National College of Ireland

Detecting Depression Over Time: Fusing Emoji and Text Representations with Transformer and CLIP Architectures

MSc Research Project
MSCAI

Amalachukwu Adaeze Atusiuba
Student ID: 23293012

School of Computing
National College of Ireland

Supervisor:SHERESH ZAHOOR

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: .....Amalachukwu Adaeze Atusiuba.....

Student ID: .....23293012.....

Programme: .....MSCAI..... **Year:** ...2025.....

Module: .....Practicum2.....

Supervisor: .....SHERESH ZAHOOR.....

Submission Due Date: .....11/08/2025.....

Project Title: .....Detecting Depression Over Time: Fusing Emoji and Text Representations with Transformer and CLIP Architectures.....

Word Count: .....6067..... **Page Count**.....22.....

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....ADAEZE.....

Date: .....09/08/2025.....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	Y
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	Y
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	Y

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Detecting Depression Over Time: Fusing Emoji and Text Representations with Transformer and CLIP Architectures

Amalachukwu Adaeze Atusiuba
23293012

Abstract

Depression is a mental disorder that can severely impact an individual's life and potentially lead to self-harm. Social media has become an avenue for bringing people together to share their opinions, and this is particularly found among people who are isolated or suffer from different forms of depression. Previous research has used different Natural Language Processing and Artificial Intelligence techniques to detect depressive disorders in people on various social media platforms, including predicting whether a text is depressed or not depressed. Current detection approaches focus on isolated textual posts, overlooking the importance of temporal context and non-verbal cues, such as emojis. To address this gap, this research introduces a longitudinal multimodal framework that analyses sequences of user posts over time, combining CLIP-based emoji embeddings with Transformer-based textual models. The framework also incorporates emoji-to-emotion sentiment mapping to enhance the emotional context of each post. Our experiments compare baseline models with progressively complex architectures, showing that the Transformer + CLIP Fusion Model outperforms others, achieving 91.3% accuracy and strong recall. This system has potential real-world applications as an early intervention tool for social media monitoring, and mental health support platforms, making it a step forward in AI-driven public health solutions.

Keywords: Depression, Context-Aware NLP, CLIP, Longitudinal and Emojis

1 Introduction

In today's fast-paced society, mental health disorders such as stress, anxiety, and depression are experiencing unprecedented growth. According to the World Health Organisation (WHO), approximately 280 million people in the world are depressed[1]. Depressive disorder is a mood disorder that results from persistent feelings of sadness and a lack of interest in carrying out activities once enjoyed[2]. Untreated depression can lead to suicide. Suicide is the fourth leading cause of death in ages 15 to 29, with more than 700,000 deaths yearly. Major depressive symptoms include weariness, irritability, difficulty focusing, feelings of hopelessness, self-hatred, social isolation and abrupt weight loss or gain[3]. Depression is treatable, but factors

[1] https://www.who.int/news-room/fact-sheets/detail/depression

[2] https://www.mayoclinic.org/diseases-conditions/depression/symptoms-causes/syc-20356007

[3] https://www2.hse.ie/conditions/clinical-depression

such as lack of access to health care, limited health care providers and social stigma prevent early detection. Early detection of depression ensures that the patient has a better outcome when they are treated. Social media platforms have become one of the most effective methods for communicating opinions, thoughts, and even feelings. Depressed individuals use social media platforms to share emotions and community.

Technology can be used to alleviate undiagnosed depression and enhance early detection. Artificial Intelligence, such as Natural Language Processing (NLP), serves as a major medium of communication between people and machines. NLP is used to identify patterns, trends, perform sentiment analysis, and detect emotions. NLP can be used in identifying depression from text (Salam and Gupta, 2018). Various Natural Language Processing (NLP), Machine Learning (ML), deep learning models (Liu, 2023) and hybrid approaches have been used to predict depression in text. Ernala et al., (2019) study shows models that are solving common limitations in depression identification through social media such as Multilingual emotion detection, sarcasm detection and time-consuming extraction. See Fig.1 below for models addressing these limitations.

To identify depression via social media, AI systems need to understand all forms of human language, not just the meaning of words, but also the emotions and tones used in the words and what they represent. Emotion detection can be tricky as textual cues are not as direct as in speech, where tone, stress, expression and pitch can be noticed (Haque et al., 2023). Emojis are digital pictograms for communication. They are one of the forms of online communication and can be used to clarify opinions in a text, such as body language or facial expressions. Emojis can be used to provide non-verbal cues in depression identification.

Figure 1: Current models addressing challenges in depression detection.

Current models often predict depression from a single post or message online, but accurate detection requires analysing multiple posts over a longer time. For example, current prediction models check a text and decide if said text is depressed or not. This paper aims to determine depression through a longitudinal model by analysing not just single messages but sequences of conversations over a period. This paper also intends to use a new model, Contrastive Language-Image Pre-training (CLIP)[4] an Open AI model that learns to understand the

[4] https://openai.com/index/clip/

relationship between images and text by creating embeddings. CLIP can be used to reduce ambiguity in text by combining text and emoji for better context.

1.1 **Research Questions**

Can the accuracy of depression detection in social media be improved by using a combination of temporal and multimodal fusion model (using CLIP embeddings) compared to traditional individual (text-only) models?

1.2 **Objectives**

The primary objectives of this research are as follows:

- To develop a time-aware depression detection model that utilises ranges of social media posts rather than isolated messages.
- To apply CLIP-based embeddings for mapping both text and emojis into a unified vector space, enabling multimodal understanding of emotional context.

This paper also aims to assess the impact of temporal information in the detection of depressive signals over time by comparing the performance of the proposed multimodal time-aware model with traditional text-only models through evaluating the effectiveness of the model using quantitative metrics such as accuracy, precision, recall, and F1-score.

1.3 **Contribution of the work**

This research contributes to the academic system by beginning the first use of CLIP for mental health NLP. CLIP is used for contextual sentiment between text and emojis, presenting CLIP as a good alternative for Bidirectional Encoder Representations from Transformers (BERT). It also lays the groundwork for a realistic prediction system for early mental health intervention in digital environments by analysing and using sequences of conversations over a period to detect users' mental state of mind.

The remainder of the paper is organised as follows: Section 2 presents the literature review of old and current contributions to depression detection. This includes traditional ML, more recent DL, and NLP-based approaches. Section 3 presents the methodology adopted in this research, such as data cleaning and feature engineering performed on the dataset. Section 4 offers the implementation details of longitudinal detection and CLIP as a new model for detection. Section 5 evaluation gives an in-depth and rigorous analysis of the results, and finally, there is the conclusion and future work.

2 Related Work

Depression has become a global concern, with social media platforms increasingly seen as both a reflection of users' emotional state and a potential source of early intervention signals. The post-COVID-19 era has given an unconventional importance to digital mental health interventions. Researchers have explored various NLP, machine learning and deep learning techniques for detecting depressive patterns in online text. This literature review analyses recent peer-reviewed studies focusing on depression detection in social media using deep learning, sentiment analysis, and multimodal features. Depression identification has gone beyond traditional models (Gupta, Singh and Kumar, 2023). Machine learning should not be used in identifying depression, as it often fails to capture intricate textual patterns, limiting its effectiveness in predicting depression in text (Garg, 2023).

2.1 Deep & Transformer-Based Detection from Text

(Philip, Iyer and Nitha L, 2024) compared various deep learning models to find the optimal outcomes in identifying depression within social media. The dataset was collected from different social media platforms. Models GRU, LSTM, BiGRU and BiLSTM were compared against the modern Transformer variant Switch Transformer, an advanced transformer variant built on Mixture of Experts (MoE) architecture. Switch Transformer had better performance over RNN-based models with an accuracy of 80.9%. (Tavchioski et al., 2023) leveraged the power of large pre-trained language models on text analysis in depression detection by fine-tuning BERT, RoBERTa, BERTweet, mentalBERT and building two ensembles (averaging and Bayesian) for depression detection to accurately label new social media posts according to the level of depression. Models were trained on both Reddit and X (formerly Twitter) using Majority Classifier, TF-IDF and Doc2Vec as baseline models. A softmax layer was applied to the top of the final hidden vector corresponding to the CLS token models to return label probabilities. Where the labels with the highest probability were used to predict the level of depression in a post. Transformer-based models and their ensembles outperformed baseline methods as they improved over single transformer-based classifiers. The averaging ensemble performed better than Bayesian ensembles, having an F1-score of 0.592 on Reddit data and an F1-score of 0.859 on the X (formerly Twitter) dataset.

(Narvaez Burbano, Caicedo Rendon and Astudillo, 2025) went a step further in introducing a novel approach called DEENT (Depression Detection Based on Encoder-only Transformer). DEENT was built on extensive dimensional data post COVID-19 by applying BERT and K-means clustering to a pre-existing X (formerly Twitter) dataset labelled for sentiment analysis. DEENT was compared against Random Forest, Support Vector Machine, XGBoost, Recurrent and Convolutional Neural Networks, and MentalBERT and performed better than all with an F1-score of 90%. (Bendebane et al., 2023) predicts mental health disorders by using deep learning to develop effective multi-class models for detecting both depression and anxiety in

social media. Sufficient multi-class models were used, and a couple of hybrids, such as CNN-LSTM and CNN-BiRNN, which are also the best in terms of performance, while CNN-RNN and CNN-GRU are the best improved by involving grid search techniques. These studies show the dominance of transformer architectures in depression detection tasks.

2.2 Emotion & Sentiment Feature Integration

To solve the limitation of traditional machine learning algorithms in capturing complex textual patterns, Raj et al. (2024) used the BERT (Bidirectional Encoder Representations from Transformers) model to analyse linguistic patterns in social media posts. BERT was compared against an autoencoder, and it outperformed the autoencoder with an F1-score of 93% and an accuracy of 91.92%, surpassing the autoencoder model by 84.84%. (Kerasiotis, Ilias and Askounis, 2024) introduced a novel neural network architecture that combines transformer-based models with metadata and linguistic markers to predict depression severity into Minimum, Mild, Moderate, and Severe. The DistilBERT model was used to extract information from the last four layers, and the learning weights are then used to create a rich text representation. Metadata and linguistic markers are then added to enhance the model's understanding of each post. The EmoRoBERTa model was also used to infer emotional patterns, and the CardiffNLP Twitter-RoBERTa model for overall sentiment. The model achieved a weighted precision of 84.26% and recall of 84.18%. Data augmentation techniques such as using BERT-based insertion were used to improve the performance of the weighted F1-score from 72.59% to 84.15%.

Similarly, Qasim et al. (2025) built a framework that leverages content-based approaches (N-grams), context-based methods (Sentence Transformers), with advanced transformer-based models (BERT-base-uncased, RoBERTa-base, and Microsoft/DeBERTa-base) to predict the severity of depression, achieving an F1 score of up to 0.91. A Fast Text Convolution Neural Network with Long Momentary Memory (FCL) was used by (Obed Matías-Cristóbal et al., 2023) to increase accuracy by extracting global features and understanding local conditions. Rizwan et al. (2022) went the intensity route using small transformer-based language models. Rizwan was able to automatically flag the intensity of a negative tweet into mild, moderate, severe. 73,355 tweets were used to conduct this research on Electra Small Generator (ESG), Electra Small Discriminator (ESD), XtremeDistil-L6 (XDL), and Albert Base V2 (ABV).

2.3 Multimodal Depression Detection

Textual data dominates depression detection research, which may overlook rich behavioural cues embedded in photos, videos, GIFs or emojis. Emoji was classified by Kuldeep Vayadande et al. (2023) using CNN. Deep learning model was trained on emoji-label pair datasets for emoji classification, text datasets for mood detection and text-emoji pair datasets for emoji generation. Liu et al. (2021) incorporated emojis as an exact feature for context-awareness in depression identification. A modified Bi-LSTM called CEmo-LSTM was used for emoji-embedding. CEmo-LSTM was used over Bi-LSTM because of performance. A Chinese online chatting

platform with 38,183,194 posts, 41% containing emojis obtained via Weibo was used for the experiment. Tenfold cross-validation was used to evaluate the impact of CEmo-LSTM and compared against Lexicon sentiment and traditional models on varoius forms such as plain text, text with emoji tags and text with emoji tags replaced with sentiment words. CEmo-LSTM outperformed the combination of Lexicon sentiment and traditional models in accuracy by 95%. The influence of multimodal in depression is further studied by Gupta, Singh and Kumar. (2023) who strengthened sentiment analysis by using emojis as strong emotional indicators and integrating them into the analytical process. To prove the effect of the study, Gupta, Singh and Kumar compared various data version of tweets on the same models and as predicted all models trained on the amalgamation of emojis and text excelled more than just text or just emojis.

Solidifying the influence of multimodal in depression detection, Ali et al. (2022) like the rest Ali et al. (2022) also blended emojis with text to show that a sentence's meaning can change or be crystal clear by exploiting emojis. CNN and LSTM were used to train defined groups with the help of a function that checks the implication of an emoji in a post. 144,196 tweets were divided into three groups of plain-text, datasets without emojis and the Emojis-text. An emoji changes the interpretation of a posted is called a Reverse-to-negative or Reverse-to-positive and vice versa. BERT-based models, enhanced with emoji decoding and PHQ-9-based lexicon features, are used to predict the likelihood of a user exhibiting depressive symptoms by (Tey et al., 2023). When compared with BERT (text only) hybrid model had a 0.98 F1 score, which was higher than BERT (text only), 0.85. Collectively, these reviews confirm that enriching models with sentiment-awareness enables deeper emotional understanding.

2.4 Temporal / Longitudinal Modelling

Longitudinal model was used by (Meng et al., 2021) in electronic health record data (EHR) of patients to better identify depression. Deep learning model, BRLTM (Bidirectional Representation Learning model) with a Transformer architecture was used to predict the future diagnosis of depression in patients by capturing not only discrete depression diagnoses but also precursor medical events. The introduction of the temporal model to EHR increased precision-recall from 0.70 at baseline to 0.76. For effective capture of multi-object tracking of targets in videos, (Gao and Wang, 2023) used MeMOTR (Long-Term Memory-Augmented Transformer for Multi-Object Tracking). MeMOTR was used to solve the limitation of traditional models that limit object features between adjacent frames, making irregular movements and long-term occlusions difficult to capture. (Yue et al., 2024) also used range in speech depression detection that captured temporal variations in spoken language patterns using a hierarchical transformer based on a dynamic window and attention merge. DWAM-Former achieved a high F1 score, which was a 7.5% improvement over previous research. Although these papers do not speak directly to depression in social media, they serve as a hypothesis that temporal/longitudinal modelling can change the way depression is identified.

Author	Methodology	Strength	Weakness/Gaps
Tavchioski et al.,2023	Used Averaging and Bayesian ensembles to fine-tune BERT, RoBERTa, BERTweet, mentalBERT	Data pre-trained on a specific domain tends to perform better when used on similar types of data.	Limited input size, which can lead to a loss of information
Narvaez Burbano, Caicedo Rendon and Astudillo, 2025	DEENT-Generic and DEENT-Bert	Large dataset built on BERT and K-means	Binary classification is limited to just text
Raj et al., 2024	BERT	The model had high precision, recall, and F1-score, meaning it is reliable in minimising false positives and negatives.	interpretability in deploying a system at scale
Kerasiotis, Ilias and Askounis, 2024	A novel method combining transformer embeddings with metadata and linguistic markers	Adds context to text	
Qasim et al., 2025	N-grams, Sentence-BERT with RoBERTa	Focused on Severity Classification, which makes it scalable and adaptable for the real world	The framework is not lightweight
Kuldeep Vayadande et al., 2023	Tokenisation and CNN	93.40% accuracy	Not applicable in the real world
Ali et al., 2022	CNN and LSTM	A combination of emoji and text provided better understanding	
Rizwan et al. 2022	ESG, ESD, XDL, and ABV were compared against DistilBERT	ESG had the highest F-score of 89%	An F1-score of 89 although higer, is not precised enough for depression

			identification
Tey et al., 2023	BERT with emoji decoding and lexicon frequencies	Introduced pre- and post-depressive states detection	Reliance on Self-Reported Diagnosis makes it difficult to ascertain the model's accuracy
Meng et al., 2021	BRLTM Model	Better Interpretability	Focused on Clinical data

Table 1: Comparative analyses of models.

In conclusion, while these papers in Table 1 have largely improved early depression detection, applying these use cases to the real world has proved difficult as its implementation remains limited. Since depression develops gradually over a period, one post or message is inadequate for recognition or detection. Therefore, analysing user histories with emoji and text data offers richer, more precise insights.

3 Research Methodology

This is a quantitative research work that follows a supervised machine learning pipeline intended to detect depression in social media posts by utilising textual content, emotional indications from emojis, and a longitudinal approach that captures sequences of conversations over time, analysing changes in human emotion over time. This paper employs a combination of experimental, exploratory, and comparative methodologies with multiple baselines to achieve its goal. This paper experiments with logistic regression and transformers on progressive model complexity. It explores patterns in emoji usage, sentiment, and textual features between depressed and non-depressed users. It goes further by comparing the impact of multimodal inputs and longitudinal structuring on model performance, and then evaluates through metrics such as accuracy, precision, recall, and F1-score.

3.1 Data collection

A secondary dataset from a public repository (Kaggle) called Mental Health Twitter[5] was used in this search. The dataset consists of user-generated English tweets, consisting of text, emojis, timestamps and more. The dataset is a categorical binary classification with 20000 rows, 11 features and a target label of depressed (1) or not depressed (0). The dataset is anonymised as it is publicly accessible, with no personally identifiable information, aligning with ethical guidelines for social media research. See Figure 2 for more details on the data structure. There are 20,000 X (formerly Twitter) posts in the dataset across 72 users with an average of 277.78 tweets per person.

[5] https://www.kaggle.com/datasets/infamouscoder/mental-health-social-media/data

	Unnamed: 0	post_id	post_created	post_text	user_id	followers	friends	favourites	statuses	retweets	label
0	0	637894677824413696	Sun Aug 30 07:48:37 +0000 2015	It's just over 2 years since I was diagnosed w...	1013187241	84	211	251	837	0	1
1	1	637890384576778240	Sun Aug 30 07:31:33 +0000 2015	It's Sunday, I need a break, so I'm planning t...	1013187241	84	211	251	837	1	1
2	2	637749345908051968	Sat Aug 29 22:11:07 +0000 2015	Awake but tired. I need to sleep but my brain ...	1013187241	84	211	251	837	0	1
3	3	637696421077123073	Sat Aug 29 18:40:49 +0000 2015	RT @SewHQ: #Retro bears make perfect gifts and...	1013187241	84	211	251	837	2	1
4	4	637696327485366272	Sat Aug 29 18:40:26 +0000 2015	It's hard to say whether packing lists are mak...	1013187241	84	211	251	837	1	1

```
dataset.shape
```

Python

```
(20000, 11)
```

Figure 2: Dataset Characteristics.

3.2 Data Preprocessing

The original data format is not usable by the model training; hence, data preprocessing is needed to transform the data into a usable format for analysis and modelling. First, the data was cleaned to ensure efficient results. Such as ensuring missing values, null values or duplicate data found in the dataset, which might impact models' training and performance, were handled. The dataset had some duplicate records; once removed dataset reduced from 20,000 to 19,488 rows. The dataset was also cleaned using text normalization, that is, removing features (columns) that do not influence the model's performance. Features such as friends, favourites, statuses, and retweets were removed. Noise reduction was achieved by removing unnecessary information in tweets such as URLs, mentions, hashtags, RTs and punctuation, leaving only valuable data. After data cleaning, a word cloud of depressed words was created; see Figure 3 for more info.

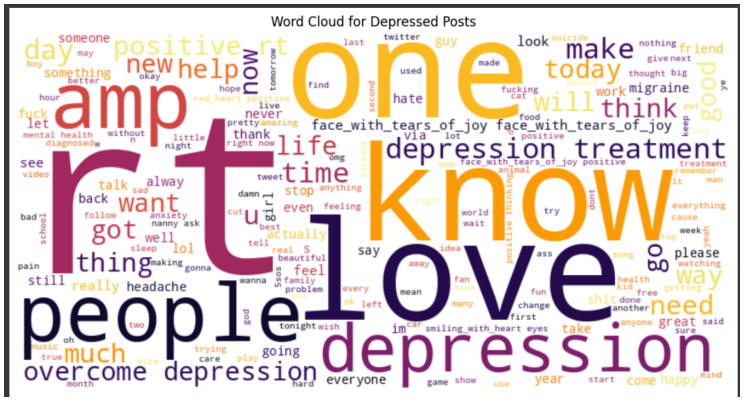

Figure 3: Word cloud of depressed words.

The original target label had an equal distribution of 10,000 depressed posts and 10,000 non-depressed posts, but during pre-processing, significant inconsistencies and data misclassification were apparent. The original dataset was labelled without context, giving depressed (1) to any tweet that had a negative phrase and vice versa. To ensure the best result, relabelling was done. The dataset was relabelled using a hybrid approach that combined

Hugging Face pre-trained emotion classification, manual emoji sentiment analysis and rule-based decision logic functions. Meaning both text and emojis were used to relabel the dataset. Random sampling validation was done to ensure accurate relabelling, and a new CSV file called “depression_dataset_with_relabel” was saved and used for sequencing work. Figure 4 below shows the dataset before and after relabelling.

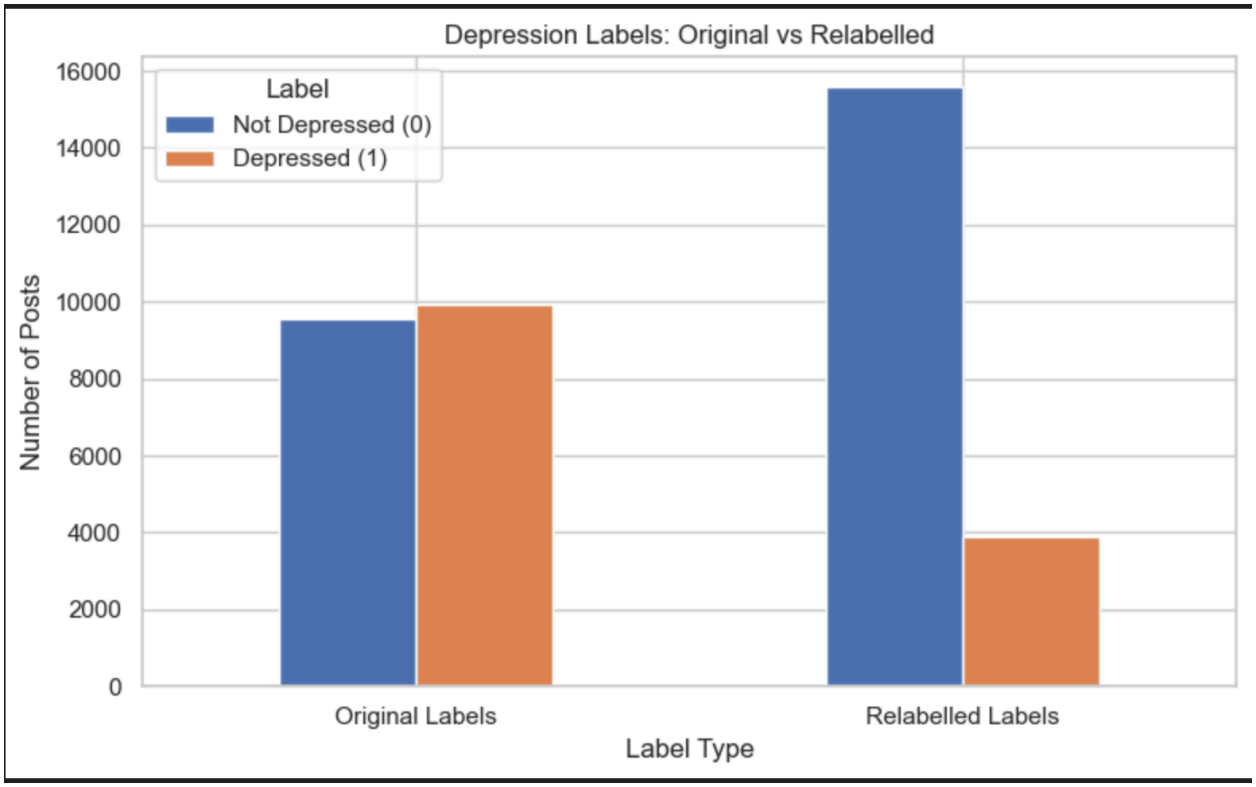

Figure 4: Depression Labels: Original vs Relabelled.

Since CLIP is used in this research for better contextual meaning between text and emojis, the Emoji Sentiment Ranking[6] dataset was used for emoji sentiment and emotion tag to extract and give meaning to emojis. Emojis are preserved in a separate column for analysing emotional cues independently, rather than as noise. Once emojis were separated, emoji demojize was used to replace emojis with textual descriptions, for example, change 😭to crying. At the end of preprocessing, the following features were added to the dataset: “clean_text_no_emojis”, “clean_text_with_emoji”, “emojis”, “emoji_sentiments” and final_text_for_model, which is a fusion of clean_text_no_emojis + emoji_sentiments. See Figure 5 below for more information.

[6] https://kt.ijs.si/data/Emoji_sentiment_ranking/

```
# Checking the final_text_for_model post_text for that specific post_id
full_text = dataset_clean.query('post_id == 637687177946734592')['final_text_for_model'].iloc[0]
print(full_text)
```

```
can't be bothered to cook, take away on the way  beaming_with_smiling_eyes  thumbs_up_medium-light_skin_tone  positive
```

Figure 5: Added features.

3.3 Data Balancing

From Figure 5 above, the relabelling of the dataset introduced imbalance in the data. Models trained on imbalanced data become biased towards the majority class, in this case, "non-depressed". To prevent bias, the dataset was balanced. Several data balancing methods were tried, including sampling, the Synthetic Minority Oversampling Technique (SMOTE), and class weight, which were all applied to baseline 1. The best approach was then used for the rest of the models.

3.4 Feature Engineering

To validate this search, different models under different complexities were applied. Scikit-learn TfidfVectorizer was used to extract important words in tweets for traditional baseline models. CLIP embedding is used for text-emoji semantics using text normalization compatible with CLIP vocabulary with vector tokens of 512-dimensional text embeddings. Longitudinal Sequence Construction was used to define sequence posts per user within defined time windows.

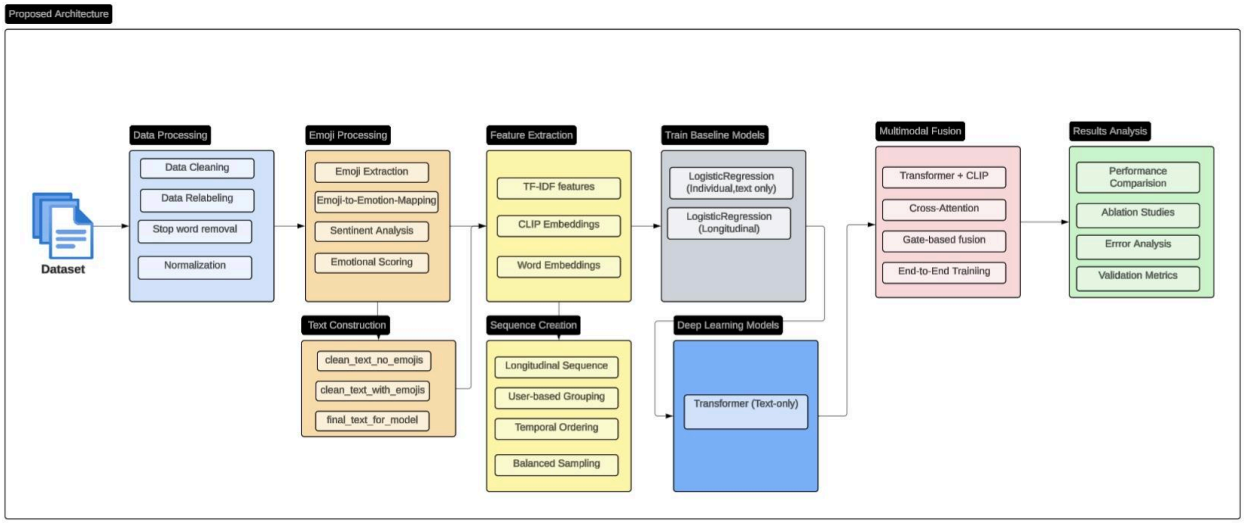

Figure 6: Structure of the proposed methodology.

3.5 Modelling Techniques

To ascertain the validity of this search, a comparison experiment was conducted on 2 models, 4 approaches using both Machine learning and Deep learning models. The dataset was trained in phrases: individual (single tweets) and longitudinal tweets (tweet sequences), both using text only (clean_text_no_emojis), then multimodal fusion consisting of Transformer + CLIP Fusion + longitudinal. This approach aims to present the current state, establish an interpretable starting

point, and subsequently demonstrate the value of longitudinal modelling in depression identification. Models were trained using an 80/20 stratified split, 80% for training and 20% for validation of model.

Logistic Regression, a machine learning model, was used as a baseline. Logistic Regression was used to train both individual and longitudinal tweets. Logistic Regression was selected as the baseline model due to its simplicity, interpretability, and proven effectiveness in text-based binary classification tasks (Hosmer Jr., Lemeshow and Sturdivant, 2013).
Transformers model a type of neural network, an advanced architecture that solves the problem of sequence transduction, or neural machine translation, is used to learn context and track relationships between sequence components[7]. Transformers are currently state-of-the-art NLP models and are considered the evolution of the encoder-decoder architecture. Transformer serves as a good baseline for this research on longitudinal modelling (Qasim et al., 2025). Transformers will be used on text-only and on multimodal fusion architecture, which is the main aim of this search; both results will be compared against the research question of this paper.

4. Implementation

The implementation of this work involved several experiments and hyperparameter tuning to obtain the best approach for depression identification. HuggingFace's CLIPTokenizer was used for tokenization to enable models analyse human language. Batch processing was done for computational efficiency. To ensure range was used to predict depression, post sequences were constructed using:

- Minimum posts per sequence: 3
- Maximum posts per sequence: 8
- Time window: 7-day sliding window.

A novel Multimodal Fusion Architecture is created using a Custom Encoder-Only Transformer + CLIP image/emoji embeddings. The transformer model was used for text processing, while Clip was used for emoji embedding processing. The transformer is combined with CLIP Fusion to identify depression in tweets and improve accuracy. To achieve this, a cross-attention mechanism was used to combine information from both text and emojis, allowing the model to contextually align emotional signals with linguistic cues. A learnable gated fusion layer is used to weigh the importance of text or emoji and learn how to prioritise each in a tweet. Residual connections are then used to retain valuable semantic and emotional features from both sources. The modal uses a joint loss function while training to optimise performance on modalities simultaneously.

[7] https://www.ibm.com/think/topics/transformer-model

Setting	Parameters	Values
Hardware Environment	GPU	Google Colab NVIDIA Tesla V100 32GB / RTX 3090 24 GB
	CPU	Intel Xeon / AMD Ryzen processors
	Memory Framework	64GB+ RAM for large sequence processing
	Framework	PyTorch 1.12, transformers 4.20, scikit-learn 1.1
Hyperparameter Optimization	Learning Rate	Grid search [1e-5, 5e-5, 1e-4, 5e-4]
	Batch Size	[8, 12, 16, 32] based on GPU memory constraints
	Epochs	Maximum 30 with early stopping (patience=8)
	Optimization	AdamW optimizer with weight decay (0.01)
	Scheduler	ReduceLROnPlateau with factor=0.5, patience=5
Regularisation and Overfitting Prevention	Dropout	0.2-0.3 across all model layers
	Gradient Clipping	Maximum norm 1.0
	Early Stopping	Validation loss monitoring
	Weight Decay	L2 regularization (λ=0.01)
Data Splitting Strategy	Train/Validation/Test Split	70%/15%/15%
	Cross-Validation	5-fold stratified cross-validation for model selection
	Class Balancing	The dataset was transformed from 80/20 to 50/50 using

		subsampling

Table 2: Training Configuration.

5. Evaluation

The performance and effectiveness of models were evaluated on 20% of the dataset. Classification evaluation metrics is used, as dataset is a classification problem (binary of depressed=0 and non-depressed=1). Accuracy was used to get the percentage of correctly predicted depression and is calculated as

$$Accuracy = \frac{TP+TN}{TP+TN+FP}$$

Precision measures the accuracy of positive detection, while percentage of actual positive predictions that were accurately predicted is calculated using recall. They are calculated with:

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

F1-score calculated as $2 \times \frac{Precision+Recall}{Precision \times Recall}$

was considered due to the imbalanced nature of mental health labels, where false negatives carry a greater social cost. Confusion Matrix is used to visualise the predictions against actual values. Ablation Study was used to understand the impact of the CLIP in research by confirming if multimodal fusion is a better way to predict depression than Text-only. If multimodal outperforms Text-only, this means that emojis provide valuable contextual information, and if there is no change or if it underperforms, we know that emojis might not be critical in detecting depression, or they need further refinement. In this section, the performance of each model and approach is validated.

5.1 BASELINE 1

Logistic Regression (Individual tweets, text-only). A balanced sub-dataset of 782 depressed and non-depressed tweets each was used to train the model on individual tweets. Baseline 1 had an accuracy of 71%. From Figure 6 below, baseline 1 correctly identified 576 out of 782 as non-depressed and 535 out of 782 as depressed. This means baseline 1 has a better ability to identify non-depressed tweets than depressed ones. Considering the importance and implications of wrong predictions, better models are needed.

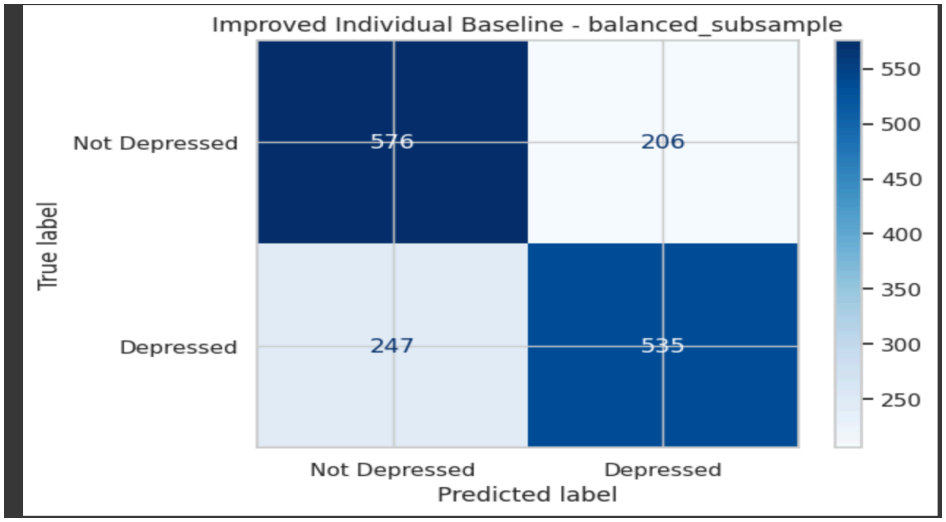

Figure 6: Confusion matrix of Baseline 1 model.

5.2 BASELINE 2

Logistic Regression (Longitudinal, text-only), baseline 2 still used Logistic Regression, but rather than individual tweets, a range of tweets is used. A total of 3294 tweets were used to train the model. Baseline 2 had an accuracy and F1-score of 86%, with better prediction of depressed than non-depressed and fewer false negatives than baseline 1, as seen in Figure 7.

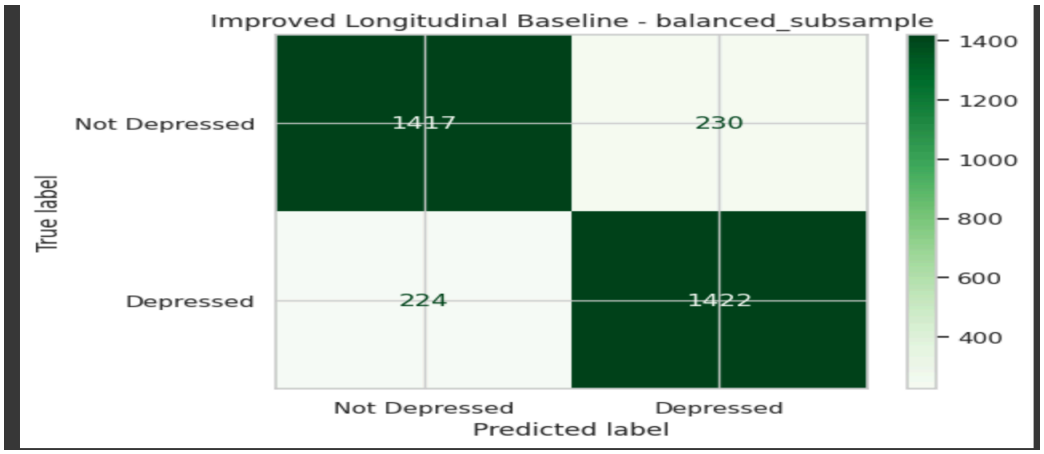

Figure 7: Confusion matrix of Baseline 2 model.

From figures 6 and 7, the following can be deduced: Baseline 2 with longitudinal tweets outperforms baseline 1 with individual tweets in accuracy by 15%, in Precision by 14%, Recall by 18% and F1-score by 16%. This implies that baseline 2 caught more depressed cases correctly and made fewer mistakes in classifying non-depressed cases as depressed. There appears to be a slight increase in false alarms in Figure 7 (206) compared to Figure 6 (230); this is acceptable.

given the much better recall and data size of Figure 6. This comparison shows that modelling the temporal progression of posts leads to a more accurate depression detection system.

5.3 BASELINE 3

Transformer Longitudinal text-only was trained on 13,169 sequences and tested on 3293 sequences. This model produced an accuracy of 91% with higher precision for non-depressed (0.95), meaning when the model says a post isn't depressed, it is 95% correct.

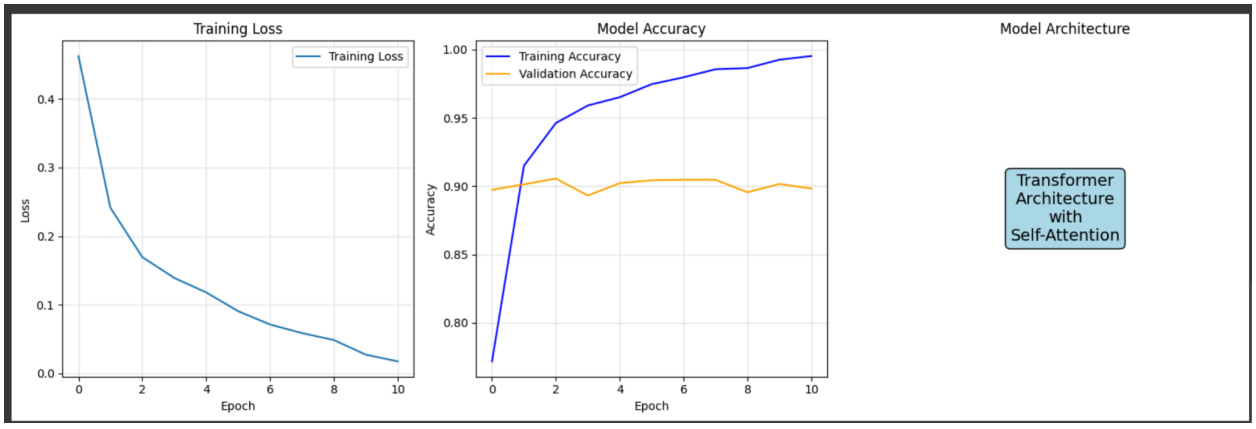

Figure 8: Loss functions of transformer longitudinal text-only model.

Training and Validation Loss is used in deep learning to identify a model's loss function, which is the difference between the predicted output of a model and the actual target value[8]. From Figure 8 above, Training accuracy improved by nearly 99% and validation accuracy was around 90%, showing a good understanding of patterns with minimal overfitting. The consistent drop in training loss indicates the model was able to learn patterns and map input features to the correct labels of depressed or non-depressed accurately.

5.4 TRANSFORMER + CLIP FUSION ARCHITECTURE

The Multimodal Fusion model was trained on a total of 16,462 sequences balanced equally (50.0%), a fusion vocabulary size of 28,237 and a CLIP embedding dimension of 512. The model produced the following result. 91.3% accuracy, 91.5% F1-score, 89.5% precision and 93.6% recall. The model is slightly better at identifying depressed posts (higher recall).

[8] https://www.geeksforgeeks.org/deep-learning/training-and-validation-loss-in-deep-learning/

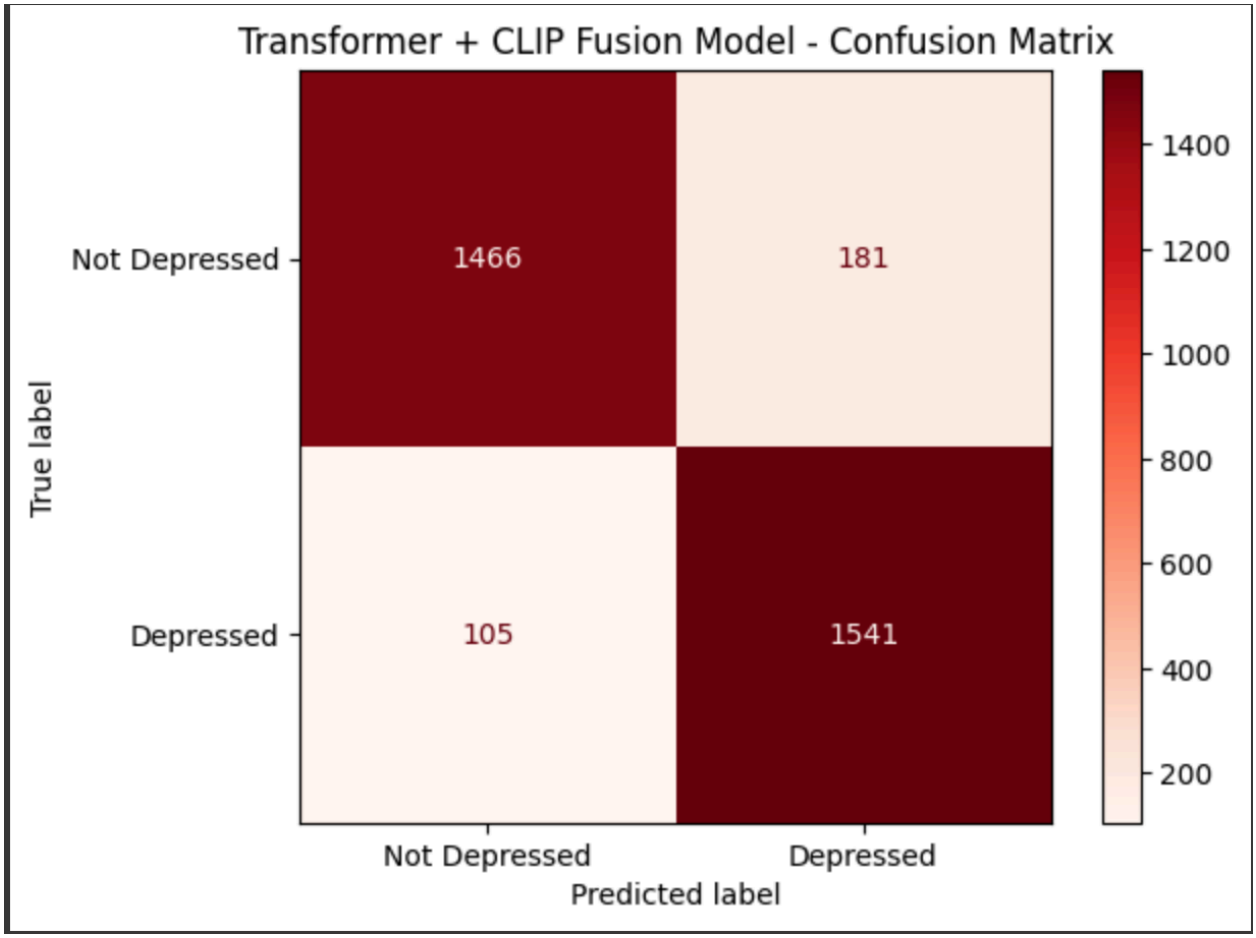

Figure 9: Confusion metric of Multimodal Fusion model.

The multimodal fusion model predicted 1466 out of 1647 non-depressed posts correctly. This means out of 1,647 posts that are non-depressed, the system accurately predicted 1,466 posts as not needing immediate mental health intervention correctly. There were 105 false alarms as the system wrongly flagged 168 non-depressed posts as depressed, which can lead to unnecessary worry. 181 tweets with depression were undetected and classified as non-depressed, which can potentially worsen mental health and lead to self-harm or suicide. The model, however, saved 1541 posts by accurately identifying depressed posts correctly.

5.5 Discussion

Although trained on the same model, Baseline 2 outperforms Baseline 1 in all metrics, Baseline 2 improved by 15% in accuracy and 16% in F1-score, which shows the impact of temporal modelling in depression detection.

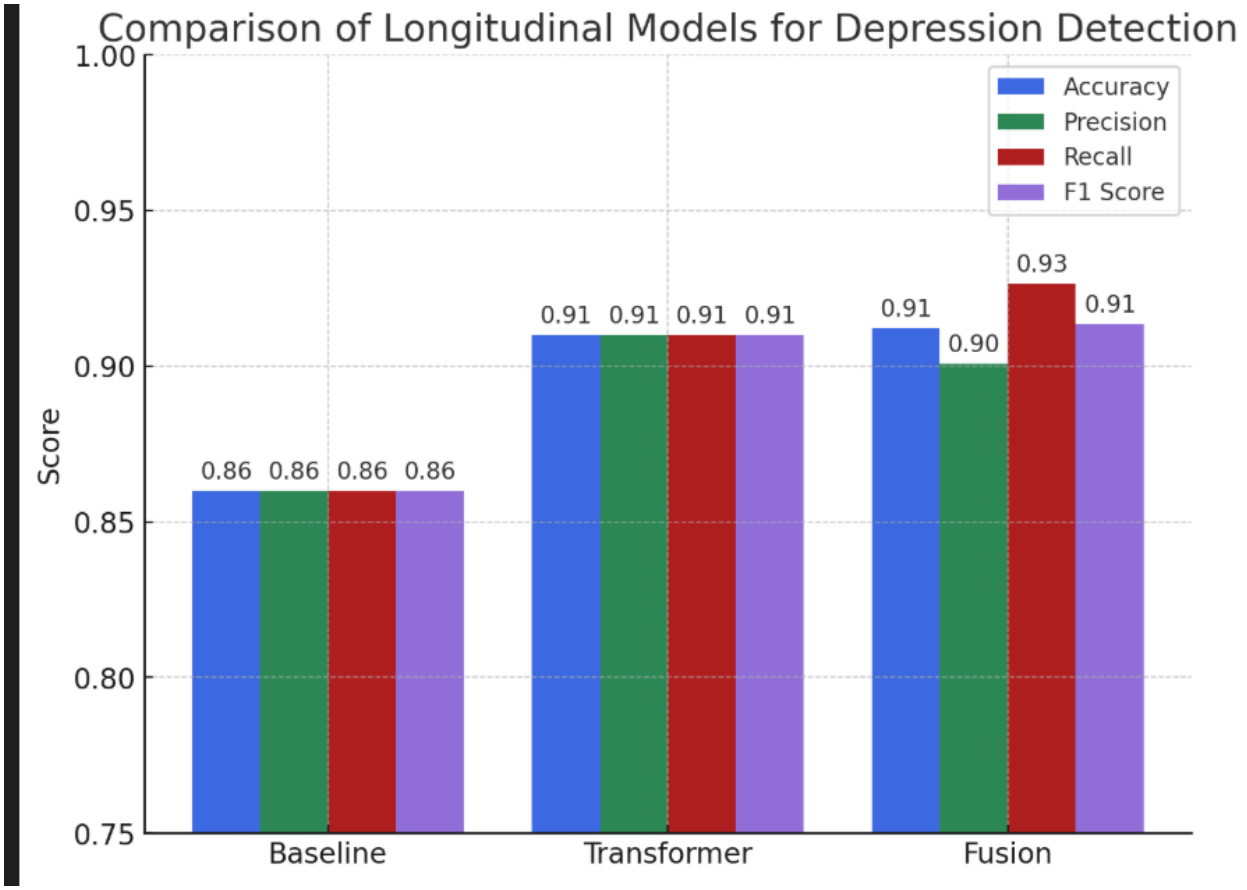

Figure 10: Comparison chart of longitudinal models

All longitudinal models outperformed individual models, from 76% accuracy of Logistic Regression (Individual tweets, text-only) to 91.3% of the multimodal fusion model. Figure 10 above, shows a 15.3% accuracy improvement from machine learning's Logistic Regression Longitudinal, text-only to deep learning transformer models. Implying transformers have attention mechanisms that understand relationships between posts over time better.

The contribution of this work multimodal fusion model slightly outperforms Baseline 3 (Transformer Longitudinal text only) in accuracy, then goes beyond accuracy by having the highest recall of 93.6% meaning fewer missed cases and identifies 3.6% more depressed individuals than model 3.

Model	Key Strengths	Observations
Longitudinal Baseline	Balanced recall/precision, simple architecture	Solid baseline, but room for improvement in

		misclassification
Transformer-Based Longitudinal	High recall for depressed class; well-learned features	Some overfitting signs between training and validation accuracy
Transformer + CLIP Fusion Model	Best balance across all metrics; multimodal inputs boost learning	Best generalisation with fewer misclassifications

Table 3: Comparison of Longitudinal Models

Figure 10 and Table 3 show that the fusion of textual and emoji (CLIP) features enhances emotional understanding. Transformer + CLIP Fusion is the best-performing model, which makes it an excellent choice for real-world deployment in early depression detection. See table 4 below for more information.

Metric	Transformer + CLIP Fusion	Transformer Longitudinal	Winner
Accuracy	91.3%	90.6%	Fusion
Precision	89.5%	89.7%	Longitudinal
Recall	93.6%	92.7%	Fusion

Table 4: Model Performance Metrics

6. Conclusion and Future Work

This research introduced a novel multimodal framework for depression detection that incorporates longitudinal text sequences with CLIP-based features, capturing the temporal progression of posts and non-verbal cues such as emojis. Unlike traditional single-post models, this longitudinal design allowed the system to observe patterns over time, which is critical for getting enough symptoms of depression. Transformer + CLIP Fusion model achieved the highest accuracy and recall compared to all baselines seen in table 4.

The higher recall in fusion model establishes that combining text, emoji, and range leads to maximising early intervention opportunities. The multimodal framework gives the best balance of accuracy and safety for real-world deployment. The results highlight the importance of multimodal and longitudinal approaches in modelling mental health indicators and demonstrate that CLIP-powered fusion architectures can offer significant advantages in this field. This work can be used as a realistic early warning system that supports healthcare systems with early identification of depression, thereby preventing worsening symptoms.

In the future, this work tends to extend to multilingualism and incorporates more forms of non-textual means of communication besides emojis.

Reference list

Bendebane, L., Laboudi, Z., Saighi, A., Al-Tarawneh, H., Ouannas, A. and Grassi, G. (2023). A Multi-Class Deep Learning Approach for Early Detection of Depressive and Anxiety Disorders Using Twitter Data. *Algorithms*, [online] 16(12), p.543. doi:https://doi.org/10.3390/a16120543.

Chauhan, S. (2024). Exploring LSTM RNNs for Depression Detection: A Deep Learning Perspective. [online] pp.536–540. doi:https://doi.org/10.1109/globalaisummit62156.2024.10947911.

Gao, R. and Wang, L. (2023). MeMOTR: Long-Term Memory-Augmented Transformer for Multi-Object Tracking. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. doi:https://doi.org/10.1109/iccv51070.2023.00908.

Garg, M. (2023). Mental Health Analysis in Social Media Posts: A Survey. *Archives of Computational Methods in Engineering*, 30. doi:https://doi.org/10.1007/s11831-022-09863-z.

Gupta, S., Singh, A. and Kumar, V. (2023). Emoji, Text, and Sentiment Polarity Detection Using Natural Language Processing. *Information*, 14(4), p.222. doi:https://doi.org/10.3390/info14040222.

Haque, R., Islam, N., Tasneem, M. and Das, A.K. (2023). MULTI-CLASS SENTIMENT CLASSIFICATION ON BENGALI SOCIAL MEDIA COMMENTS USING MACHINE LEARNING. *International Journal of Cognitive Computing in Engineering*. doi:https://doi.org/10.1016/j.ijcce.2023.01.001.

Hosmer Jr., D.W., Lemeshow, S. and Sturdivant, R.X. (2013). *Applied Logistic Regression*. *Wiley Series in Probability and Statistics*. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:https://doi.org/10.1002/9781118548387.

Kerasiotis, M., Ilias, L. and Askounis, D. (2024). Depression detection in social media posts using transformer-based models and auxiliary features. *Social Network Analysis and Mining*, 14(1). doi:https://doi.org/10.1007/s13278-024-01360-4.

Meng, Y., Speier, W., Ong, M.K. and Arnold, C.W. (2021). Bidirectional Representation Learning From Transformers Using Multimodal Electronic Health Record Data to Predict Depression. *IEEE Journal of Biomedical and Health Informatics*, 25(8), pp.3121–3129. doi:https://doi.org/10.1109/jbhi.2021.3063721.

Narvaez Burbano, R., Caicedo Rendon, O.M. and Astudillo, Carlos.A. (2025). An Encoder-Only Transformer Model for Depression Detection from Social Network Data: The DEENT Approach. *Applied Sciences*, 15(6), p.3358. doi:https://doi.org/10.3390/app15063358.

Obed Matías-Cristóbal, Jesús Padilla-Caballero, Gonzales-Rivera, R., Benavente-Ayquipa, R., Segundo Pérez-Saavedra and Frans Cardenas-Palomino (2023). Enhancing Sentiment Analysis In Text Of Social Media Texts Using Hybrid Deep Learning Model And Natural Language Processing. *6th International Conference on Contemporary Computing and Informatics*, 8(DOI: 10.1109/IC3I59117.2023.10397710), pp.1776–1780. doi:https://doi.org/10.1109/ic3i59117.2023.10397710.

Philip, A.T., Iyer, R.R. and Nitha L (2024). Analysing Depression in Social Media: A Study of Basic Deep Learning Algorithms and Transformer Model with a Comparative Approach. *Third International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)* , [online] (DOI: 10.1109/ICSTSN61422.2024.10670781), pp.1–6. doi:https://doi.org/10.1109/icstsn61422.2024.10670781.

Qasim, A., Gull Mehak, Hussain, N., Gelbukh, A. and Sidorov, G. (2025). Detection of Depression Severity in Social Media Text Using Transformer-Based Models. *Information*, [online] 16(2), pp.114–114. doi:https://doi.org/10.3390/info16020114.

Raj, A., Ali, Z., Chaudhary, S., Bali, K.K. and Sharma, A. (2024). Depression Detection Using BERT on Social Media Platforms. *2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, (979-8-3503-8969), pp.228–233. doi:https://doi.org/10.1109/iicaiet62352.2024.10730329.

Tavchioski, I., Robnik- Sikonja, M. and Pollak, S. (2023). Detection of depression on social networks using transformers and ensembles. doi:https://doi.org/arXiv%20preprint%20arXiv:2305.05325,%202023.

Tey, W.-L., Goh, H.-N., Lim, A.H.-L. and Phang, C.-K. (2023). Pre- and Post-Depressive Detection using Deep Learning and Textual-based Features. *International Journal of Technology*, 14(6), pp.1334–1334. doi:https://doi.org/10.14716/ijtech.v14i6.6648.

Yue, X., Zhang, C., Wang, Z., Yu, Y., Cong, S., Shen, Y. and Zhao, J. (2024). Hierarchical transformer speech depression detection model research based on Dynamic window and Attention merge. *PeerJ Computer Science*, 10, p.e2348. doi:https://doi.org/10.7717/peerj-cs.2348.