Investigating the Effectiveness of Data Mining Techniques in Identifying Early Signs of Mental Health Issues through Multimodal Social Media Data

By,

Debolina Das

Student ID: 2873143

Supervisor: Dr. Mubashir Ali

A thesis submitted to the University of Birmingham for the degree of MSc in Data Science School of Computer Science University of Birmingham, Birmingham, UK

September 2025

Declaration

I hereby certify that the work presented in this project is my own, except where otherwise indicated. Generative AI tools were used transparently and responsibly to support my work. Specifically:

Code development was assisted by Claude Sonnet 3.7.

Debugging and optimisation were supported by ChatGPT 4.0.

Report structuring and editing for clarity, grammar, and formatting were supported by ChatGPT 4.0.

All outputs from these tools were reviewed, tested, and edited by me to ensure accuracy and to reflect my own understanding. The project's direction, structure, and contributions remain my own. A record of Generative AI usage, including example prompts and my critical reflections, is provided in the Appendix.

Acknowledgements

I would take this opportunity to express my sincere gratitude to my supervisor, Dr. Mubashir Ali, for his guidance, constructive feedback, and support throughout this project. His incessant encouragement and valuable technical support have been of immense help in realizing this project. His guidance gave the environment to enhance my knowledge, skills and to reach the pinnacle with sheer determination, dedication and hard work. I am grateful to the academic and support staff of the School of Computer Science, University of Birmingham. I extend my thanks to my friends and family for their unwavering believe and support.

Abstract

This study investigates the efficacy of multimodal data mining approaches in detecting early indicators of mental health issues through social media posts. Focusing on the three modalities of data (textual, visual, and emoji) extracted from Reddit posts, advanced deep learning models are implemented that incorporate BiLSTM, LSTM, attention mechanisms, and fusion layers. An important and key contribution is the integration of the mismatch features that capture emotional incongruence across all the modalities. Experimental results indicate that modeling these inconsistencies significantly enhances recall and PR-AUC, establishing the model's potential for practical early intervention applications.

Contents

List of Figures:

List of Tables

Chapter 1

Introduction

As per the latest statistics of the World Health Organization (WHO), 1 in 8 people approximately live with a mental disorder. Among these the most common are depression and anxiety, where estimated 5% of adults are suffering from depression [1]. New data from the WHO Regional Office for Europe shows a steep rise in problematic social media use among young generation, with rates increasing from 7% in 2018 to 11% in 2022. Nonetheless, 12% of adolescents are at risk of problematic gaming, raises urgent concerns about the impact of digital technology on the mental health and well-being of young people [2]. The increasing involvement of social media into our daily lives have led to significant research in this field and its impact on mental health. Social media being such a prevalent part of today's life people often tend to express their emotions or feeling through social media posts. [3]

This increasing risk of mental health disorders like depression and anxiety, calls for immediate need in developing tools for early detection and intervention. Social media platforms, such as Reddit, provide a rich source of user-generated data which can help in reflecting emotional and behavioural patterns.[4] Data mining techniques helps to discover patterns in vast and unstructured data sources which have been used to analyze linguistic patterns, detect emotional cues, and classify psychological conditions based on user-generated content. While traditional models primarily focus is on textual content, the integration of visual and symbolic data—such as emojis and images—remains underexplored, although it is used widely in the online communication.

As noticed in today's trend in a social media post, user often combines text, emojis, and images to convey emotions more effectively which opens the way for Multimodal deep learning model which will include different modalities of data sources to improve model robustness and interpretability. Recent studies show that multimodal fusion can improve emotion detection and sentiment classification.

This research helps in identifying that a mismatch between different modalities—for example, cheerful images combined with negatively worded text—can be an implicit signal of underlying mental health issues. Traditional unimodal approaches may overlook such nuanced

contradictions. By modelling both the content and the congruence between modalities, it is possible to uncover patterns that more closely align with real-world psychological distress.

1.1 Aim:

There are 2 questions which this research aims to answer:

1. Can multimodal features (images, text, and emojis) from mental health-related Reddit posts improve the accuracy of early mental health issue detection compared to unimodal features?
2. Can a mismatch between the text, emojis, and images in social media posts help us detect early signs of mental health issues?

These questions are meant to do two things: first, see how well different types of input work, and second, see if differences between them may be used to make inferences about mental health.

.

1.2 Objective:

This study includes Reddit posts that contain a combination of text, emojis, and linked image URLs. Unimodal, multimodal, and multimodal with mismatch features models are developed and compared on their ability to predict early signs of mental health issues. Label refinement strategy to address potential noise in the ground-truth annotations is applied to the models which evaluates all models using metrics that prioritize sensitivity (recall) and robustness under class imbalance (PR-AUC).

1.3 Contribution:

- Multimodal deep learning pipeline development combining text (via LSTM and BiLSTM), image (via CLIP), and emoji features.
- Use mismatch features to measure emotional incongruence across different modalities.
- Impact of label refinement strategies on the model performance
- Evaluation of model performance : 5-fold cross-validation, with a focus on the PR-AUC and recall metrics.

Chapter 2

Literature Review

Early Detection of Mental Health Issues Using Social Media Posts, Saeed, Q.B., and Ahmed, I. (2025)

This study investigates the effectiveness of a deep learning framework in detecting early signs of mental health issues via Reddit posts. Traditional machine learning algorithms, such as support vector machines and random forests, rely heavily on manually selected characteristics and do not accurately reflect the quality and context of language [9]. This study presents a multi-modal framework that incorporates BiLSTM networks for the analysis of linguistic and temporal features, which is augmented by a cross-modal attention mechanism [4]. The key advantage cross-modal attention mechanism is that it enables understanding of the context between features like language and post timing.

This study provides strong performance metrics (accuracy: 74.55%, F1: 0.7376), which show the importance of multi-modal learning for mental health detection and provides a baseline for further improvements by using more advanced attention mechanisms and other data modalities [4]. However, the study primarily focuses on text and temporal data and lacks consideration for visual or symbolic modalities such as images or emojis.

On the Complementarity of Images and Text for the Expression of Emotions in Social Media, Khlyzova, A., Silberer, C., and Klinger, R. (2022)

This paper shows the relation between text and images in expressing emotions in multimodal Reddit posts. Here, a dataset is developed that labels the emotional relationship between image and text as complementary, illustrative, or opposing. It used a pretrained RoBERTa model [10] and a residual neural network [11] to create classification models for the prediction of each of the three classes and found that multimodal analysis improved the understanding of emotional expressions.

The model is then evaluated on predicting emotions, text–image relations, and emotion stimuli using unimodal and multimodal models, based on the F1 measure. The study also highlights that certain emotions, such as anger and sadness, benefit significantly from joint modelling of text and visual data.

Emotion Fusion for Mental Illness Detection from Social Media: A Survey, Zhang, T., Yang, K., Ji, S. and Ananiadou, S. (2023)

This paper is a comprehensive survey that involves how different emotion fusion techniques are used to detect mental illnesses via social media. It reviews various fusion strategies likely early fusion, late fusion, and hybrid approaches, that combine text, visual, audio, or other modality features. It focuses on strategies for fusing textual emotion information to support mental illness detection. Fusion strategies based on both conventional machine learning algorithms [12] and emerging deep learning approaches [13] are studied. The effects and challenges of different fusion strategies are studied and highlight the promising new research directions. The study stresses the importance of effective alignment techniques, highlights dataset limitations, and points out challenges with interpretability.[6]

It provides valuable insights into the following 3 different aspects:

- Effects of feature engineering-based methods and deep learning-based methods;
- Effects of emotion fusion;
- Effects of different fusion strategies.

DepressionEmo: A Novel Dataset for Multilabel Classification of Depression Emotions, Rahman, A.B.S. et al. (2024)

This work presents a dataset called DepressionEmo, which is created for multilabel emotion categorisation related to depression symptoms. The dataset includes Reddit posts, which are annotated for eight emotion categories like helplessness, guilt, worthlessness, etc., using a hybrid of human annotation, zero-shot models, and LLM-generated labels.

This dataset was created through a majority vote over inputs by zero-shot classifications from pre-trained models and validating the quality by annotators and ChatGPT, which shows an acceptable level of interrater reliability between annotators. This correlation between emotions distribution over time, and linguistic analysis is conducted on DepressionEmo. Several text classification methods are classified into two groups: machine learning methods, such as SVM, XGBoost, and Light GBM; and deep learning methods, such as BERT, GAN-BERT, and BART. The pretrained BART model, bart-base, has the highest F1- Macro of 0.76.[7]

CLIP: Learning Transferable Visual Models From Natural Language Supervision, Radford, A. et al. (2021)

This is study by OpenAI introduces CLIP (Contrastive Language–Image Pretraining), a model trained on over 400 million image–text pairs. CLIP learns a shared embedding space for images and their corresponding textual descriptions using contrastive learning. It shows strong zero-shot transfer capabilities, making it suitable for emotion and sentiment alignment tasks across modalities.

It considers two different architectures for the image encoder. Firstly, it uses ResNet-50 [14] as the base architecture for the image, which is then modified using rect-2 blur pooling. The global average pooling layer is replaced with an attention pooling mechanism. The attention pooling is implemented as a single layer of "transformer-style" multi-head QKV attention where the query is conditioned on the global average-pooled representation of the image [8]. The second architecture includes a Vision Transformer (ViT) [15], which closely follows the implementation with only minor modifications of adding a normalization layer to the combined patch and positioning the embeddings before the transformer, and uses a different initialization scheme.[8]

Key Differences and Contributions of This Work

The key areas where the proposed model is different from the existing literature are:

- Integrating four distinct modalities: text, emojis, images (via CLIP), and temporal metadata.
- Adding elements like valence gaps and sign flips that can help forecast mental health problems.
- A cross-attention fusion mechanism is put that aligns the modalities contextually.
- Applying label refinement and safe label filtering, improving robustness in scenarios with annotation noise.
- Using 5-fold stratified cross-validation for evaluation, with an emphasis on recall and PR-AUC, ensuring reliability in detecting high-risk individuals

Chapter 3

Methodology

It includes steps such as obtaining and cleaning data, enhancing and crafting features, training methods, including the adoption of appropriate pre-processing steps, such as feature generation for mismatched information and model architecture development, planning training methods, and defining evaluation strategies. The overall pipeline of the methodology is shown in the image below:

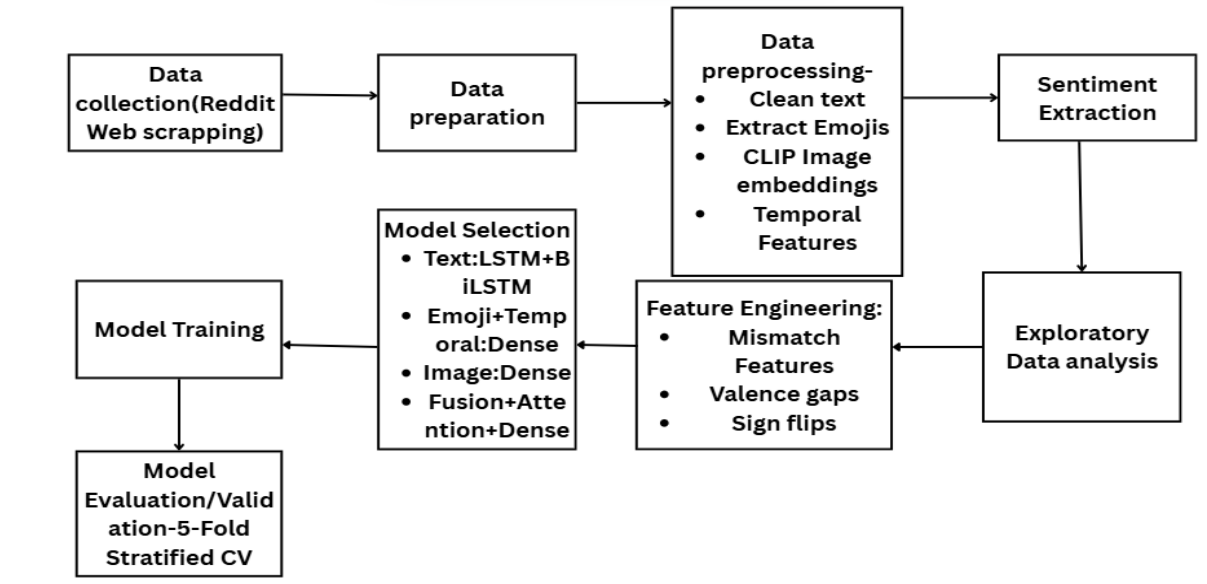

Fig. 3.1 Model Pipeline

3.1 Data Collection

The initial step involves data collection which is done from Reddit using the open-source python package called Python Reddit API Wrapper (PRAW) which is simple and efficient in interacting with the Reddit's API. It allows developers and researchers to programmatically access, retrieve, and analyse Reddit content, including posts, comments, user information, and metadata. Data from four main subreddits such as r/depression_memes, r/anxietymemes, r/sad, and r/depression is targeted to collect posts that contained textual content, emojis, and image links. The important key variables obtained using PRAW are:

Title and Text: This includes the body and title of the posts, which incorporates the textual data.

Emojis_Combined: This contains the emoji used either in the text or title field, along with standalone emojis which was used to do the emoji analysis.

Image_URLs: This contains the image URLs which was obtained from a post in Reddit and were processed using CLIP for image analysis.

Timestamp: This column forms the temporal features of the data, which indicates the time of the day when the post is made.

3.2 Data Preprocessing:

Data preprocessing involves adequate processing applied to raw data to prepare it for further analysis [17]. It is important to do effective preprocessing of multimodal Reddit data for critical of building a robust machine learning pipeline.

3.2.1 Textual Data Preprocessing:

The text data is prepared using various NLP techniques which includes NLTK, VADER, regex, and OpenAI's CLIP model. The step-by-step procedure is as below::

- The data present in the text fields is converted to lower case to maintain the data consistency.
- With the help Regex-based filtering punctuation and special characters is removed
- The NLTK stopwords list is to focus on meaningful words and remove common stopwords such as 'is','are'etc..
- Tokenization: The cleaned text was split into individual tokens using nltk. word_tokenize().
- To ensure there is consistent input size for LSTM layer, padding ofthe word_tokens is done where the sequence of data is padded into a fixed length of tokens.
- The tokens generated are the converted to Vector form using GloVe

3.2.2 Emoji Data Preprocessing

The emoji data present in the posts is initially extracted using the Unicode-aware expressions. The emotional tone of the emojis is assessed as:

- Regex based filtering method is used to filter out the unicode emojis characters from the title and text body
- Sentiment Scoring: Each emoji was mapped to a sentiment score using the VADER lexicon.

- These sentiment scores obtained is later used to calculate valence gap and sign flip, which helped in capturing the mismatch between emojis and text/image sentiment.

3.2.3 Image Data Preprocessing

CLIP model of Open AI is used to process the Reddit posts having images as URL:

- Image URL Validation: Only posts with valid image links (ending in .jpg, .png, etc.) were considered.
- The images from the image URL is downloaded using the python libraries.
- The downloaded image is then passed through a CLIP model which results in 512-dimensional dense vectors that helps in capturing the semantic content of the image.
- Normalization: Embeddings were normalized and standardized for downstream input into Dense layers.

3.2.4 Temporal Data Features

The dataset contains temporal data such as timestamp which is pre-processed to introduce the temporal dynamics:

- Timestamps: Each post's timestamp (in UTC) was converted into features such as time of day (e.g., morning, afternoon, night).

3.3 Exploratory data analysis:

Exploratory Data Analysis (EDA) was performed to gain insights into the structure, distribution, and relationships within the cleaned dataset.

- Sentiment Distribution plots created for Text and Emoji sentiment.
- PCA analysis of the Image sentiments
- Word clouds and TF-IDF identify emotionally salient words in mental health posts
- Temporal features were visualized to get their effect on mental health by analysing the time of the post.

3.4 Multimodal Feature Extraction

Each modality was represented as follows:

Modality	Representation
Text	Pre-trained word embeddings are passed through BiLSTM → LSTM layers
Emoji + Temporal	Emoji sentiment and temporal features are passed through Dense layers
Image	512-dimensional CLIP embeddings are fed into Dense layers

3.5 Mismatch Feature Engineering

The emotional incongruence is modelled by introducing the "mismatch features":

- Valence Gap: The absolute difference between sentiment polarity scores across modalities (e.g., between text and image).
- Sign Flip: Binary indicator that captures if the sentiment polarity of one modality opposes the others (e.g., positive text + negative emoji).

These mismatch features aim to detect subtle cognitive dissonance or masking, which may signal emotional distress.

Chapter 4

Data Exploration

The dataset used in this evaluation was sourced from Reddit using PRAW, which contains 800 posts from 2017 to 2025, including image data, textual data, and Emoji data.

4.1 Text and Emoji Sentiment Analysis:

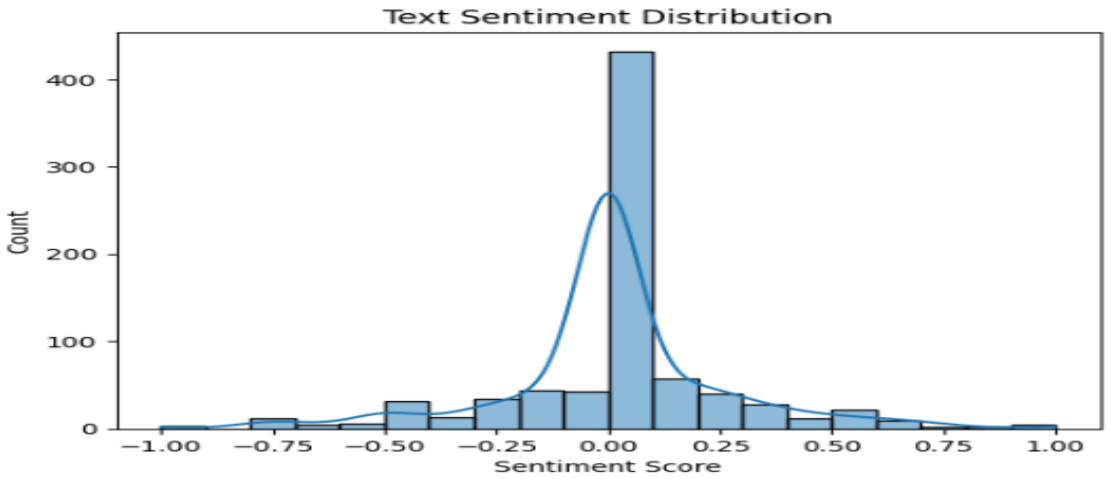

Fig. 4.1 Distribution of text sentiment scores derived using the VADER sentiment analyzer.

The text sentiment scores are largely centred around 0, indicating a prevalence of neutral or mixed emotional content in the dataset. The normal distribution of the data indicates that the sentiment values are well-balanced which is beneficial for the downstream modelling because it reduces the skewness in the input features.

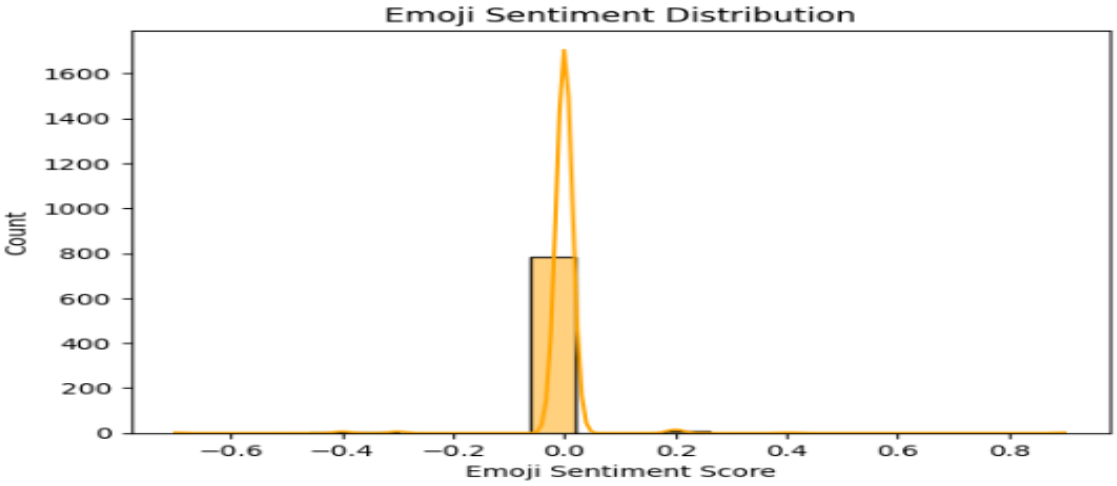

Fig. 4.2 Distribution of emoji sentiment scores computed using VADER mappings.

The emoji sentiment distribution is centred around 0 which indicates that the emojis present in the data convey neutral sentiment. Compared to textual sentiment, emoji sentiment appears less varied and more condensed, with very few emojis expressing extreme positivity or negativity

4.2 Correlation Matrix of Multimodal Features

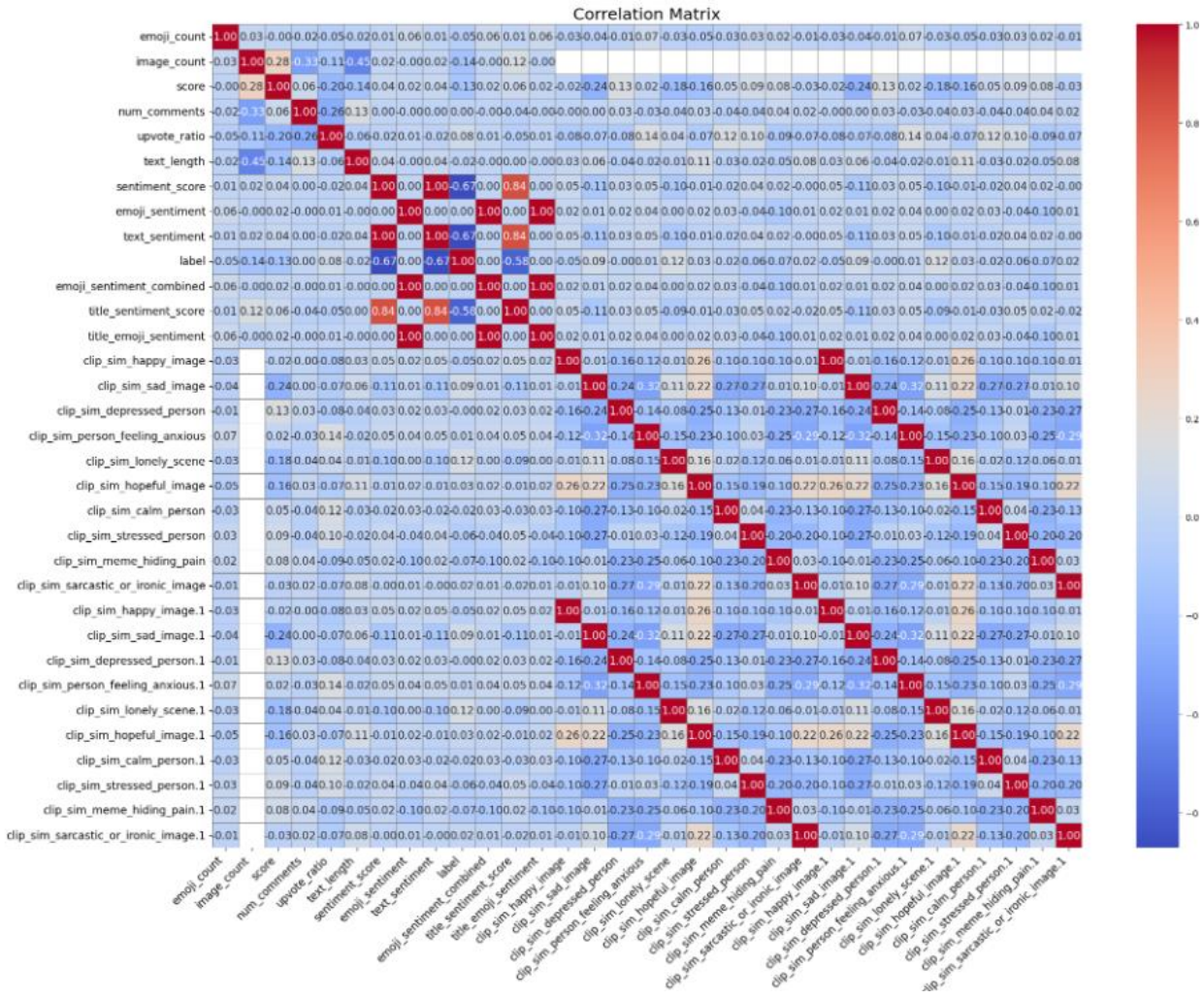

Fig 4.3 Correlation heatmap of engineered features.

The correlation matrix analysis shows that there is limited linear correlation between sentiment signals from different modalities. For example, the text body may convey neutral sentiment but the included emojis or images suggestive of distress. These cross-modal mismatch are subtle but are very beneficial in identifying the early signs of mental health issues. To make use of this insight, the model incorporates mismatch features such as valence gaps and polarity flips, which explicitly quantify disagreement across modalities.

Additionally, the low inter-modality correlation supports the use of a cross-attention mechanism in the model architecture. Cross-attention mechanism makes the network able to

learn dependencies and contextual interplay between modalities dynamically, allowing the model to assign greater importance to the most informative or incongruent signal at a given time.

4.3 PCA of Image Sentiment Features by Subreddit

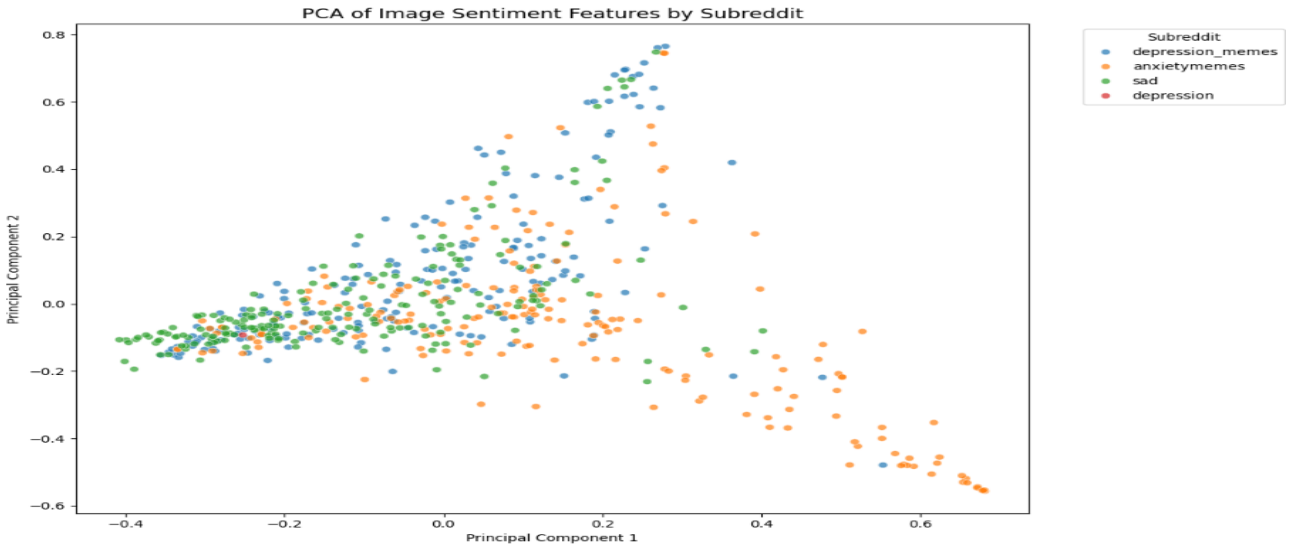

Fig 4.4 PCA plot of CLIP-derived image sentiment embeddings, coloured by source subreddit.

Overlapping Clusters can be found for subreddits depression_memes and anxietymemes, suggesting that the visual content from these communities shares more common sentiment features. The depression subreddit shows greater spread in the principal components, reflecting a broader emotional expression in images—ranging from highly neutral to strongly negative sentiment cues.

4.4 Textual Content Overview via Word cloud

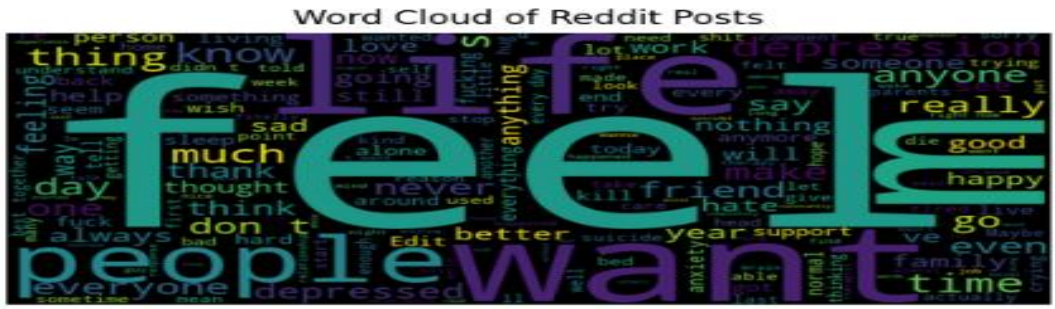

Fig 4.5 Word cloud for Reddit post

To gain an initial understanding of the textual content present in the Reddit posts, a word cloud was generated, which emphasizes the most frequently used words, where word size

corresponds to its relative frequency across the dataset. As per the figure the term “feel” is the most dominant, suggesting that users often tend to express their internal emotional states.

4.5 Temporal Activity Patterns of Reddit Use

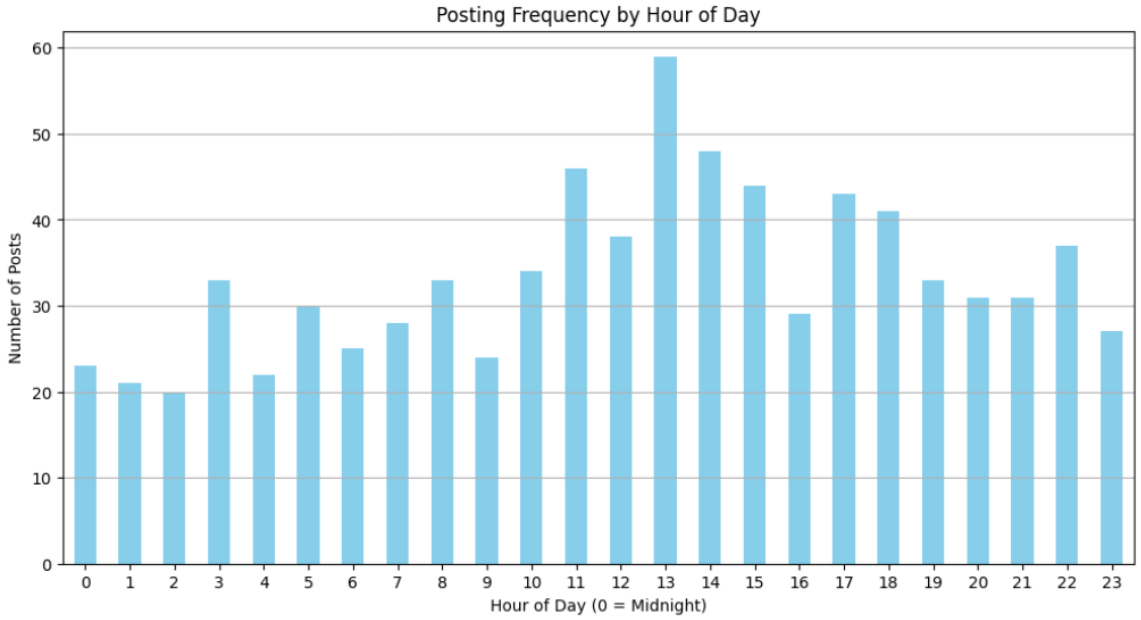

Fig 4.6 Temporal Activity Patterns of Reddit Use

As seen in the figure there is a spike around 1 PM , which indicates this to be the most active period for users share their thoughts and feelings online. This may suggest about midday stress or reflection during breaks. There is relatively high posting frequency in the evening hours compared to the early morning or late-night hours.

Chapter 5

Data Modelling

The data modelling involves design and development of a multimodal deep learning model which processes and integrates information from text, emojis with temporal features, and image content and provides a binary output. As the social media data is complex and diverse, the architecture is structured to fuse diverse inputs through modular branches—each optimized for a specific modality—followed by a cross-attention mechanism that facilitates dynamic interaction among the features. Each modality has its own input branch with the outputs later fused to make a final prediction.

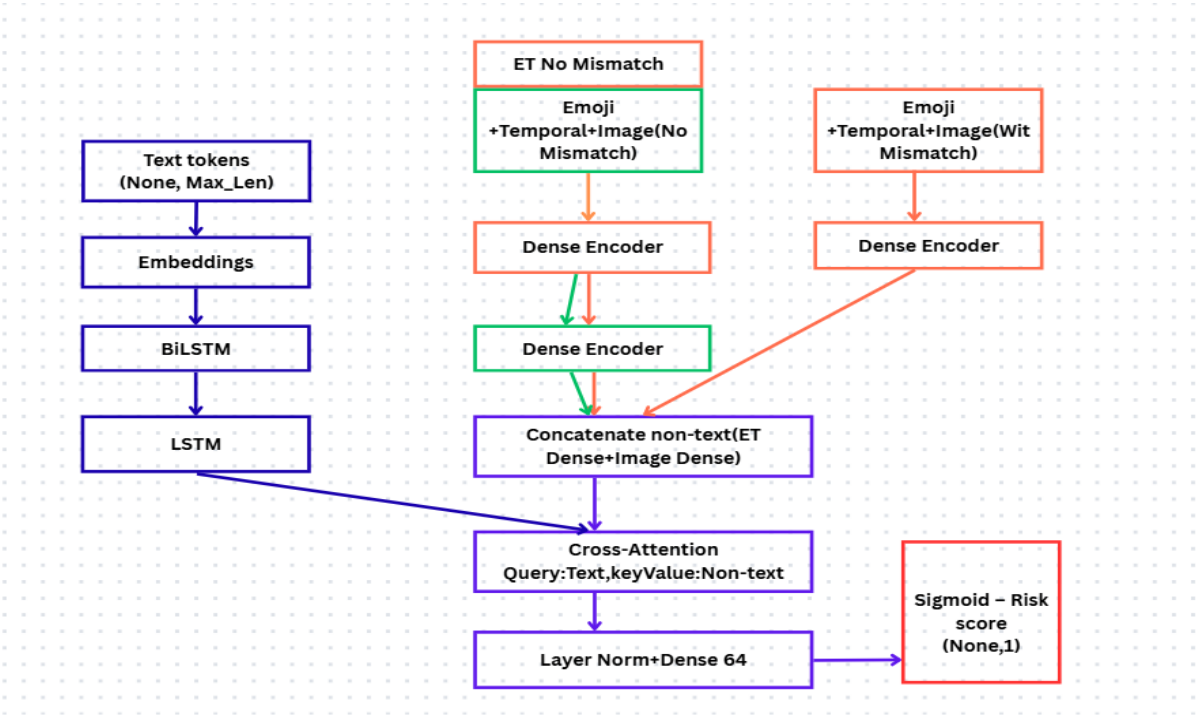

Fig 5.1 Model Architecture

As illustrated in Figure 8, the model consists of three main branches:

1. Textual Branch – Bidirectional LSTM + LSTM
2. Emoji + Temporal Branch – Fully Connected Dense Layers
3. Visual Branch (Image CLIP Embeddings) – Dense Layer

These outputs are concatenated and passed through a Cross-Modal Attention module before entering the final classification layer.

5.1 Text Processing (BiLSTM + LSTM Block)

Reddit posts often contain emotionally nuanced language that indicates a person's emotions that can evolve across the length of a sentence. Such dynamics have been captured by passing the tokenized and embedded text inputs to a Bidirectional Long Short-Term Memory (BiLSTM) layer followed by a unidirectional LSTM.

- Bidirectional long short-term memory (BiLSTM) networks process input sequences both forward and backward, allowing the model to collect more contextual information.[4] These forward and backward directions of the sequence ensure that important emotional cues aren't missed, regardless of their position in the sentence.
- LSTM is capable of learning long-term dependencies in sequential data [4]. The LSTM layer created in the model helps to distil this information into a compact representation, ideal for downstream fusion with other modalities.
- Max Pooling is applied across the sequence of hidden states to obtain a fixed-size embedding, denoted as z_text.

$$h_t = BiLSTM(x_t, h_{t-1}) \quad (1)$$

$$z_{text} = max(H) \quad (2)$$

Where:

- h_t represents the hidden state at time step t produced by the BiLSTM layer.
- Input x_t is the word at position t in the text.
- h_{t-1} is the hidden state from the previous time step — it's part of how the LSTM maintains memory across a sequence.
- z_{text} is a fixed-length vector that contains the most prominent features from the entire text

5.2 Emoji and Temporal Feature Processing

Emojis often reflect the emotional intent of a post and, when combined with temporal metadata (such as time of posting, day of the week, and work vs. weekend hours), can help uncover behaviour patterns associated with mental health states.
These features are combined into a numeric vector and passed through a Dense layer of 64 units with ReLU activation. A second Dense layer refines the embedding, outputting a vector

z_emoji. This structure allows the model to learn abstract representations and interactions between emoji sentiment and posting behaviour.

5.3 Image Processing Using CLIP Embeddings

CLIP (Contrastive Language-Image Pretraining) understands the relationship between text and images with its correct textual description and rejects incorrect descriptions. This allows CLIP to embed both images and text into the same feature space, where:

- Similar meanings = closer vector representations
- Mismatch/incongruence = farther apart vectors

To extract meaningful representations from image URLs, CLIP embeddings is used. It provides a 512-dimensional vector of semantic representation of each image which is then processed through a Dense layer with 64 units ..

$$z_{image} = Dense(CLIP\_embedding) \quad (3)$$

This output is the visual sentiment of each post.

5.4 Fusion and Cross-Attention

The outputs from the text, emoji-temporal, and image branches is combined using a Fusion mechanism to form a unified representation of each post.

$$Z = [z_{text}, z_{emoji}, z_{image}] \quad (4)$$

A cross-attention mechanism is applied which enhances the interaction between these modalities. The model's ability to weigh the relevance of features across different sources is enabled by this mechanism. For example, to identify inconsistencies between a cheerful emoji and negative textual sentiment.

Cross-attention ensures that the model does not treat each modality in isolation but instead interprets them in relation to one another—an essential feature when assessing emotional congruence or mismatch in social media posts.

5.5 Classification Layer

The fused attention output then goes through a final Dense layer which has a Sigmoid activation function which results in a output which makes binary prediction (mental health risk or not.).

$$\hat{y} = \sigma(WZ + b) \quad (5)$$

Where:

- W is the weights that learn the importance of each feature
- Z is the combined feature representation
- b is the bias, which is the trainable offset that helps the model fit the data better
- σ is the sigmoid function that converts the output to be binary 0 or 1.

5.6 Justification for Layer sizes:

A Dense layer consisting of 64 neurons is used following the feature extraction steps. This number strikes a balance between model expressiveness and computational efficiency. The number of neurons taken into consideration is adequate for the understanding of complex patterns but small enough to avoid overfitting which is beneficial while working with multimodal data and a relatively limited dataset. Performance evaluations which involved AUC and Recall, confirmed that this size was sufficient for achieving strong results.

5.7 Training Strategy and Optimization:

For effective training of the multimodal neural network which will ensure stable convergence different optimization and regularization techniques were used:

5.7.1 Optimizer – Adam

The model was trained using the Adam (Adaptive Moment Estimation) optimizer, which is widely recognized for its efficiency in handling sparse gradients and non-stationary objectives. It automatically adjust the learning rate for each parameter during training which makes it very effective in deep learning tasks such as this multimodal setup

5.7.2 Loss Function – Binary Cross-Entropy

As the output of this model is a binary classification which detects if a post shows any signs of mental health issue or not, Binary Cross-Entropy is used as the loss function. In this the distance between the predicted probabilities and the true binary labels is measured and incorrect classifications are penalized more when the model is confident but wrong.

5.7.3 Regularization Techniques

To handle the overfitting issue which ensures that the model generalizes well to unseen data, the following regularization methods were applied:

- **Dropout**: In the designing of the refined model a dropout rate of 0.3 which randomly disables 30% of the neurons during each training iteration. This prevents the network to become too dependent on the overly reliant features.
- **EarlyStopping**: EarlyStoping with a patience value of 5 is used to monitor the training performance. Training is halted early if the validation performance did not show any improvement over five consecutive epochs, which helps in avoid unnecessary training cycles, hence reducing overfitting.

5.8 Label Refinement Strategy

A common challenged faced with the real-world dataset is the label noise which may occur due to incorrect annotation of the posts. This discrepancy happens because of the subjective nature of the mental health interpretation or if the labels are created only taking the sentiments from a single modality into consideration. To address this a label refinement strategy is also incorporated in the model training workflow:

5.8.1 Refinement Using Model Confidence

After the initial model is trained, it is used as a pseudo-labelling agent to reassess the dataset.

- The model predicted the probability of each sample belonging to the positive class.
- A label is revised if the sample's predicted probability is above 90% confidence although its original label disagreed with the prediction.

This process allowed the model to correct potential mislabels, which enhanced the signal-to-noise ratio in the dataset hence improving learning performance for the training data.

5.8.2 Safe Label Refinement

A more cautious variant, termed Safe Label Refinement, is also explored, where only samples with very high agreement across multiple modalities were considered for label changes which minimized the risk of wrongly flipping correct labels due to overfitting or bias.

Chapter 6

Evaluation and Results

This chapter offers a thorough analysis of the experimental results from evaluating the proposed multimodal model. It starts by describing the evaluation metrics used, then provides a detailed overview of quantitative performance outcomes, visual assessment through loss and precision-recall curves, a comparative analysis of model variants, and a discussion interpreting the findings in relation to the research objectives.

6.1 Evaluation Metrics

To assess model performance effectively, multiple evaluation metrics were employed, each offering distinct insights into model behaviour:

- **Accuracy**: It measures the overall proportion of correctly classified samples among the total samples.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

  Where:

  - **TP** = True Positives
  - **TN** = True Negatives
  - **FP** = False Positives
  - **FN** = False Negatives

- **Precision**: It is defined as the ratio between true positives and the sum of true and false positives. A higher precision indicates the model's ability to make fewer false alarms.

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

- **Recall (Sensitivity)**: It is defined as the ratio between true positives and the sum of true and false negatives.

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

  For the scenarios where mental health detection is considered, recall is a very important, as the aim is to minimize missed cases...

- **F1 Score**: It is the harmonic means of precision and recall, which provides a balanced metric when false positives and false negatives are equally critical.

$$F1 = 2.\frac{Precision.Recall}{Precision+Recall} \quad (9)$$

- **ROC-AUC (Receiver Operating Characteristic - Area Under Curve)**: he ability of a model to distinguish between classes which determines its discriminative power is indicated by the ROC-AUC curve. AUC 1 denoted strong discriminative power.
- **PR-AUC (Precision-Recall AUC)**: This is the primary metric which is chosen when there is class imbalance as this curve puts importance on model performance on the positive class (mental health concern), which is often underrepresented. It calculates the area under the Precision-Recall Curve.

6.2 Quantitative Results:

Performance of Multimodal Models Before Label Refinement:

Multimodal models before the label refinement strategy is applied, had the below performances:

Multimodal Model 1 without Early Stopping and Dropout:

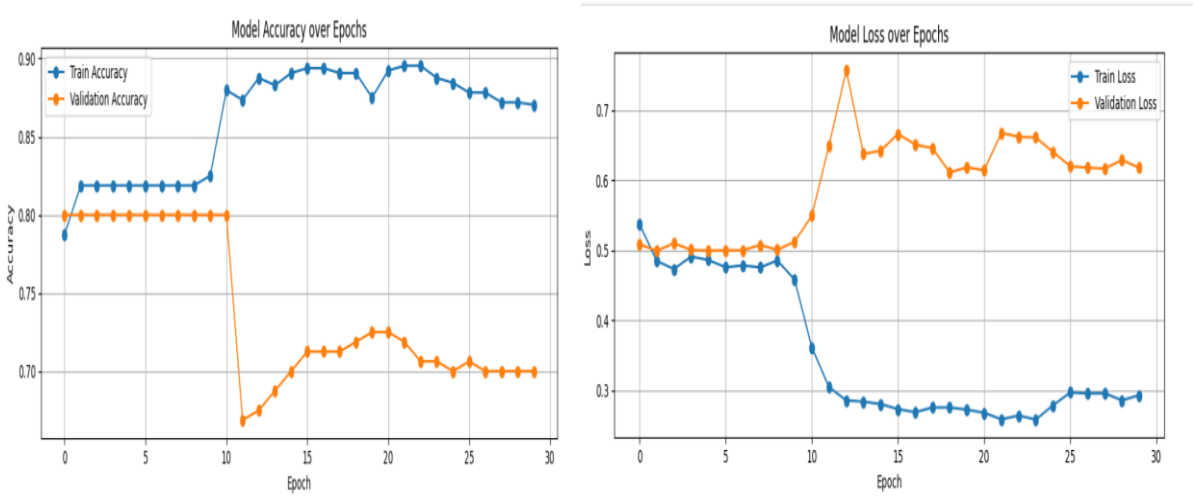

Fig. 6.1 Training vs. Validation Accuracy and Loss over 30 Epochs for one of the baseline models (before label refinement).

The accuracy and loss plot for both training and validation datasets for 30 epochs is shown in the above figure where the training accuracy steadily increases and stabilizes around 88–

89%, whereas the validation accuracy plateaus around 79% and drops briefly near the 10th epoch.

The difference between training and validation loss indicates of potential overfitting, where the model memorizes training data patterns but fails to generalize effectively to unseen data. This performance pattern justified the later use of label refinement techniques to enhance data quality and reduce noise in weakly labelled samples. **Multimodal Model 1 with Early Stopping and Dropout:**

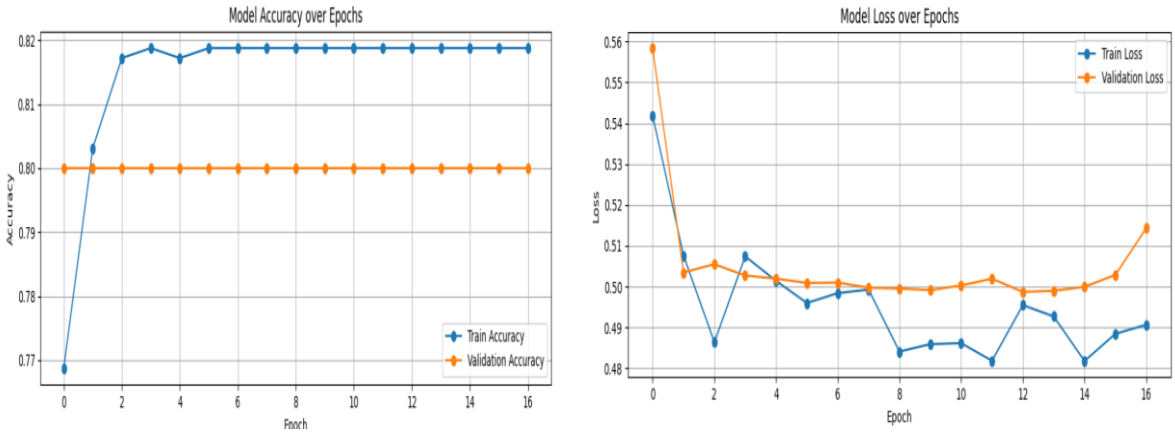

Fig 6.2. Training and Validation Performance After Regularization

To avoid the overfitting and for the enhancement of the generalization, Dropout and EarlyStopping were integrated into the training. A drop-out rate on 0.3 is applied which randomly deactivates neurons to avoid co-adaptation of features. EarlyStopping helped in monitoring the validation loss and is halted after 5 consecutive epochs which led to a stable and synchronized training pattern. After the regularization the training accuracy ranged at approximately 82%, and validation accuracy is of 80%, which remained constant across epochs. This consistency indicates that the model is not overfitting and is learning generalizable patterns from the data.

As per the loss curve after an initial decline, both training and validation losses stabilized, showing no erratic spikes or signs of divergence. Compared to earlier configurations without these regularization measures, this model demonstrates more robust learning and convergence behaviour.

Research Question 1: Can multimodal features (images, text, and emojis) from mental health-related Reddit posts improve the accuracy of early mental health issue detection compared to unimodal features?

To address RQ1, multiple models were developed using distinct feature modalities — including text-only, image-only, emoji+temporal signals, and two multimodal combinations (before label refinement). The comparative performance of all these different models is shown in the below figure, which presents the Receiver Operating Characteristic (ROC) curves along with the Area Under Curve (AUC) scores.

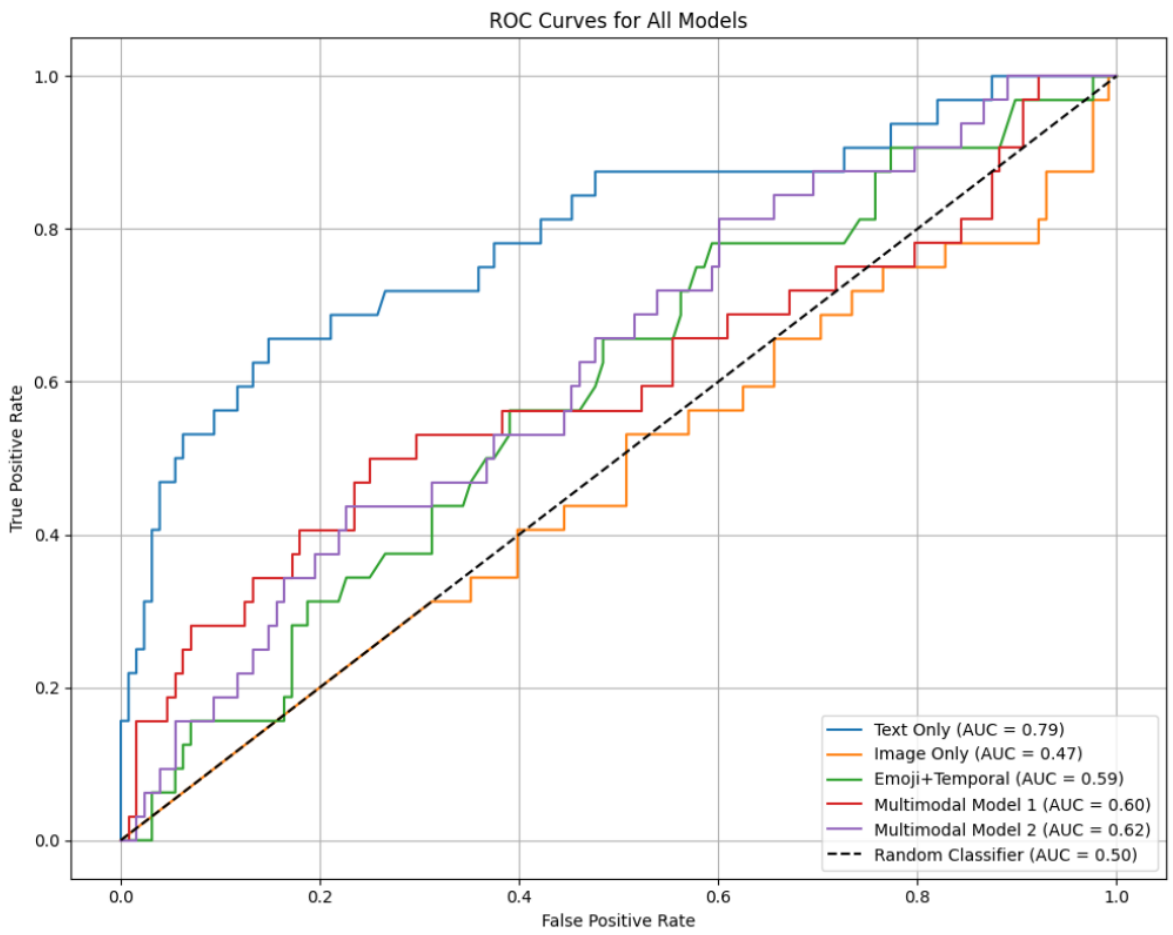

Fig 6.3. ROC Curves of Pre-Refinement Models across Modalities.

The figure above is the representation of the ability of the model to distinguish between positive and negative classes across various thresholds. In an imbalanced dataset like this, a higer AUC indicates better classification capability from the model.

Key Observations:

- The Text-only model outperformed all others with an AUC of 0.79, suggesting that linguistic cues are highly informative for early mental health signal detection.
- The Image-only model showed the lowest AUC of 0.47, which denotes that image alone cannot be sufficient in the signalling.
- Emoji + Temporal model showed AUC of 0.59, indicating that behavioral posting patterns and emoji sentiment contribute useful but insufficient signal on their own
- Multimodal Models 1 and 2 (with various fusion strategies) showed AUCs of 0.60 and 0.62, respectively. While these scores exceed the baseline of the emoji+temporal model, they did not outperform the text-only approach in the pre-label-refinement stage.

The insights derived from this analysis is that multimodal learnig hold the potential of further improvement and is better in performance than the baseline unimodal Image and Emoji models

Label Confidence & Refinement

Figure 6.4 shows the histogram of probabilities predicted from the model against the original labels for both positive (Label = 1) and negative (Label = 0) samples. The predicted probability of a post to be related with a mental health concern is denoted in the X-aixs and the y-axis shows the frequency of posts at each probability bin..

The plot reveals two clear clusters:

- Left Cluster (Low Probability ~ 0.05–0.10): This indicates that although maximum of the samples are confidently predicted as negative cases (Label = 0) which is correct, but some true positive samples (Label = 1) fall into this range which indicates possible mislabeling.
- Right Cluster (High Probability ~ 0.6): A smaller number of posts are confidently predicted as positive (Label = 1), with some original labels marked as negative (Label = 0). These high-confidence disagreements suggest annotation noise or semantic mismatch across modalities (e.g., an image conveying distress, but the text appears neutral).

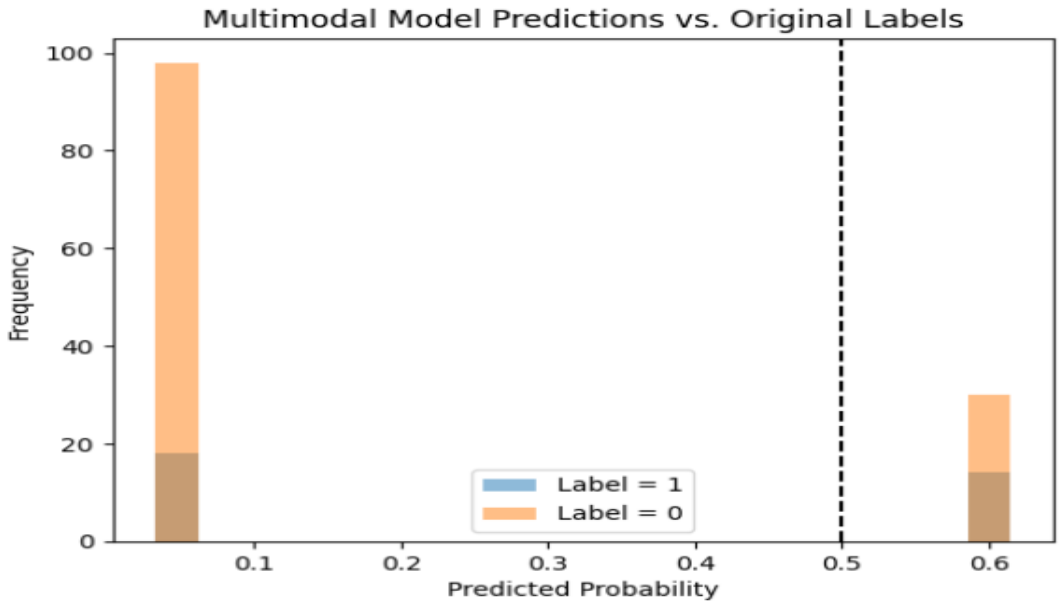

Fig 6.4. Multimodal Model Predictions vs. Original Labels

The dashed vertical line at 0.5 represents the classification threshold. Posts falling to the right but originally labelled as 0 are strong candidates for label refinement. These are the cases where the model exhibits high confidence against the original annotation.

Performance of Multimodal Models After Label Refinement:

Before the label refinement the data is split into 3 splits: Training, Testing and Validation.

Multimodal Model 1 without Early Stopping and Dropout:

The multimodal model 1 was trained on the adjusted dataset, where low-confidence or noisy labels were corrected using a high-confidence prediction threshold (≥ 90%).

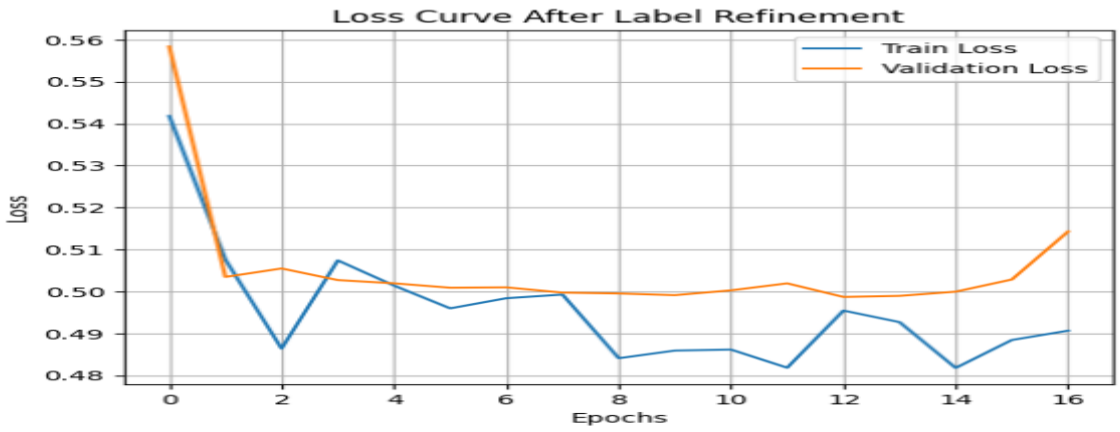

Fig 6.5 Loss curve after applying safe label refinement

At the outset, both training and validation losses are relatively high (above 0.55), which is expected due to the complexity of the multimodal data. However, after the initial epochs there is rapid and steady decline in both curves which denotes a strong improvement in the model's learning stability. The training loss eventually converges below 0.49, and the validation loss stabilizes around 0.50, without significant divergence between the two — suggesting effective generalization and reduced overfitting.

This behaviour is notably different from earlier experiments without refinement, where validation loss tended to rise sharply. The consistency between training and validation performance here confirms that the model has learned more coherent and reliable patterns from the cleaned labels, emphasizing the importance of addressing label noise in real-world social media data.

Refined Model performance with Regularization:

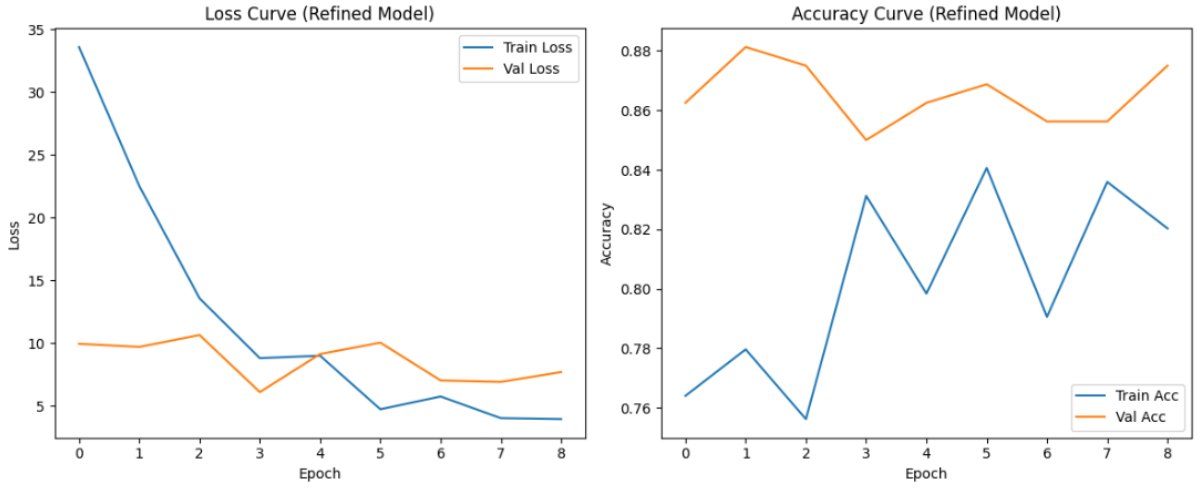

Fig 6.6. Loss and Accuracy Curve of Refined Model

Training and Validation Loss

A steep decline in training loss can be observed which starts from 30 and gradually decreases to below 5 which indicates the model's capability to learn from meaningful representations in the refined dataset. The behaviour of the training loss which is lower and stable without any spikes indicates generalization and less overfitting. The narrowed gap between the training and validation losses is a testament to the improved data quality after label refinement.

Training and Validation Accuracy

The right panel shows the Training and Validation Accuracy . The validation accuracy of the model ranges between 86% to 88%. There is gradual improvement in the training accuracy. Although there are minor fluctuations due to the absence of any overfitting pattern in which training accuracy exceeds validation by a wide margin the model demonstrates its robustness.

A safer label refinement strategy was also used where the confidence interval of 95% was used and the model performance is shown below:

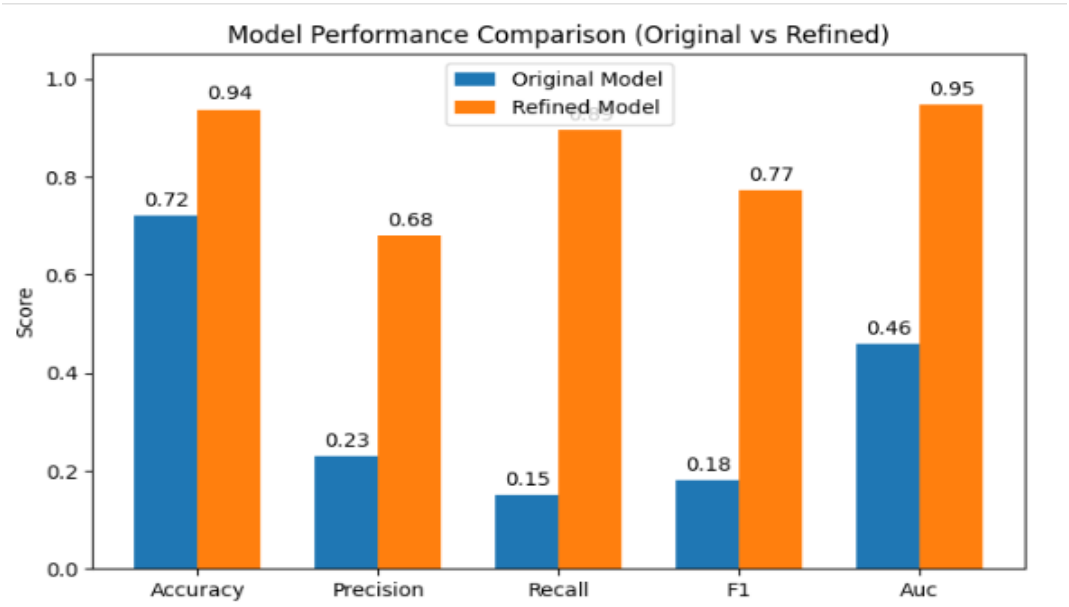

Fig 6.7 Model performance comparison

Label refinement directly tackled overfitting by improving label quality, which in turn stabilized the training process.

Final Test Set Evaluation:

The final set evaluation was done on the unseen data to gain deeper insights into the model's generalization capability. The higher number of true negatives (126) and very few true positives (2), indicates that the model is highly skewed towards predicting the majority class (no issue). The low recall for the mental health class with a AUC score of 0.48 is below the random guessing threshold of 0.50 which indicated that the refined model after label refinement had limited discriminaive ability. These indicate that the model performs well on the dominant class, but is ineffective in capture the patterns associated with early signs of mental health concerns when a new unseen data in introduced which opens the future scope for better label refinement strategies.

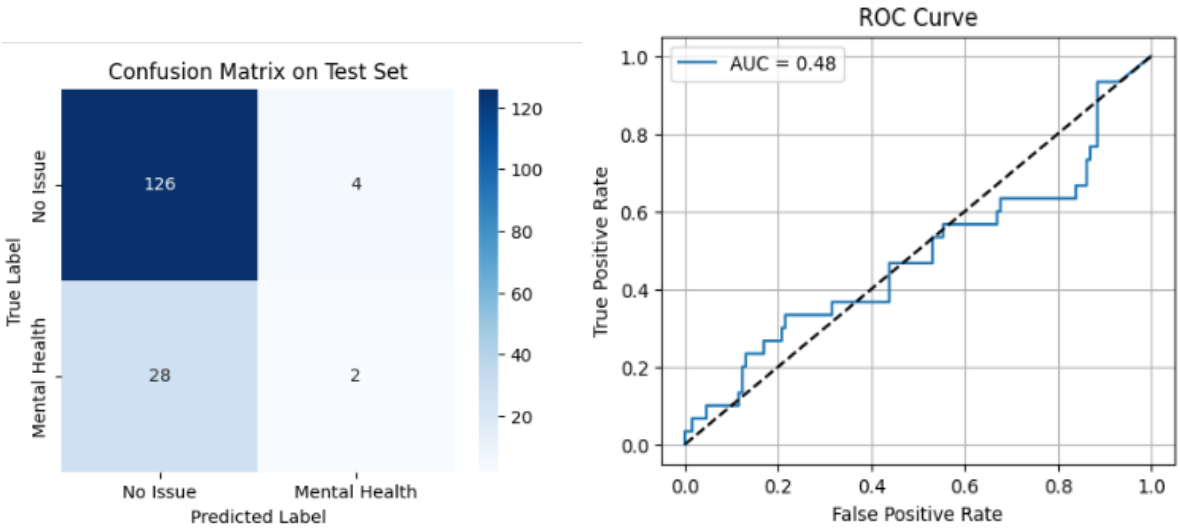

Fig 6.8 Confusion matrix and ROC curve for the Final test set

Research Question 2: Can a mismatch between the text, emojis, and images in social media posts help us detect early signs of mental health issues?

Mismatch features were introduced into the model in the form of Valence Gaps and Sign flips which may help to identify the hidden indications in a post. The PR curve is a below:

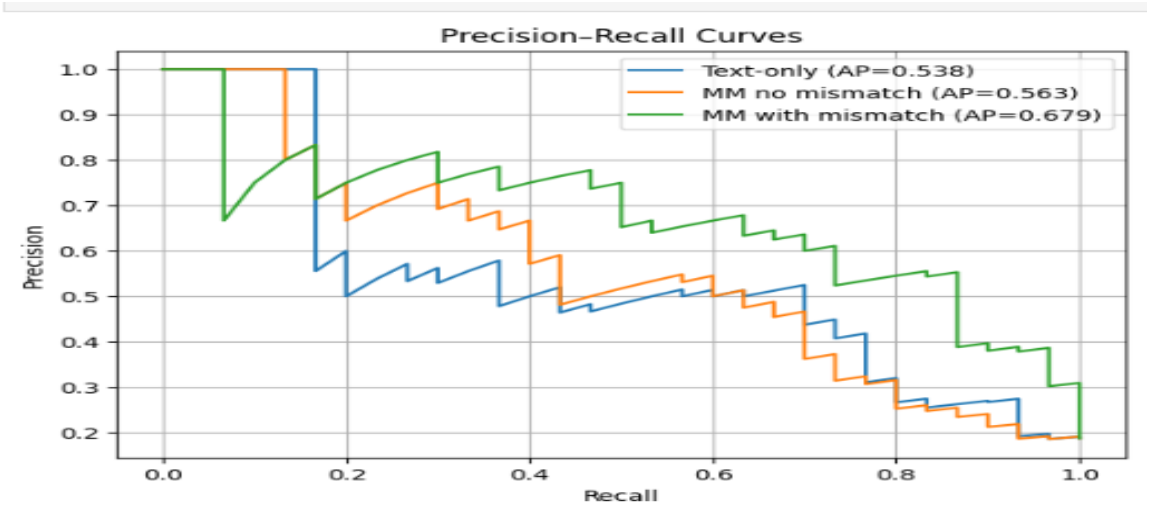

Fig 6.9 Precision Recall Curves with mismatch

We compared three model variants: a text-only model, a multimodal model without mismatch features, and a multimodal model with mismatch signals explicitly included. The Precision-Recall (PR) curve reveals that incorporating mismatch features significantly improves model performance in identifying positive (mental health-related) cases.There is an increase in the

average precision from 0.538 in text only model which is our baseline model to 0.679 in multimodal model with mismatch features included which clearly demonstrates that semantic incongruence between modalities (e.g., cheerful emojis paired with distressing text) can act as a subtle, yet powerful indicator of underlying mental health concerns. The results of elevated precison across all levels of recall helps in validating the hypothesis that cross-modal inconsistencies can enhance early detection accuracy, especially in nuanced or ambiguous posts.

Fold CV summary:

Model	Accuracy	Precision	Recall	F1	ROC AUC	PR AUC
Text-only	0.825	0.550	0.512	0.516	0.790	0.564
Image-only	0.660	0.261	0.453	0.331	0.620	0.250
Emoji-only	0.588	0.211	0.446	0.284	0.541	0.225
MM (no mismatch)	0.790	0.452	0.512	0.474	0.759	0.521
MM (with mismatch)	0.790	0.535	**0.771**	**0.706**	**0.907**	**0.705**

Table 6.1 5-Fold CV summary

- The text only model helped as a strong unimodal baseline denoting good accuracy and AUC but the recall was moderate which indicates that only text data may miss may subtle nuances which can be captured with modelling the mismatch features and other data modalities.
- The image and emoji only model performed comparatively poor across all metrics with low precision and recall indicating that only images and emoji's data do not have the capability to convey emotional state without the text data.
- Multimodal without Mismatch: Combining modalities (text + emoji + image) modestly improved overall performance, especially in recall. This supports the notion that multi-source data helps in identifying psychological states better than any single modality.
- Multimodal with Mismatch Feature: This version significantly outperformed all others in recall (0.77) and AUC (0.907). This mismatch feature helps the models decision to distinguish between emotional tone in different modalities. For an example, if a post with cheerful emoji is paired with a deeply negative caption or dark image is a possible trigger of emotional incongruence. Capturing such incongruities proves vital in early mental health detection.

This performance validates Research Question 2 with the improvement in recall (+26%) and PR-AUC (+0.18) when the mismatch feature is used clearly demonstrates its value.

Chapter 7

Future Scope

This study demonstrated the feasibility of using multimodal content—including text, images, emojis, and their semantic mismatches—to identify early indicators of mental health issues on social media platforms. While the results are promising it widens up the scope of improvement to broaden itself in real-world scenario in the following ways:

1. Advanced Label Refinement Strategies:

The label refinement strategy applied in this study is a thresholding mechanism which uses the model confidence score from the initial epochs which may be is not sufficient to capture the complex mental health signal. Better label refinement strategies like below can be explored to increase the PR-AUC score:

- Use of Bayesian Neural Networks
- Soft labelling techniques, where labels are refined a probability value instead of binary

2. Advanced Mismatch Modelling:

The current mismatch feature includes sign flips and valence gaps between different sentiment scores from all modalities which is rule based.

- Neural networks can be trained to learn mismatch patterns directly from multimodal embeddings.
- Explore multimodal contrastive learning to better capture discordance between emotional signals.

3. Incorporation of Audio and Video Modalities

The current approach includes image, emoji and text data, future scope can include other modalities of data such as Audio and Video derived from Social media platforms such as TikTok, Instagram, and YouTube and below analysis can be done:

- Voice tone analysis
- Facial emotion recognition

Chapter 8

Conclusion

This study involved exploring the fusion of the multimodal data such as Text, emoji and image into a single fusion model and exploring if the data mining techniques mentioned are effective in detecting the early signs of mental health. A novel and innovative concept of mismatch features was also introduced to capture conflicting emotional signals across modalities (e.g., sad text but happy emoji) could signal distress.

The results showed that in some scenarios where the label refinement and mismatch features introduced the model performed in a different way. In many multimodal models outperform unimodal image and emoji+temporal ones, especially when mismatch features were included. These mismatches proved valuable in capturing subtle emotional contradictions that might otherwise be missed.

However, a post-processing technique called label refinement, meant to improve label quality using model confidence, led to a performance drop. This revealed that automatic label correction, while promising, must be used cautiously—especially in sensitive domains like mental health.

Hence, we can conclude from the study that,

- The performance of multimodal approaches are better than unimodal (image, emojis) ones but robust label refinement strategies are required to further improve the model's performance.
- Hidden and deeper emotional incongruence can be captured using mismatch features.
- Label refinement strategies must be carefully designed.

These insights build the way for building more empathetic and intelligent tools for early mental health detection online.

CHAPTER A

APPENDIX A: Training logs of different models

- Multimodal model without regularization:

```
20/20 [==============================] - 17s 223ms/step - loss: 0.5375 - accuracy: 0.7875 - val_loss: 0.5087 - val_accuracy: 0.8000
Epoch 2/30
20/20 [==============================] - 2s 100ms/step - loss: 0.4849 - accuracy: 0.8188 - val_loss: 0.4994 - val_accuracy: 0.8000
Epoch 3/30
20/20 [==============================] - 3s 138ms/step - loss: 0.4733 - accuracy: 0.8188 - val_loss: 0.5101 - val_accuracy: 0.8000
Epoch 4/30
20/20 [==============================] - 4s 204ms/step - loss: 0.4911 - accuracy: 0.8188 - val_loss: 0.5006 - val_accuracy: 0.8000
Epoch 5/30
20/20 [==============================] - 3s 147ms/step - loss: 0.4861 - accuracy: 0.8188 - val_loss: 0.4995 - val_accuracy: 0.8000
Epoch 6/30
20/20 [==============================] - 3s 144ms/step - loss: 0.4758 - accuracy: 0.8188 - val_loss: 0.5001 - val_accuracy: 0.8000
Epoch 7/30
20/20 [==============================] - 3s 168ms/step - loss: 0.4784 - accuracy: 0.8188 - val_loss: 0.5001 - val_accuracy: 0.8000
Epoch 8/30
20/20 [==============================] - 4s 180ms/step - loss: 0.4755 - accuracy: 0.8188 - val_loss: 0.5076 - val_accuracy: 0.8000
Epoch 9/30
20/20 [==============================] - 3s 138ms/step - loss: 0.4855 - accuracy: 0.8188 - val_loss: 0.5006 - val_accuracy: 0.8000
Epoch 10/30
20/20 [==============================] - 3s 148ms/step - loss: 0.4583 - accuracy: 0.8250 - val_loss: 0.5123 - val_accuracy: 0.8000
Epoch 11/30
20/20 [==============================] - 3s 164ms/step - loss: 0.3614 - accuracy: 0.8797 - val_loss: 0.5493 - val_accuracy: 0.8000
Epoch 12/30
20/20 [==============================] - 3s 136ms/step - loss: 0.3044 - accuracy: 0.8734 - val_loss: 0.6499 - val_accuracy: 0.6687
Epoch 13/30
20/20 [==============================] - 3s 131ms/step - loss: 0.2856 - accuracy: 0.8875 - val_loss: 0.7571 - val_accuracy: 0.6750
Epoch 14/30
20/20 [==============================] - 2s 117ms/step - loss: 0.2838 - accuracy: 0.8828 - val_loss: 0.6375 - val_accuracy: 0.6875
Epoch 15/30
20/20 [==============================] - 3s 127ms/step - loss: 0.2803 - accuracy: 0.8906 - val_loss: 0.6424 - val_accuracy: 0.7000
Epoch 16/30
20/20 [==============================] - 3s 129ms/step - loss: 0.2732 - accuracy: 0.8938 - val_loss: 0.6663 - val_accuracy: 0.7125
Epoch 17/30
20/20 [==============================] - 3s 138ms/step - loss: 0.2687 - accuracy: 0.8938 - val_loss: 0.6510 - val_accuracy: 0.7125
Epoch 18/30
20/20 [==============================] - 3s 132ms/step - loss: 0.2754 - accuracy: 0.8906 - val_loss: 0.6459 - val_accuracy: 0.7125
Epoch 19/30
20/20 [==============================] - 3s 140ms/step - loss: 0.2754 - accuracy: 0.8906 - val_loss: 0.6114 - val_accuracy: 0.7188
Epoch 20/30
20/20 [==============================] - 3s 136ms/step - loss: 0.2724 - accuracy: 0.8750 - val_loss: 0.6185 - val_accuracy: 0.7250
Epoch 21/30
20/20 [==============================] - 3s 138ms/step - loss: 0.2679 - accuracy: 0.8922 - val_loss: 0.6145 - val_accuracy: 0.7250
Epoch 22/30
20/20 [==============================] - 3s 136ms/step - loss: 0.2585 - accuracy: 0.8953 - val_loss: 0.6676 - val_accuracy: 0.7188
Epoch 23/30
20/20 [==============================] - 3s 136ms/step - loss: 0.2639 - accuracy: 0.8953 - val_loss: 0.6622 - val_accuracy: 0.7063
Epoch 24/30
20/20 [==============================] - 3s 125ms/step - loss: 0.2582 - accuracy: 0.8875 - val_loss: 0.6610 - val_accuracy: 0.7063
Epoch 25/30
20/20 [==============================] - 3s 135ms/step - loss: 0.2786 - accuracy: 0.8844 - val_loss: 0.6405 - val_accuracy: 0.7000
Epoch 26/30
20/20 [==============================] - 3s 129ms/step - loss: 0.2969 - accuracy: 0.8781 - val_loss: 0.6201 - val_accuracy: 0.7063
Epoch 27/30
20/20 [==============================] - 3s 136ms/step - loss: 0.2955 - accuracy: 0.8781 - val_loss: 0.6181 - val_accuracy: 0.7000
```

Table A.1 Training log of Multimodal without regularization

- Multimodal Model with regularization:

```
Epoch 1/30
20/20 [==============================] - 22s 319ms/step - loss: 0.5418 - accuracy: 0.7688 - val_loss: 0.5584 - val_accuracy: 0.8000
Epoch 2/30
20/20 [==============================] - 3s 130ms/step - loss: 0.5076 - accuracy: 0.8031 - val_loss: 0.5035 - val_accuracy: 0.8000
Epoch 3/30
20/20 [==============================] - 3s 140ms/step - loss: 0.4864 - accuracy: 0.8172 - val_loss: 0.5055 - val_accuracy: 0.8000
Epoch 4/30
20/20 [==============================] - 3s 143ms/step - loss: 0.5074 - accuracy: 0.8188 - val_loss: 0.5027 - val_accuracy: 0.8000
Epoch 5/30
20/20 [==============================] - 3s 131ms/step - loss: 0.5014 - accuracy: 0.8172 - val_loss: 0.5019 - val_accuracy: 0.8000
Epoch 6/30
20/20 [==============================] - 3s 150ms/step - loss: 0.4960 - accuracy: 0.8188 - val_loss: 0.5009 - val_accuracy: 0.8000
Epoch 7/30
20/20 [==============================] - 3s 146ms/step - loss: 0.4984 - accuracy: 0.8188 - val_loss: 0.5010 - val_accuracy: 0.8000
Epoch 8/30
20/20 [==============================] - 3s 130ms/step - loss: 0.4993 - accuracy: 0.8188 - val_loss: 0.4997 - val_accuracy: 0.8000
Epoch 9/30
20/20 [==============================] - 2s 121ms/step - loss: 0.4841 - accuracy: 0.8188 - val_loss: 0.4995 - val_accuracy: 0.8000
Epoch 10/30
20/20 [==============================] - 3s 148ms/step - loss: 0.4859 - accuracy: 0.8188 - val_loss: 0.4991 - val_accuracy: 0.8000
Epoch 11/30
20/20 [==============================] - 3s 131ms/step - loss: 0.4862 - accuracy: 0.8188 - val_loss: 0.5003 - val_accuracy: 0.8000
Epoch 12/30
20/20 [==============================] - 2s 122ms/step - loss: 0.4818 - accuracy: 0.8188 - val_loss: 0.5019 - val_accuracy: 0.8000
Epoch 13/30
20/20 [==============================] - 3s 143ms/step - loss: 0.4955 - accuracy: 0.8188 - val_loss: 0.4987 - val_accuracy: 0.8000
Epoch 14/30
20/20 [==============================] - 3s 147ms/step - loss: 0.4927 - accuracy: 0.8188 - val_loss: 0.4990 - val_accuracy: 0.8000
Epoch 15/30
20/20 [==============================] - 3s 135ms/step - loss: 0.4818 - accuracy: 0.8188 - val_loss: 0.5000 - val_accuracy: 0.8000
Epoch 16/30
20/20 [==============================] - 3s 135ms/step - loss: 0.4884 - accuracy: 0.8188 - val_loss: 0.5029 - val_accuracy: 0.8000
Epoch 17/30
20/20 [==============================] - ETA: 0s - loss: 0.4907 - accuracy: 0.8188Restoring model weights from the end of the best epoch: 13.
20/20 [==============================] - 3s 159ms/step - loss: 0.4907 - accuracy: 0.8188 - val_loss: 0.5143 - val_accuracy: 0.8000
Epoch 17: early stopping
```

Table A.2 Training log of Multimodal with regularization

➢ Refined model without regularization

```
20/20 [==============================] - 3s 145ms/step - loss: 0.2080 - accuracy: 0.8828 - val_loss: 0.1763 - val_accuracy: 0.8625
Epoch 2/30
20/20 [==============================] - 3s 141ms/step - loss: 0.1962 - accuracy: 0.8672 - val_loss: 0.1730 - val_accuracy: 0.8750
Epoch 3/30
20/20 [==============================] - 3s 137ms/step - loss: 0.2037 - accuracy: 0.8703 - val_loss: 0.1720 - val_accuracy: 0.8750
Epoch 4/30
20/20 [==============================] - 3s 131ms/step - loss: 0.2028 - accuracy: 0.8703 - val_loss: 0.1719 - val_accuracy: 0.8750
Epoch 5/30
20/20 [==============================] - 3s 135ms/step - loss: 0.2009 - accuracy: 0.8734 - val_loss: 0.1712 - val_accuracy: 0.8750
Epoch 6/30
20/20 [==============================] - 3s 137ms/step - loss: 0.2204 - accuracy: 0.8766 - val_loss: 0.1630 - val_accuracy: 0.9250
Epoch 7/30
20/20 [==============================] - 3s 138ms/step - loss: 0.2787 - accuracy: 0.9109 - val_loss: 0.1904 - val_accuracy: 0.9375
Epoch 8/30
20/20 [==============================] - 3s 136ms/step - loss: 0.2443 - accuracy: 0.9172 - val_loss: 0.1739 - val_accuracy: 0.9438
Epoch 9/30
20/20 [==============================] - 3s 136ms/step - loss: 0.2358 - accuracy: 0.9187 - val_loss: 0.1572 - val_accuracy: 0.9438
Epoch 10/30
20/20 [==============================] - 3s 132ms/step - loss: 0.2238 - accuracy: 0.9234 - val_loss: 0.1695 - val_accuracy: 0.9312
Epoch 11/30
20/20 [==============================] - 3s 130ms/step - loss: 0.2247 - accuracy: 0.9250 - val_loss: 0.1666 - val_accuracy: 0.9375
Epoch 12/30
20/20 [==============================] - 3s 132ms/step - loss: 0.2218 - accuracy: 0.9234 - val_loss: 0.1735 - val_accuracy: 0.9312
Epoch 13/30
20/20 [==============================] - 3s 130ms/step - loss: 0.2096 - accuracy: 0.9266 - val_loss: 0.1638 - val_accuracy: 0.9375
Epoch 14/30
20/20 [==============================] - 3s 133ms/step - loss: 0.2076 - accuracy: 0.9266 - val_loss: 0.1636 - val_accuracy: 0.9375
Epoch 15/30
20/20 [==============================] - 3s 129ms/step - loss: 0.1980 - accuracy: 0.9281 - val_loss: 0.1845 - val_accuracy: 0.9187
Epoch 16/30
20/20 [==============================] - 3s 130ms/step - loss: 0.2055 - accuracy: 0.9250 - val_loss: 0.2055 - val_accuracy: 0.9000
Epoch 17/30
20/20 [==============================] - 3s 133ms/step - loss: 0.1963 - accuracy: 0.9250 - val_loss: 0.2043 - val_accuracy: 0.9000
Epoch 18/30
20/20 [==============================] - 3s 134ms/step - loss: 0.1989 - accuracy: 0.9266 - val_loss: 0.1826 - val_accuracy: 0.9187
Epoch 19/30
20/20 [==============================] - 2s 122ms/step - loss: 0.1916 - accuracy: 0.9344 - val_loss: 0.1854 - val_accuracy: 0.9187
Epoch 20/30
20/20 [==============================] - 3s 136ms/step - loss: 0.1780 - accuracy: 0.9391 - val_loss: 0.1771 - val_accuracy: 0.9250
Epoch 21/30
20/20 [==============================] - 3s 134ms/step - loss: 0.1709 - accuracy: 0.9359 - val_loss: 0.1882 - val_accuracy: 0.9187
Epoch 22/30
20/20 [==============================] - 3s 132ms/step - loss: 0.1775 - accuracy: 0.9375 - val_loss: 0.1844 - val_accuracy: 0.9187
Epoch 23/30
20/20 [==============================] - 3s 132ms/step - loss: 0.1650 - accuracy: 0.9375 - val_loss: 0.1874 - val_accuracy: 0.9187
Epoch 24/30
20/20 [==============================] - 3s 133ms/step - loss: 0.1663 - accuracy: 0.9375 - val_loss: 0.1804 - val_accuracy: 0.9250
Epoch 25/30
20/20 [==============================] - 3s 134ms/step - loss: 0.1662 - accuracy: 0.9359 - val_loss: 0.1715 - val_accuracy: 0.9312
Epoch 26/30
20/20 [==============================] - 3s 128ms/step - loss: 0.1649 - accuracy: 0.9391 - val_loss: 0.1869 - val_accuracy: 0.9187
Epoch 27/30
20/20 [==============================] - 3s 132ms/step - loss: 0.1656 - accuracy: 0.9391 - val_loss: 0.1809 - val_accuracy: 0.9250
Epoch 28/30
20/20 [==============================] - 3s 132ms/step - loss: 0.1622 - accuracy: 0.9391 - val_loss: 0.1739 - val_accuracy: 0.9312
Epoch 29/30
```

Table A.3 Training log of refined without regularization

➢ Refined model with regularization

```
Epoch 1/30
20/20 [==============================] - 2s 22ms/step - loss: 33.6099 - accuracy: 0.7641 - val_loss: 9.9387 - val_accuracy: 0.8625 - lr: 0.0010
Epoch 2/30
20/20 [==============================] - 0s 9ms/step - loss: 22.5226 - accuracy: 0.7797 - val_loss: 9.6912 - val_accuracy: 0.8813 - lr: 0.0010
Epoch 3/30
20/20 [==============================] - 0s 15ms/step - loss: 13.5610 - accuracy: 0.7563 - val_loss: 10.6366 - val_accuracy: 0.8750 - lr: 0.0010
Epoch 4/30
20/20 [==============================] - 0s 9ms/step - loss: 8.7937 - accuracy: 0.8313 - val_loss: 6.0872 - val_accuracy: 0.8500 - lr: 0.0010
Epoch 5/30
20/20 [==============================] - 0s 8ms/step - loss: 8.9878 - accuracy: 0.7984 - val_loss: 9.1234 - val_accuracy: 0.8625 - lr: 0.0010
Epoch 6/30
20/20 [==============================] - 0s 6ms/step - loss: 4.7152 - accuracy: 0.8406 - val_loss: 10.0313 - val_accuracy: 0.8687 - lr: 0.0010
Epoch 7/30
14/20 [====================>.........] - ETA: 0s - loss: 5.4712 - accuracy: 0.7969
Epoch 7: ReduceLROnPlateau reducing learning rate to 0.0005000000237487257.
20/20 [==============================] - 0s 7ms/step - loss: 5.7356 - accuracy: 0.7906 - val_loss: 7.0158 - val_accuracy: 0.8562 - lr: 0.0010
Epoch 8/30
20/20 [==============================] - 0s 8ms/step - loss: 4.0061 - accuracy: 0.8359 - val_loss: 6.8995 - val_accuracy: 0.8562 - lr: 5.0000e-04
Epoch 9/30
20/20 [==============================] - 0s 7ms/step - loss: 3.9305 - accuracy: 0.8203 - val_loss: 7.6853 - val_accuracy: 0.8750 - lr: 5.0000e-04
```

Table A.4 Training log of refined with regularization

CHAPTER B

APPENDIX B: Generative AI prompt usage

Table B.1: Examples of Generative AI Prompts Used in This Dissertation

Purpose	Prompt	Usage
Literature Review Support	Summarize recent literature (2020–2024) on Cross modal or multimodal fusions for mental health	Helped me get a overview of the research paper and helped check the algorithms
Sentence Refinement	Help me rewrite the abstract of my dissertation to be more concise and policy-oriented	Helped me draft some sections in a efficient way
Methodology Clarification	Help me understand how PR-AUC and Recall are related and its impact	Helped in better understanding of the evaluation metrics
Coding and Debugging	Why am I getting tensor flow keras not loading error	Helped handle the libraries and dependencies effectively

CHAPTER C

APPENDIX C: GITLAB REPOSITORY

The full project (code, data, and dashboards) is hosted on https://git.cs.bham.ac.uk/projects-2024-25/dxd444

Repository contents

1. Mental_health_code_UPDATED_with_test_set.ipynb— main Jupyter notebook with the analysis and figures.

3. Reddit_Data_Final_Raw,csv — primary dataset used in the analysis.
4. README.md — basic project instructions

Runtime notes

- Python 3 environment (e.g., 3.8+). Use Jupyter Notebook/Lab to run Mental_health_code_UPDATED_with_test_set.ipynb
- . Install the libraries listed in the notebook/README before execution

References

1. Global Mental Health Commission, n.d. *Mental health statistics*. [online] Available at: https://globalmentalhealthcommission.org/mental-health-statistics/#:~:text=Approximately%201%20in%208%20people%20globally%20%28about%20970,disorders%20affected%20about%205.1%25%20of%20the%20global%20population. [Accessed 1 Sept. 2025].

2. World Health Organization (WHO), 2024. *Teens, screens and mental health*. [online] 25 September. Available at: https://www.who.int/europe/news/item/25-09-2024-teens--screens-and-mental-health [Accessed 1 Sept. 2025].

3. Sage Journals, 2025. *Article in Journal of Constructivist Psychology*. [online] Available at: https://journals.sagepub.com/doi/abs/10.1177/10664807251346978 [Accessed 1 Sept. 2025].

4. ArXiv, 2025. *Paper ID: 2503.07653*. [pdf] Available at: https://arxiv.org/pdf/2503.07653 [Accessed 1 Sept. 2025].

5. ArXiv, 2022. *Paper ID: 2202.07427*. [pdf] Available at: https://arxiv.org/pdf/2202.07427 [Accessed 1 Sept. 2025].

6. ArXiv, 2023. *Paper ID: 2304.09493*. [pdf] Available at: https://arxiv.org/pdf/2304.09493 [Accessed 1 Sept. 2025].

7. ArXiv, 2024. *Paper ID: 2401.04655*. [pdf] Available at: https://arxiv.org/pdf/2401.04655 [Accessed 1 Sept. 2025].

8. ArXiv, 2021. *Paper ID: 2103.00020*. [pdf] Available at: https://arxiv.org/pdf/2103.00020 [Accessed 1 Sept. 2025].

9. Baron, R.M. and Kenny, D.A., 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), pp.1173–1182. Available at: https://www.researchgate.net/publication/281274059_The_moderator-mediator_variable_distinction_in_social_psychological_research_Conceptual_strategic_and_statistical_considerations [Accessed 1 Sept. 2025].

10. Mohammadi, M., Al-Fuqaha, A., Sorour, S. and Guizani, M., 2019. Deep learning for IoT big data and streaming analytics: A survey. *Neural Computing and Applications*, 32, pp.459–

495. Available at: https://link.springer.com/article/10.1007/s11063-019-10035-7 [Accessed 1 Sept. 2025].

11. He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*. [pdf] Available at: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf [Accessed 1 Sept. 2025].

12.ScienceDirect, n.d. *Machine learning*. [online] Available at: https://www.sciencedirect.com/topics/computer-science/machine-learning [Accessed 1 Sept. 2025].

13.ScienceDirect, n.d. *Learning approach*. [online] Available at: https://www.sciencedirect.com/topics/computer-science/learning-approach [Accessed 1 Sept. 2025].

14. Radford, A. et al., 2018. Language models are unsupervised multitask learners. *ArXiv preprint arXiv:1812.01187*. [pdf] Available at: https://arxiv.org/pdf/1812.01187 [Accessed 1 Sept. 2025].

15. Brown, T. et al., 2020. Language models are few-shot learners. *ArXiv preprint arXiv:2010.11929*. [pdf] Available at: https://arxiv.org/pdf/2010.11929 [Accessed 1 Sept. 2025].

16. Boe, B., 2023. *PRAW: The Python Reddit API Wrapper*. [online] Available at: https://praw.readthedocs.io [Accessed 28 Aug. 2025].

17. DataCamp, n.d. *Data preprocessing: What it is, why it matters, and how to do it*. [online] Available at: https://www.datacamp.com/blog/data-preprocessing [Accessed 1 Sept. 2025].

18. Monash University, 2023. *Ethical implications of AI*. [pdf] Available at: https://researchmgt.monash.edu/ws/portalfiles/portal/347545354/347545058_oa.pdf [Accessed 1 Sept. 2025].

19. Sinha, A., 2023. How AI learns: Demystifying loss functions & the Adam optimizer. *Medium*. [online] Available at: https://medium.com/@akankshasinha247/how-ai-learns-demystifying-loss-functions-the-adam-optimizer-ed29862e389c [Accessed 1 Sept. 2025].

20. Sokolova, M. and Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), pp.427–437.

21. Powers, D.M.W., 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), pp.37–63.

22. Chinchor, N., 1992. MUC-4 evaluation metrics. In: *Proceedings of the 4th Conference on Message Understanding*. pp.22–29.

23. Fawcett, T., 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), pp.861–874.

24. Davis, J. and Goadrich, M., 2006. The relationship between Precision-Recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning*. pp.233–240.

25. Udurume, M., Caliwag, A., Lim, W. and Kim, G., 2022. Emotion recognition implementation with multimodalities of face, voice and EEG. *Journal of Information and Communication Convergence Engineering*, 20(3), pp.174–180. https://doi.org/10.56977/jicce.2022.20.