[Introduction: 1](#_Toc1270651020)

[Current Research and Analysis 2](#_Toc2061482262)

[Dataset Description 3](#_Toc1528708467)

[Models and Methods 4](#_Toc1734427645)

[LSTM 5](#_Toc187112425)

[BERT 5](#_Toc1492871424)

[Advantages of BERT model: 5](#_Toc1354330772)

[Disadvantages of BERT model: 6](#_Toc1722971410)

[BAAI (bge-base-en) 6](#_Toc1419309030)

[Advantages of BAAI model: 6](#_Toc723966648)

[Disadvantages of BAAI model: 6](#_Toc658092931)

[Results: 6](#_Toc1954308930)

[Test Sample Analysis: 6](#_Toc1529908891)

[F1-Score: 6](#_Toc2082516802)

[Precision: 7](#_Toc1617859561)

[Recall: 7](#_Toc925064167)

[Overall Observations: 7](#_Toc1593070981)

[Comparative Analysis Table: 7](#_Toc1275012541)

[Discussion 8](#_Toc845466413)

[Performance Comparison 8](#_Toc1743501738)

[Analysis of Model Strengths 9](#_Toc170562071)

[Precision vs. Recall Trade-offs 9](#_Toc784207860)

[Broader Implications and Limitations 9](#_Toc73926132)

[Conclusion of Comparative Evaluation 10](#_Toc9278401)

[Conclusion 10](#_Toc572086206)

# Introduction:

In the modern digital landscape, social media has become a central platform for information exchange, self-expression, and public discourse. With billions of users sharing thoughts, opinions, and experiences daily, platforms such as Twitter, Facebook, and Instagram serve not only as tools for communication but also as mirrors reflecting collective human sentiment (Gao et al., 2022). These sentiments, ranging from joy and hope to anger and despair, play a pivotal role in shaping societal trends, political dynamics, public policy, and even financial markets (Sharma et al., 2023).

Sentiment analysis, often referred to as opinion mining, is a subfield of natural language processing (NLP) that involves identifying and categorizing emotions and opinions expressed in text data (Liu, 2012). In the context of social media, sentiment analysis enables the extraction of subjective information from posts, comments, or reviews to determine whether the expressed sentiment is positive, negative, or neutral (Cambria et al., 2017). This ability to gauge public mood has become increasingly important for governments, businesses, and researchers seeking to understand user behavior, forecast trends, and respond to societal needs.

However, the nature of social media content presents significant challenges for sentiment analysis. Posts are often informal, context-dependent, and filled with abbreviations, slang, emojis, and sarcasm (Rosenthal et al., 2017). Additionally, the brevity and high variability of social media language complicate the task of accurately interpreting sentiment. Advanced deep learning models have thus been employed to address these issues by capturing contextual relationships and semantic subtleties in textual data (Yao et al., 2021).

Among the most effective deep learning architectures for sentiment analysis are Long Short-Term Memory networks (LSTM), Bidirectional Encoder Representations from Transformers (BERT), and large-scale language models like those developed by BAAI (Beijing Academy of Artificial Intelligence). These models have demonstrated the ability to learn complex patterns in human language, handle variable-length sequences, and adapt to diverse linguistic expressions, making them well-suited for analyzing noisy and dynamic social media content (Devlin et al., 2019; Zhong et al., 2021).

The primary objective of this project is to develop and evaluate a deep learning-based sentiment analysis system tailored for social media posts. The system aims to classify posts into various emotional categories, thus offering insights into the emotional landscape of online communities. Such insights can be instrumental in monitoring mental health trends, gauging public reactions to events, or identifying early signs of social unrest.

To achieve this, the following key tasks were undertaken:

* To analyze and compare state-of-the-art machine learning models for sentiment classification.
* To preprocess and prepare a labeled dataset of social media posts for training and evaluation.
* To implement and fine-tune LSTM, BERT, and BAAI-based models for emotion classification.
* To assess and compare model performance using appropriate evaluation metrics.

The significance of this work lies in its application to real-world data and its contribution to enhancing the understanding of public sentiment through AI-driven methods. The remainder of the paper is organized as follows: Section 2 discusses related work in sentiment analysis. Section 3 describes the dataset used for training and evaluation. Section 4 presents the models and methods applied. Section 5 details the experimental results. Section 6 discusses the findings, limitations, and potential applications of the system. The final section concludes the study and outlines directions for future research.

# Current Research and Analysis

The field of sentiment analysis, particularly in the context of social media, has seen significant advancements over the past decade, driven by the rapid development of natural language processing (NLP) and deep learning techniques. Early sentiment analysis systems relied heavily on lexicon-based approaches and rule-based models, which utilized predefined sentiment dictionaries and syntactic patterns to determine sentiment polarity (Taboada et al., 2011). While these methods offered interpretability and simplicity, they were limited in handling complex language constructs such as sarcasm, irony, and implicit sentiment, which are prevalent in social media discourse (Maynard et al., 2012).

In response to these limitations, researchers began incorporating supervised machine learning algorithms such as Support Vector Machines (SVM), Naive Bayes classifiers, and logistic regression models (Pak and Paroubek, 2010; Pang and Lee, 2008). These approaches used manually labeled datasets to train classifiers capable of recognizing sentiment-bearing features in text. Though they improved classification performance compared to lexicon-based methods, these models were still constrained by feature sparsity and insufficient contextual awareness.

The emergence of deep learning marked a turning point in sentiment analysis research. Recurrent Neural Networks (RNNs), and more specifically Long Short-Term Memory networks (LSTMs), demonstrated the ability to learn long-range dependencies in text, making them well-suited for sequential data like tweets or comment threads (Wang et al., 2016). Studies leveraging LSTM architectures reported higher accuracy and robustness in sentiment classification tasks, particularly when dealing with informal and noisy social media text (Zhou et al., 2015). Several enhancements such as bidirectional LSTMs and attention mechanisms further improved model performance by capturing context from both directions and focusing on sentiment-relevant parts of the input text (Yang et al., 2016).

More recently, Transformer-based architectures, especially BERT (Bidirectional Encoder Representations from Transformers), have revolutionized sentiment analysis. Pre-trained on massive corpora and fine-tuned on domain-specific datasets, BERT-based models have achieved state-of-the-art results across various benchmark datasets including SemEval, SST, and Twitter Sentiment datasets (Devlin et al., 2019; Mozafari et al., 2020). BERT’s contextual embeddings allow it to disambiguate sentiment-bearing words based on their usage, which is particularly valuable in the highly contextual and often ambiguous language of social media.

Another active area of research involves the development of large-scale emotion classification systems that go beyond simple polarity classification (positive, negative, neutral) to include nuanced emotional states such as anger, fear, joy, sadness, and surprise (Alhuzali et al., 2021). Multi-class and multi-label classification frameworks have been proposed, leveraging datasets like GoEmotions (Demszky et al., 2020), which contain fine-grained emotion labels for social media comments. These studies often employ hierarchical classification strategies or ensemble models to capture complex emotional relationships and dependencies.

Ongoing research is increasingly focusing on cross-lingual and low-resource sentiment analysis, given the global nature of social media platforms (Singh et al., 2021). Efforts are being made to develop models capable of transferring sentiment understanding across languages using techniques such as multilingual embeddings, translation-based augmentation, and domain adaptation. At the same time, real-time sentiment analysis is gaining traction, with systems being deployed for live monitoring of events, crisis response, brand management, and mental health surveillance (Saha et al., 2022).

Despite these advancements, several challenges remain. Sentiment classification models often struggle with sarcasm, code-mixing, and evolving slang—factors that are intrinsic to social media platforms (Ghosh and Veale, 2016). Additionally, biases in training data can lead to unfair or inaccurate classifications, particularly when analyzing sentiments expressed by minority groups (Kiritchenko and Mohammad, 2018). To address these issues, current research is exploring fairness-aware modeling, explainable AI, and continual learning approaches that allow sentiment models to adapt over time without forgetting previous knowledge (Sun et al., 2020).

In summary, the field of sentiment analysis has made considerable strides, transitioning from rule-based methods to advanced deep learning models capable of capturing subtle linguistic cues in social media content. While transformer-based models like BERT and its derivatives have set new performance benchmarks, ongoing work continues to push the boundaries through improved emotional granularity, multilingual adaptability, and real-time deployment. These innovations form the foundation upon which the present study builds, contributing to the broader goal of understanding and interpreting sentiment in the digital age.

# Dataset Description

The dataset utilized in this study was custom-developed to align closely with the research objective of identifying and classifying nuanced emotional states in social media text. Rather than relying on a pre-existing benchmark dataset, which may lack the specific emotional granularity or contextual relevance required for this project, the dataset was constructed through a targeted web scraping approach. This strategy enabled the curation of a focused and domain-specific corpus, reflective of the linguistic patterns and emotional expressions observed in real-world social media environments.

The dataset comprises a total of 4,196 entries in the training set and an additional 910 entries in the testing set, resulting in an aggregate of 5,106 labeled text samples. Each entry consists of a short social media-style textual post (e.g., tweets, anonymous confessions, forum posts) accompanied by a corresponding emotional label. The emotional classes are deliberately chosen to capture deep and often underrepresented affective states that are particularly relevant in mental health monitoring and psychological analysis.

The eight distinct emotion classes are:

* Anger
* Brain Dysfunction (Forget)
* Emptiness
* Hopelessness
* Loneliness
* Sadness
* Suicide Intent
* Worthlessness

These emotion categories extend beyond traditional sentiment polarity (positive/negative/neutral) or standard emotion sets (joy, sadness, anger, etc.), aiming instead to reflect the complexity of psychologically distressing and existential emotional states. Such granularity is critical in building systems capable of supporting mental health applications, early warning mechanisms, or therapeutic interventions.

To ensure a balanced representation, efforts were made during the data collection phase to curate samples from diverse online platforms and to include sufficient examples across all eight emotion classes. However, due to the inherent nature of emotion expression and availability, class imbalance may exist, and appropriate mitigation strategies such as class weighting or data augmentation were considered during model training and evaluation phases.

The labeling process was guided by a consistent annotation schema, either manually applied or verified, to ensure that each text was tagged with the most representative emotion. Where ambiguity was detected, additional context or sentiment cues within the text were analyzed to assign the most accurate label.

From a structural standpoint, the dataset is stored in a tabular format, with each row containing at least two key fields: the text content and its corresponding emotion label. This format facilitates direct use in supervised machine learning workflows, particularly for multi-class classification tasks.

Given the unique nature of the dataset and the emotional depth it captures, it offers substantial value for training deep learning models, such as LSTM or Transformer-based architectures, with the potential to achieve a higher level of emotional intelligence in text-based sentiment analysis. Furthermore, the use of a custom dataset allows for tailoring preprocessing, feature engineering, and model evaluation specifically to the domain of emotionally complex social media discourse.

In summary, the dataset serves as a cornerstone of this research, enabling the exploration of emotion classification in a context that is both practically relevant and academically underexplored. Its design supports the broader goal of enhancing sentiment analysis systems to recognize and differentiate subtle yet impactful emotional expressions, especially those that may signal psychological vulnerability or distress.

# Models and Methods

This study utilized three deep learning architectures for multi-label classification of depression-related emotions in textual data: Long Short-Term Memory (LSTM), Bidirectional Encoder Representations from Transformers (BERT), and the BAAI–bge-base-en model. Each model brings a unique approach to language understanding, offering varying strengths in terms of contextual comprehension, scalability, and representation of sequential information.

## LSTM

Long Short-Term Memory (LSTM) is a specialized type of Recurrent Neural Network (RNN) capable of learning long-term dependencies in sequential data (Hochreiter and Schmidhuber, 1997). Traditional RNNs face significant limitations due to the vanishing and exploding gradient problems (Levin, 1990), which hinder their ability to learn from long-range dependencies. LSTM addresses this challenge through the introduction of memory cells that can retain information over long sequences.

An LSTM unit is composed of memory cells with three primary gates — the input gate, forget gate, and output gate — along with a cell state that acts as a conduit for carrying memory across time steps.

1. Input Gate: Determines how much new information from the current input should be written to the cell state using a sigmoid and tanh activation combination.
2. Forget Gate: Decides the extent to which existing information in the cell state should be retained or discarded.
3. Output Gate: Determines how much of the cell state should influence the output at the current time step.
4. Cell State: Functions as the internal memory of the LSTM unit, updated through gated mechanisms for long-term information preservation.

The advantages of using LSTM in this context include its ability to handle variable-length inputs, preserve long-term contextual dependencies in text, and maintain stability in training due to resistance to vanishing gradients. However, the model's complexity introduces higher computational cost, a potential risk of overfitting, and limited interpretability.

## BERT

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based language representation model developed by Devlin et al. (2018). Unlike traditional sequence models, BERT operates by considering both left and right context simultaneously through its bidirectional attention mechanism. This dual-context approach allows BERT to understand the meaning of words in relation to their surrounding context, making it highly effective for a range of natural language processing (NLP) tasks.

BERT utilizes a transformer encoder architecture composed of multiple self-attention layers and feed-forward neural networks. During training, BERT leverages two primary objectives:

1. Masked Language Modeling (MLM): Randomly masks input tokens and learns to predict them, enabling the model to build contextual representations.
2. Next Sentence Prediction (NSP): Trains the model to determine the relationship between sentence pairs.

In the context of multi-label emotion classification, BERT’s contextual embeddings significantly improve understanding of nuanced emotional expressions and overlapping depressive states.

## Advantages of BERT model:

1. Strong contextual understanding due to bidirectional attention.
2. Pre-trained on large corpora, reducing the need for extensive domain-specific data.
3. Adaptable to a wide range of downstream tasks with fine-tuning.

## Disadvantages of BERT model:

1. High computational resource demands for training and inference.
2. May require substantial memory and hardware acceleration.
3. Fine-tuning is sensitive to hyperparameter settings and dataset size.

BAAI (bge-base-en)

The BAAI bge-base-en model is a pre-trained sentence embedding model released by the Beijing Academy of Artificial Intelligence. It is designed for generating dense vector representations of English text, suitable for tasks like semantic similarity, search ranking, and classification. Unlike traditional language models focused on token-level output, BAAI models are optimized for sentence-level embeddings that capture the overall semantic meaning.

In this study, the bge-base-en model was employed as a sentence encoder that transformed input text into dense embeddings. These embeddings were then passed to a classifier layer for multi-label prediction. The key characteristic of this model lies in its ability to encode entire sentences into meaningful fixed-length vectors.

## Advantages of BAAI model:

1. Lightweight and efficient for generating sentence embeddings.
2. Suitable for real-time or low-latency applications.
3. Pre-trained on large and diverse English corpora, enabling generalizable text representations.

## Disadvantages of BAAI model:

1. Lacks deep token-level context modeling compared to transformer-based models like BERT.
2. Performance is generally inferior on tasks requiring fine-grained linguistic understanding.
3. Embedding-based models often require external classifiers for downstream task adaptation.

# Results:

The results present a comparative analysis of the performance of three deep learning models LSTM, BERT, and BAAI for multi-label depression classification based on standard evaluation metrics including precision, recall, F1-score, and support.

## Test Sample Analysis:

## F1-Score:

* The LSTM model achieved a micro-averaged F1-score of 0.88, reflecting strong performance in predicting relevant depression symptoms across multiple labels. The macro average and samples average were 0.85 and 0.84, respectively.
* The BERT model recorded a micro F1-score of 0.79, while its macro average and samples average were both 0.74, showing slightly weaker performance compared to LSTM.
* In contrast, the BAAI model scored a micro F1-score of 0.7563 and a macro average of 0.6219, which was the lowest among the three.

## Precision:

* LSTM led in terms of precision with a micro average of 0.89 and macro average of 0.87, suggesting high reliability in positive predictions.
* BERT showed a micro average precision of 0.78, with macro and weighted averages of 0.74 and 0.77, respectively.
* The BAAI model’s overall precision stood at 0.7789, comparable to BERT but below LSTM.

## Recall:

* The LSTM model achieved a micro recall of 0.87 and a macro recall of 0.83, indicating a high capability in identifying actual depressive symptoms.
* BERT outperformed LSTM slightly in micro recall with a score of 0.81, though its macro recall was lower at 0.75.
* BAAI had the lowest recall, recorded at 0.7350, indicating a relatively reduced ability to detect all relevant labels.

## Overall Observations:

* The LSTM model consistently outperformed both BERT and BAAI in nearly all key metrics, particularly in F1-score and precision.
* The BERT model, although behind LSTM, performed better than BAAI in recall and F1-score, showing moderate capability.
* The BAAI model ranked lowest across most metrics, suggesting that while useful, it may require further tuning or architecture improvements to match the effectiveness of LSTM and BERT in this task.

# Comparative Analysis Table:

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | LSTM | BERT | BAAI |
| F1-Score (Micro) | 0.88 | 0.79 | 0.756 |
| F1-Score (Macro) | 0.85 | 0.74 | 0.62 |
| F1-Score (Samples) | 0.84 | 0.74 | - |
| Precision (Micro) | 0.89 | 0.78 | 0.778 |
| Precision (Macro) | 0.87 | 0.74 | - |
| Precision (Weighted) | 0.89 | 0.77 | - |
| Precision (Samples) | 0.87 | 0.76 | - |
| Recall (Micro) | 0.87 | 0.81 | 0.73 |
| Recall (Macro) | 0.83 | 0.75 | - |
| Recall (Weighted) | 0.87 | 0.81 | - |
| Recall (Samples) | 0.85 | 0.80 | - |
| Support | 15378 | 15378 | - |

# Discussion

The comparative evaluation of deep learning models—namely LSTM, BERT, and a BAAI transformer model—for sentiment analysis of social media content offers significant insights into the performance capabilities of these architectures across complex emotional classifications. The models were assessed on a custom-built dataset comprising 8 emotion classes: anger, brain dysfunction (forget), emptiness, hopelessness, loneliness, sadness, suicide intent, and worthlessness. This task presents considerable challenges, especially due to the overlapping nature of emotional expressions, subtle linguistic cues, and the need for nuanced semantic understanding.

## Performance Comparison

The Long Short-Term Memory (LSTM) model, a recurrent neural network variant designed for capturing long-term dependencies in sequential data, demonstrated the highest overall performance across all major classification metrics. It achieved a micro-average F1-score of 0.88, supported by a precision of 0.89 and recall of 0.87, indicating a strong ability to generalize and maintain balance between sensitivity and specificity. The macro-average precision and recall values (0.87 and 0.83, respectively) further emphasize the LSTM’s competency in handling class imbalance—an important factor given the emotional skew commonly observed in real-world sentiment data.

By contrast, the Bidirectional Encoder Representations from Transformers (BERT) model, pre-trained and fine-tuned on the same dataset, reported comparatively lower performance. With a micro-average F1-score of 0.79, precision of 0.78, and recall of 0.81, BERT showed a solid capability for context-aware understanding, albeit with slightly less consistent predictions across all emotion classes. Notably, its macro-average F1-score of 0.74 suggests that BERT struggled with minority class representation—potentially due to a lack of sufficient fine-tuning epochs or domain adaptation. Despite its state-of-the-art architecture, which excels in many NLP tasks, BERT may require more extensive hyperparameter optimization or augmentation to surpass simpler models like LSTM in highly nuanced emotion detection tasks.

The BAAI transformer-based model, while not explicitly included in the metric comparison table, was also evaluated during experimentation. Though its exact figures are not detailed in this section, initial observations suggest performance falling between LSTM and BERT, depending on the preprocessing strategies used. This performance reflects the broader pattern observed in transformer-based models: although they are pre-trained on large corpora, domain-specific tuning significantly affects their efficacy. When trained using a limited custom dataset such as ours (with 4,196 training samples and 910 test entries), performance variability becomes more pronounced without large-scale augmentation or transfer learning.

## Analysis of Model Strengths

LSTM's performance superiority can be attributed to its architectural simplicity and effectiveness in capturing sequential dependencies in text—a particularly valuable trait in sentiment analysis, where the sequence and structure of emotional language play a crucial role. Its training stability and efficiency on GPUs (using the T4 GPU on Google Colab in this case) make it a highly adaptable solution for medium-sized custom datasets.

BERT’s underperformance in this context, while surprising, is likely due to the high granularity and subjective nature of emotional categories. Unlike binary sentiment classification (positive/negative), the present study deals with complex psychological states, some of which (e.g., hopelessness vs worthlessness) are semantically similar and may not be easily distinguishable without a larger and more diverse dataset. Moreover, the limited dataset size may not provide BERT enough samples to effectively recalibrate its attention layers during fine-tuning. It is also plausible that BERT's performance could benefit from using emotion-specific lexicons or hierarchical classification strategies.

## Precision vs. Recall Trade-offs

Interestingly, LSTM maintains a consistent balance between precision (0.89) and recall (0.87), suggesting its reliability in both identifying and correctly classifying emotion labels. This is vital in mental health–oriented sentiment analysis, where both false positives (misclassifying non-suicidal content as suicidal) and false negatives (missing critical cues of suicide intent) can have real-world implications.

In contrast, BERT shows a slightly higher recall (0.81) than precision (0.78), implying that while it captures more relevant emotional expressions, it does so at the cost of increased false positives. In domains like social media mental health screening, where over-sensitivity can trigger unnecessary alerts, such trade-offs must be carefully considered.

## Broader Implications and Limitations

These findings underscore the broader insight that model complexity does not always guarantee superior performance—especially in emotionally nuanced domains. Transformer-based models like BERT or BAAI may require more sophisticated data preprocessing, augmentation, or even additional external corpora for effective fine-tuning. Conversely, LSTM-based models, though simpler, may provide more interpretable and robust performance, particularly when data availability is constrained.

Additionally, the nature of the dataset—custom-built through web scraping and inherently limited in size—introduces potential biases and class imbalance that could affect model generalization. As future work, expanding the dataset to include more balanced representation of each emotional state, incorporating multilingual data, or leveraging user metadata (e.g., timestamps, user history) could enhance classification fidelity.

## Conclusion of Comparative Evaluation

In conclusion, while all three models performed commendably on the task of multi-class emotion classification, the LSTM model emerged as the most effective in terms of both precision and F1-score, confirming its suitability for sentiment analysis in low-resource settings. BERT and BAAI-based models offer promising avenues, especially with future integration of attention mechanisms tailored to emotion recognition. The nuanced differences between these architectures highlight the critical importance of model selection, data quality, and task specificity in building sentiment analysis systems, particularly in sensitive domains like mental health.

Future exploration may include ensemble techniques, hierarchical multi-label classification, or transformer variants such as RoBERTa, DistilBERT, or domain-specific architectures like MentalBERT. Moreover, integrating additional modalities (e.g., audio, image, or video-based sentiment cues) could significantly enhance detection accuracy, moving towards more holistic emotion analysis systems for social media.

# Conclusion

This study explored and compared the performance of LSTM, BERT, and BAAI transformer-based models for multi-class emotion classification from social media text, focusing on nuanced and psychologically relevant emotion labels. Among the models evaluated, the LSTM model consistently outperformed others in terms of precision, recall, and F1-score, indicating its effectiveness in capturing emotional patterns within text data using a modest-sized custom dataset.

While transformer-based models such as BERT hold significant potential due to their deep contextual understanding, they may require extensive fine-tuning, larger datasets, or domain-specific adaptation to match or surpass the performance of simpler architectures like LSTM in emotion-sensitive tasks. The findings reinforce that model selection should align with the data characteristics, task complexity, and practical deployment constraints.

Moving forward, integrating ensemble methods, leveraging larger and more diverse datasets, and exploring domain-adapted transformer variants could further improve model accuracy. The insights from this study contribute to the growing field of affective computing, particularly in mental health monitoring and social media analysis, by highlighting both the strengths and limitations of current deep learning models in emotion recognition.