JOURNAL OF
EMERGING INVESTIGATORS

Article

Evaluating key factors in emotion detection models for AI-driven personalized bibliotherapy

Pransh Dalal[1], Arpan Dalal[1]
[1] Emerald High School, Dublin, California

SUMMARY

Mental health challenges are becoming a serious problem globally, showcasing the need for more accessible and personalized forms of treatment. Bibliotherapy, the use of books to support one's emotional state, is a promising solution. However, traditional methods often lack personalization in their approach, using static reading lists that do not adapt to the reader. This study evaluates the potential of natural language processing (NLP) models in an emotion-driven bibliotherapy framework. We fine-tuned several models on an emotion-labeled dataset containing English X (formerly Twitter) messages and assessed their accuracy across key emotional categories, including joy, sadness, anger, fear, surprise, and love. We hypothesized that more parameters and layers, an increased number of attention heads and larger hidden sizes, greater feed-forward network (FFN) dimensions and pretraining corpora, and larger model vocabularies would enhance performance as measured by standard classification metrics, including accuracy, precision, recall, and F1-score. Pearson correlation analyses linked these metrics to transformer architectural and pretraining characteristics. Our results partially supported this hypothesis. While the number of attention heads showed a moderate and statistically significant positive correlation with all performance metrics, other factors showed weak or negligible correlations and were not statistically significant. Notably, the number of layers exhibited a strong negative correlation with performance metrics, suggesting that deeper models do not lead to better outcomes in emotion classification. Additionally, corpus size was negative for most emotions but positive for love, indicating that the benefits of larger corpora may depend on the type of emotion being modeled. Overall, these findings will help build stronger models for emotion-driven bibliotherapy applications.

INTRODUCTION

Mental health disorders are one of the leading causes of disability across all age groups, affecting nearly 970 million people worldwide (1). Despite the increasing spread of depression, anxiety, and other mental health disorders, access to mental health care is extremely limited due to factors such as cost, stigma, and shortage of mental health professionals (2). Thus, there has been a massive rise in interest in alternative approaches that can be more accessible. One such method is bibliotherapy, the use of books and reading to support one's psychological well-being (3,4). This approach is traditionally used in both clinical and self-help contexts, leading to improvements in stress management and aiding individuals with depression and anxiety (3). However, many bibliotherapy approaches use static reading lists that are not personalizable to address unique readers (5).

Recently, there have been many advances in natural language processing (NLP) and machine learning. NLP is a subfield of artificial intelligence and machine learning focused on designing algorithms and models that allows computers to analyze, understand, represent and generate human language. NLP models have already been applied to digital mental health tools for sentiment analysis and development of therapeutic solutions. They have become essential tools for detection of mental health disorders, monitoring of risk behaviors, and support for therapeutic interventions (6-9). Applications span from characterizing the language of bipolar disorder and predicting mood episodes to large-scale analyses of social media for depression, suicidal ideation, and other conditions (7-8). Emerging work shows the potential of real time NLP analysis for crisis detection through the deployment of these models (9).

NLP models are able to classify and interpret human emotions from text inputs with impressive accuracy (10). Transformer based architectures use self-attention mechanisms to model contextual relationships between all tokens in a sequence, allowing for a richer understanding of language. These models have improved the accuracy of sentiment and emotion detection compared to earlier approaches, such as support vector machines, recurrent neural networks, and convolutional neural networks (11-14). This improvement is significant because in digital mental health applications, accurate emotion detection is critical, as even small improvements in classification accuracy can allow for a more precise personalization of interventions. For this reason, we chose to focus this study solely on transformer-based models.

Larger models, characterized by a greater number of parameters and layers, have an increased capacity to capture patterns and structures within the text data they are exposed to (26). This allows them to develop a more nuanced understanding of relationships in text, which is crucial for interpreting emotions in text. More attention heads enable the model to focus on different aspects of the input sequence, forming diverse contextual representations, while a larger hidden size and expanded FFN provide a richer space for these representations to develop and transform nonlinearly (27-28). Additionally, a larger and more diverse pretraining

https://doi.org/10.59720/25-129

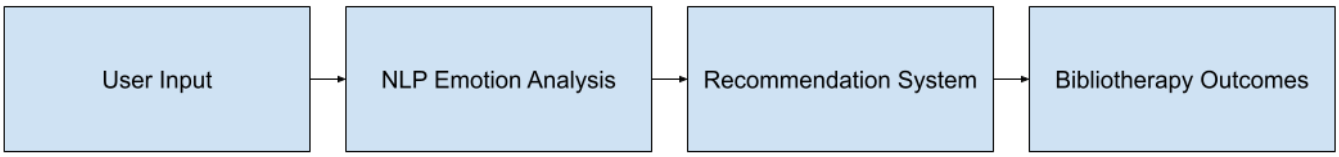

Figure 1: The AI-bibliotherapy framework. In this framework, natural language processing (NLP) can be integrated into a bibliotherapy application. User input is analyzed through NLP emotion classification, which feeds into a recommendation system that suggests books based on emotion.

corpus exposes the model to a wider range of linguistic styles, topics, and emotional expressions, facilitating the learning of more generalizable features. Finally, a larger model vocabulary minimizes out-of-vocabulary (OOV) words, ensuring the model can process and understand a broader set of terms, including specific emotional words, without relying on less precise sub-word tokenization (29). Together, these factors contribute to a deeper and more comprehensive language understanding, which we predict will improve performance in emotion classification tasks.

Using NLP emotion detection with book recommendation algorithms could change how readers discover new books (15). This NLP integration in an emotion-driven bibliotherapy framework would contain NLP as the interpretive layer within a bibliotherapy application (**Figure 1**). In this framework, the system begins with the user input, such as journal entries, that is then analyzed by an NLP model for emotional content. The emotion classification results would then feed into a recommendation algorithm trained on a large corpus of literature tagged with emotional metadata. This framework would allow a system to not only recommend books by topic or genre, but by their emotional resonance with the reader's mood.

In this study, we investigated the role of transformer-based NLP models in enabling emotion-driven bibliotherapy through accurate emotion classification. We fine-tuned fifteen pretrained transformer models on a standardized emotion-labeled dataset and evaluated their performance across six core emotions: joy, sadness, anger, fear, surprise, and love (16-25). We hypothesized that architectural and pretraining factors, including number of attention heads, hidden size, FFN, vocabulary size, and pretraining corpus size, would positively influence classification performance as measured by accuracy, precision, recall, and F1-score. Our results partially supported this hypothesis: while the number of attention heads showed a moderate and statistically significant positive correlation with all performance metrics, most other factors exhibited weak or negligible effects. Notably, the number of layers demonstrated a strong negative correlation with performance, suggesting that deeper models do not necessarily improve emotion classification. These findings provide insight into which transformer design choices most effectively balance accuracy and efficiency for emotion-driven bibliotherapy applications.

RESULTS

To evaluate different models and identify significant factors influencing model effectiveness, we conducted a two-part experimental study. The first component analyzed the key metrics of emotion detection using 15 NLP models, while the second applied statistical analyses to uncover correlations between hypothesized factors and evaluation metrics.

We selected 15 pre-trained NLP models. Each model was fine-tuned on a standardized emotion-labeled dataset, which consisted of raw text sequences corresponding to emotional statements. Model performance was assessed using the standard classification metrics of accuracy, precision, recall, and F1-score. In the context of this experiment, a positive prediction refers to the model predicting that an input text belongs to a specific emotion category (e.g., predicting "joy" for a text excerpt). Because the dataset contains six labeled emotions — joy, sadness, anger, fear, surprise, and love — the model makes one positive prediction per excerpt by selecting the most likely emotion class. Precision refers to the proportion of excerpts for which the model accurately predicted a given emotion. Recall measures the proportion of all excerpts labeled with a given emotion that the model correctly identified. The F1-score represents the harmonic mean of precision and recall and reflects how reliably the model identifies each emotion class.

The models' accuracy, F1-scores, precision, and recall were variable (**Table 1**). RoBERTa-base achieved the highest accuracy at 93.00%, along with an F1-score of 93.03%, precision of 93.08%, and recall of 93.00%. Close contenders included Albert-base-v2 (accuracy 92.95%, F1-score 92.90%, precision 92.93%, recall 92.95%) and Bert-base-cased (accuracy 92.85%, F1-score 92.76%, precision 92.81%, recall 92.85%). Other strong performers were Bert-base-uncased (accuracy 92.65%) and DistilBERT-base-uncased (accuracy 92.60%), demonstrating consistent performance across metrics.

Overall, model accuracy ranged from 29.05% (Google MobileBERT-uncased) to 93.00%, averaging 88.63%. F1-score varied from 13.08% to 93.03%, with a mean of 87.73%. Precision spanned 8.44% to 93.08%, averaging 87.82%, while recall ranged from 29.05% to 93.00%, with an average of 88.63% (**Table 1**). RoBERTa-base and Albert-base-v2 were the top-performing models across classification metrics, with lighter models like DistilBERT and MiniLM also achieving competitive performance with shorter training times.

We computed Pearson correlation coefficients (r) between each factor and each performance metric to quantify linear relationships (**Figure 2**). The architectural factors tested were number of parameters, number of layers, hidden size, FFN dimension, number of attention heads, vocabulary size, and pretraining corpus size. We assessed statistical significance using p-values ($p < 0.05$ considered significant). Pearson correlation analysis was performed across all 15 models, using each model as one data point for each architectural factor. In other words, correlations were computed based on

https://doi.org/10.59720/25-129

Model Name	Accuracy	Precision	Recall	F1-Score
roberta-base	0.93	0.931	0.93	0.93
albert-base-v2	0.93	0.929	0.93	0.929
bert-base-cased	0.928	0.928	0.928	0.928
electra-base-discriminator	0.928	0.928	0.928	0.927
distilbert-base-uncased	0.926	0.927	0.926	0.926
bert-base-uncased	0.926	0.929	0.926	0.926
mpnet-base	0.926	0.928	0.926	0.926
roberta-base-squad2	0.926	0.927	0.926	0.926
xlnet-base-cased	0.922	0.924	0.922	0.923
distilroberta-base	0.919	0.922	0.919	0.92
electra-small-generator	0.898	0.899	0.898	0.894
all-MiniLM-L6-v2	0.891	0.894	0.891	0.894
electra-small-discriminator	0.876	0.88	0.876	0.87
camembert-base	0.824	0.812	0.824	0.807
mobile-bert-uncased	0.29	0.084	0.29	0.131

Table 1:Performance comparison of models across evaluation metrics. Performance of 15 pretrained NLP models on emotion classification. Models are ranked by F1-score, with corresponding values for Accuracy, Precision, and Recall. All models were trained for three epochs, assessing the following emotional categories: joy, sadness, anger, fear, surprise, and love.

the individual metric values for each model, not on averaged values or values aggregated by architecture type. This analysis enabled us to determine which architectural features were meaningfully associated with improvements or declines in model performance.

Pearson correlation coefficients for each factor and each performance metric are summarized below (**Figure 2)**. Weak positive correlations were observed between the number of parameters and the F1-score (r=0.3442), precision (r=0.3344), and recall (r=0.3465), though none of these correlations were statistically significant (p>0.20), suggesting that simply increasing the number of model parameters does not necessarily lead to proportional improvements in performance. In contrast, the number of layers exhibited a strong and statistically significant negative correlation across all performance metrics (F1-score: r=−0.8041, p=0.0003; precision: r=−0.8077, p=0.0003; recall: r=−0.8018, p=0.0003), indicating that adding more layers tends to significantly decrease F1-score, precision, and recall. The pretraining corpus size showed negligible and non-significant correlations with F1-score (r=−0.1664, p=0.6248), precision (r=−0.1939, p=0.5678), and recall (r=−0.1754, p=0.6060), suggesting that the size of the pretraining corpus does not have a linear relationship with model performance. Similarly, vocabulary size presented very weak and non-significant positive correlations with F1-score (r=0.1703, p=0.5440), precision (r=0.1642, p=0.5588), and recall (r=0.1711, p=0.5421), indicating that variations in vocabulary size are not significant predictors of changes in performance metrics. Hidden size also displayed weak and non-significant positive correlations with F1-score (r=0.2312, p=0.4071), precision (r=0.2163, p=0.4389), and recall (r=0.2367, p=0.3957), suggesting that increasing the hidden layer size does not significantly improve performance on this task. Next, the number of attention heads showed a moderate and statistically significant positive correlation with performance metrics (F1-score: r=0.5299, p=0.0422; precision: r=0.5167, p=0.0486; recall: r=0.5344, p=0.0401), indicating that increasing the number of attention heads is meaningfully associated with improvements in F1-score, precision, and recall, highlighting its importance in model architecture for this task. Finally, the FFN dimension size displayed weak and non-significant positive correlations with performance metrics (F1-score: r=0.2367, p=0.3957; precision: r=0.2163, p=0.4389; recall: r=0.2312, p=0.4071), suggesting that increasing the FFN size does not yield statistically reliable improvements in classification performance, indicating that this factor is less critical compared to other architectural parameters.

When analyzed by emotion, similar patterns emerged with some variation in effect sizes (**Figure 3**). For sadness, F1-score was positively correlated with attention heads (r=0.524, p=0.0449) and moderately with parameters (r=0.337, p=0.2195), hidden size (r=0.224, p=0.4213), and feed-forward dimension (r=0.224, p=0.4213), while layers were strongly negatively correlated (r=−0.804, p=0.0003). Corpus size showed a moderate negative association (r=−0.293, p=0.3823). Joy followed a similar pattern, with attention heads (r=0.497, p=0.0593) positively associated and layers (r=−0.811, p=0.0002) negatively associated with performance; parameters (r=0.325, p=0.2376), hidden size (r=0.194, p=0.4896), and feed-forward dimension (r=0.194, p=0.4896) showed weaker contributions, while corpus size showed a weak negative correlation (r=−0.109, p=0.7501). Anger and fear exhibited strong positive correlations with attention heads (r=0.534, p=0.0403; r=0.518, p=0.0478, respectively) and parameters (r=0.330, p=0.220), and strong negative correlations with layers (r=−0.790, p=0.0005; r=−0.807, p=0.0003), with corpus size weakly negative in both cases (r=−0.162, p=0.6337; r=−0.249, p=0.4603). Love showed the highest sensitivity to attention heads (r=0.597, p=0.0189), with moderate positive contributions from parameters (r=0.420, p=0.1189), hidden size (r=0.313, p=0.2563), and FFN dimension (r=0.313, p=0.2563), while layers remained negatively correlated (r=−0.789, p=0.0005) and corpus size showed a weak positive association (r=0.199, p=0.5580). Notably, Love was the only emotion where the correlation with pretraining corpus size showed a positive correlation, suggesting it may require more diverse semantic context than other categories. Surprise demonstrated slightly weaker correlations overall, with positive contributions from attention heads (r=0.542, p=0.0369), hidden size (r=0.370, p=0.1750), and feed-forward dimension (r=0.370, p=0.1750), a moderate negative correlation with layers (r=−0.487, p=0.0655), and a weak negative relationship with corpus size (r=−0.216, p=0.5229). These results indicate that attention heads consistently provide the most substantial positive impact across all emotions, whereas deeper models and larger corpora generally reduce emotion-specific performance. Parameters, hidden size, and FFN dimensions show weaker

https://doi.org/10.59720/25-129

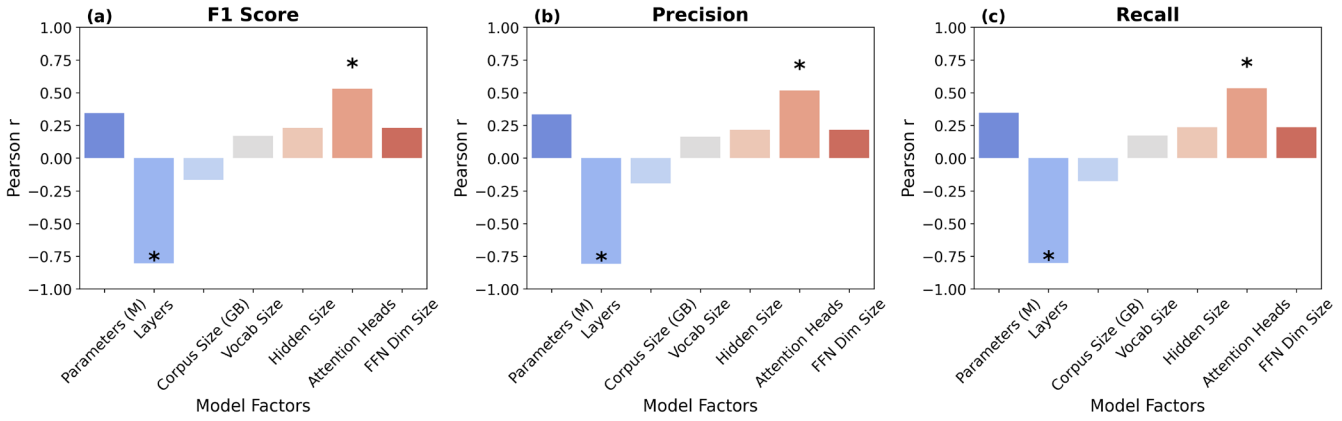

Figure 2: Overall metrics correlations. Pearson correlation coefficients (r) between various architectural and training-related factors (Number of Parameters in Millions, Number of Layers, Pretraining Corpus Size, Vocabulary Size, Hidden Size, and Number of Attention Heads, and Feed-Forward Network Dimension Size) and overall model performance metrics: (A) F1-score, (B) Precision, and (C) Recall). Each model was trained for three epochs. The height of each bar indicates the strength and direction of the linear relationship, offering insights into which factors positively or negatively influence general model effectiveness. * indicates p <0.05, Pearson Product-Moment Correlation.

but generally positive associations, suggesting that careful tuning of these architectural factors can further enhance performance depending on the emotion being detected.

DISCUSSION

This study evaluated the performance of 15 transformer models for emotion classification and analyzed how specific architectural and pretraining factors influence evaluation metrics. Most models achieved strong classification performance, with RoBERTa-base being the top performer across all metrics, followed by ALBERT-base-v2 and BERT-base-cased. Notably, lighter models, such as DistilBERT-base-uncased and MiniLM, also achieved high performance, suggesting that efficient models can still provide high accuracy for emotion classification tasks. These findings are consistent with previous research; RoBERTa's improved pretraining objective and dynamic masking strategy have been shown to substantially enhance contextual representation, while ALBERT's parameter sharing enables efficiency without sacrificing accuracy (17, 21). Similarly, DistilBERT and MiniLM have demonstrated that compressed architectures retain strong downstream performance, supporting the observation that smaller models can effectively capture emotion-laden linguistic cues (18, 23).

Our findings partially support the initial hypothesis. Specifically, the number of attention heads had a moderate and statistically significant positive correlation with evaluation metrics, indicating that an increased number of heads improves the model's ability to capture complex relationships within language. This observation aligns with the original transformer architecture, which emphasized multi-head attention as a mechanism to capture diverse linguistic dependencies (30). Later work has shown that while not all attention heads are equally useful, additional heads can contribute to nuanced contextualization (31,32). In contrast, the number of layers exhibited a strong and statistically significant negative correlation with performance metrics. Prior studies have suggested that deeper models often require larger datasets and careful optimization to realize performance gains, which may explain why layer depth in this study corresponded with overfitting or optimization challenges rather than improved generalization (33,34). Other factors, including total number of parameters, hidden size, feed-forward dimension, pretraining corpus size, and vocabulary size, did not show statistically significant correlations, echoing findings that architectural scaling does not always yield proportional benefits for downstream emotion recognition (35).

The emotion-specific analysis further contextualizes these results. Attention heads consistently provided a strong positive influence across all emotions, with the effect most pronounced for love, anger, and fear. Conversely, the negative association with layer depth was evident across all emotions, suggesting that deeper models may struggle with generalization when tasked with highly subjective and context-sensitive categories. The moderate contributions of parameters, hidden size, and FFN dimension also reflect earlier findings that model complexity interacts differently across emotion categories, with certain classes, such as surprise or love, showing greater sensitivity to representational capacity. Interestingly, corpus size showed emotion-dependent patterns, weakly negative for most categories but weakly positive for love, suggesting that certain emotions may benefit from broader exposure, while others require focused domain adaptation, in line with research on domain sensitivity in pretrained language models (36).

Several hypotheses may explain why NLP models performed differently despite the similarity of their architectures. RoBERTa and ALBERT variants' outperformance of other models can be attributed to differences in pretraining objectives, tokenization strategies, and layer normalization techniques, all of which enhance contextual representation of emotion-laden text (17, 21). The relatively poor performance of MobileBERT may reflect its optimization for mobile efficiency rather than maximum contextual understanding, resulting in a trade-off between speed and nuanced representation (22). Variability among other models may also be influenced by

https://doi.org/10.59720/25-129

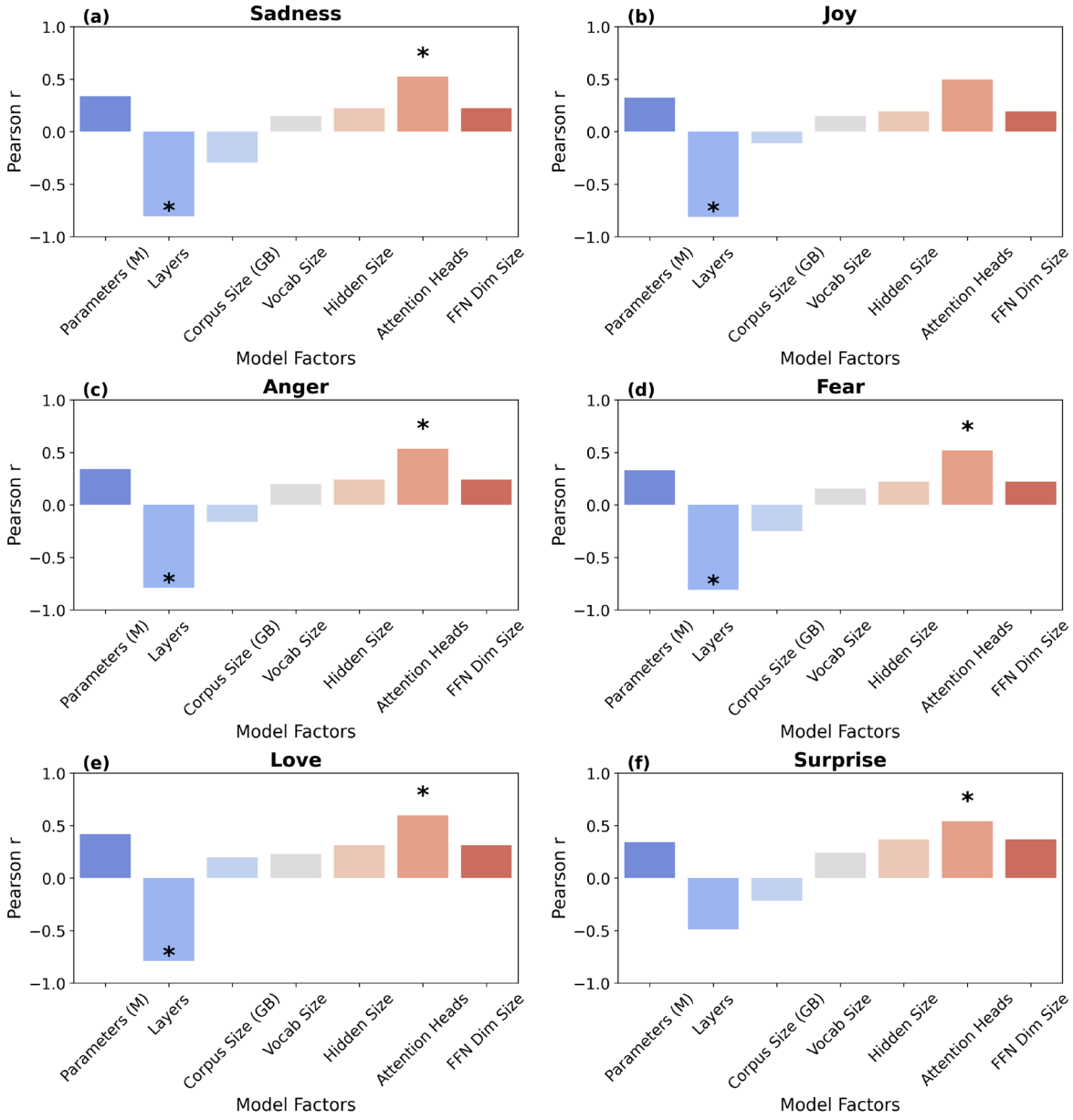

Figure 3: Emotion-specific correlations. Pearson correlation coefficients (r) between architectural and training factors (Number of Parameters in Millions, Number of Layers, Feed-Forward Network Dimension Size, Pretraining Corpus Size, Vocabulary Size, Hidden Size, and Number of Attention Heads) and emotion-specific model F1-scores. Each model was trained for 3 epochs. The chart includes the emotion categories sadness (a), joy (b), anger (c), fear (d), love (e), and surprise (f). * denotes $p < 0.05$, Pearson Product-Moment Correlation.

pretraining corpora and tokenization methods, as WordPiece and Byte-Pair Encoding represent linguistic variation differently, affecting generalization across subtle emotional expressions (37,38).

Several limitations may have influenced these results. This study was conducted on a single dataset, which may restrict generalizability, and the architectural factors tested may not fully capture all design choices that influence evaluation metrics. Additionally, correlations observed may reflect dataset-specific optimization challenges rather than inherent limitations of deeper models, as prior scaling-law research suggests larger models often underperform when trained on limited data (33).

Despite these limitations, the results offer actionable insights for AI-driven personalized bibliotherapy. High-performing models, such as RoBERTa-base and ALBERT-base-v2, were incorporated into a publicly available bibliotherapy application, enabling real-time analysis of user text inputs to detect emotional states and provide targeted reading recommendations (39). By identifying which architectural features, such as attention heads, improve model sensitivity to emotional content, developers can optimize NLP pipelines for both efficiency and accuracy. Moreover, these findings caution against assuming that larger or deeper models will always improve performance, underscoring the importance of designing computationally

https://doi.org/10.59720/25-129

efficient architectures for therapeutic applications.

Future research should replicate these findings across larger and more diverse emotion datasets to assess generalizability. Studies could further explore why layer depth negatively correlated with performance, investigating whether alternative regularization, initialization, or architectural modifications mitigate this effect. Practical experiments incorporating top-performing models into bibliotherapy systems could also evaluate real-world clinical outcomes. Additionally, hybrid or multimodal architectures that integrate complementary data types may enhance emotion classification. Comparative studies that evaluate different pretraining strategies on emotion-rich, user-generated text will be essential for contextualizing these findings within the broader NLP field.

MATERIALS AND METHODS

A fine-tuning pipeline was implemented using the Hugging Face Transformers library (40). A broad range of 15 pretrained transformer models were evaluated on emotion classification (**Table 1**). The dataset used in this task was the "dair-ai/emotion" dataset (41). The dataset contains 20,000 text samples with six emotions: sadness, joy, love, anger, fear, and surprise (**Table 2**). The dataset was pre-divided into three sets - training, validation, and testing following an 80/10/10 split.

The input to each model was the raw text from the dataset, which was processed into numerical format using model-specific tokenizers (Table 3). The tokenizer first breaks the text into subword units called tokens, then converts these into numerical IDs. To ensure a uniform input size for the models, we used a maximum sequence length of 128 tokens, applying truncation to cut off longer texts and padding to add special tokens to shorter texts. The DataCollatorWithPadding class was used to batch the inputs during training, which pads the samples in each batch to the longest sequence length within that batch.

Each model was fine-tuned for three epochs using the Hugging Face Trainer API (42). A learning rate of 2 * 10^-5 was used, a standard value for fine-tuning transformer models to ensure the model's weights were updated appropriately. A per-device training batch size of 16 and per-device evaluation batch size of 16 were chosen to balance memory usage and training efficiency. A weight decay of 0.01 was applied to the AdamW optimizer to act as a form of regulation to penalize large weights and prevent overfitting. To simulate a larger batch size, gradient accumulation step of two was used, which means the gradients were accumulated over two batches before an optimization step was performed. An early stopping callback with a patience of two was implemented. This technique monitors the model's performance on the validation set every 500 steps. If the validation performance does not improve for two consecutive evaluation steps, the training process is stopped to save time and prevent overfitting. Fp16 was enabled and 16-bit floating-point was used to speed up training and reduce memory usage. Checkpoints were saved every 500 steps, and the best model was automatically loaded at the end of training based on validation F1 score.

Model performance was evaluated on the test set using several standard metrics for classification tasks. Accuracy, precision, recall, and F1-score were used as evaluation metrics. In this study, a positive prediction refers to the model selecting one of the six emotion labels - joy, sadness, anger, fear, surprise, or love - as the predicted class for an input text. Precision, therefore, measures how often the model's predicted emotion label matches the true label. Recall measures how many excerpts belonging to a given emotion the model successfully identifies. Because the task is multi-class classification, one positive prediction is made per text input, allowing us to compute per-class precision, recall, and F1-scores in addition to overall weighted averages. Variable names used in analysis (e.g., f1_score, precision, recall, Params_M, Hidden_Size) correspond directly to the metrics and architectural factors defined in the Results section.

These overall metrics were computed and a generated per-class classification report to analyze performance on each individual emotion label. All experiments and statistical analysis were conducted using Python 3.12.3 (43). Key libraries used include Hugging Face Transformers and Datasets, Scikit-learn, Pandas, NumPy, and PyTorch (44-49). The code and scripts used for experiments and analyses in this study are available at the following GitHub repository: https://github.com/PranshDalal/EvaluatingModels/tree/main. This repository includes the model training script and results referenced in the manuscript.

Received: May 15, 2025
Accepted: November 25, 2025
Published: April 27, 2026

REFERENCES

1. "Mental Disorders." World Health Organization, www.who.int/news-room/fact-sheets/detail/mental-disorders. Accessed day month year
2. Modi, Hemangi, et al. Exploring Barriers to Mental Health Care in the U.S. 30 Mar. 2023, https://doi.org/10.15766/rai_a3ewcf9p.
3. Arslan, Gökmen, et al. "Benefits of Positive Psychology-Based Story Reading on Adolescent Mental Health and Well-Being." Child Indicators Research, vol. 15, no. 3, 6 Jan. 2022, pp. 763–782. https://doi.org/10.1007/s12187-021-09891-4.
4. Heath, Melissa Allen, et al. "Bibliotherapy: A Resource to Facilitate Emotional Healing and Growth." School Psychology International, vol. 26, no. 5, Dec. 2005, pp. 563–580. https://doi.org/10.1177/0143034305060792.
5. Correll, Andrew B, et al. "Literary Prescriptions: Applying Bibliotherapy in a Psychotherapeutic Context." Innovations in Clinical Neuroscience, vol. 21, no. 7-9, Sept. 2024, p. 15, pmc.ncbi.nlm.nih.gov/articles/PMC11424070/.
6. Le Glaz, Aziliz, et al. "Machine Learning and Natural Language Processing in Mental Health: Systematic Review." Journal of Medical Internet Research, vol. 23, no. 5, 4 May 2021, p. e15708. https://doi.org/10.2196/15708.
7. Harvey, Daisy, et al. "Natural Language Processing Methods and Bipolar Disorder: Scoping Review." JMIR Mental Health, vol. 9, no. 4, 22 Apr. 2022, p. e35928. https://doi.org/10.2196/35928.
8. Montejo-Ráez, Arturo, et al. "A Survey on Detecting Mental Disorders with Natural Language Processing: Literature Review, Trends and Challenges." Computer Science Review, vol. 53, 1 Aug. 2024, p. 100654. https://doi.org/10.1016/j.cosrev.2024.100654.

https://doi.org/10.59720/25-129

9. "Using NLP to Detect Mental Health Crises." Stanford Institute for Human-Centered AI, 8 Jan. 2024, hai.stanford.edu/news/using-nlp-detect-mental-health-crises.
10. Lin, Shangyue. "Text Emotional Analysis in Natural Language Processing." Applied and Computational Engineering, vol. 36, no. 1, 22 Jan. 2024, pp. 163–172. https://doi.org/10.54254/2755-2721/36/20230440.
11. Yao, Xiuzhen, et al. "Emotion Classification Based on Transformer and CNN for EEG Spatial–Temporal Feature Learning." Brain Sciences, vol. 14, no. 3, 11 Mar. 2024, pp. 268–268, https://doi.org/10.3390/brainsci14030268. Accessed 27 Sept. 2024.
12. Arriaga, Octavio, et al. "Real-Time Convolutional Neural Networks for Emotion and Gender Classification." arXiv, 20 Oct. 2017, https://doi.org/10.48550/arxiv.1710.07557.
13. Majumder, Navonil, et al. "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations." arXiv, 1 Nov. 2018, https://doi.org/10.48550/arxiv.1811.00405.
14. Kumar, Gowda, et al. "Transformers in Sentiment Analysis: A Paradigm Shift in Deep Learning Research." Journal of Information Systems Engineering & Management, vol. 10, no. 5s, 24 Jan. 2025, pp. 262–280, https://doi.org/10.52783/jisem.v10i5s.612.
15. Berbatova, Melania. "Overview on NLP Techniques for Content-Based Recommender Systems for Books." Proceedings of the Student Research Workshop Associated with RANLP 2019, 15 Sept. 2019, pp. 55–61. https://doi.org/10.26615/issn.2603-2821.2019_009.
16. Devlin, Jacob, et al. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." arXiv, 11 Oct. 2018, https://doi.org/10.48550/arXiv.1810.04805.
17. Liu, Yinhan, et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv, 26 July 2019, https://doi.org/10.48550/arXiv.1907.116922.
18. Sanh, Victor, et al. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter." arXiv, 2019, https://doi.org/10.48550/arXiv.1910.01108.
19. Yang, Zhilin, et al. "XLNet: Generalized Autoregressive Pretraining for Language Understanding." arXiv, 19 Jun. 2019, https://doi.org/10.48550/arxiv.1906.08237.
20. Song, Kaitao, et al. "MPNet: Masked and Permuted Pre-Training for Language Understanding." arXiv, 2 Nov. 2020, https://doi.org/10.48550/arxiv.2004.09297.
21. Lan, Zhenzhong, et al. "ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations." arXiv, 8 Feb. 2020, https://doi.org/10.48550/arxiv.1909.11942.
22. Sun, Zhiqing, et al. "MobileBERT: A Compact Task-Agnostic BERT for Resource-Limited Devices." arXiv, 14 Apr. 2020, https://doi.org/10.48550/arxiv.2004.02984.
23. Wang, Wenhui, et al. "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers." arXiv, 5 Apr. 2020, https://doi.org/10.48550/arxiv.2002.10957.
24. Clark, Kevin, et al. "ELECTRA: Pre-Training Text Encoders as Discriminators rather than Generators." arXiv, 23 Mar. 2020, https://doi.org/10.48550/arxiv.2003.10555.
25. Martin, Louis, et al. "CamemBERT: A Tasty French Language Model." arXiv, 21 May 2020, https://doi.org/10.48550/arxiv.1911.03894.
26. Li, Zhuohan, et al. "Train Large, Then Compress: Rethinking Model Size for Efficient Training and Inference of Transformers." arXiv, 26 Feb. 2020, https://doi.org/10.48550/arxiv.2002.11794.
27. Vu, Binh, et al. "A Systematic Approach to Fine-Tuning Transformers for Emotion Detection on the Empathetic Dialogues Benchmark." International Journal of Information Technology, 19 July 2025, https://doi.org/10.1007/s41870-025-02645-3.
28. "T5 (Language Model)." Wikipedia, Wikimedia Foundation, 10 Dec. 2024.
29. Calderón-Fajardo, Víctor, et al. "From Words to Visuals: A Transformer-Based Multi-Modal Framework for Emotion-Driven Tourism Analytics." Information Technology & Tourism, 22 July 2025, https://doi.org/10.1007/s40558-025-00334-2.
30. Vaswani, Ashish, et al. "Attention Is All You Need." arXiv, 12 June 2017, https://doi.org/10.48550/arxiv.1706.03762.
31. Michel, Paul, et al. "Are Sixteen Heads Really Better than One?" arXiv, 4 Nov. 2019, https://doi.org/10.48550/arxiv.1905.10650.
32. Kovaleva, Olga, et al. "Revealing the Dark Secrets of BERT." arXiv, 11 Sept. 2019, https://doi.org/10.48550/arxiv.1908.08593.
33. Kaplan, Jared, et al. "Scaling Laws for Neural Language Models." arXiv, 22 Jan. 2020, https://doi.org/10.48550/arxiv.2001.08361.
34. Raffel, Colin, et al. "Exploring the Limits of Transfer Learning with a Unified Text-To-Text Transformer." Journal of Machine Learning Research, vol. 21, no. 140, 2020, pp. 1–67, https://www.jmlr.org/papers/v21/20-074.html.
35. Wilcox, Ethan Gotlieb, et al. "Bigger Is Not Always Better: The Importance of Human-Scale Language Modeling for Psycholinguistics." Journal of Memory and Language, vol. 144, 23 May 2025, p. 104650, https://doi.org/10.1016/j.jml.2025.104650.
36. Gururangan, Suchin, et al. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." arXiv, 5 May 2020, https://doi.org/10.48550/arxiv.2004.10964.
37. Wu, Yonghui, et al. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." arXiv, 2016, https://doi.org/10.48550/arxiv.1609.08144.
38. Sennrich, Rico, et al. "Neural Machine Translation of Rare Words with Subword Units." arXiv, 2015, https://doi.org/10.48550/arxiv.1508.07909.
39. "Get AI-Powered Book Recommendations Based on Your Mood." MoodReads, 2025, https://www.moodreads.ai/. Accessed 1 Sept. 2025.
40. Wolf, Thomas, et al. "HuggingFace's Transformers: State-of-The-Art Natural Language Processing." arXiv, 11 Feb. 2020, https://doi.org/10.48550/arxiv.1910.03771.
41. "Dair-Ai/Emotion · Datasets at Hugging Face." Hugging Face, 23 Mar. 2023, https://wwww.huggingface.co/datasets/dair-ai/emotion.
42. "Trainer." Hugging Face, https://www.huggingface.co/docs/transformers/en/main_classes/trainer.
43. Python Software Foundation. "3.7.3 Documentation." Python.org, 2019, https://www.docs.python.org/3/.
44. "Transformers." Hugging Face, huggingface.co/docs/transformers/en/index.
45. "Datasets." Hugging Face, https://www.huggingface.co/

https://doi.org/10.59720/25-129

docs/datasets/en/index.
46. Scikit-learn. “Scikit-Learn: Machine Learning in Python.” Scikit-Learn.org, 2024, https://www.scikit-learn.org/stable/.
47. Pandas. “Pandas Documentation — Pandas 1.0.1 Documentation.” Pandas.pydata.org, 2024, https://www.pandas.pydata.org/docs/.
48. NumPy. “NumPy Documentation.” Numpy.org, 2022, https://www.numpy.org/doc/.
49. PyTorch. “PyTorch Documentation — PyTorch 2.7 Documentation.” Pytorch.org, 2024, https://www.docs.pytorch.org/docs/stable/index.html.