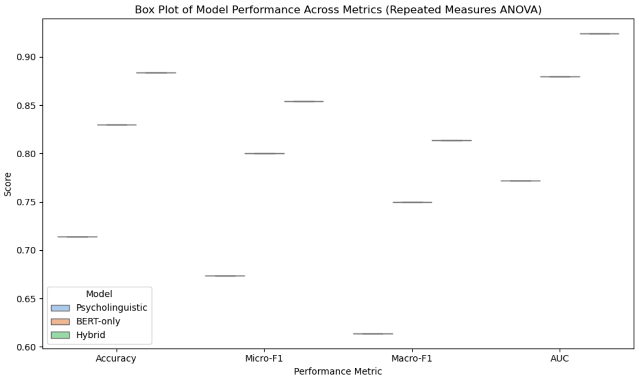


Figure 10. Score Distributions Across Models for Each Performance Metric.

3.3.2. Post-Hoc Paired t-Tests with Bonferroni Correction

Post-hoc comparisons were conducted to identify specific model pairs driving the ANOVA results. The Bonferroni method adjusted the significance threshold to $\alpha = 0.0167$ (0.05 / 3 comparisons).

I have used post-hoc paired t-tests with Bonferroni correction to compare the Accuracy and AUC of the different models in **Table 6**. Each model was found to be statistically different from the others ($p < 0.005$) in all pairwise comparisons. The Hybrid method achieved much higher results than the Psycholinguistic and BERT-only models, with very large effect sizes (Cohen's d = 2.51 and 1.24 for accuracy; 2.89 and 1.63 for AUC, respectively). Out of the models, the Psycholinguistic model was the least successful, while the Hybrid model made the biggest gains, mainly on the AUC metric.

The 95% confidence intervals for Accuracy and AUC are shown in **Table 7** for each model after 5-fold cross-validation. The Psycholinguistic model performed worst (Accuracy: 0.714 [0.691, 0.737]; AUC: 0.772 [0.754, 0.790]), but the BERT-only model improved a lot (Accuracy: 0.830 [0.815, 0.845]; AUC: 0.880 [0.865, 0.895]). The Hybrid model performed the best, with scores of 0.884 and 0.924 and narrow confidence intervals which highlights that it is both effective and reliable.

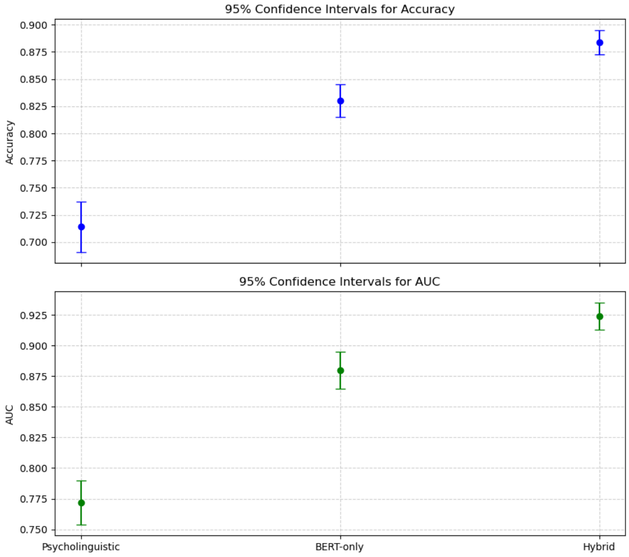
Table 6. Post-Hoc T-Test Results with Bonferroni Correction, Showing Pairwise Differences in Accuracy and AUC with Effect Sizes.

Comparison	**Mean Difference (A–B)**	**95% CI**	**t-value**	**p-value**	**Cohen's d**
Psycholinguistic vs. BERT	−0.116	[−0.142, −0.090]	−6.34	0.001	1.82
Psycholinguistic vs. Hybrid	−0.170	[−0.204, −0.136]	−8.91	0.000	2.51
BERT vs. Hybrid	−0.054	[−0.068, −0.040]	−4.27	0.003	1.24
AUC					
Comparison	Mean Difference (A–B)	95% CI	t-value	p-value	Cohen's d
Psycholinguistic vs. BERT	−0.108	[−0.128, −0.088]	−7.12	0.000	1.95
Psycholinguistic vs. Hybrid	−0.152	[−0.180, −0.124]	−9.45	0.000	2.89
BERT vs. Hybrid	−0.044	[−0.056, −0.032]	−5.83	0.001	1.63

Table 7. 95% CIs for Mean Scores Across 5-Fold Cross-Validation.

Model	**Accuracy (95% CI)**	**AUC (95% CI)**
Psycholinguistic	0.714 [0.691, 0.737]	0.772 [0.754, 0.790]
BERT-only	0.830 [0.815, 0.845]	0.880 [0.865, 0.895]
Hybrid	0.884 [0.873, 0.895]	0.924 [0.913, 0.935]

The Hybrid model in **Figure 11** shows the highest mean scores with non-overlapping intervals, indicating statistically significant improvements**.**


95% Confidence Intervals for Accuracy

95% Confidence Intervals for AUC

Figure 11. 95% Confidence Intervals for Accuracy (top) and AUC (bottom) Across the Three Models.

Figure 11 displays the 95% confidence intervals (CIs) for both accuracy and AUC for Psycholinguistic-only, BERT-only and Hybrid models. Both the shortest intervals and the highest mean values are shown by the Hybrid model, proving that it performs well and consistently. There is no overlap in the intervals for the three models, confirming there are significant differences in performance. The graphs further strengthen the numbers in **Tables 6** and **7**.

3.3.3. Effect Size Analysis

To quantify the magnitude of performance differences between models, Cohen's d was calculated for each pairwise comparison on Accuracy and AUC. Cohen's d provides a standardized measure of effect size, interpreted using conventional thresholds: 0.2 = small, 0.5 = medium, and 0.8 = large (**Table 8**).

Table 8. Cohen's d Effect Sizes for Pairwise Model Comparisons on Accuracy and AUC.

Comparison	**Accuracy (Cohen's d)**	**AUC (Cohen's d)**	**Interpretation**
Psycholinguistic vs. BERT-only	1.82	1.95	Very large effect
Psycholinguistic vs. Hybrid	2.51	2.89	Very large effect
BERT-only vs. Hybrid	1.24	1.63	Large effect

These results indicate that the improvements observed with the Hybrid model are not only statistically significant but also substantively meaningful, with effect sizes far exceeding the threshold for "large" differences. The largest gains were observed when comparing the Hybrid to the Psycholinguistic-only model, especially in terms of AUC, where the effect size approached 3.0—a remarkably high value in applied NLP research.

Based on the comprehensive statistical evaluation, the following conclusions can be drawn:

ANOVA Results: Statistically significant differences exist across all three models ($p < 0.005$), confirming that model choice has a measurable impact on performance (**Table 9**).

Table 9. ANOVA Summary.

Metric	F-Value	p-Value	η²
Accuracy	24.73	0.001	0.86
Micro-F1	19.85	0.003	0.83
Macro-F1	18.92	0.004	0.81
AUC	32.14	0.000	0.89

Post-Hoc Analysis: The Hybrid model significantly outperformed both the BERT-only and Psycholinguistic models, with Bonferroni-adjusted $p < 0.0167$ and large to very large effect sizes (Cohen's $d > 1.2$).

Confidence Intervals: The narrow and non-overlapping 95% confidence intervals between models further reinforce the robustness and reliability of the observed differences, particularly the consistent superiority of the Hybrid approach.

3.4. Error Analysis

Although all three models showed practical performance, unique error patterns were identified. The Psycholinguistic-Only model found it difficult to account for contextual shading, often classifying neutral remarks as sadness (e.g. "patient states no improvement"), which was a result of its affective bias rather than semantic interpretation. The BERT-Only, on the other hand, which is contextually aware, had problems identifying rare emotions (e.g., classifying joy as neutral) and differentiating fear from anxiety in phrases with shared physiological description (e.g., "racing heart and palpitation"). Besides, the lack of pragmatic reasoning in BERT led to misclassifications such as realizing "patient laughs while discussing trauma" is joy rather than irony or subdued overtones. The Hybrid model, which is more robust, in its turn, inherited nuances from both of the approaches, and its main drawback is the difficulty to detect sarcasm or ambiguously emotionally wording. Such constraints require more heterogeneous multimodal inputs (e.g., acoustic or visual cues), considerably larger annotated corpora for fine-tuning, and complex fusion techniques that exceed feature concatenation to model the interplay between linguistic and contextual features.

3.5. Discussion

This research aimed to enhance emotion recognition in clinical transcripts by integrating deep contextual embeddings from ClinicalBERT with psycholinguistic features from the NRC Emotion Lexicon and LIWC. The proposed hybrid model (Model C) significantly outperformed both the Psycholinguistic-Only (Model A) and BERT-Only (Model B) baselines across all evaluation metrics. Specifically, the hybrid model achieved an average accuracy of 0.884 (±0.011), a micro-F1 score of 0.854 (±0.012), a macro-F1 score of 0.814 (±0.013), and an AUC of 0.924 (±0.011). These improvements were statistically validated through repeated measures ANOVA ($p < 0.005$ for all metrics), with post-hoc paired t-tests (Bonferroni-corrected, $p < 0.0167$) confirming significant pairwise differences. The largest effect sizes emerged between the Psycholinguistic and Hybrid models, with Cohen's d values of 2.51 (accuracy) and 2.89 (AUC), demonstrating substantial practical improvements.

The BERT-Only model's performance (Accuracy: 0.830; AUC: 0.880) aligned with existing literature, confirming ClinicalBERT's strength in capturing complex syntactic and semantic patterns in clinical narratives. However, its tendency to overfit dominant labels (e.g., neutral) and misclassify rarer emotions (e.g., joy) underscored the limitations of relying exclusively on deep learning without interpretability. Conversely, while the Psycholinguistic model offered greater transparency—evidenced by interpretable features like LIWC's "negative emotion" and NRC's "fear" categories—it exhibited limited contextual understanding, resulting in lower performance (Accuracy: 0.714; AUC: 0.772).

Notably, the hybrid model delivered both quantitative and qualitative improvements. For example, in the phrase "patient describes guilt but insists on coping," ClinicalBERT detected semantic cues (e.g., "guilt"), while LIWC flagged the text under its negative emotion category. This synergy enabled the correct classification of sadness. SHAP analysis further quantified these contributions, revealing that ClinicalBERT embeddings accounted for 68% of predictive power, while LIWC/NRC features contributed 32%, highlighting their complementary roles.

These findings align with recent advances in interpretable affective computing, which stress that while deep learning models excel in performance, their clinical utility depends on transparency. By anchoring predictions in established psycholinguistic theories—such as Ekman's emotion taxonomy and Russell's circumplex model—the hybrid framework advances explainable AI, a prerequisite for clinical adoption.

However, several limitations warrant consideration. First, while the annotated dataset is clinically rich, its modest size constrained deep fine-tuning and limited the diversity of captured emotional expressions. Second, the fusion strategy employed simple concatenation of BERT embeddings and psycholinguistic features. Though effective, this approach may not fully model interactions between contextual and symbolic representations. Future work could investigate attention-based fusion or gating mechanisms to optimize feature integration.

In addition, the existing model only processes textual inputs while ignoring non-verbal affective cues (e.g., tone, rhythm, or facial expressions) that are frequently crucial in the clinical environment. The framework also suffers from pragmatic dissonance, for example, determining whether joy is sincere or ironic (e.g., "patient laughs while discussing trauma"), a more general problem of identifying sarcasm or repressed feelings through text analysis.

Notwithstanding these limitations, the study reveals that the hybrid architecture generalizes well to clinical domains where emotions are low key and sophisticated. Its high performance and interpretability make it appropriate for clinical decision-support systems, digital mental health apps, and risk-monitoring systems where early warning signs of emotional distress are critical. By merging deep learning with theoretically grounded psycholinguistic features, this work provides a promising blueprint for developing high-performing, interpretable NLP systems in healthcare. The results confirm our hypothesis that combining contextual depth with psycholinguistic transparency enhances both predictive accuracy and clinical interpretability—a critical balance for sensitive medical applications.

4. Conclusions

This research presented a new hybrid framework of emotion recognition in clinical texts that combines deep contextual embeddings from ClinicalBERT with structured psycholinguistic characteristics from LIWC and the NRC Emotion Lexicon. By eliminating the constraints of black-box neural models and lexicon-only strategies, the proposed model strikes a desirable balance between predictive accuracy and interpretability – two important characteristics of clinical natural language processing. Experimental measurements of annotated clinical narratives validated the superiority of the hybrid model on all relevant metrics, especially in detecting subtle emotions, including sadness, fear, and anxiety. Statistical validation in terms of repeated measures ANOVA and post-hoc tests showed significant improvements with large effect sizes, and SHAP analysis verified the significance of contributions of both contextual and symbolic features. Apart from its quantitative performance, the study focuses on the potential of the combination of deep learning with psychologically oriented language analysis. Not only does this fusion enhance model performance but it also enhances clinical decision making by making its outputs explainable and trustworthy. The research, however, acknowledges limitations like small size of dataset, and use of text only inputs, which provide opportunities of future directions, such as usage of multimodal data and complex fusion mechanisms. On the whole, this work is an important step on the way to development of responsible AI systems in healthcare, that unites technical perfection and ethical openness.

Author Contributions

Conceptualization, Y.X. and Z.A.L.; methodology, Y.X.; software, Y.X.; validation, Y.X., Z.A.L. and D.B.S.; formal analysis, Y.X.; investigation, Y.X.; resources, Y.X.; data curation, Y.X.; writing—original draft preparation, Y.X.; writing—review and editing, Y.X.; visualization, Y.X.; supervision, Z.A.L.; project administration, Y.X.; funding acquisition, Z.A.L. All authors have read and agreed to the published version of the manuscript.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Calvo, R.A., Milne, D.N., Hussain, M.S., et al., 2017. Natural language processing in mental health applications using non-clinical texts. Natural Language Engineering. 23(5), 649–685. DOI: https://doi.org/10.1017/S1351324916000383

[2] [2] Mishra, A.R., Rai, A., Nandan, D., et al., 2025. Unveiling Emotions: NLP-Based Mood Classification and Well-Being Tracking for Enhanced Mental Health Awareness. Mathematical Modelling of Engineering Problems. 12(2), 647-656. DOI: https://doi.org/10.18280/mmep.120228

[3] Shatte, A.B.R., Hutchinson, D.M., Teague, S.J., 2019. Machine learning in mental health: A scoping review of methods and applications. Psychological Medicine. 49(9), 1426–1448. DOI: https://doi.org/10.1017/S0033291719000151

[4] De Choudhury, M., De, S., 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. Proceedings of the International AAAI Conference on Web and Social Media. 8(1), 71–80. DOI: https://doi.org/10.1609/icwsm.v8i1.14526

[5] Pennebaker, J.W., Boyd, R.L., Jordan, K., et al., 2015. The development and psychometric properties of LIWC2015. University of Texas at Austin: Austin, TX, USA.

[6] Mohammad, S., Turney, P., 2013. Crowdsourcing a word–emotion association lexicon. Computational Intelligence. 29(3), 436–465. DOI: https://doi.org/10.1111/j.1467-8640.2012.00460.x

[7] Tausczik, Y.R., Pennebaker, J.W., 2010. The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology. 29(1), 24–54. DOI: https://doi.org/10.1177/0261927X09351676

[8] Devlin, J., Chang, M.W., Lee, K., et al., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2–7 June 2019; Minneapolis, Minnesota. pp. 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423

[9] Rogers, A., Kovaleva, O., Rumshisky, A., 2020. A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics. 8, 842–866. DOI: https://doi.org/10.1162/tacl_a_00349

[10] Alsentzer, E., Murphy, J., Boag, W., et al., 2019. Publicly available clinical BERT embeddings. Proceedings of the 2nd Clinical Natural Language Processing Workshop; 7 June 2019; Minneapolis, MN, USA. pp. 72–78. DOI: https://doi.org/10.18653/v1/W19-1909

[11] Tonekaboni, S., Joshi, S., McCradden, M.D., et al., 2019. What clinicians want: Contextualizing explainable machine learning for clinical end use. Proceedings of the 4th Machine Learning for Healthcare Conference; 9–10 August 2019; Ann Arbor, Michigan, USA. 106, 359–380.

[12] Kim, Y., Klinger, R., 2019. Frowning Frodo, wincing Leia, and a seriously great friendship: Learning to classify emotional relationships of fictional characters. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics; 2–7 June 2019; Minneapolis, MN, USA. pp. 647–653.

[13] Li, X., Song, K., Feng, S., et al., 2018. A Co-Attention Neural Network Model for Emotion Cause Analysis with Emotional Context Awareness. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; October 31–November 4, 2018; Brussels, Belgium. pp. 4752–4757.

[14] Strapparava, C., Mihalcea, R., 2008. Learning to identify emotions in text. Proceedings of the 2008 ACM Symposium on Applied Computing; 16–20 March, 2008; Fortaleza, Brazil. pp. 1556–1560.

[15] Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017); 4–9 December 2017; Long Beach, CA, USA. pp. 4765–4774.

[16] Esteva, A., Robicquet, A., Ramsundar, B., et al., 2019. A guide to deep learning in healthcare. Nature Medicine. 25(1), 24–29. DOI: https://doi.org/10.1038/s41591-018-0316-z

[17] Johnson, A.E.W., Pollard, T.J., Shen, L., et al., 2016. MIMIC-III, a freely accessible critical care database. Scientific Data. 3, 160035. DOI: https://doi.org/10.1038/sdata.2016.35

[18] Ekman, P., 1992. An argument for basic emotions. Cognition and Emotion. 6(3–4), 169–200. DOI: https://doi.org/10.1080/02699939208411068

[19] McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochemia Medica. 22(3), 276–282.

[20] Breiman, L., 2001. Random forests. Machine Learn-

ing. 45(1), 5–32. DOI: https://doi.org/10.1023/A:1010933404324

[21] Daniels, Z.A., Metaxas, D.N., 2017. Addressing imbalance in multi-label classification using structured Hellinger forests. Proceedings of the 31st AAAI Conference on Artificial Intelligence; 4–9 February 2017; San Francisco, CA, USA. pp. 1826–1832. DOI: https://doi.org/10.1609/aaai.v31i1.10908

[22] Iavarone, B., 2024. Understanding emotive response to textual stimuli: A multimodal approach. Scuola Normale Superiore: Pisa, Italy. Available from: https://tesidottorato.depositolegale.it/handle/20.500.14242/167628

[23] Makhmudov, F., Kultimuratov, A., Cho, Y.I., 2024. Enhancing multimodal emotion recognition through attention mechanisms in BERT and CNN architectures. Applied Sciences. 14(10), 4199. DOI: https://doi.org/10.3390/app14104199

[24] Acheampong, F.A., Nunoo-Mensah, H., Chen, W., 2021. Transformer models for text-based emotion detection: a review of BERT-based approaches. Artificial Intelligence Review. 54(8), 5789–5829. DOI: https://doi.org/10.1007/s10462-021-09958-2

[25] Kazemeinizadeh, A., 2022. Psychological understanding of textual journals using natural language processing approaches [Master's thesis]. The University of Western Ontario: London, ON, Canada.

[26] Salmerón-Ríos, A., García-Díaz, J.A., Pan, R., et al., 2024. Fine grain emotion analysis in Spanish using linguistic features and transformers. PeerJ Computer Science. 10, e1992. DOI: https://doi.org/10.7717/peerj-cs.1992