# Multimodal Depression Detection: Leveraging Textual and Image-Based Social Media Data

by

MD. Tazbid Tahshin
21201044
Fardeen Mohammad Monayem
21301350
Vaskor Debnath
21301211
Md. Tasrif Khan
21301351

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
June 2025

# Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

**Student's Full Name & Signature:**

| | |
|---|---|
| MD. Tazbid Tahshin<br>21201044 | Fardeen Mohammad Monayem<br>21301350 |
| Vaskor Debnath<br>21301211 | Md. Tasrif Khan<br>21301351 |

# Approval

The thesis/project titled "Multimodal Depression Detection: Leveraging Textual and Image-Based Social Media Data" submitted by

1. MD. Tazbid Tahshin (21201044)
2. Fardeen Mohammad Monayem (21301350)
3. Vaskor Debnath (21301211)
4. Md. Tasrif Khan (21301351)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 28, 2025.

**Examining Committee:**
Supervisor:
(Member)

**Dr. Jannatun Noor Mukta**
Associate Professor
Department of Computer Science and Engineering
Director, Bachelor of Science in Data Science
United International University

Program Coordinator:
(Member)

**Dr. Golam Rabiul Alam**
Professor
Department of Computer Science and Engineering
School of Data and Sciences
Brac University

Head of Department:
(Chair)

**Dr. Sadia Hamid Kazi**
Associate Professor
Department of Computer Science and Engineering
School of Data and Sciences
Brac University

# Ethics Statement

This is to declare that the following research work has been conducted within all ethical guidelines. Confidentiality of users will strictly be maintained to ensure privacy of their personal data. As a result, the ethical aspects of research will be closely adhered to in this study. Above all, we promised to avoid plagiarizing at all costs.

# Abstract

Depression is a prevalent and serious mental health condition characterized by persistent feelings of sadness, loss of interest, and significant impairment in daily functioning. The increasing global burden of depression highlights the urgency of developing more effective and timely detection methods. Traditional diagnostic methods face limitations such as subjective assessments, limited access to professional care, societal stigma, and delays in early detection, underscoring the necessity for innovative and accessible approaches. To address these challenges, our goal was to develop a robust multimodal machine learning framework to predict early depression interventions through textual & visual data obtained from social media platforms. By integrating user-generated textual posts with visual information from shared images, this study leverages advanced computational models, particularly CLIP, Time2Vec, BLIP-2, and Cross-Modal model, to capture intricate emotional and behavioral patterns associated with depressive symptoms. CLIP's capability to align textual and visual representations, combined with Time2Vec's efficiency in capturing temporal dynamics, significantly enhances the accuracy and depth of depression detection. Also, BLIP-2 frozen state time-sensitive effectiveness and cross-modal integration of fundamental embedding techniques helped us to compare and evaluate the model's effectiveness. This multimodal approach effectively combines insights from both data modalities, overcoming the limitations of single-modality analyses and providing a better comparative understanding of users' mental states. This research thoroughly discusses and addresses various challenges related to data quality, ethical considerations, and complexities associated with real-world multimodal datasets. Ethical handling of sensitive personal data, maintaining user privacy, and confirming transparency in algorithmic prediction decision giving which are critical aspects that have been meticulously considered throughout this study. The findings underscore the potential of social media as an effective platform for proactive mental health monitoring, early intervention, and the promotion of mental well-being. Further research is recommended to incorporate additional modalities like video and speech or audio, enhance real-time decision-making capabilities and ensure linguistic and cultural inclusivity. Addressing these future directions will significantly increase the practical applicability and generalizability of depression detection systems, paving the way for broader societal benefits and improved mental health outcomes.

**Keywords:** Depression detection, Multimodal machine learning, social media, CLIP, Time2Vec, Cross Modal, BLIP-2, Bi-directional, Advance Cross Modal, ethical concerns, proactive intervention.

# Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.
Secondly, to our supervisor, Dr. Jannatun Noor Mukta ma'am, for her kind support and advice in our work. She helped us whenever we needed help.
And finally, to our parents, without their support, it may not be possible. With their kind support and prayer, we are now on the verge of our graduation.

# Table of Contents

**Declaration** i

**Approval** ii

**Ethics Statement** iii

**Abstract** iv

**Acknowledgment** v

**Table of Contents** vi

**List of Figures** 1

**1 Introduction** 2
- 1.1 Research Problem . . . 3
- 1.2 Reseach Objectives . . . 4
- 1.3 Scope and Challenges . . . 7
  - 1.3.1 Scope . . . 7
  - 1.3.2 Challenges . . . 7
  - 1.3.3 Contribution . . . 8

**2 Literature Review** 9
- 2.1 Image-based Research . . . 9
- 2.2 Text-based Research . . . 12
- 2.3 Video-based Research . . . 21
- 2.4 Speech-based Research . . . 24
- 2.5 Multimodal Research . . . 27

**3 Background** 31
- 3.1 Data Acquisition . . . 31
- 3.2 Dataset Structure . . . 31
  - 3.2.1 Positive and Negative User Categories . . . 31
  - 3.2.2 User Folders . . . 32
- 3.3 Dataset Preprocessing . . . 34
- 3.4 BERT (Bidirectional Encoder Representations from Transformers) . . . 37
  - 3.4.1 BERT Architecture . . . 37
  - 3.4.2 Mechanism . . . 38
- 3.5 CLIP (Contrastive Language-Image Pre-training) . . . 39
  - 3.5.1 Model Architecture . . . 39

3.5.2 Loss Function: Contrastive Loss . . . 40
3.5.3 Classifier: Softmax Activation . . . 41
3.6 BLIP-2: Bootstrapped Language-Image Pretraining Version 2 . . . 41
3.6.1 BLIP-2 Architecture . . . 42
3.6.2 Data Flow Process . . . 43
3.6.3 Mathematical Framework . . . 43
3.6.4 Mechanism . . . 43
3.7 Time2Vec . . . 44
3.7.1 Model Architecture . . . 45
3.7.2 Mechanism . . . 46
3.7.3 Dimensionality Reduction in Time2Vec . . . 46
3.7.4 Comparison with Traditional Time Embeddings . . . 46
3.8 GRU (Gated Recurrent Unit) . . . 47
3.8.1 GRU (Gated Recurrent Unit) Architecture . . . 47
3.9 Lime Xai . . . 48
3.9.1 LIME Xai Working Procedure: . . . 48

**4 Research Methodology 49**
4.1 CLIP Model Methodology: . . . 49
4.1.1 Working Procedure . . . 51
4.1.2 Computation . . . 51
4.1.3 LIME Xai Integration . . . 52
4.2 Cross-Modal Model . . . 53
4.2.1 Textual Data Processing via BERT . . . 53
4.2.2 Visual Data Processing via CLIP . . . 54
4.2.3 Cross-Modal Attention . . . 54
4.2.4 Temporal Encoding . . . 54
4.2.5 Model Computation . . . 56
4.3 BLIP-2 . . . 56
4.3.1 Data Flow . . . 57
4.3.2 Computation . . . 59
4.3.3 Hyperparameters . . . 60
4.4 Time2Vec Model . . . 61
4.4.1 System Flowchart Description . . . 62
4.4.2 Working Procedure . . . 62
4.4.3 CNN integration: . . . 63
4.4.4 LSTM integration: . . . 63
4.4.5 Data Flow . . . 64
4.4.6 Computation . . . 65

**5 Experimental Evaluation 66**
5.1 Performance Metrics Benchmarking . . . 66
5.2 Metrics Comparison and Discussion . . . 66
5.2.1 CLIP . . . 66
5.2.2 BLIP-2 . . . 68
5.2.3 Time2Vec . . . 69
5.2.4 Cross-Modal . . . 70
5.2.5 Advanced Cross-Modal . . . 71
5.3 Discussion . . . 72

5.3.1 Insufficient Regularization and Data Sparsity . . . . . 72
5.3.2 Ineffective Multimodal Fusion . . . . . 73
5.3.3 Data Quality . . . . . 73
5.3.4 Training Regime and Optimization . . . . . 73
5.3.5 Vanishing Gradient in Temporal Encoding . . . . . 73

**6 Limitation and Future Work** **75**
6.1 Limitation . . . . . 75
6.2 Future Work . . . . . 75

**7 Conclusion** **77**

**Bibliography** **82**

# List of Figures


3.1 Total Number of images: Depressed vs Non-depressed . . . . . 32
3.2 Total Number of Tweets: Depressed vs Non-depressed . . . . . 33
3.3 Tweet Frequency Over Time (Depressed vs Non-depressed Users) . . 33
3.4 Dataset Preprocessing Pipeline . . . . . 35
3.5 Preprocessed Dataset . . . . . 36
3.6 BERT Architecture[15] . . . . . 38
3.7 CLIP Architecture[32] . . . . . 39
3.8 BLIP-2 Architecture[53] . . . . . 42
3.9 Time2Vec Architecture[61] . . . . . 45

4.1 CLIP Model Architecture . . . . . 50
4.2 Cross-Modal Model Architecture . . . . . 55
4.3 BLIP-2 Model Architecture . . . . . 58
4.4 Time2Vec Model Architecture . . . . . 61

5.1 CLIP model Metrics Graph . . . . . 67
5.2 LIME with CLIP . . . . . 68
5.3 BLIP model Metrics Graph . . . . . 69
5.4 Time2Vec Metrics Graph . . . . . 70
5.5 Cross-Modal Metrics Graph . . . . . 71
5.6 Advanced Cross-Modal Metrics Graph . . . . . 72

# Chapter 1

# Introduction

Depression is a mental health problem with extensive consequences for both individual well-being and the overall functioning of society. Over the past several years, the explosion of social media sites has added complexity to the etiological landscape of mental illness, including depression. Although social media provides a venue for connection, self-expression, and community, it has also been linked to the cultivation of symptoms of loneliness, anxiety, and lower self- esteem. Such adverse effects are particularly concerning among susceptible groups, such as adolescents, who are even more exposed to social comparison, cyberbullying, and pressure to uphold idealized virtual selves.

It is increasingly important to understand the relationship between using social media and depression in this age, in which online activities are deeply integrated into everyday life. While depression diagnosis and treatment have conventionally depended on clinical environments, social media has opened up new possibilities for the identification and exploration of indicators of the disorder. Social media websites like Twitter are now mainstream outlets where people share their opinions, feelings, and lives as they unfold. Social media has thus emerged as a treasure trove for researchers exploring depression expression and incidence. Tweets—image tweets and text tweets—are thus potentially a window into the internal worlds of subjects, including those who have not yet received professional attention or an official diagnosis.

The relationship between using social media and depression is complicated and cannot be explained in terms of a cause-and-effect relationship. Social media does contribute to depression but is mostly a platform where people freely share emotional pain. Most people with depressive symptoms use social media platforms for communicating sadness, loneliness, or hopelessness. The fact that these postings are public and, in many cases, unedited means that patterns of behavior or linguistic choice that are indicative of underlying mental illness can be identified. Liu et al. (2022)[41], for example, recognize that an additional hour spent on social media links to a 13% increase in the likelihood of depression among adolescents. This result highlights the value of social media content, either self-reported or observed by others, which acts as a window into the user's emotional and psychological state. Social media in this case offers an innovative opportunity for early exploration and detection of depressive symptoms from the digital footprint of users.

The impact of depression, however, reaches far beyond the individual, touching on larger societal elements such as social relationships, public health, and economic stability. Depression can significantly hinder one's capability to have healthy relationships, moving forward further with sensations of isolation. The persistent stigma regarding mental health still discourages many from getting the help they require, and too many endure in silence. At work, depression is among the top causes of disability, resulting in decreased productivity, absenteeism, and overall poorer performance. At societal levels, un-treated depression imposes a significant economic burden, raising healthcare utilization and decreasing quality of life. Most alarmingly, perhaps, the disorder is also heavily linked with suicide; Simon and VonKorff (1998) [1] found that approximately 15% of individuals who have been diagnosed with depression later die by suicide, highlighting the potentially fatal consequences of inadequate treatment.

Here, the significance of social media as an indicator of depression cannot be emphasized enough. Through tracking trends in the online behavior and textual content of its users, mental health researchers and practitioners are able to better understand the lived experience of depressed individuals. For instance, patterns of frequent negative language, expressions of hopelessness, or even changes in posting behavior can be markers of depression (H Shannon et al., 2022) [42]. As social media is now the prevailing means of communication, its use as an indicator of mental health is a field of growing significance. The utilization of these internet sites for early detection could reach individuals who would otherwise remain unidentified and result in early intervention and assistance.

## 1.1 Research Problem

While depression is a very serious concern in contemporary health care, a social stigma of mental illness persists, making people hold back from consulting health care professionals, with serious consequences. Akbari et al.(2016) [4] reported that over 70% of patients with incipient depression are hesitant to consult a mental health practitioner.

Rising global rates of depression, especially among adolescents, emphasize the urgency for effective detection and intervention. Social media platforms, while providing a space for connection, also allow individuals, particularly adolescents, to express feelings of sadness, hopelessness, and isolation, which can be indicative of depression. These platforms offer an opportunity for early intervention, as identifying depressive symptoms through social media posts can help mitigate the condition's severity. Moreover, early intervention improves treatment outcomes and reduces the economic burden of untreated depression. Kuramoto-Crawford [6] reported that 42% of adults with mental illness in the USA cannot afford treatment, while Abdin et al.(2023) [48] highlighted the additional cost of $3938.9 per person annually due to mental health issues.

The challenge of depression detection is compounded by the time-consuming and subjective nature of traditional diagnostic methods. Social media provides a real time, behavioral indicator of a person's emotional state, offering a means to detect

depression early. The absence of diagnosis and treatment significantly elevates suicide risk, with Tang et al. (2022) [45] noting that many individuals who commit suicide have never been diagnosed with a mental illness. Rehm and Shield (2019) [19] found that one out of every eight people suffers from mental health conditions, and Bachmann (2018) [10] demonstrated that individuals with depression have suicide rates ten times higher than the general population.

Untreated depression also severely impairs daily functioning, reducing productivity by up to 35% as noted by Chowdhury and Sumiya (2022) [36]. Grazier 2019) [16] estimated the global economic loss from depression and anxiety to be $1 trillion annually. To mitigate risk and improve individuals' quality of life early detection performs an essential role.

Most studies focus on a single disorder, but depression often coexists with other mental health conditions. Harris (2023) [52] argued that depression is frequently a primary cause of other disorders, which emphasizes the need for a comprehensive approach. While mental health detection much like 81% of it uses social media posts, most of it relies solely on text, which does not fully capture the complexity of depression. Khoo et al. (2024) [57] suggested that combining text, images, and user interactions is crucial for effective depression detection. To achieve more accurate understanding to analyze one's mental heath, use both textual and visual content can not denied.

Adolescents, especially those active on social media, are particularly vulnerable to depression. Social media platforms often exacerbate feelings of inadequacy, making early detection even more crucial.

To address these challenges, the study utilizes multimodal data, including text and images, to detect depression more effectively. By integrating both textual and visual content from social media posts, the study aims to improve detection accuracy and provide real-time, proactive intervention, ultimately enhancing depression management and treatment outcomes.

## 1.2 Reseach Objectives

The study seeks to quantify the mental health of users of social media through the conduct of multimodal analysis of data posted by users. Previous studies have established the basis for enhancing depression detection models to be more efficient and precise with the addition of multiple forms of data to facilitate earlier diagnoses of depressive illnesses. Although 81% of studies on detection of mental disorder are based mainly on social media text data, users nowadays tend to post a multi-modal collection of data types mostly texts but also images, audio and video clips to show their mood and psychological status. This provides a huge loophole in the complete evaluation of mental wellness because present models tend to ignore the important information contained in non-text data.

Depression is likely to begin subtly, and individuals will tend to express their sadness, hopelessness, and isolation in subtle and not necessarily overt means. These early expressions can be communicated through both textual and visual social media posts—anything from the tone of a caption to the subject of a photograph or video. But the majority of current detection models look only at textual information, excluding these faint, multimodal signals. Consequently, early signs of depression might be overlooked, losing chances for early treatment and support.

This study aims to analyze the gap between the convention contemporary method and fill it by creating a more integrated model that incorporates both textual and visual data. By examining not just the words users post, but also the visual information they convey: facial expressions, color, and affective tone of images or videos—a multimodal model provides a more complete and accurate insight into a person's state of mind. It heightens the likelihood of early detection by capturing the indirect and multi- layered way people convey their distress on the internet.

Second, incorporating this model on social media sites could allow instant analysis of user-generated content, with subtle warning flags being detected as and when they occur. Computer systems based on these flags can highlight potential signs of depression, sending timely alerts to users or mental health workers—without violating user privacy or personal space. As Akbari et al. (2022) [4] discuss, mental health conditions such as depression remain undiagnosed at the onset since symptoms are too easily discounted—particularly if they are communicated indirectly or non-verbally. Indeed, a diagnosis is overlooked in close to 70% of instances. By addressing this vital deficiency, the current study aims to consolidate early detection mechanisms and enable individuals to get the help required before the condition becomes more severe.

**The aims of the current research are as follows:**
The research aims to assess the mental health of social media users by using a multimodal analysis of user-posted data. Prior research has laid the foundation for improving the efficiency and reliability of depression detection models by incorporating various types of data, thus enabling earlier diagnoses of depressive disorders. Although 81% of research on mental disorder detection primarily relies on social media textual data, modern social media users frequently share multimodal data formats like image, text, video as well as audio, to express their emotional and mental states. This creates a significant gap in the comprehensive assessment of mental health, as current models often overlook the valuable insights embedded in non-textual data.

Depression, in particular, is often subtle in its early stages, with individuals expressing their feelings of hopelessness, sadness, and isolation indirectly through both text and visuals. Current models that rely solely on textual data fail to capture these nuanced expressions, leading to missed opportunities for early intervention. By integrating both visual and textual content, this research seeks to bridge this gap, offering a more holistic and accurate understanding one's mental state. This presented multimodal approach enhances the ability to detect depression by analyzing

both the words users choose and the visual cues they share, such as images, facial expressions, or even the mood conveyed in videos.

Moreover, the integration of this model into social media platforms could enable analysis of users' posts by real-time monitoring, allowing for timely interventions. Automated analysis could identify early warning signs of depression, alerting individuals or health professionals to the potential onset of depression without compromising user privacy or violating personal space. This approach respects the need for confidentiality while providing a proactive mechanism for depression detection. Akbari et al. [4] noted that mental illness, including depression, often goes undiagnosed in its early stages because its warning signs are subtle and easily overlooked. This results in missed diagnoses approximately 70% of the time, especially when the signs are expressed through indirect or non-textual means. By addressing this gap, the research aims to improve the early detection and intervention strategies for depression, potentially reducing the severity and impact of the condition by providing earlier support to individuals in need.

**The objectives of this research are as follows:**

1. **Enhance Depression Detection System by Combining Text and Visual Information**

   Enhance the precision and efficiency of depression detection by integrating textual and visual content of social media posts. The multimodal approach aims to represent more complete understanding of one's midset, improving early detection and intervention strategies.

2. **Contrast the performance of various Machine Learning Multimodal Models for Depression Detection**

   Various machine learning models— such as CLIP, BLIP2, and ViLT— are compared for depression detection from Twitter data. Compare each model on parameters such as accuracy, computational efficiency, and scalability to find the most effective model that can be utilized for real- time analysis.

3. **Recognize Early Warning Signs of Depression to Avoid Escalation**

   Identify precursory symptoms of depression in social media messages, noting word-based and image-based cues that could represent emerging depressive symptoms. It is in a bid to avert escalation to more harmful outcomes like suicide or self-harm.

4. **Social Media as a Screening Tool for Depression**

   Assess the efficacy of social media data as a non- invasive screening process to determine individuals who are at risk of depression. This study will determine if social media analysis can augment conventional screening processes and offer an extra level of proactive intervention.

5. **Facilitate Ethical Management of User Information Without Compromising Privacy**

   Examine the ethical consequences of utilizing social media data for depression identification, specifically highlighting how confidentiality and privacy of the users will be protected during the study. The study will take into consideration best practices in handling sensitive mental health data to ensure trust and fulfill privacy requirements.

Finally, by spanning the divide between technology and mental health, this study aims to transform how depression is diagnosed and treated so that it is identified early when it can make a real difference in people's lives and overall well-being.

## 1.3 Scope and Challenges

### 1.3.1 Scope

This research is interested in investigating the possibility of utilizing multimodal social media data—Twitter, in this case—to identify early indications of depression. Through the integration of visual and text information, the study hopes to enhance the precision and timeliness of the depression detection system. Primary interest will lie in experimenting and evaluating multimodal learning models, namely CLIP, BLIP-2, Time2Vec, Cross-Modal, and Advanced Cross-Modal, to ascertain how well they can interpret social media posts for signs of depression.

Moreover, the research entails examining if social media data may serve as an effective, non-intrusive early warning system for detecting individuals at risk of depression with a scalable approach to early detection. The research faced significant ethical issues while working with social media data for mental health surveillance, with privacy and confidentiality being ensured as a prime consideration in the research.

Furthermore, the potential for incorporating these depression detection systems within current social media sites will be explored based on ongoing, real-time monitoring without an intrusion into users' privacy. Finally, this study will offer perspective on how best technology can be employed to detect, prevent, and intervene in depression at the earliest point of incidence in order to optimize mental health consequences and reduce the wider societal effect of depression going untreated.

### 1.3.2 Challenges

There are a number of serious challenges that must be overcome to successfully utilize social media data for depression detection. Among the most significant challenges is the reliability and quality of the data itself. Social media data is generally subjective, unstructured, and biased. Individuals' postings can be vague, sarcastic, or contradictory, and therefore hard to decipher in terms of actual emotional state.

In addition, people may knowingly or unknowingly conceal or fake their emotions, resulting in false positives or neglect of depression indicators. This variability is challenging to create robust models with good performance on heterogeneous populations and properly detect depression.

Ethical issues also have a key function in such challenges. While social media is rich with information, the use of this information for depression detection brings about grave concerns of user privacy, consent, and data protection. It is absolutely essential that mental health information is processed in the most ethical manner possible and in strict adherence to privacy laws. The potential for abuse or misuse of data also highlights the necessity for transparency and accountability while developing and implementing such systems.

Finally, scalability is another significant challenge. While this research deals with Twitter data, depression detection models need to be scalable to other social media sites where user behaviors and content differ. Ensuring that the system can process enormous volumes of data from diverse sites without affecting the accuracy will be crucial for the system to be successful.

### 1.3.3 Contribution

Most of the works done on detecting depression only detected depression without showing any type of explanation.

1. In the thesis LIME XAI was used to see which part of the text and which part of the image contributed to detecting depression by running the multimodal model. This would assist the medical professionals to detect depression more efficiently.

2. Besides, BLIP-2 was not used in depression detection, even though its previous version, BLIP, was used.

3. Moreover, we fine-tuned the CLIP model for better results.

4. Additionally, the dataset was preprocessed in a different way to remove biases from the dataset as the dataset was very biased.

# Chapter 2

# Literature Review

In the age of science, Social media has profoundly shaped people's lives. People tend to use social media to escape from reality, where they can express their feelings and thoughts without thinking about the rest of the world. Here they can find their true selves. So, evaluating one's mental condition through how they express themselves is easier. Naslund et al. (2020) [24] stated that individuals who are suffering from one or more mental disorders 70% of middle-aged and older individuals, to upwards of 97% of younger generations, overuse social media platforms. Birnbaum et al. (2017) [5] showed in their report that adolescents and young adults aged 12 to 21 with psychotic disorders overuse social media platforms, exceeding 2.5 hours per day. Berryman et al. (2018) [11] conducted a study that examined 467 young adults monitoring their usage of social media, and tendency to engage in vague booking (intentionally vague but highly personal and emotional) to prove the correlation between the usage of social media and mental health conditions. They also concluded that the outcome was considered to be symptoms of general mental health conditions such as suicidal ideation, loneliness, and social anxiety. Rehm and Shield [19] stated that according to a report by the World Health Organization (WHO), one out of eight people suffers from one or more mental health conditions. So, a vast number of studies are being conducted in this sector. Zhang et al. [47] in their review revealed that 81% of research in this area on mental illness detection is based on social media posts.

## 2.1 Image-based Research

Li et al. (2023) [54] introduced a Hybrid Multi-Head Cross Attention Network (HMHN) for depression detection from facial images. The model uses Deep Feature Fusion (DFF) and Grid-Wise Attention (GWA) for the extraction of low-level features and Multi-Head Cross Attention Block (MAB) and Attention Fusion Block (AFB) for the interaction of high-level features. The approach effectively extracts depression features from facial areas in parallel and has good performance on not only AVEC 2013 but also AVEC 2014 datasets, where the RMSE values is 7.38 and MAE is 6.05 (AVEC 2013) and RMSE of 7.60 and MAE of 6.01 (AVEC 2014). The model has some limitations. It relies heavily on facial expressions, which may not fully represent depression in patients with mild or no visible signs. Small sample sizes of datasets limit generalizability, and model complexity may influence their

real-time application in clinical practice.

Kong et al. (2021) [39] proposed a Deep Convolutional Neural Network (CNN) model to identify depression from facial images. It used a 3D CNN architecture to learn spatial and temporal facial features and obtained 98.23% accuracy for depression detection. It identified significant facial features like eye movements and mouth expressions, which are extensively associated with depression. However, the research had some limitations. The model was trained from a small set of clinical facial images, and the generality of the model to a larger population can be undermined. Second, relying only on facial expressions can fail to detect cases where depression is not visible, and model interpretability can be an issue of clinical acceptability.

Reece and Danforth (2017) [7] predicted depression based on 43,950 photographs of 166 individuals using features such as color (hue, saturation, brightness), filter use, faces per photograph, and posting frequency from Instagram photographs. This study developed a logistic regression model of depression prediction from visual characteristics of Instagram photos with particular focus on color and texture patterns. The model was roughly 70% accurate in detecting depression through image characteristics like low brightness and high saturation in the photos of depressed individuals. The study had some limitations, including a small sample of 166 participants, lack of multimodal data (e.g., text and interaction data), and low interpretability of results. Despite these limitations, the study identified a promising potential for using social media photos in detecting depression.

Guo et al. (2022) [37] proposed an automatic depression detection model based on solely visual data through a Temporal Dilated Convolutional Network (TDCN) for capturing long-range temporal features (e.g., head pose, facial landmarks) of video sequences, which also employs a Feature-Wise Attention (FWA) module for multi-stream feature representation fusion. Tested on the DAIC-WOZ dataset, this vision-only model established state-of-the-art performance with an F1 of around 0.80, surpassing previous vision-only approaches. Good as it was, the limitations of this model are its use of controlled interview data (DAIC-WOZ) that may limit generalizability to real-world applications, possible overfitting to certain patterns in the data, and not leveraging other modalities such as audio or text that could still contribute to depression detection accuracy.

Nepal et al. (2024) [58] presented MoodCapture, a smartphone-based depression detection system using over 125,000 front-camera selfies of 177 participants who were clinically diagnosed along with PHQ-8 scores. They used 711 facial landmark features mapped with OpenFace and featured a Random Forest classifier that yielded 75% to 91% balanced accuracy across different subgroups and an MCC of 0.70. For regression, it correctly predicted PHQ-8 scores with low error, though the exact MAE was not given. A deep learning approach based on EfficientNet-B0 (fine-tuned from blocks 6–7) was also attempted but was not as good as the Random Forest, highlighting the advantage of structured facial features over raw image data in this task. SHAP analysis showed that depression prediction was influenced by some 3D facial landmarks and head pose changes. Limitations in spite of strong performance included a clinically homogeneous and modest-sized sample, privacy concerns re-

lated to selfie taking in everyday life, and reduced generalizability to everyday or non-clinical environments.

Table 2.1: Image-based Research Comparison

| Ref | Paper Title | Model/ Framework | Result (based on dataset) | Limitations |
|---|---|---|---|---|
| [54] | A facial depression recognition method based on hybrid multi-head cross-attention network | CNN and HMHN | AVEC 2013 outperformed with RMSE=7.38 and MAE=6.05 | Limited on the diversity of facial expressions due to less availability of datasets. |
| [39] | Automatic Identification of Depression Using Facial Images with Deep Convolutional Neural Network | FCN, VGG11, VGG19, ResNet50, Inception-V3 | FCN performed best here with 98.23% accuracy and precision of 98.11% | Small Dataset with clinical facial images used |
| [7] | Instagram photos reveal predictive markers of depression | Bayesian Logistic Regression, Random Forest Classifier | The model correctly identified 70% of all target class cases with 60% accuracy in Bayesian Logistic Regression and 70% accuracy for Random Forest | Due to privacy concerns out of 509 around 221 which is 43% of the ratio didn't want to share their data. Which caused a lack of data here. |
| [29] | Sch-net: a deep learning architecture for automatic detection of schizophrenia | CNN, CBAM | Sci-net performed 97.68% accuracy and even got to 99.52% on LANNA children speech database. | Indicators like hallucination, forgetting, movement disorder for Schizophrenia were not explored here. Limitation of assessing disease severity. Datasets not containing psychological/neurological disorders. |

| | | | | |
|---|---|---|---|---|
| [58] | MoodCapture: Depression Detection Using In-the-Wild Smartphone Images | Random Forest Classifier, EfficientNet-B0 | FCN performed best here with 98.23% accuracy and precision of 98.11% | Small dataset, privacy issues, and Limited generalizability issues. |

Image-based research papers shown in Table 2.1, convolutional architectures dominate: Li's [54] Hybrid Multi-Head Cross-Attention Network fine-tunes CNN backbones with hierarchical attention, Kong [39] selects a 3-D CNN to dig out spatiotemporal information, and Guo [37] uses a Temporal Dilated CNN to spread temporal context—highlighting a community-wide presumption that facial dynamics are more informative than appearance. Simpler baselines exist as well: Reece [7] uses logistic regression over hand-designed Instagram colour and texture statistics, and Nepal [58] shows that a Random Forest on 3-D front-facing facial-landmark features can outperform an EfficientNet, a reminder that traditional models are still in the running when feature engineering is sophisticated. Reported performance varies from 70% accuracy [7] to 98% [39], with Li's [54] reporting RMSE 7 and MAE 6 on AVEC, Guo's [37] F1 0.80 on DAIC-WOZ, and Nepal's [58] system's balanced accuracy up to 91% + MCC 0.70. But any given study has common limitations: small or homogeneous cohorts limit external validity; reliance on overt facial information may miss mild or masked depression; and privacy or acceptability concerns loom over camera-based approaches. Most models also ignore complementary modalities (speech, text, physiology) that may disambiguate ambiguous visual cues. And complex attention blocks and deep 3-D convolutions pose deployment problems to real-time or resource-constrained environments, highlighting a trade-off between accuracy and practicality that the next wave of multimodal, efficiency-focused research must overcome.

## 2.2 Text-based Research

Jaman et al. (2021) [31] addressed the pressing issues of clinical mental illness such as bipolar disorder, PTSD, depression and schizophrenia. The study proposed a comparison of ML, DL and GRU(RNN) models based on NLP to accurately predict mental health state. The study also included the societal stigma regarding mental illness and pointed out the limited access to professional help. The authors described the detrimental effects of undiagnosed mental illnesses which was their main motivation. Due to societal stigma, individuals are reluctant to seek help bringing severe consequences to themselves and others. The study also highlighted the significance of social media platforms which enabled individuals to talk about their insecurities and feelings maintaining confidentiality of the user. So the authors used a dataset of Reddit users who expressed their mental health conditions along with self-reported diagnosis. The dataset included posts from both affected individuals and a control group of mentally healthy users. The pre-processing of the data was done by removing stopwords, tokenization, lemmatization and TF-IDF Vectorization. In order to improve overall performance metrics of the existing models a comparison of Logistic Regression, GRU, Support Vector Machine (SVM) and BERT was performed to pre-processed dataset. The study achieved Key Performance Indicator (KPI)

among all models, especially SVM. With the help of better hardware and use of BERT LARGE and LSTM with embedding techniques on existing models, future work can be done to produce more promising results.

Sun and Wu (2021) [33] focused on using NLP to identify mental health problems using social media posts due to the increasing importance of early detection to avoid foreseen risk factors. The research aims to study how advanced NLP with Deep Learning Models could improve effectiveness of existing system. This study emphasizes collecting and preprocessing data as the model performance deeply relies on the quality of the dataset. TalkLife and Dreaddit were the two existing datasets used in this study. Besides, the authors collected data from Twitter and Reddit which are directly connected with mental illness. The process was done manually by experts to preserve accuracy. The preprocessing of the dataset is done by conversion to lowercase, removal of unwanted and stopwords, and expansion of contractions. Five models including Random predictor model, sentence2vec, Deep learning LSTM model, pre-trained BERT, fine-tunes BERT were implemented on the dataset. The LSTM model with embedding layer used tokenization to predict mental health state. The pretrained and Fine-tuned BERT model used word embedding technique with the weight assigned to different words. These three techniques served promising results in terms of accuracy in the detection of mental health. Among all the models, fine-tuned BERT outperformed other models as it uses task-specific labeled data enabling specialized performance. The authors proposed of improvement of the fine-tuned BERT model with a larger training dataset to accumulate higher accuracy and provide real-time analysis using other indicating factors of mental illness.

In this study of Vasha et al.(2023) [56] they have examined the dataset by at first applying data mining to categorize whether it was depressive or non-depressive. Furthermore, they have collected the data in Bengali language from Facebook posts and comments. Then labeled them in TF & IDF vectorizer format to apply in ML algorithms. The TF & IDF vectorizers basically used to convert the string keywords to numeric value.This section will explore how sentiment analysis, a related NLP technique, is used to understand user opinions and emotions expressed in social media text. We will discuss how topic detection complements sentiment analysis by providing a deeper understanding of the underlying themes within social media conversations. Moreover, a hybrid classification made of two classifiers (SVM and NB) has been used for the betterment of study by getting more precise and perfect accuracy. Here an extra column was created to erase the duplicate data mainly. Lastly, to bring the research above with a future vision ML algorithms were significantly used.

The study here of Gupta et al.(2021) [30] analyzes the bio signal data of individuals from IoMT devices and compares with the collected dataset from social media dataset. The majority of the information contained in bio-signals taken from the human body is invisible to the naked eye and even to specialists. These bio-signals are captured by implanted and wearable IoMT devices for the diagnosis of mental health conditions. Psycholinguistic hints in the written word at the same time. Using RNN (recurrent neural network) shows the most accurate ratio here compared to other models in the IoMT dataset. On the other hand, algorithms like SEA,

BNet & CNN+BiLSTM have been applied on social media tweets having the highest accuracy here on this social media dataset. Additionally, the whole research was conducted on three states- suicidal ideation, disturbed mental state & emotional mental state. From the detection of mental health conditions the ratio for social media dataset was 5:3:1 in the earlier mentioned three states whereas for the IoMT dataset the ratio was mainly carried out on the disturbed mental state. The main issue to achieve real time healthcare analysis was the lack of privacy for the users of social media.

The study here was conducted on text from social media using supervised and unsupervised learning ML approaches where SL was more adopted than UL(Liu et al., 2022) [40]. Here SL uses classifications to forecast the presence of psychiatric disorders like depression and anxiety. On the contrary, UL forecasts sets of unlabeled data of clustering and compression phenomena. Additionally, UL was used to analyze the underlying psychological dimensions whereas for higher-level analyses SL used the existing information in the feature dataset. The SL approach included both regression and classification. Where it was found that the behavioral aspects compared to texts achieved the highest F1 score. In UL MDL(multimodal depressive dictionary learning) obtained the best performance having the highest F1 score of 85%. The research has been considered upon the privacy and ethical issues about data protection. Lastly, using a deep learning approach over other classifiers it gained over 94% of accuracy on depression detection.

According to Babu and Kanaga(2021) [35] Using multi class classification with deep learning algorithms it shows the highest precision value during the sentiment analysis. Here based on two polarities positive and negative classes the sentiment was classified (also named as binary classification). But in ternary classification three polarity classes were used like positive, negative and neutral class. Besides, they have implemented the data of the users who often use emojis or emoticons to express their feelings in social media. And to preprocess them NLP and tokenizer to separate the emoji's from text was used here. Moreover, they have collected the data by utilizing the hash tags. Although Binary and Ternary classifications are good, they have used multi-class classification utilizing deep learning algorithms for better precise classification results in case of texts and emoticons. Also, to make the data polarized into sentiment classes deep learning and sundry machine learning algorithms were used. Lastly, having applied the combination of CNN-LSTM algorithms on the depression datasets gave the higher precision value over here.

Since most of the people do not express their depression emotions in social media directly(Chiong et al., 2021) [28]. In this research both single and ensemble models of ML classifiers predict extracted input features from the text itself using textual features process. Furthermore, BOW was used to decompose into single words from the entire text. To handle the imbalanced data dynamic over and under sampling process was implemented. Besides all the approaches were done using the CV (10-fold Cross Validation) approach and each experiment was repeated 10 times. Finally, it was concluded that Linear Regression outperformed the optimum performance with both approaches. However, comparing for better results of depression detection Random Forest showcased results better compared to others using general

text.

In this study of Islam et al.(2018) [12] they have used an online public source in order to collect the facebook data. The approach here was supervised learning techniques, mainly LIWC (linguistic inquiry and word count). The main motive of their research was to detect depression on emotional terms. It was shown that the hourly depression patterns with the highest rate were AM rather than PM based on the average trend for all comments. In addition, on a seasonal trend the highest depression detection was observed during both August and September in the AM. Lastly, upon all the classifiers results from midnight to midday it was found to be between 60% to 80%.The deficiency here in depression analysis was needed while extracting various types of emotional features. The study also does not identify sufferers who actually are.

The study over here of Wshah et al.(2019) [20] about PTSD was conducted on nonlinear ML techniques, mainly SVMs with nonlinear kernels and RF. Additionally, to combine predictions from multiple models ensemble techniques were used over here. As a result, the main aim of the study is to support better early interventions by developing computational methods to moreaccurately identify at-risk patients. The dataset used here was on the basis of critical care service individuals of 90 persons having 11 main features and a clinical cutoff of PTSD for a score of 33. From the results compared to the other single classifiers the ensemble methods of SVM with a Gaussian kernel got the highest over them having an AUC of 0.85 over 1 month of elevation. Here the nonlinear ensemble method models surpassed single linear models. Lastly, the research will provide better results for real clinical settings for at-risk patients. Besides additional data like demographic variables, PTSD symptoms (baseline).

According to Orabi et al.(2018) [13], the detection of depression from Twitter users is done using the RNNs and CNNs (deep learning techniques). Word embedding optimization and comparative evaluation processes were used for this purpose. Word embedding optimization focuses on classification to identify the users suffering from depression from their tweets from a dataset CLPsych2015. They also focus on comparative evaluation to compare the performance of deep learning architectures used on NLP tasks to detect the mental health problems.

López-Úbeda et al.(2019) [18] focuses on anorexia detection from twitter messages using NLP technologies and machine learning. It was seen that due to using these technologies early detection of mental disorders were possible and these techniques can be used for anorexia detection. But anorexia detection was done on spanish messages where the dataset was trained by providing answers to the BDI questionnaire. Hence, using NLP with ML models can help in early detection of mental health disorders. Furthermore, we'll explore methods like Bag-of-Words (BoW) feature engineering and word embeddings to capture the relationships between words in social media text. These features are crucial for extracting the most meaningful insights from the data.

According to the study of Asad (2019) [14], The mental health problem depression

was detected from the Twitter posts of the user. For this purpose data was collected from two of the most famous social media platforms which are facebook and twitter. A semi-structured interview method was used for this purpose. Then the machine learning method was used to process the data from the users from social networking sites. NLP(Natural Language Processing) was used along with SVM(Support Vector Machines) and NB(Naive Bayes) theorem was conducted to detect depression in the most effective way. Among the SNS users, the tweets were collected using a python package called Beautiful Soup. The first username is fetched from the command line and then a request is sent to the twitter page where an HTML tag enclosed the tweets. Inside the ¡p¿ tag of HTML the actual tweet lies which contains the paragraph. On the other hand, for the facebook data of the users from the past one year was collected using the JSON format. Then they were converted to CSV from JSON files. For data pre-processing NLP is used. The reason is that all the data which was collected were in string format. But to train machine learning models we need a numerical feature. For this reason tokenization was used. But the limitation of this study is that the depression detection has been done in only one language. Depression detection from other languages cannot be done from this study.

The study of Bae et al.(2021) [26] suggests that detection of Schizophrenia is possible using ML approaches along with NLP(Natural Language Processing) techniques. For this study the database was collected from Reddit consisting of six non-mental health related subreddits. For data preprocessing NLP was used. All the words were converted to lowercase letters and to split the sentences into words the posts were tokenized. After that the specific keywords like "schizophrenia" and words beginning with "schizo" were removed.Only the original posts were included and the comments were excluded. Then machine learning models like Support Vector Machines(SVM), LR, NB and RF were used. The performance of the models were evaluated using the evaluation metrics called accuracy, precision, recall and F1 score. But the limitation of this study is that schizophrenia is detected from only reddit posts. It does not detect schizophrenia from facebook and twitter type popular social media platforms.

According to Chen et al. (2023) [50], depression is a mental health concern that is widely spread globally, affecting about 3.8% of the population(Detecting Reddit Users With Depression Using a Hybrid Neural Network SBERT-CNN, 2023). People are increasingly using social media platforms such as Reddit to share their mental health problems like depression and seek community help online. The use of deep learning methods gained popularity due to its effectiveness, user-friendliness and high performance in different NLP tasks. Then, this study's hybrid deep learning model combines a sentence BERT (SBERT) which has been pre trained with a convolutional neural network (CNN) to identify people with depression from their individual posts on Reddit. SBERT captures semantic information in posts while CNN converts embeddings further and temporally identifies user behavioral patterns with accuracy 0.86 and F1 score 0.86 surpassing previous machine based models for depression detection . The findings demonstrate that the hybrid model can effectively recognize depression individuals through various social media information, having implications for wider text classifications and clinical applications. These questions ask you quickly: what the prevalence of depression is and how the hybrid model works.

Keyvanpour et al.(2022) [38] stated that most new neural network-based methods tend to choose a linear rectifying activation function for hidden layers, as emphasized by Bottou in 2012. Convolutional operator layers of the proposed approach make use of this activator. Convolutional operator layers employ a linear rectified activator which plays an important role in feature extraction in neural networks . Random gradient descent has been discussed by the author as one way to iteratively optimize the objective function. Among other things, this optimization technique involves calculating the minimum of a function using only randomly chosen parts of the training data . Random gradient descent provides a randomized descent approximation of the gradient decrement which is the edge toward finding minimum or optimal solution to train neural networks efficiently.

The article of Bouarara (2021)[27] is about how social media affects mental health negatively and is based on a British study that focused on this matter amongst persons from 14 to 35 years . It aims to identify psychological disorders in users of social platforms to help them overcome illness like anxiety, phobia, depression and paranoia. In this research tweets are analyzed using RNN (recurrent neural network) in order to find possible threats, risks of suicide, loneliness as well as other related problems with an accuracy of 85% which is performing better in comparison to other techniques such as social cockroaches, decision tree or naïve Bayes . The results achieved through the RNN method were verified by means of experimental measures like f-measure, recall, precision, entropy and accuracy . This paper stresses the necessity of making use of modern technologies like RNNs aimed at combating mental health issues relating to the use of social platforms thereby demonstrating the efficiency of these methods in detecting and possibly reducing online mental health problems.

The study of Apoorva et al.(2023) [34] discussed about applying social media for detecting depression in young adults early, especially Twitter, and the importance of reporting symptoms to self for this mental health condition. This work utilizes deep learning models such as RNN (Recurrent Neural Network) and LSTM (Long Short-Term Memory) to analyze linguistic markers in tweets to detect depressive users. In this case, the LSTM model performed better than the simple RNN model with an accuracy of 96.21%. It is argued that Twitter can be used to accumulate a lot of data through which efficient deep learning models can be trained to understand an individual's state of mind far ahead of traditional methods . That is why both RNN and LSTM models were trained on sentiment140 dataset and Twitter dataset (obtained via twitter API). By doing so, it was expected that these models would improve their ability to accurately recognize depressed individuals based on their tweet texts.

Sumathy et al. (2022) [44] have covered in their study some of the basic terms like Opinion mining, sentiment analysis which play an essential role for the prevalent approaches of NLP to analyze text-based emotions. Opinion mining is also described as identifying subjective information within a piece of text - it relies on the text characteristics such: subjectivity, opinion target, opinion owner and the particular time the opinion was created on. Generally, sentiment analysis is a type of text clas-

sification, which classifies emotions in the free text. Several types of methods exist for text classification such as machine learning methods, lexical methods, or hybrid methods , and usually the choice depends on annotated data available. Importance of valence lexicons, corpora, and dictionaries for sentiment analysis is reflected with English being a predominant language, mainly because of specific resources available Challenges exist in sentiment analysis for languages other than English, such as Tamil, due to the lack of resources like lexicons and corpora, requiring the development of new language resources to improve sentiment analysis accuracy.

Table 2.2: Text-based Research Comparison

| Ref | Paper Title | Model/ Framework | Result (based on dataset) | Limitations |
|---|---|---|---|---|
| [30] | Real-Time Mental Health Analytics Using IoMT and Social Media Datasets: Research and Challenges | RNN | XGBoost technique with 95.71% and based on questionnaires 83.87% | Challenging to find datasets to detect suicidal ideation where most of the studies were based on disturbed and non disturbed mental state. |
| [50] | Detecting Reddit Users with Depression Using a Hybrid Neural Network SBERT-CNN | SBERT-CNN | F1 score of 0.86 similar to the exact accuracy result. | Limited to validate only reddit posts for detection. |
| [40] | Detecting and Measuring Depression on Social Media Using a Machine Learning Approach: Systematic Review | Supervised Learning(SL), Multimodal Depressive Dictionary Learning(MDDL) | F1 score of 85% | Lack of Textual data from users on social media |
| [33] | Early detection of mental disorder via social media posts using deep learning models | Senetence2vec, LSTM, BERT | 84% for fine-tuned BERT model | Biased data, Depends only on text, privacy issues |
| [38] | Deep learning-based text analysis for depression illness detection on social media posts | Deep, CNN and LSTM | This Deep method identifies with 73% F1, 78% precision and 70% recall score. | Lack of extracting temporal and sequential nature of posts from social networks |

| | | | | |
|---|---|---|---|---|
| [31] | Utilization of Machine Learning Classifiers to Predict Different Forms of Mental Illness: Schizophrenia, PTSD, Bipolar Disorder and Depression | SVM, BERT and GRU(RNN) models based on NLP | SVM (Depression: 82, Bipolar: 83, Schizophrenia: 80, PTSD: 87), BERR (Depression: 72, Bipolar: 75, Schizophrenia: 73, PTSD: 77), GRU (Depression: 62, Bipolar: 61, Schizophrenia: 56, PTSD: 60) | Lack of high-end gear |
| [56] | Depression Detection in Social Media Comments Data using Machine Learning Algorithms | SVM and NB, TF-IDF Vectorizer applied on multiple classifiers | SVM 75.15% and NB 72.19% accuracy rate | Lack of using deep learning methods and shortage of datasets due to the overwhelming privacy policy of social media. |
| [35] | Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review | Multiclass Classifier, NLP, Tokenizer, CNN-LSTM algorithms | CNN-LSTM with 92.06% of highest accuracy | Withal of emoticons or emojis without utilizing it to have sentiment score values. |
| [28] | A Textual-Based Featuring Approach for Depression Detection using Machine Learning Classifiers and Social Media Texts | BOW, 10-fold CV | RF with the highest rate of accuracy 96.11%. | Limitation to work on Unsupervised classifiers or unlabelled datasets. |
| [14] | Depression detection by analyzing social media posts of user | SVM and NB classifier | Questionnaire results are between 20-30% range and NB precision got 1 as the highest result and accuracy result of 74% in NB. | Limitation of language problem (only one language and except english no other language has been tested) |

| | | | | |
|---|---|---|---|---|
| [20] | Predicting Post-traumatic Stress Disorder Risk: A Machine Learning Approach | Ensemble methods of SVM & RF | Ensemble methods of SVM outperformed having AUC of 0.85 over others | Lack of additional data like- trauma histories, symptoms of baseline PTSD etc |
| [12] | Depression Detection from Social Network Data using Machine Learning Techniques | KNN, DT, SVM | KNN, SVM, Precision, Recall and DT results are between 60-80% | Extract more types of emotional features in depression analysis. |
| [13] | Deep Learning for Depression Detection of Twitter Users | RNN, CNN | CNN model reports with 87.957% and RNN with 83.236% performance | Lack of sufficient annotated training data. Models unoptimized can not give much accurate performance. |
| [18] | Detecting Anorexia in Spanish Tweets | Six classification models. MLP and SVM | Both MLP and SVM are the same as 0.938 with the default settings in the Sci-kit-learn | Limited to textual information didn't work on rhetorical figures(sarcasm or irony). |
| [27] | Recurrent Neural Network (RNN) to Analyze Mental Behaviour in Social Media | RNN | RNN accuracy of 85% compared to other techniques. | Limited identification into only psychological disorder problems. |
| [26] | Schizophrenia detection using machine learning approach from social media content | ML(SVM,NB,RF) and NLP | RF got the highest 0.969 then SVM 0.907, LR 0.896 and lastly NB 0.868 | Limited to only one reddit post and doesn't detect schizophrenia from popular social media platforms like Facebook and Twitter. |
| [34] | Depression Detection on Twitter Using RNN and LSTM Models | RNN and LSTM model | LSTM here outperformed with 96.21% having 0.1077 validation loss | Limited to tweet data information only. |
| [44] | Retracted: Machine Learning Technique to Detect and Classify Mental Illness | Naive Bayes, SVM | 47.94% for Naive Bayes and 58.96% for SVM | Unclear evaluation protocols, domain constraints, language constraints |

As shown in Table 2.2, across this diverse collection of text-based mental-health studies, some modeling trends and traps reappear. Shallow classic classifiers such as SVM, logistic regression and Naïve Bayes are still ubiquitous—again, often as baselines—but now recurrent nets (GRU, vanilla RNN, LSTM) and transformer models (BERT, fine-tuned BERT, SBERT) are the leading reused architectures. Hybrids of pretrained sentence-level embeddings augmented with CNN or BiLSTM layers yield the best balance between context-sensitivity and efficiency, with F1 0.85 to 0.96 on Reddit or Twitter datasets. Even more simple feature-engineered pipelines, though—e.g., TF-IDF vectors to SVM or a Random-Forest—can compete with deep nets when data are scarce or noisy, e.g., Bengali Facebook update or IoMT biosignal–text comparisons. Common traps are: small or language-limited datasets constrain generalisability; reliance on self-reported labels is high, so ground-truth bias risk is high; sparsity and class imbalance require ad-hoc sampling; and privacy or ethics restrict real-time deployment, especially where peoples' timelines are trawled without consent. Finally, most studies are unimodal, ignoring multimodal cues audio, images, sensor streams that might enhance robustness, and few focus on model interpretability beyond token-level saliency, so clinicians do not trust black-box predictions.

## 2.3 Video-based Research

Addressing the issue of depression Meshram et al. (2023) [55] proposed Convolutional Neural Network (CNN) by analyzing facial visual frames extracted from recorded interviews to detect patterns that can be mapped to depression detection. The system used KNN for textual analysis and random forest for dimensionality reduction and regression. The proposed multimodal approach improved by 2.7% compared to existing frameworks. Multimodal analysis using visual and textual data provided a comprehensive assessment and diagnosis accuracy. However, the quality of the assessment heavily relies on image data extracted from videos that may biased impact on overall prediction. If the continuity of the frames can be preserved, the system can more accurately capture facial expressions and increase robustness of the system.

Singh and Kumar (2022) [43] focused on their study about SAD detection from video surveillance using ResNet-101, a deep convolutional neural network (CNN). Their study demonstrated that video surveillance can be effectively used for early detection of mental health disorders like stress, anxiety, and depression by analyzing facial expressions. The dataset used in this study included facial images from video surveillance, processed to identify and classify emotional states associated with SAD. They applied advanced image preprocessing techniques, such as Kalman filtering and bilateral filtering, to improve facial image quality, addressing challenges like lighting and motion blur. Moreover, the study explored the use of nonlinear ensemble methods, including SVM with a Gaussian kernel, which outperformed traditional linear models, achieving an AUC of 0.85 after one month (Singh and Kumar, 2022). This research highlights the potential of combining deep learning with video surveillance for real-time, unobtrusive monitoring of mental health conditions. However, challenges such as reliance on facial expressions alone and environmental factors affecting image quality remain, suggesting a need for multimodal approaches

to further enhance model accuracy.

Ashraf et al. (2020) [22] present a review of the utility of machine learning (ML) for depression detection in static image and video data. They identify the use of facial expressions, eye movements, and posture as visual characteristics from which to determine if any signs of depression could be obtained. ML techniques are proposed with various algorithms, such as CNN and SVM, which provide feature extraction and improve the detection accuracy. The review highlights the importance of using diverse datasets and standardized measures for the evaluation of multimodal processes to ensure reliable outcomes. It proposes that multimodal data usage, for example, a video for emotion with a speech, text, or combination of physiological signals, may improve a model's predictive ability. While improvements have been made, Ashraf et al. (2020) report that issues of data quality should be considered in addition to generalizability in developing a model for machine learning, and further research is needed to allow for its general use in real-world contexts.

Zhu et al. (2019) [21] study how physiological responses can reveal users' emotional responses to videos depicting depression. Their study demonstrated that while users' conscious recognition or recollection of depression was low (27% accuracy), features from physiological signals galvanic skin response and pupil dilation, yielded a 92% classification accuracy using neural networks. This study suggests the possibility of a physiological automated tool for depression detection, to be a more consistent way of obtaining data than subjective narration. The study also highlighted the challenge of conscious recognition and suggested fusion of multimodal data could enhance the effectiveness of the system.

Table 2.3: Video-based Research Comparison

| Ref | Paper Title | Model/ Framework | Result (based on dataset) | Limitations |
|---|---|---|---|---|
| [43] | Detection of Stress, Anxiety, and Depression Using ResNet-101 in Video Surveillance | ResNet-101, CNN, SAD | Accuracy 98.4% in detecting SAD from facial expressions, SVM Gaussian kernel ensemble method AUC of 0.85 after one month of monitoring | The study's reliance on facial expressions alone may overlook other mental health indicators and be affected by environmental factors. |
| [46] | Exploring Hybrid and Ensemble Models for Multiclass Prediction of Mental Health Status on Social Media | BERT and RoBERTa, BiLSTM neural networks | RoBERTa outperforms BERt by 1.8% and F1 score here with a maximum of 77% in stress detection obtained | Limitation to only textual features, not incorporating with the behavioral activity data on mental health status |

Table 2.4: Speech-based Research Comparison

| Ref | Paper Title | Model/ Framework | Result (based on dataset) | Limitations |
|---|---|---|---|---|
| [29] | Sch-net: a deep learning architecture for automatic detection of schizophrenia | CNN, CBAM | Sci-net performed 97.68% accuracy and even got to 99.52% on LANNA children speech database. | Indicators like hallucination, forgetting, movement disorder for Schizophrenia were not explored here. Limitation of assessing disease severity. Datasets not containing psychological/neurological disorders. |
| [21] | Detecting emotional reactions to videos of depression | CNN and RNN | Physiological responses were able to classify 92% of the emotional reactions to videos of depression, but only had a low level of conscious recognition (27%) | Low conscious recognition of depression and imbalanced Dataset with less interdependency between classes. |

From the Table 2.3, the papers reviewed for depression detection from videos all used deep learning models, with Convolutional Neural Networks (CNNs) being most common across studies. The models were most commonly used to extract and analyze face features from video data, supplemented in some studies with approaches like K-Nearest Neighbors (KNN), Support Vector Machines (SVM), or ensemble approaches for additional accuracy in classification. Multimodal solutions combining textual or physiological signals with visual cues performed better than unimodal systems, as evidenced by higher detection accuracy. Singh and Kumar (2022) [43] demonstrated the potential of high-resolution facial data from surveillance video, when used in combination with nonlinear models such as SVM using Gaussian kernels, to perform very well with an AUC of 0.85. However, a chronic issue in studies continues to be with the heavy dependence on video and facial expression data quality, which can be affected by environmental factors such as lighting or motion blur. In addition, models that depend on facial features alone may miss less conspicuous symptoms of depression. Ashraf et al. (2020) [22] suggested heterogeneous datasets and standardized metrics of evaluation for enhancing generalizability. Notably, Zhu et al. (2019) [21] demonstrated the potential of physiological signals such as galvanic skin response and pupil dilation with a classification accuracy of 92%, thereby offering an alternative that is more objective and stable to facial analysis based on subjective inference. Overall, although present solutions are encouraging, work needs to be continued to address issues of data quality as well as to develop robust,

multimodal systems that can be practically deployed in real-world applications.

## 2.4 Speech-based Research

The study conducted by Fu et al. (2021) [29] shows the automatic detection of Schizophrenia using a deep learning architecture. A convolutional neural network(CNN) was designed which works for schizophrenic speech detection. The Sci-net architecture uses two components which are skip connections and convolutional block attention module (CBAM). The part with skip connections enhance the information of the features. CBAM sheds light on the effective features. The dataset which was used for this study consisted of 28 patients with Schizophrenia and 28 patients who are healthy. It was seen that Sci-net performed really well on the dataset and has an accuracy of 97.68%. The same system when tested on LANNA database gave an accuracy of 99.52%. The motive of this paper is to prevent the misdiagnosis of schizophrenia and show that impaired speech can be effective for the diagnosis of schizophrenia. However, the problem of this study is that they have worked mainly on speech impairment. There are also other indicators for Schizophrenia like hallucination, forgetting, movement disorder etc. These indicators were not explored in this study.

The study of Zanwar et al.(2022) [46] investigates using social media data the task of automatic mental health detection, which has attracted a lot of attention recently. However, earlier studies have mostly treated MHD as a binary classification problem; however, this article indicates that there is an importance of multiclass classification to account for slight variations over specific mental health disorders detected through speech patterns. The study examines how anxiety, attention deficit post-traumatic stress disorder (PTSD), bipolar disorder, hyperactivity disorder (ADHD), depression and psychological stress can give predictions using social media posts on Reddit. To this end, the paper compares hybrid and ensemble models based on transformer models (BERT and RoBERTa) and BiLSTM neural networks trained with different linguistic features such as syntactic complexity, lexical sophistication, sentiment lexicons and emotion lexicons. For each specific mental health condition being examined here there are feature ablation experiments aimed at finding out the most indicative types of features that could help predict these conditions resulting into a deeper understanding about the role played by linguistic features in mental health status prediction.

Wei, L. and Mei, Z. (2024) [60] studied how speech as a mode of communication can be leveraged to detect the signs of depression. This study mentioned about speech recognition technologies to analyze emotional and cognitive states. Pitch, speech rate, and word choice are the features that are processed Support Vector Machines (SVM) and neural networks. The author put emphasis on the modality of early diagnosis and personalized methodology for diagnosis due to the challenges mentioned in the paper, such as variable factors like age, gender, cultural background and dialect. Also, ethical concerns regarding data privacy and model interpretability are also a question that arises here. Lastly, the author emphasized the need interdisciplinary collaboration in clinical application for early diagnosis.

Liu, Z. et al. (2015) [3] presented through their study the correlation between speech features such as speech rate, jitter, pitch and depression which was studied using a self-made dataset of 300 subject( 100 controlled, 100 depressed and 100 high-risk individuals) by interviewing them. Their study focused on the differentiability between depressed and not depressed by comparative analysis and follow-up study. Sequential Forward Floating Selection (SFFS) and minimal-redundancy-maximal-relevance (mRMR) are the techniques that has been used for Feature selection to identify which feature influence over depressive states. And later use of SVM and k-NN method to evaluate the features but the study did not any numeric metric result as the study being on-going.

Vázquez-Romero and Gallardo-Antolín (2020) [25] introduced on Convolutional Neural Networks (CNNs) based ensemble learning system using log-spectrograms derived from speech signals which is effective on detecting depression. To improve accuracy and ensure robust and reliable classification system this study focused on ensemble method combines multiple CNN models. The depressed dataset was build upon collecting the AVEC-2016 dataset which included samples of individuals who are diagnosed with depression before. The results demonstrated in this study significant improvements in depression detection performance compared to traditional methods. Log-spectrogram utilization in the system as input helped to capture temporal and frequency aspects of speech which presented critical angles in detecting patterns associated with depression. Fo more comprehensive study the authors suggested on combining deep learning architectures like CNNs with ensemble learning which can effectively address challenges in speech-based depression detection.

Table 2.5: Speech-based Research Comparison

| Ref | Paper Title | Model/ Framework | Result (based on dataset) | Limitations |
|---|---|---|---|---|
| [3] | Detecting Depression Using an Ensemble LR Model Based on Multiple Speech Features | k-NN, SVM, feature selection methods (mRMR, SFS, SFFS) | SVM scored 82% in F1 with 88% accuracy | Speech variability for dialect and pronunciation, ethical considerations (data privacy, interpretability), lack of comparison method |
| [46] | Exploring Hybrid and Ensemble Models for Multiclass Prediction of Mental Health | BERT and RoBERTa, BiLSTM neural networks | RoBERTa outperforms BERT by 1.8% and F1 score of 77% in stress detection obtained | Limitation to only textual features, not incorporating with the behavioral activity data on mental health status |
| [60] | Detecting depression through dialogue: A review of speech recognition technologies | Support Vector Machine(SVM), Neural networks | 90.2% for LSTM, 77% for Support Vector Machine(SVM) | Ethical concerns, inaccurate model interpretability |

| | | | | |
|---|---|---|---|---|
| [29] | Sch-net: a deep learning architecture for automatic detection of schizophrenia | CNN, CBAM | Sci-net performed 97.68% accuracy and even got to 99.52% on LANNA children speech database. | Indicators like hallucination, forgetting, movement disorder for Schizophrenia were not explored here. Limitation of assessing disease severity. Datasets not containing psychological/neurological disorders. |
| [25] | Automatic Detection of Depression in Speech Using Ensemble Convolutional Neural Networks | Ensemble CNNs, log-spectrograms, ensemble averaging | The ensemble CNN achieved an F1-score of 0.65 (depressed) and 0.80 (non-depressed), with an accuracy of 0.74, outperforming the baseline SVM | Requires a large dataset, potential bias in dataset selection |

As shown in the Table 2.5, across these oral-based mental-health investigations, CNN-style architectures dominate. Fu et al.'s (2021) [29] Sci-Net adds skip connections and a CBAM attention module to a spectrogram CNN to forecast 97.7% accuracy on a 56-speaker corpus and 99.5% on LANNA. Vázquez-Romero and Gallardo-Antolín (2020) [25] also leverage spectral cues with a log-spectrogram CNN ensemble outperforming standard pipelines. Standard classifiers persist: SVMs reappear in both Wei and Mei's (2024) [60] depression-screening prosody-based screen and Liu et al.'s (2015) [3] jitter-and-pitch study, where sequential feature selection directs handcrafted metrics into SVM/k-NN appraisers. Transformer-based text–speech hybrids appear in Zanwar et al. 's (2022) [46] investigation, whose BERT/RoBERTa + BiLSTM ensembles pose a seven-way Reddit classification challenge revealing the weakness of binary labelling. Albeit with diverse front ends, three weaknesses pervade: sample sizes are small or homogenous (e.g. 28 + 28 speakers for schizophrenia, single-language corpora), limiting external validity; most systems address one condition only, omitting comorbid symptoms such as anxiety or cognitive impairment; and none properly address ethical issues of voice surveillance, dialect bias or clinical interpretability. Accuracy figures in excess of 95% with controlled conditions indicate promise, but an expansion to real-world deployment will require larger, linguistically diverse datasets, multimodal fusion, and explainable models clinicians—and patients—can rely on.

## 2.5 Multimodal Research

Lin et al. (2020) [23] proposed a multimodal approach using visual and textual data from social media. The system combines CNN-based classifier and BERT model to extract features from images and TEXT from Twitter. The system design consists of three parts- offline training, online detection and analysis. During offline training, features are extracted from tweets and images posted by the user using CNN and BERT models. The features extracted from profile and posted pictures using CNN-based classifier were used to train the system and generate vectors for analysis. Also, BERT uses a fixed-size vector that represents the user's textual data. These features are used to classify users from depressed and non-depressed subclasses. For online detection, the system predicts the user's mental health condition using user images and tweets. The multimodal fusion combined approach improved prediction compared to other methods and provides detailed analysis and insights for intervention for treatment. Although the study was successful in providing better accuracy, mental health does not depend on only one indicator. So only depression analysis can not fully capture the mental health condition of a person.

Shen et al. (2017) [8] addresses the issues of depression apart from the traditional diagnosis methods. With the rise of the use of social media, the authors proposed a novel approach leveraging social media data using a multimodal depressive dictionary learning model for early detection. Using their own three dataset features and preprocessing it, feature extraction was focused on six feature groups: user profile features, social network features, topic-level features, visual features, emotional features, and domain-specific features related to depression. The system combined Uni-modal Dictionary Learning and Multimodal Joint Sparse Representation. The UN-modal dictionary learning focuses on learning sparse representation for each feature modality separately where multi-modal joins uni-modal dictionaries to capture common patterns to enhance the effectiveness of the model. The multimodal approach increases the accuracy of identification and provides fresh perspectives on how depressed people behave online. Although the increased overall performance metrics of the model, it could not represent the overall population especially less active users which can impact model's effectiveness. A comprehensive methodology combining advanced deep learning techniques and the usage of video and audio data of users may open paths for improvement for future works.

Fang et al. (2023) [51] proposed a multimodal analysis to detect depression to address the issue where most studies focus on only one data type either textual or visual. The authors integrated multiple data modalities to improve system accuracy to prevent the issue of misdiagnosis. In the study author used multimodal fusion model with multi-level attention mechanism which operates on two stages- feature learning and fusion stage. Using bidirectional LSTM the system learns features from audio, visual and text data which Is later integrated to reduce dependency on single data. The study demonstrated superior performance over existing models. The use of advanced feature extraction and fusion techniques presented a more robust model in the area of depression detection. However, improving the model's capacity to recognize and differentiate between various emotional states using sentiment analysis could improve performance metrics of this system.

According to Orabi et al.(2018) [13], the detection of depression from Twitter users is done using the RNNs and CNNs (deep learning techniques). Word embedding optimization and comparative evaluation processes were used for this purpose. Word embedding optimization focuses on classification to identify the users suffering from depression from their tweets from a dataset CLPsych2015. They also focus on comparative evaluation to compare the performance of deep learning architectures used on NLP tasks to detect the mental health problems.

Bhuvaneswari and Prabha (2023) [49] has addressed the issue of depression detection through using social media data, emphasizing on computational approaches as a result of language variations and spelling mistakes in user-generated content on social media platforms. Earlier studies have examined deep learning methods to identify problems like as depression from social media data. This article presents an alternative model for detecting depression that incorporates some novel deep learning mechanisms and hybrid feature selection. The proposed model combines Term Frequency Inverse Document Frequency integrated Modified Information Gain (TFIDF-MIG) and the Improved Elephant Herding Algorithm (IEHA). Using this hybrid feature selection approach, it is intended that the classification accuracy would be enhanced while reducing the total number of features selected for this model. The present study proposes a new deep learning method called attention included improved ReLU-based Convolution Neural Network with Long Short-Term Memory (AIRCNN-LSTM), which is an advancement over existing models used in depression detection.

Gui et al. (2019) [17] proposed a cooperative multimodal approach to detect depression from Twitter by combining textual and visual content using historical user posts. Their COMMA model, based on multi-agent reinforcement learning with actor-critic architecture and shaped rewards, enables cooperative selection of depression-indicative texts and images. The model achieved strong performance with an accuracy of 83.6%, precision of 84.3%, recall of 81.5%, and F1-score of 82.9%, significantly outperforming previous baselines with over 30% error reduction. Despite its effectiveness, the model's limitations include computational complexity and the lack of post-level annotations, which hinder detailed interpretability.

Sadeghi et al. (2024) [59] explored a multimodal model for detecting depression by combining Large Language Models (LLMs) and facial expression analysis. The model combines text-based user post features and facial expressions from 3D facial landmarks to approximate depression severity. The use of LLMs helps in the extraction of content and sentiment from text-based input, whereas facial expression analysis helps in extracting the non-verbal signals of depression. The approach achieved high performance, outperforming traditional unimodal approaches, and gained a deeper understanding of depression by fusing the two modalities. The work has some limitations. Using facial expressions and text alone may not capture other critical features of depression, e.g. voice tone or bodily well-being. Data imbalance in current depression datasets may impact the model performance, and there is an interpretability issue in clinical applications in the real world.

Table 2.6: Multimodal Research Comparison

| Ref | Paper Title | Model/ Framework | Result (based on dataset) | Limitations |
|---|---|---|---|---|
| [23] | SenseMood: A deep multimodal framework integrating BERT-based text and CNN-based visual features fused with low-rank tensor learning. | SenseMood using BERT for text, CNN for images, and low-rank tensor fusion. | Accuracy 88.4%, precision 90.3%, recall 87.0%, F1-score 93.6%; outperformed state-of-the-art. | Relies on self-declared labels and noisy social content; limited cross-platform robustness. |
| [17] | A cooperative multimodal approach to detect depression from Twitter using historical posts with actor-critic architecture and shaped rewards. | COMMA: Cooperative Multi-modal depression detection via actor-critic reinforcement learning. | Achieved 83.6% accuracy, 84.3% precision, 81.5% recall, 82.9% F1; reduced error by over 30% compared to baselines. | High computational complexity and lack of post-level annotations limit interpretability. |
| [8] | Multimodal framework using Twitter data and six groups of features including social, visual, and linguistic indicators for depression detection. | MDL: Multimodal Depressive Dictionary Learning with sparse user representation. | Accuracy 86%, precision 85%, recall 84%, F1-score 85%; outperformed baselines by 3–10%. | Dependent on heuristic keyword-based labeling; noisy and less generalizable. |
| [13] | Deep Learning for Depression Detection of Twitter Users | RNN, CNN | CNN model reports with 87.957% and RNN with 83.236% performance | Lack of sufficient annotated training data. Models unoptimized can not give much accurate performance. |
| [49] | A deep learning approach for the depression detection of social media data with hybrid feature selection and attention mechanism | AIRCNN-LSTM | Hybrid FS approach here obtained higher classification accuracy. | Limitation on selected datasets |

| | | | | |
|---|---|---|---|---|
| [59] | Harnessing multimodal approaches for depression detection using large language models and facial expressions | LLMs and SVR | MAE scored 4.22 and RMSE scored 5.08 | Limited to videos and not on audio as the image data is only retrieved from the videos. Only worked on text converting the facial features by PHQ-8. |
| [51] | A multimodal fusion model with multi-level attention mechanism for depression detection | Multimodal fusion using LSTM (audio/visual) and Bi-LSTM with attention (text), with multi-level attention mechanisms | Achieved 78.3% accuracy in MFM-Att model on DAIC-WOZ dataset | Limited to DAIC-WOZ dataset; lacks real-world validation. Feature sparsity and class imbalance remain concerns |

From the Table 2.6, it is found that a critical analysis of the papers assessed on multimodal depression detection displays a shared reliance on deep learning models such as CNNs, RNNs, BERT, and LSTMs. Multimodal integration—particularly of text and image—remains the most preferred approach in few of the studies. Lin et al. (2020) [23] uses BERT and CNNs in combination for image and tweet feature extraction, while Fang et al. (2023) [51] enhanced this with three-modal integration of audio, visual, and text data coupled with a multi-level attention mechanism, with a significant reduction in detection error. Dictionary learning was employed in Shen et al. (2017) [8] and standard RNN/CNN models in Orabi et al. (2018) [13], though Shen et al.'s [8] model was plagued by generalizability due to restricted use with less active users. Gui et al.'s (2019) [17] COMMA model employed multi-agent reinforcement learning, with satisfactory performance at the cost of increased computational complexity and lower interpretability due to the lack of post-level annotations. Bhuvaneswari and Prabha (2023) [49] pushed deep learning methods towards hybrid feature selection technique (TFIDF-MIG with IEHA) and a novel AIRCNN-LSTM model, addressing the challenge of noisy social media data. Sadeghi et al. (2024) [59] sought integrating large language models with facial expressions, where the benefit of viewing both verbal and non-verbal cues is highlighted, although their model still failed to incorporate voice and physiological cues. In general, while multimodal solutions never fail to outperform unimodal baselines, shortcomings remain in data imbalance, lack of interpretability, and incomplete use of broader behavioral cues.

In summary, it is seen from all the researched that was reviewed that most of them used traditional Machine learning algorithms, Convolutional Neural Networks(CNNs), Recurrent Neural Networks(RNNs) etc. Another common problem is that the dataset that was used for the studies was not publicly available and they were very limited. Even if the dataset was found it was seen that the dataset was very limited and small. Furthermore, most of the studies worked only on textual data. In this study, we try to solve this problem as we are working on a big dataset. Moreover, we are not only using traditional algorithms but also doing multimodal analysis which denotes that we are working on both textual and visual data and using attention fusion which is not used in most of the studies.

# Chapter 3

# Background

## 3.1 Data Acquisition

For this study, the dataset was a publicly available dataset which is provided by Gui et al. (2019) [17]. The dataset organization has labels for depression detection with social media texts and images of users. The dataset was chosen because it aligns closely with the objectives of this research, as it consists of both textual and visual data, a key focus of the current study. The data was freely accessible and obtained from the research paper "Cooperative Multimodal Approach to Depression Detection in Twitter" by Gui et al. (2019) [17], ensuring that ethical standards were maintained throughout the data acquisition process.

## 3.2 Dataset Structure

Our dataset is distributed user wise. Here 1402 users are positive and 1402 users are negative. Positive means that the user is depressed and negative means that the user is not depressed. Every user has a timeline.txt file with their corresponding images if there are any. The dataset structure as described:

### 3.2.1 Positive and Negative User Categories

The two main splits of the dataset: `positive` and `negative`. Each folder contains 1,402 user subfolders, representing users identified as either depressed or non-depressed.
Each tweet in the `timeline.txt` file has a unique **ID**, and if the tweet contains an image, it is saved with the same ID in the form of an image file (e.g., `idA.jpg`, `147.jpg`).
Some image files may be empty or corrupted (size 0 bytes). These images are due to invalid URLs and can be ignored during analysis without affecting the dataset's integrity.
This structure enables the comprehensive analysis of not only textual but also visual content in tweets, facilitating the detection of depression through a multimodal

approach. The dataset is carefully organized to allow easy access to each user's data for model training and evaluation.

### 3.2.2 User Folders

Each user has a subfolder named after them (e.g., `user1`, `user2`, etc.). Inside each user folder:

- `timeline.txt`: This file contains the user's tweets, with each tweet accompanied by metadata such as the timestamp, tweet text, and a unique tweet ID.
- **Image Files**: If a tweet includes an image, the corresponding image is stored in the same user folder, named according to the unique tweet ID (e.g., `idA.jpg`, `idB.jpg`, etc.).

It should be noted that dataset is imbalanced, even though the users labeled depressed and not depressed are equal, but the distribution of depression-related content across users may not be uniform, potentially impacting the performance of detection models.

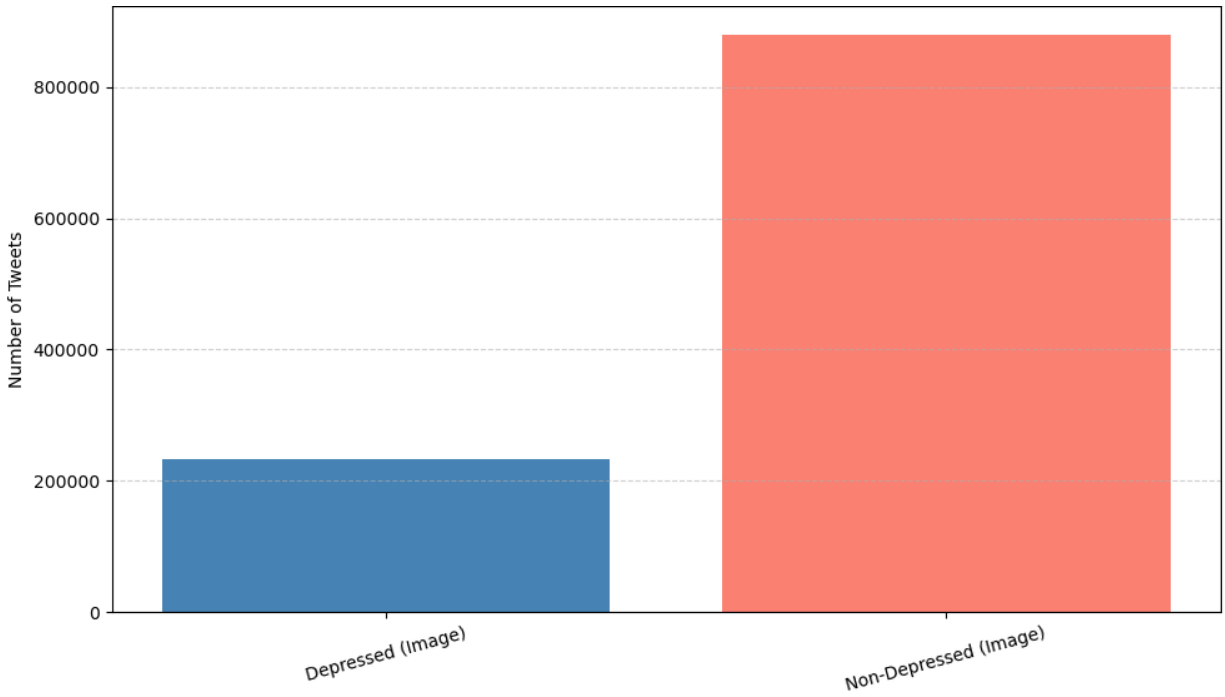


Figure 3.1: Total Number of images: Depressed vs Non-depressed

- **Image Frequency:** Depressed users posted fewer images compared to non-depressed users, as shown in the graph.
- **Tweet Count:** Additionally, depressed users tweeted less frequently than non-depressed users, a trend depicted in the graph.

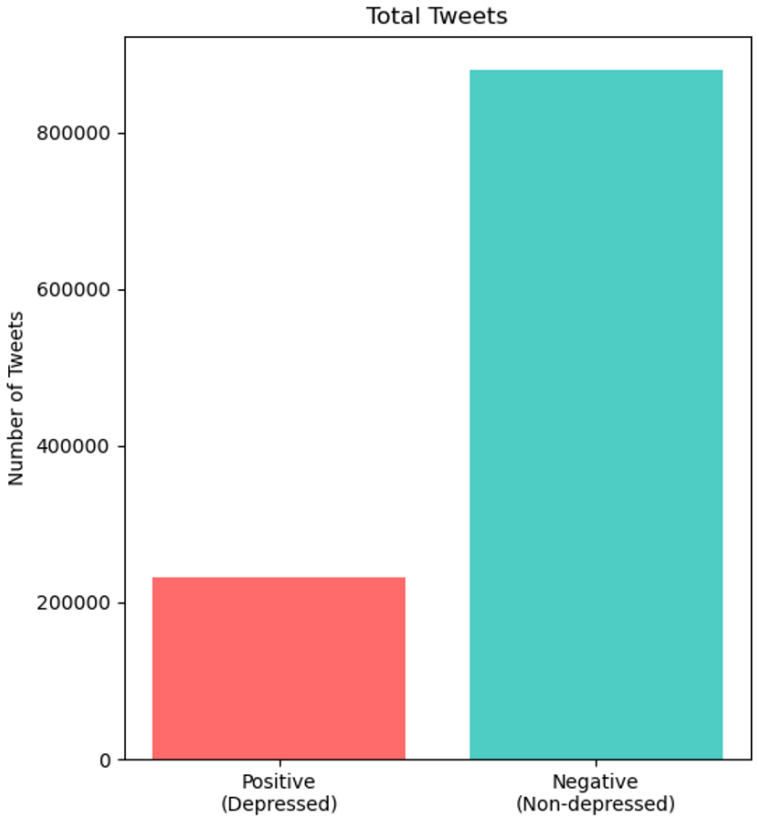


Figure 3.2: Total Number of Tweets: Depressed vs Non-depressed

From these graphs we can understand that non-depressed users are much more active compared to depressed users in social media.

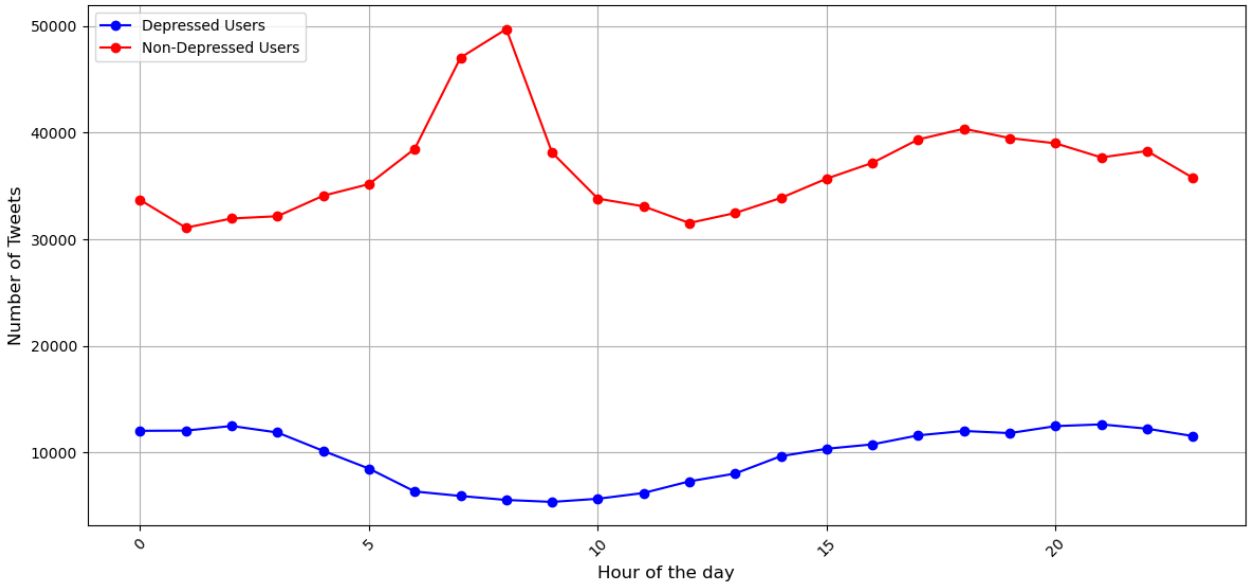


Figure 3.3: Tweet Frequency Over Time (Depressed vs Non-depressed Users)

This graph shows the depressed and non depressed users tweet frequency over time. The number of tweets from non-depressed users has surpassed depressed users according the graph. Considering the timeframe tweet count for non-depressed users is highest between hours 5 to 10 whereas for depressed users between hours 5 to 10 is the lowest.

## 3.3 Dataset Preprocessing

The dataset underwent extensive cleaning to address various issues. First, empty folders having neither image nor text files were deleted. Then, any corrupted images were removed for users where such issues were identified. Duplicate users were skipped for working efficiently. Additionally, any text or images that were in languages other than English were removed. Below is the flowchart of the data-cleaning process.

### Preprocessing Step:

- **Empty Folders**
  - **Without Preprocessing:** Empty folders would remain in the dataset, potentially causing the model to load unnecessary data or attempt to process empty files. This could lead to wasted resources and errors during data loading.
  - **After Preprocessing:** There were many folders in the data. They were removed as they were not useful. It is done so that data can be preprocessed efficiently and working with the data becomes easier.
- **Corrupted Images**
  - **Without Preprocessing:** Corrupted images would give error in preprocessing. The model may also not be able to learn properly because of corrupted images.
  - **After Preprocessing:** Corrupted images are removed so that model can be trained properly.
- **Duplicate Users**
  - **Without Preprocessing:** Duplicate users would contribute to the overrepresentation of certain data points. This would lead to data imbalance, where the model overemphasizes learning from some users while ignoring others. The model may also overfit to repeated patterns from the duplicates by reducing unseen data generalization capability.
  - **After Preprocessing:** By skipping duplicate users, it is ensured that the model observes a diverse set of unique samples and learns from it. This improves the generalization ability of the model, making it more


Figure 3.4: Dataset Preprocessing Pipeline

robust and reducing the risk of overfitting. It leads to a more balanced training set.

- **Non-English Language Text**
  - **Without Preprocessing:** Including non-English text could introduce noise into the training data. The model could get confused, leading to incorrect predictions, poor performance, or difficulties in processing.
  - **After Preprocessing:** By removing non-English text, the dataset is made consistent with the model's language expectations (likely English in this case). The model is now trained with relevant data, improving the accuracy and performance for the intended use case.
- **Stratified Sampling**
  - **Without Preprocessing:** At the initial the dataset size for both positive and negative users were the same having a balanced type.
  - **After Preprocessing:** After preprocessing them the ratio in the negative had the upper hand of 69.9% here and the positive one got 30.1%. Here biasness occurred in the dataset as negative got almost 4 times bigger than positive one. However, we made stratified sampling in order to balance it. Since undersampling removes a big chunk of data from the majority class randomly which would have caused a big data loss and inefficiency for the model to give better results. On the contrary, stratified sampling at first makes groups and subgroups of the same pattern type data. Then take data from every group equally, making sure no important data is lost.

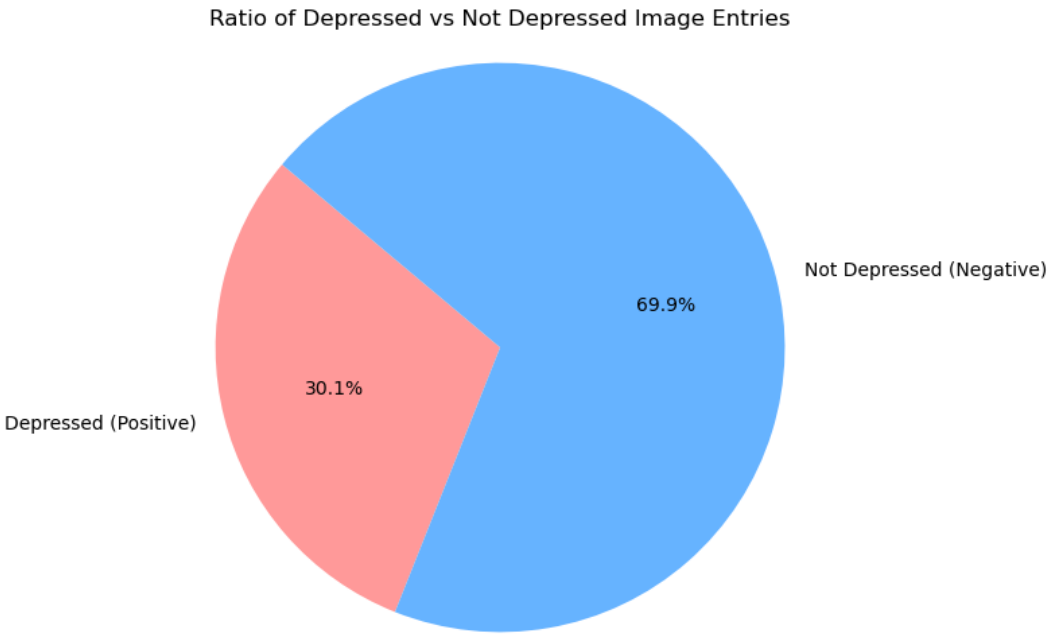


Figure 3.5: Preprocessed Dataset

The bar chart shows the class distribution in the dataset with 37,787 samples (69.9%) labeled as Negative and 16,263 samples (30.1%) labeled as Positive.

Each of these steps helps clean the dataset. After completing all steps, it is ensured that the data used for training is valid, relevant, and aligned with the model's expected input. This contributes to better model performance and more accurate predictions.

## 3.4 BERT (Bidirectional Encoder Representations from Transformers)

BERT is considered to be a highly advanced language model that has the capability of comprehending and processing human language. The previous model could only work in one direction, whereas BERT reads words both preceding and following a word, thereby being able to grasp the entire context more accurately. This bidirectional approach allows BERT to build more informative, contextualized word representations, which have achieved remarkable gains across a range of language tasks—including named entity recognition, sentiment analysis and question answering. (Devlin et al., 2019) [15].

### 3.4.1 BERT Architecture

BERT has a transformer architecture which was first described and introduced by Vaswani et al. (2017)[9]. There is a bidirectional encoder which looks at sentence from both directions. Just like other transformers BERT has self- attention mechanisms.

BERT tokenizes text using WordPiece tokens. For this reason BERT can handle words which are not fanilaiar very effectively. CLS is used for start of sentence in case of classification tasks and SEP ius used for different sentences.

BERT learns language through two main pre-training tasks: Masked Language Modeling (MLM), where it predicts missing words in a sentence. And also it figures out if one sentence logically follows another known as Next Sentence Prediction (NSP). These tasks help BERT develop a deep understanding of language, both within single sentences and across multiple sentences.

- **Bidirectional Encoder**: BERT processes input text using both directions, which helps it to better understand word meanings within their full context.
- **WordPiece Tokenization**: WordPiece tokenization method uses split text to subword units so thatthe model can process and contextualize out of vocabulary works.
- **Pre-training Tasks**: **MLM** and **NSP** help to recognize relationships between sentence pairs and learns contextual representations of words.

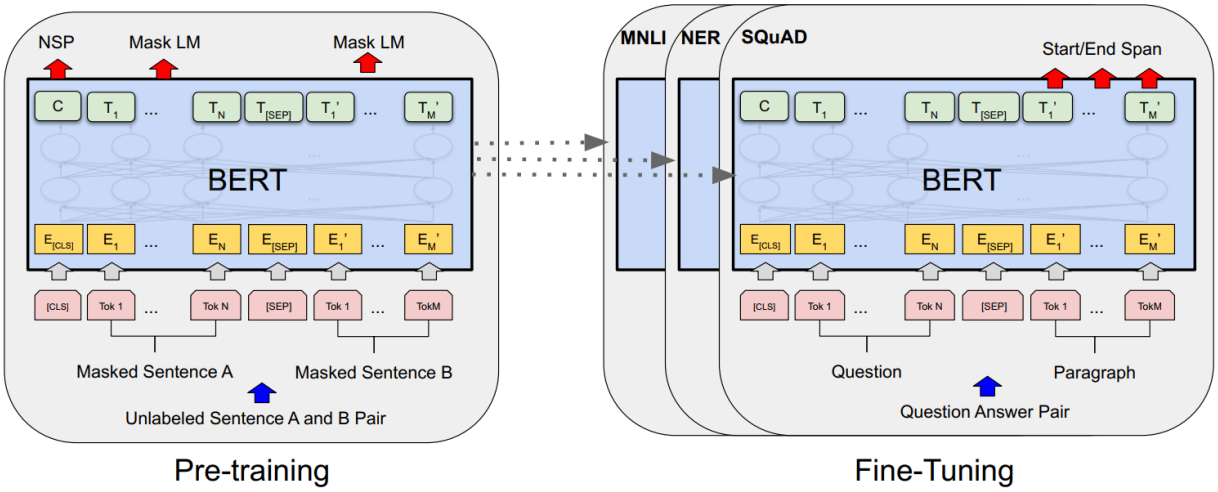


Figure 3.6: BERT Architecture[15]

### 3.4.2 Mechanism

BERT operates through the self-attention mechanism. Self-attention computes attention scores by analyzing the relation of one word with rest of the sentence. It helps BERT to capture both local and global context with input text by weighing the importance of each word relative to others, capturing both local and global context within the input text. The attention score between a token $t_i$ and token $t_j$ is by scaled dot-product attention formula which is:

$$\text{Attention Score}(i, j) = \frac{\mathbf{Q}_i \cdot \mathbf{K}_j^T}{\sqrt{d_k}}$$

where token $t_i$, $\mathbf{Q}_i$ the query vector , $d_k$ is the dimension of the key vector and $\mathbf{K}_j$ is the key vector for token $t_j$. After the calculation token can focus on the most relevant context understanding it's importance withing the entire sequence.

BERT uses three embedding techniques **position embeddings**, **token embeddings** and **segment embeddings**. These embeddings are formed into input representation. Then it passes through Transformer encoder consisting multiple layers to generate contextualized token representations. NSP enables BERT to learn the relationship between sentence pairs which can be used for question answering, sentiment analysis and more. (Devlin et al., 2019)[15].

- **Self-Attention Mechanism**: Relation between tokens are observed which helps to capture contextual relationships.
- **Masked Language Modeling (MLM)**: BERT randomly takes 15% of the tokens and masks them to predict the masked words using the surrounding context.
- **Next Sentence Prediction (NSP)**: Relation between two sentences are observed through this mechanism which is essential for tasks like natural language inference, captioning and question answering.

The bidirectional attention mechanism combined with transformer architecture helped BERT to set new standards in NLP performance gained recognition as a influential model in recent advancements in the field.

## 3.5 CLIP (Contrastive Language-Image Pre-training)

The **Contrastive Language-Image Pre-training in short CLIP** model is a multimodal deep learning architecture designed to handle both image and text data. CLIP uses a shared high dimensional space where mapped images and textual data are used to help the model learn effectively, which enables the model to perform tasks such as text-based image retrieval, image classification and vice versa. CLIP utilizes a contrastive loss method that trains it to associate relevant images and texts, enhancing its capability across a wide range of domains to perform zero-shot tasks.(Radford et al., 2021) [32].

### 3.5.1 Model Architecture

The architecture of CLIP consists of two primary components where one encoder is used for text processing and another one for processing images. Both encoders map their respective inputs into embeddings, which are then compared to measure their similarity in the shared space.

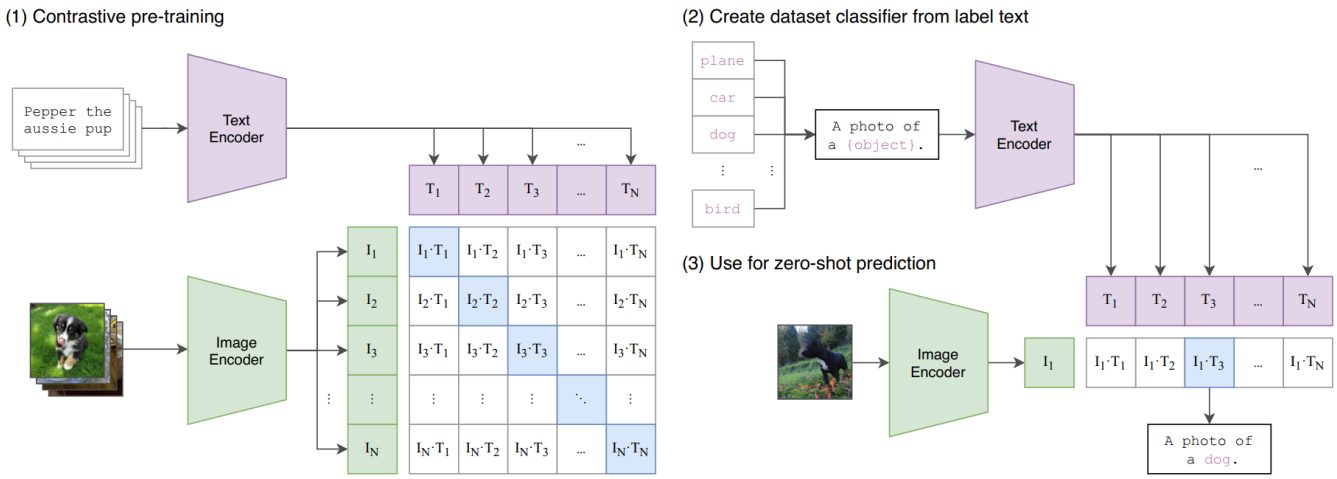


Figure 3.7: CLIP Architecture[32]

- **Image Encoder**: The CLIP image encoder used here is basically based on a Vision Transformer (ViT), although other architectures like ResNet can also be used. The encoder processes input images and generates embeddings that represent the content of the image.
- **Image Preprocessing**: Before feeding the image into the encoder, the image is preprocessed by:

  - **Resize**: The image is resized to a fixed resolution, typically 224x224 pixels.
  - **Normalization**: Here in [0, 1] or [-1, 1] the pixel values are normalized.

- **Image Embedding**: At first the image is divided into patches then each passed into a linear projection to obtain an embedding. These embeddings are processed by the Vision Transformer (ViT). The output is a feature vector, which is then normalized using the L2 norm:

$$\mathbf{v}_{\text{image}} = \frac{\mathbf{v}_{\text{image}}}{\|\mathbf{v}_{\text{image}}\|_2}$$

  where: - $\mathbf{v}_{\text{image}}$ is the image feature vector (embedding), - $\|\mathbf{v}_{\text{image}}\|_2$ is the L2 norm of the image feature vector.

- **Text Encoder**: The CLIP text encoder is based on a pre-trained transformer architecture like GPT-2 or BERT. It processes tokenized text and generates text embeddings.

  - **Text Preprocessing**: The text input is tokenized using Byte Pair Encoding (BPE), converting the text into a sequence of subword tokens. The tokenized text is then transformed into embeddings. To capture the order of the tokens positional encodings are added here:

$$\mathbf{v}_{\text{tokens}} = \mathbf{E}_{\text{tokens}} + \mathbf{E}_{\text{positions}}$$

    where: - $\mathbf{E}_{\text{tokens}}$ are the embeddings of individual tokens, - $\mathbf{E}_{\text{positions}}$ are the positional embeddings.

  - **Text Embedding Normalization**: The text embeddings are normalized using the L2 norm:

$$\mathbf{v}_{\text{text}} = \frac{\mathbf{v}_{\text{text}}}{\|\mathbf{v}_{\text{text}}\|_2}$$

    where: - $\mathbf{v}_{\text{text}}$ is the text feature vector (embedding).

### 3.5.2 Loss Function: Contrastive Loss

A contrastive loss function is used to train the CLIP model this makes it to assign lower similarity scores to non-matching pairs (negative pairs) and higher similarity scores to matching image-text pairs (positive pairs). The contrastive loss for a pair of image and text embeddings is defined as:

$$\mathcal{L}_{\text{contrastive}} = -\log\left(\frac{\exp(\text{Cosine Similarity}(\mathbf{v}_{\text{image}}, \mathbf{v}_{\text{text}}))}{\sum_{j=1}^{N}\exp(\text{Cosine Similarity}(\mathbf{v}_{\text{image}}, \mathbf{v}_{\text{text}j}))}\right)$$

where: - $\mathbf{v}_{\text{image}}$ and $\mathbf{v}_{\text{text}}$ are the embeddings for the image and text, respectively, - $\mathbf{v}_{\text{text}j}$ represents the embeddings of the negative text samples (non-matching text), - $N$ is the total number of samples in the batch.
This loss function forces the model to the embeddings farther apart of non-matching image-text pairs while bringing the embeddings of matching pairs closer in the shared space.

### 3.5.3 Classifier: Softmax Activation

In cases of supervised fine-tuning, CLIP can be extended with a classifier to perform image classification and some other specific tasks. The output of the CLIP encoder (concatenation of the image and text embeddings) is passed followed by a softmax activation function through a fully connected neural network to predict the class probabilities:

$$\mathbf{y} = \text{Softmax}(\mathbf{W} \cdot \mathbf{v}_{\text{combined}} + \mathbf{b})$$

where: - $\mathbf{y}$ is the output probability vector, - $\mathbf{W}$ is the weight matrix of the classifier, - $\mathbf{v}_{\text{combined}}$ is the concatenated vector of image and text features, - $\mathbf{b}$ is the bias term. The cross-entropy loss is applied during training to calculate the loss values and minimize the error between ground truth labels and predicted class probabilities.

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$$

where: - $y_i$ is the true label where 1 is for the positive class, 0 for the negative class, - $p_i$ is the predicted probability for the positive class.
The model's ability to work on both text and image representations alignly in a space shared without the need for specific task assigned architectures, allows it to perform various multimodal tasks smoothly which makes it highly versatile for applications requiring cross-modal understanding.
Through its powerful contrastive learning framework and flexible architecture, CLIP has established itself as a strong model for text-based image retrieval tasks and zero-shot image classification. By learning joint representations of both text and images, CLIP enables effective interaction between the two modalities.

## 3.6 BLIP-2: Bootstrapped Language-Image Pre-training Version 2

BLIP-2: Bootstrapped Language-Image Pretraining Version 2 is a deep multimodal model for vision and text information understanding and processing. In contrast to the majority of current models that treat images and text in isolation, BLIP-2 introduces a new cross-attention mechanism for enabling effective vision-language interaction. This allows the model to generate joint representation of the two modalities, significantly improving the performance on image and text understanding tasks, including multimodal classification and caption generation.

BLIP-2 adapts the fundamental principles of the Vision ViT(Vision Transformer) for image processing and employs a frozen pre-trained language model like BERT- for text understanding. Combining the two stable components with the Q-Former for query construction and cross-modal fusion allows the BLIP2 model to facilitate well across a mass range of practical applications for example- depression detection from multimodal inputs (such as tweets and images).

### 3.6.1 BLIP-2 Architecture

BLIP-2 has been designed with multimodal inputs in mind, using a varied set of modules to combine visual inputs along with textual content. The elements include:

- **Vision Encoder:** A Vision Transformer encodes image data to produce representations from which visual outputs can be generated. The vision encoder transforms raw image data into a structured format that is understandable and can be processed by the later parts of the model.
- **Q-Former:** A light-weight query transformer generates queries from visual features, which guide language model attention on significant text features. The Q-Former is a bridge for both visual and textual modalities and sets up meaningful correspondence between language understanding and image content.
- **Frozen Language Model:** The language model which is pre-trained on a corpus (such as GPT, BERT, or T5) is given the questions output by the Q-Former and positions them in context following the text it is provided as input. With this component kept frozen, BLIP-2 has the excellent language understanding capabilities of well-known models but tailors them for multimodal processing.
- **Cross-Modal Fusion:** The fusion module integrates textual as well as visual representations and outputs a common joint representation, which is a foundation for downstream processing. This section is crucial in creating a joint understanding for both modalities. The issue of matching text with visual features is stressed by Li et al. (2022) [53], who propose the Q-Former as a bridge for extracting visual features useful for the text.

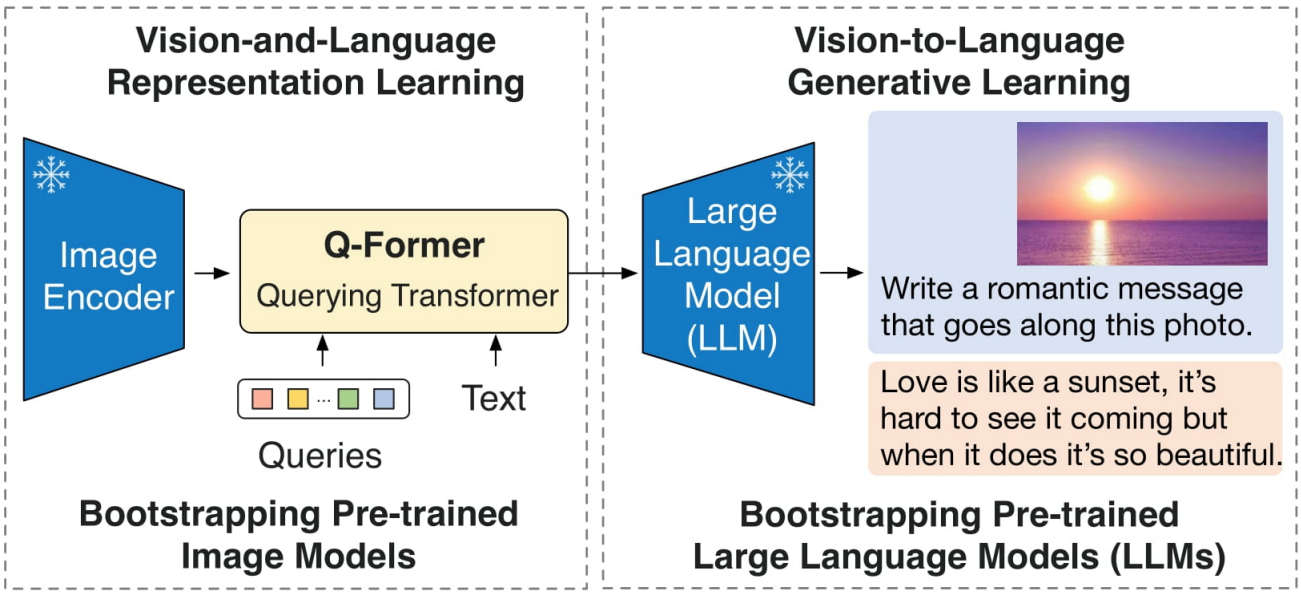


Figure 3.8: BLIP-2 Architecture[53]

### 3.6.2 Data Flow Process

The flow of data through the model follows these sequential steps:

- **Image Processing:** The image input is passed into the Vision Encoder, which generates a set of visual features.
- **Query Generation:** These inputs are fed into the Q-Former in order to produce queries.
- **Text Processing:** The frozen language model processes these questions in order to concentrate on input text.
- **Cross-Modal Integration:** The representations are integrated into a shared representation for downstream processing.

### 3.6.3 Mathematical Framework

**Vision Encoder Processing**
Mathematically, the Vision Encoder processes the image patches as follows:

$$E(x_i) = W_p \cdot x_i + b_p$$

where $E(x_i)$ is the embedding of the $i$-th patch from an image, and $W_p$ is a patch embedding projection matrix that is learned. The image features are next passed through self-attention layers in order to capture patch relationships.
**Q-Former Query Generation**
The Q-Former generates queries from the visual features:

$$Q_{\text{queries}} = \text{Q-Former}(E_{\text{image}})$$

**Language Model Processing**
They're then passed on to the frozen language model, which outputs the contextualized text features:

$$T_{\text{text}} = \text{LLM}(Q_{\text{queries}})$$

**Cross-Modal Fusion**
The output of the language model is combined with the image features using a cross-attention mechanism in order to create a joint representation:

$$F_{\text{joint}} = \text{Fusion}(F_{\text{image}}, F_{\text{text}})$$

This is then used for making classifications or final predictions.

### 3.6.4 Mechanism

**Cross-Attention Architecture**
Cross-attention is BLIP-2's main operating functionality for vision-language fusion. The attention operation in this process allows the model to focus on salient image and word areas during processing. The cross-attention mechanism, for example,

computes attention between image and word representations, allowing for a conditioning of a modality on the other.

**Attention Score Calculation**
Cross-attention mechanism can be expressed in terms of the scaled dot-product attention formula

$$\text{Attention Score}(i, j) = \frac{(Q_i \cdot K_j^T)}{\sqrt{d_k}}$$

where:

- $Q_i$ is the query vector for token $i$ (from the image or text),
- $K_j$ is the key vector for token $j$ (from the opposite modality),
- $d_k$ is the dimension of the key vectors.

**Feature Projection and Alignment**
Projection layers are used for dimensionally compressing image as well as text features, which are now prepared for fusion

$$F_{\text{text}} = \text{Projection}(T_{\text{text}})$$

They help align closer together feature spaces of both image and text in a way that they can be easily integrated by the model.

**Training Objective**
We use cross-entropy loss as our loss function during training. It is expressed by:

$$L_{\text{CE}} = -\sum_i y_i \log(p_i)$$

where:

- $y_i$ is the true label (e.g., depressed or non-depressed),
- The probability for each class as predicted is $p_i$.

This loss function is optimized during training in order for the model's predictions to be more accurate.

## 3.7 Time2Vec

Traditional time representation methodologies, like raw timestamps or one-hot representation, are not benefiting from the inherent cyclical properties of time (e.g., months within a year, hours within a day). The need for developing a Time2Vec arose out of this limitation, as it aimed for a more robust feature representation that would include periodic as well as non-periodic information. A unique feature of this model is its combination of a sine wave to capture periodic features with a linear transformation to handle non-periodic qualities. This combination allows the network to understand temporal dynamics within both cyclical contexts (e.g., days or hours) and continuous situations (e.g., years or ages).

### 3.7.1 Model Architecture

Time2Vec is a model-agnostic method of temporal data encoding proposed by Kazemi et al. (2019) [15]. The main goal of this method is to efficiently capture periodic and non-periodic temporal patterns. Time2Vec encodes raw temporal data into a vectorized representation, making it directly usable in multiple machine learning models and enhancing these models' ability to work with temporal data efficiently. The Time2Vec encoder is formulated as:

$$t2v(t) = [\text{Linear}(t); \sin(\omega_1 t + \phi_1); \sin(\omega_2 t + \phi_2); \ldots; \sin(\omega_k t + \phi_k)]$$

- $\text{Linear}(t) = w_0 t + b_0$ represents the non-periodic component using a linear transformation.
- $\sin(\omega_i t + \phi_i)$ is used for modeling periodic terms by a sine wave with isoipalearmable parameters $\omega_i$ (frequency) and $\phi_i$ (phase shift).
- The $[;]$ refers to the concatenation of the linear and sinusoidal components, leading to the final time vector.

Linear segments are used to model non-periodic phenomena, like steady growth or decrease, while sinusoidal segments are used to model periodic phenomena, like day/night or seasonal variations. The combination of both approaches allows Time2Vec to capture a broad range of temporal dynamics.

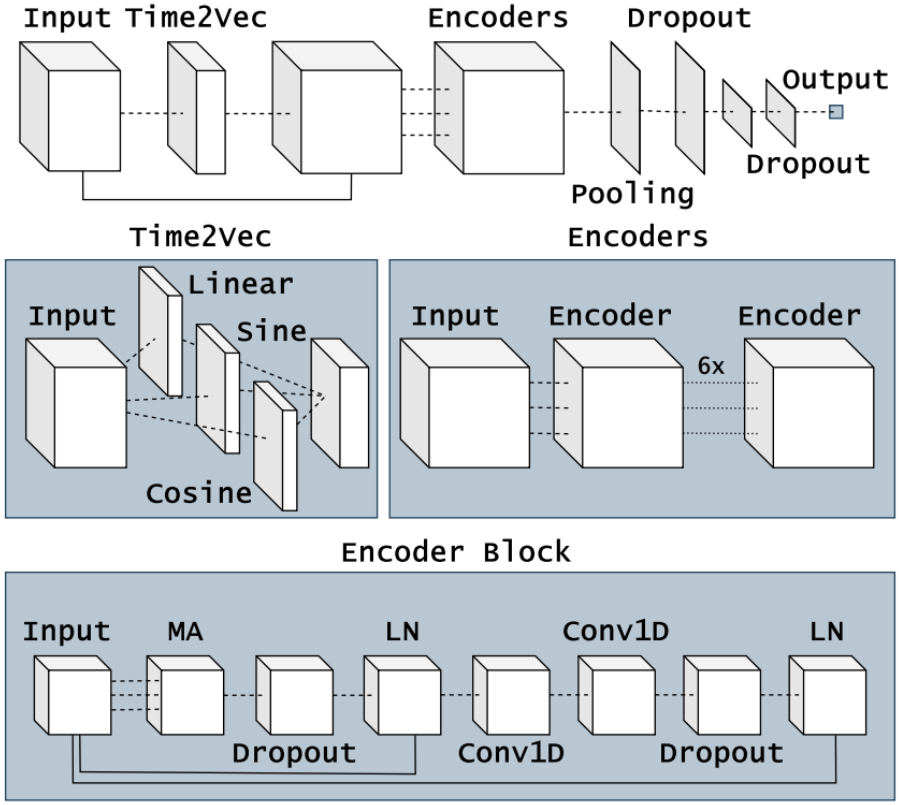


Figure 3.9: Time2Vec Architecture[61]

### 3.7.2 Mechanism

The encoding mechanism in Time2Vec is designed to be highly expressive and efficient to use in deep learning. It becomes so by combining a linear transformation with a periodic transformation of the input time variable, as explained in the sections that follow.

- **Linear Component:** The first dimension of the output of Time2Vec is a linear transformation of the input time, which is given by $w_0t+b_0$, where $w_0$ and $b_0$ are learning parameters. The linear component helps capture nonperiodic, monotonic trends in time, e.g., linear growth or decline.

- **Periodic Components:** The remaining dimensions consist of sinusoidal functions, which include a learnable frequency and phase for each of them. That is, for each $i$-th periodic component, the encoding is $\sin(\omega_i t + \phi_i)$, where $\omega_i$ and $\phi_i$ are learnable parameters for representing frequency and phase shift, respectively. These enable representing repeating, cyclical patterns in temporal data, for instance, daily or seasonal cycles.

- **Concatenation:** The output vector for Time2Vec is generated by concatenation of both linear and periodic components. This combined vector covers non- periodic as well as periodic patterns in time and hence is more expressive than simple one-hot or raw timestamp encodings.

- **Learnable Parameters:** All weights, biases, frequencies, and phase shifts of the Time2Vec layer are all learnable, which allows the network to automatically find out the optimal temporal features for the target task during training. This feature enables Time2Vec to provide a robust and flexible representation of time, thereby enriching machine learning models' awareness of time in a variety of sequence modeling cases.

### 3.7.3 Dimensionality Reduction in Time2Vec

Time2Vec achieves dimensionality reduction by converting the raw temporal feature into a compact vector representation. Unlike the one-hot encodings that produce high- dimensional and sparse representations, which are computationally expensive, Time2Vec uses a fixed-size vector that better captures relevant temporal information. Not only does this achieve dimensionality reduction, but it also increases the generalization capability of the model over capturing temporal patterns.

### 3.7.4 Comparison with Traditional Time Embeddings

Traditional time embeddings typically leverage fixed attributes or simple positional encodings, which might be insufficient to cover complex temporal relationships in some situations. By comparison, the flexible aspect of the parameters of Time2Vec improves its expressive power and flexibility, with the ability to be specially designed for a wide range of temporal patterns. Such inherent flexibility with Time2Vec greatly improves its utility in generating varied forms of time series data in a broad

range of application domains. The integration of Time2Vec improves model performance in applications where understanding temporal patterns is required.

## 3.8 GRU (Gated Recurrent Unit)

The **Gated Recurrent Unit (GRU)** is a notable contribution to neural networks with the aim of addressing the shortcomings of basic RNNs to learn the long-term dependencies in sequential data. First proposed by Cho et al. in 2014 [2], the architecture is a more efficient version of Long Short-Term Memory (LSTM) networks while having the same effectiveness in different kinds of tasks. The GRU uses a gating mechanism allowing the model to decide what fragments of a particular sequence are vital to remember and what parts to forget. By allowing these gating mechanisms, the GRU successfully achieves long-distance dependencies without the extra complexity usually involved with LSTMs.

### 3.8.1 GRU (Gated Recurrent Unit) Architecture

The GRU design has two primary components: reset and update gates. The gates regulate information flowing in the network, allowing for better capture of sequential data dependencies when comparing GRUs against typical RNNs. The reset gate has control of the degree to which the hidden state which was there before is ignored. This property is important because it permits the model to update the hidden state when needed, eliminating unnecessary information and focusing on the latest input. On the other hand, the update gate controls the amount of the previous hidden state that should remain in order to include it in the next hidden state. This helps enable the network to preserve important long-range information yet still updates the state using new information from inputs.

The GRU processes a sequential input. The network updates the hidden state at each time step using the previous hidden state and a current input. The update is controlled by the two gate components reset and update gates, which determine the amount of details to remember and forget at each step. The reset gate controls the amount of the previous hidden state should be ignored. The update gate decides the amount of the previous hidden state that should be carried over to the recent hidden state. The target hidden state is calculated based on the current input and the modified previous hidden state(using the reset gate). The final hidden state is a weighted combination of the previous hidden state and the target hidden state, with the update gate controlling the balance.

The GRU is known for learning long-range dependencies while processing sequential data. Due to its design and gating scheme, it only processes parts of the input sequence and thus ends up making it very useful for machine translation, speech processing, as well as sentiment analysis. Due to its efficiency and simplicity, GRU has gained large popularity in the deep learning community, and it is a high-performance alternative to more elaborate designs such as LSTMs without a loss in performance.

So, for sequence modeling tasks use of GRU is considered to be one of the suitable and preferable techniques.

## 3.9 Lime Xai

LIME, or Local Interpretable Model-agnostic Explanations, is a technique devised for explaining black-box predictions in a manner that is comprehendible for human beings. LIME is designed for providing local interpretability by approximating a complex model's behavior using a simple, understandable model in a small part of the input space.

### 3.9.1 LIME Xai Working Procedure:

- **Instance Selection:** The model first recognizes a single instance (e.g. a text-image combination) that acts as the foundation for explaining its prediction
- **Perturbation:** LIME creates an altered version of the input instance by making small changes (like blurring parts of an image or replacing words in a sentence).
- **Prediction Generation:** During this step, opaque predictive model known as CLIP is applied on every altered instance to get its respective prediction.
- **Surrogate Model:** LIME builds an interpretable surrogate model, either in the form of a decision tree or linear model, from the perturbed dataset. The surrogate model is designed to mimic the behavior of the black-box model in a localized region.
- **Local Explanation:** The surrogate model provides an understandable reason for the model's prediction on a particular instance, often expressed in terms of feature importance scores (e.g. what areas of the image or words in the text had the largest impact).

# Chapter 4

# Research Methodology

## 4.1 CLIP Model Methodology:

This work introduces a novel depression prediction model that has been tailored specifically to examine the complex and subtle interplay between text and image in social media data. This model is based on the Contrastive Language-Image Pre-training (CLIP) approach, which is characterized by its excellent performance in combining text and image data into a unified semantic space. This unified representation allows for a holistic examination of user-generated content, thus providing a more rigorous emotional assessment than is possible using either source of data alone. Additionally, the use of CLIP is warranted by its higher zero-shot generalization capability, which allows the model to detect and understand novel slang, memes, and image trends without the need for ongoing retraining—a critical advantage in the rapidly evolving internet landscape. From raw data to advanced predictive results is achieved through a structured and carefully crafted pipeline. As a first step, the process includes the fusion of a multimodal dataset, where class weights are carefully computed to counteract the threats of bias caused by data imbalance, to ensure that signs of depression are properly prioritized. Next, the data is stratified into training and testing subsets, which allows for firm and unbiased evaluation of how good the model is. The original CLIP-ViT model is then thoroughly fine-tuned on this prepared dataset, thus re-aligning its vast, pretrained knowledge with the specialized and subtle task of identifying digital markers that predict depression. Introducing regularization adds a penalty term to the loss function, which prevents unnecessary complexity within the model. For CLIP fine-tuning, this is expressed as:

$$L_{\text{total}} = L_{\text{contrastive}} + \lambda \cdot R(\theta)$$

A cornerstone of this model is its profound commitment to transparency, achieved through the integration of Local Interpretable Model-agnostic Explanations (LIME). Recognizing that a prediction without explanation is of limited value in sensitive domains, the model does not merely deliver a classification. Instead, LIME provides a clear deconstruction of the reasoning behind each individual prediction. It generates human-readable outputs that highlight the specific words within a text and pinpoint the most influential regions of an image that guided the model's conclusion. By illuminating the "why" behind the "what," this approach fosters essential trust and provides a transparent framework for an AI system designed to operate in the critical area of mental health analysis.

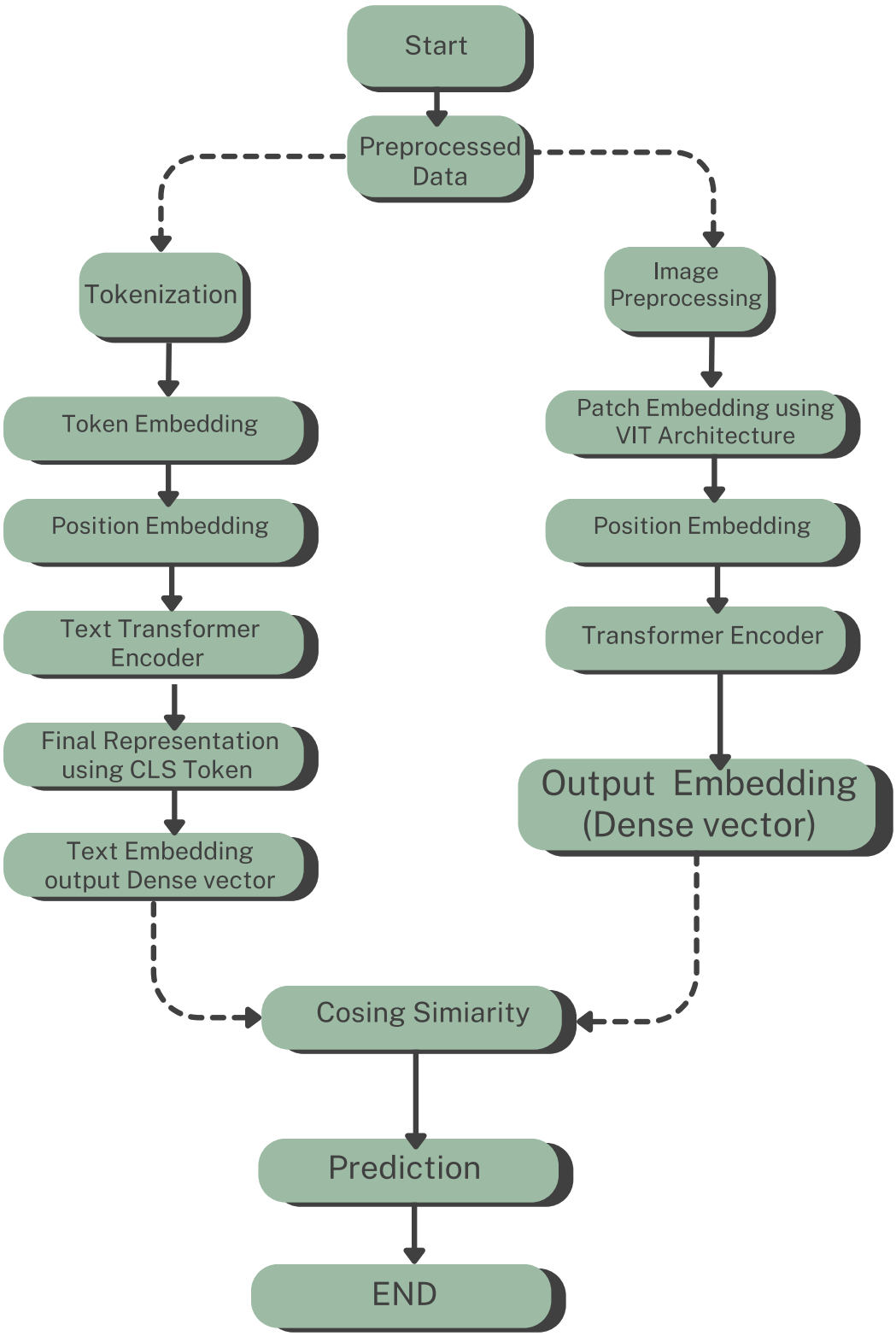


Figure 4.1: CLIP Model Architecture

### 4.1.1 Working Procedure

At first the dataset is passed through the clip model. Text and images are processed separately for the dataset. First the text is tokenized using BPE tokenizer. Then token embedding is done. In the next step position embedding of the tokens is done and then the text is passed to the transformer encoder. After that a final representation is created using a CLS token aggregating the information of the entire sequence of tokens and a text embedding output is created which is a dense vector. For the image first image preprocessing is done. Images are resized, normalized and truncated in this process.Then patch embedding is done using ViT architecture. Similar to the text embedding process in the next step, position embedding is used and passed to the transformer encoder and an output embedding is created which is a dense vector just like the output of text embedding. Then cosine similarity is used so that similarity is found between the vectors of text and images and a final prediction is created. This is how the model is working in the code.

### 4.1.2 Computation

In this research, the **CLIP (Contrastive Language-Image Pre-training)** model is utilized to predict depression based on data from social media. It leverages both textual and visual information, generating embeddings for each modality and aligning them in a shared semantic space. The following steps outline the key computational components involved in the CLIP model, including the use of loss functions and optimization.

The primary loss function in CLIP is contrastive loss, which attempts to pull corresponding pairs of image-text closer in a space which has both image and text embeddings and non- matching ones farther apart. The cosine similarity between the embeddings is calculated by the model, and contrastive loss is maximized for pairs which are matched and minimized for pairs which do not match.

This contrastive loss function leads the model to capture image and text representations that are meaningfully aligned in the same semantic space. With this, it is possible for the model to compare and analyze both modalities well, which is crucial in tasks like depression detection, where textual and visual cues both play a role.

In addition to simple contrastive loss, a symmetric contrastive loss is employed. That is, the model computes contrastive loss in both directions:

- From **image-to-text**: Making sure that image embedding aligns well with text embedding.
- From **text-to-image**: Ensuring that the text embedding is consistent with image embedding.

By utilizing symmetric contrastive loss, it is more effective in learning to represent both image and text in the shared space in a more robust fashion, benefiting multimodal analysis in general.

To improve generalization and make sure that the model does not overfit, a **regularization loss** is applied. Regularization discourages the model from becoming overly complex by adding a penalty term for the loss function that penalizes large weights in the model. This helps the model to focus on features which are the most important and avoid overfitting to noise in the training data.
The regularization term ensures that the model maintains simplicity and generalizes well to new, unseen data, which is critical when working with real-world datasets like social media posts.

For classification tasks, such as predicting depression, **cross-entropy loss** is used during the fine-tuning phase. This loss function evaluates how effectively the model's predicted class probabilities match the class labels which are true (e.g., depressed or non-depressed). Cross-entropy loss encourages the model to lessen the difference between predicted probabilities and the proper labels, improving the performance of the model in classification tasks.

### 4.1.3 LIME Xai Integration

LIME XAI is integrated into the code with respect to the CLIP model:

- **Text and Image Inputs:**
  - Text input is extracted from the test sample and passed to the **LimeTextExplainer.**
  - Image input (sample['image']) is passed to the **LimeImageExplainer**.
- **Prediction Wrapper for LIME:**
  - The code defines a prediction function that converts images and texts into embeddings using the CLIP model.
  - The text and image embeddings are computed separately using the CLIP model's encoders (for image and text).

    image embeddings and text embeddings are combined and passed through the CLIP model's classifier to make predictions.
- **Text Explanation with LIME:**
  - The **LimeTextExplainer** is used to explain the importance of various text features for the given text input.
  - LIME perturbs the text (removes or alters words) and measures the change in the prediction to see which words were most influential in determining whether the post was classified as "Depressed" or "Non-Depressed".
  - Text importance is visualized using a bar chart that shows the most influential features (words).
- **Image Explanation with LIME:**

- The **LimeImageExplainer** is used to explain the importance of image regions for prediction.
- **Perturbations of the image** (masking parts of the image) are generated, and the effect on the prediction is observed.
- **Positive contributing features** are highlighted using LIME's segmentation and mask techniques. The positive and negative features are visualized side-by-side.

- **Visualizations:**
  - The text explanation is displayed as a bar chart, showing which words contributed to the depression prediction.
  - The image explanation is displayed in a subplot that includes:
    * The original image.
    * **Positive contributing features** (regions of the image that influenced the model's prediction).
    * **Negative features** (regions of the image that did not contribute positively to the prediction).

## 4.2 Cross-Modal Model

The cross-modal model combines textual and image data for depression detection by incorporating natural language process methodologies. The suggested model utilizes **BERT** to analyze text information and **CLIP** to analyze the visual data. A **Cross-Modal Attention Mechanism** is used to fuse these two modalities, which is followed by a **Temporal Encoder** aimed at extracting the temporal dynamics related to depressive symptoms. This architecture complements ability of the model to learn and train on complex interactions between text and images which improves its efficiency and effectiveness in depression detection.

### 4.2.1 Textual Data Processing via BERT

User input is handled using BERT, a language model based on transformer architecture, known for its ability to capture contextual signals in language. The text is first tokenized by a WordPiece Tokenizer found in BERT, which breaks down textual inputs into smaller units, or subwords. Rare words, for example, are tokenized into recognized subwords so that a large vocabulary spectrum is tolerated by the model. Each word or subword token is then mapped on a distinct ID using pre-trained vocabulary for the model.

It is passed through a number of self-attention layers in the BERT model, after it is tokenized. The self-attention enables BERT to attend not only to word level information but also to how each word is interacting in a given sentence in a contextual way. The word “bank” will be in a unique representation when it is in “a bank of a river” as opposed to when it is in “a financial bank.”

BERT processes tokens in a bidirectional manner, receiving left and right contextual information for each word simultaneously. This helps the model to widely understand each word's meaning from its surrounding words in the sentence, resulting in dense, contextualized embeddings for each token. The [CLS] token, which represents the entire sequence, is utilized as a foundation for developing a single embedding which encapsulates the full interpretation of the input text. The embeddings are then produced for further processing in the model.

### 4.2.2 Visual Data Processing via CLIP

The images from user posts are processed by CLIP, a multimodal model developed by OpenAI. CLIP utilizes a Vision Transformer (ViT) when processing images. The model breaks images into patches, which are processed in a way just like word processing in a text, and each patch is encoded as a feature vector of information in the image. The ViT applies self-attention on patches, enabling the model to discern relationships between image components, such as objects, textures, or scenes. The feature vector of the image is embedded in a shared feature space where it is comparable across textual embeddings.

### 4.2.3 Cross-Modal Attention

Once textual and image data have gone through BERT and CLIP respectively, they undergo a Cross-Modal Attention Mechanism. The mechanism is designed in a manner where it integrates the text and image embeddings in a joint space so that the model is able to grasp how both modalities interact with one another. Besides, this mechanism is composed of a transformer-like structure where each word and image patch attend to all other entities in the merged sequence. Moreover, it enables the model to focus on meaningful text-image relations. For instance, when the text describes a ”sad dog” and a dog appears in the image, this mechanism will enable the model to focus on the relation between the textual word 'sad' and the dog in the image, thus enabling better detection of depression.

### 4.2.4 Temporal Encoding

This model relies mainly on Temporal Encoding as it encodes for temporal sequential patterns in user behavior. Since depression symptomatology gradually progresses, it is helpful for it to be able to capture temporal progression in emotional states from post- to-post. A GRU-based Temporal Encoder is utilized in this project for modeling for this temporal dimension. The GRU allows the model to capture long-term dependencies between posts but does so in a highly efficient way even when it

is given sequences of mixed lengths.

All of the posts, along with corresponding text and image embeddings, are given as inputs to the GRU, which integrates each of the new inputs into its hidden state. The GRU is composed of two gates: one is an update gate and another one is a reset gate, these control information flow and allow the model to retain important information but not forget irrelevant information. Through a series of posts, processing enables the temporal encoder to retain long- term shifts in emotional tone, which sheds light on the emerging emotional state of the user. This is crucial for depression detection, as a symptom is more commonly a gradual change and not a distinct episode.

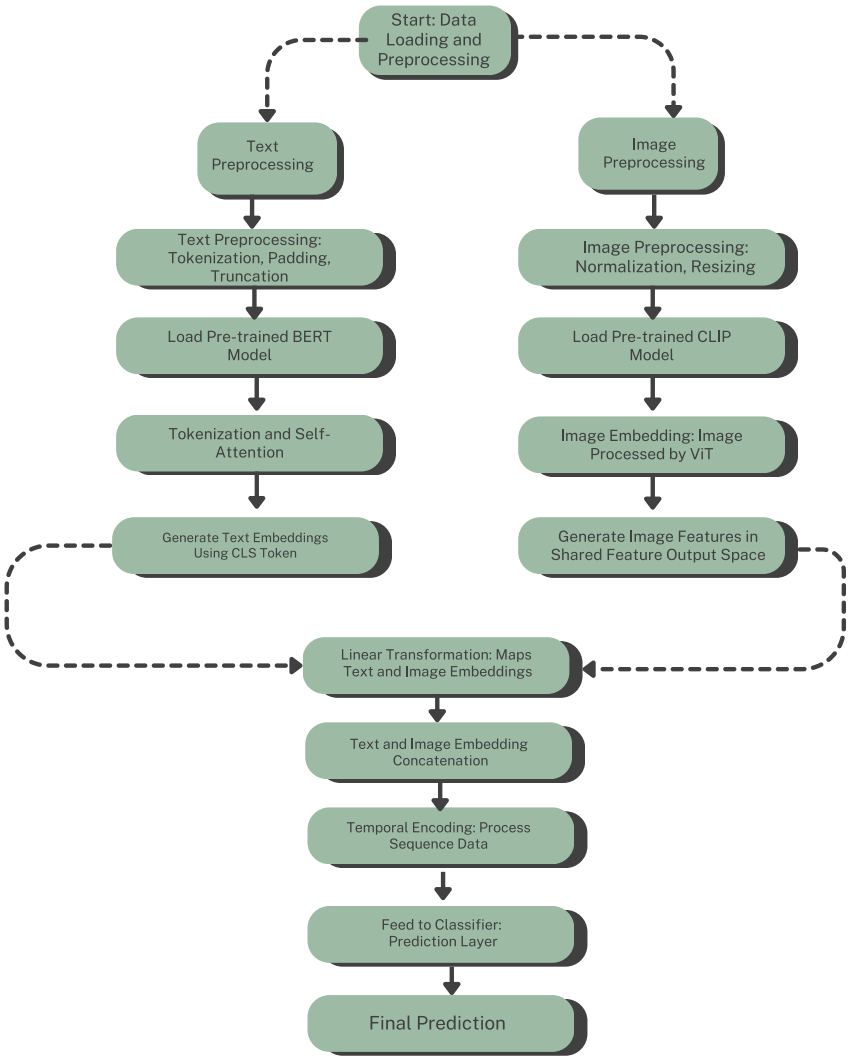


Figure 4.2: Cross-Modal Model Architecture

### 4.2.5 Model Computation

The model is optimized using cross-entropy loss, which calculates the difference between the probability distribution predicted by the model and true labels. Cross-entropy loss is a popular loss for classification tasks and is useful in helping optimize the model's parameters so that it minimizes error in its predictions.

**AdamW**, a version of the Adam optimizer, is employed for optimization. AdamW is well-recognized for its superior effectiveness in coping with sparse gradients as well as for being capable of quickly converging. AdamW utilizes a personalized strategy to adjust the learning rate of every individual parameter, hence enabling more effective updates during training. In addition, the inclusion of a weight decay feature in AdamW helps counteract the possibility of overfitting by imposing penalties on overly large model parameters, an important feature when modeling multimodal data with complex models.

During the training process, the model parameters are updated iteratively via back-propagation. This is done by computing the loss gradients with respect to these parameters, which are then used to adjust the weights. As training continues, the model improves its representation of images and text, and thus its capacity to predict users' depression status by continually examining their posts and images. The combination of temporal encoding with a cross-modal attention mechanism is a major development in the field of multimodal depression detection. The use of a GRU-based encoder for temporal encoding enables the model to study sequences of posts at different time frames. Depression is described as an unstable condition where symptoms increasingly develop over time. By taking into account these temporal fluctuations, the model is able to effectively track changes in a person's emotional state, which could shed light on the development of depressive symptoms.

The cross-modal attention mechanism strengthens the ability of a model to analyze multimodal data effectively. Differently from processing image data and text data in a separate way, the attention mechanism enables a model to process both modalities in a common way, learning both how they collaborate and how they influence each other. This is particularly important in depression detection, where apparent cues such as meme or face expression may aid text content. Paying attention to correct areas of both text and image, the attention mechanism increases a model's ability for making a prediction in a more precise way.

The temporal encoding preserves the emotional trajectory of the user, and the cross-modal attention allows the model to pay attention to text-image interaction.

## 4.3 BLIP-2

BLIP-2 is used to detect depression in order to extract rich semantic features from both textual and image data through its multimodal encoder-decoder architecture. By aligning visual and textual representations, it enhances the model's ability for detection of depressive indicators across social media posts or related content.

- **BLIP-2 Vision Model:** The Vision model is based on a pretrained transformer-based vision backbone (such as Vision Transformer, ViT) that extracts deep visual features from input images. The vision model is freezed, which means its weights are not updated during training, allowing us to leverage the pretrained knowledge of the model for feature extraction while minimizing computational costs.

- **BLIP-2 Language Model:** Language model, which is a pretrained transformer model, extracts contextual textual characteristics from a sample text, say tweets. As in vision model, during training, a language model is frozen in a manner that it retains its excellent ability for text processing and text interpretation without any further refinement.

- **Vision Projection Head:**Linear projection head is applied on those visual features which are obtained by BLIP-2 Vision. ReLU activations and dropout layers are applied here for reducing dimensions of those features which are obtained and introducing non-linearity for improved performance.

- **Text Projection Head:** Similarly, in vision projection head, text projection head reduces textual feature dimensionality from BLIP-2 Language model-derived features. It also comprises ReLU activations along with dropout layers for efficient training and avoiding overfitting.

- **Feature Concatenation:** The projected visual and textual features are concatenated into a single vector. This concatenated feature vector represents the multimodal information extracted from both the image and the text, which is crucial for downstream tasks like depression classification.

- **Classifier Head:** The concatenated feature vector goes through a fully connected neural network, which outputs a classification for depression detection which is binary (Depressed vs. Non-Depressed). This classifier head is the last step in the architecture of the model and learns to map the fused multimodal features to the depression label.

- **Frozen BLIP-2 Backbone:** Both the BLIP-2 Vision and Language models are frozen during training, which decreases the number of trainable parameters and the cost of computation. Only the projection heads and the classifier head are trainable.

### 4.3.1 Data Flow

- **Data Loading:** The dataset is loaded, ensuring class balance and random sampling for generalization.
  - Images are resized to 224x224, optionally augmented (e.g., flipped), and normalized to standardize pixel values.
  - Texts are cleaned (removal of URLs, mentions, hashtags), tokenized, and truncated to a maximum length to ensure consistent input size.

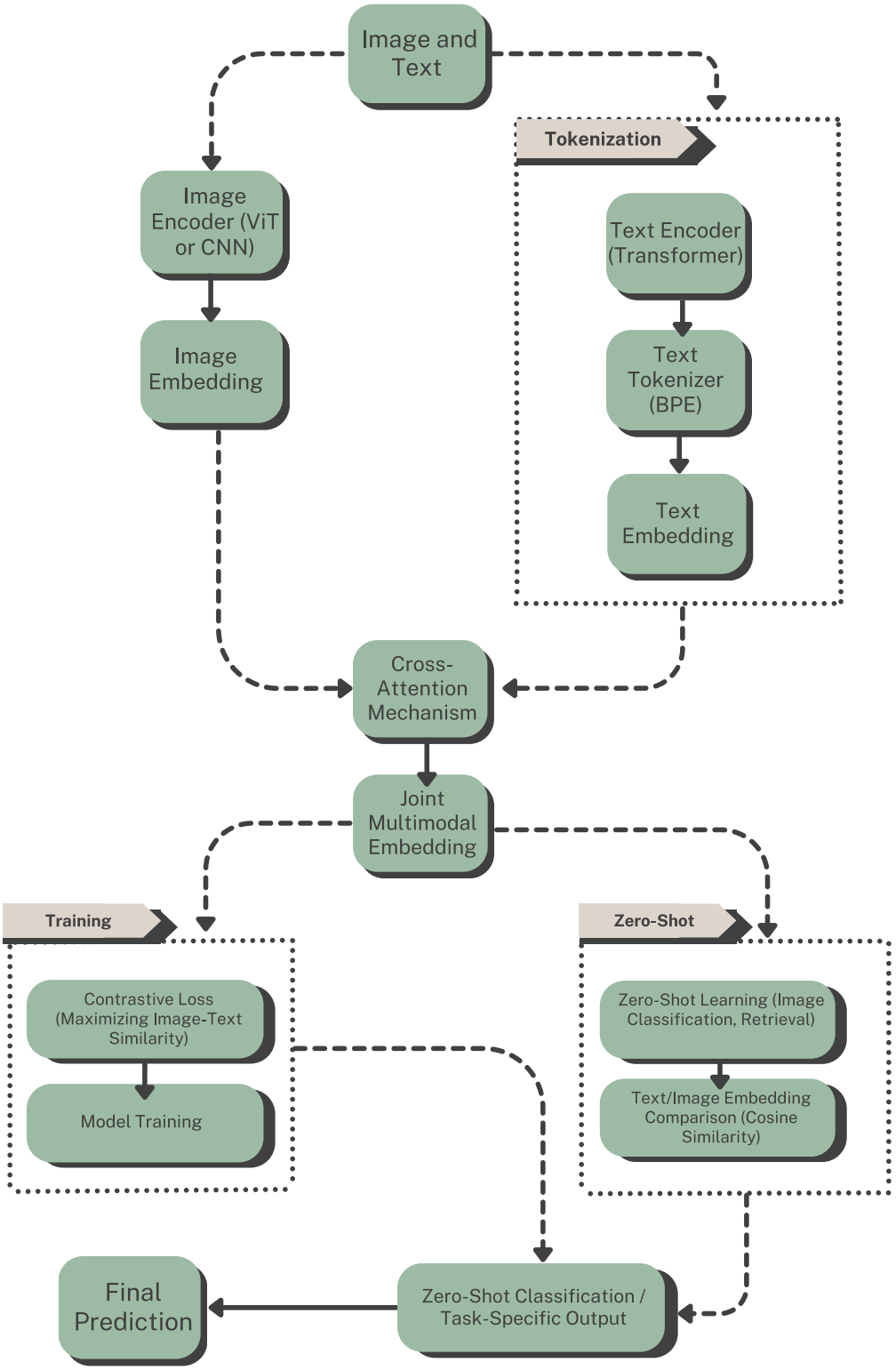


Figure 4.3: BLIP-2 Model Architecture

- **BLIP-2 Feature Extraction**:
  - Images are passed through the BLIP-2 Vision model, extracting visual features.
  - Text is passed through the BLIP-2 Language model, extracting textual features.
- **Feature Projection:** The projected features are created by reducing the dimensionality of both the visual and textual features via linear layers with ReLU and dropout for regularization.
- **Concatenation:** The projected features from both modalities are concatenated into a single vector, representing the combined multimodal information.
- **Classification:** The concatenated vector is fed into the Classifier head, which outputs the final prediction for depression classification.

### 4.3.2 Computation

- **Image Preprocessing:**The image is resized to 224x224, converted to RGB format, and augmented (e.g., random horizontal flipping) to improve generalization.
- **Text Preprocessing:** Text is cleaned by removing noise (e.g., URLs, mentions) and truncated to ensure uniform input size. The text is then tokenized into word pieces.
- **Feature Extraction:**
  - BLIP-2 Vision model computes visual embeddings (with frozen weights).
  - BLIP-2 Language model computes textual embeddings (with frozen weights).
- **Feature Projection:** Linear layers reduce the feature size and apply nonlinearity (ReLU) and dropout for regularization.
- **Feature Fusion:** The projected vision and text features are concatenated to form a multimodal representation.
- **Classification:** The concatenated features are passed through the **fully connected layers** that output logits for binary depression classification.
- **Training:** The model is trained using **weighted cross-entropy loss** (to address class imbalance), the **AdamW optimizer**, learning rate scheduling, and **early stopping**.
- **Evaluation:** The model's performance is evaluated using metrics such as **accuracy**, **F1 score** (weighted and macro), **precision** and **recall**.

### 4.3.3 Hyperparameters

- **Batch Size:** The size of the batch is set to 32 for efficient training. This is a commonly chosen value as it balances memory consumption and computational efficiency without causing excessive memory overhead.
- **Learning Rate:** The initial rate of learning is set to 0.0005. This value was chosen based on prior experimentation and literature for similar models. A lower rate of learning helps in fine-tuning the model effectively without causing extreme updates to the weights, which could harm convergence.
- **Dropout Rate:** The rate of dropout is set to 0.3 in both the projection heads and classifier head to prevent overfitting and promote generalization.

**Vision Model (ViT)**
We chose Vision Transformer as it outperformed other models in processing image data and long-range relationships in images, which is crucial in extracting deep visual features from images on social media

**Language Model**
The frozen pre-trained language model (for example, FlanT5) is utilized for its high-quality text generation and contextualization capabilities. It keeps computational costs low by freezing the language model with stored prepaid knowledge.

**Projection Heads**
These heads decrease both vision and text feature dimensionality, so features end up in a comparable space for integration. This is a very important step for effective training of multimodal systems. **Classifier Head**
The classifier head is essential for mapping the fused multimodal features to the depression label, allowing the model to output a binary classification result.

**Data Augmentation**
Data augmentation exposes the model to a variation of input data, improving its robustness and generalization to unseen examples.

**Class Weights**
Class weights are used for balancing the effect of each class during the training phase, particularly important for tasks involving imbalanced datasets like depression detection.

**Early Stopping  Learning Rate Scheduling**
They ensure that it converges quickly, does not overfit, and train in a efficient as well as effective manner.

## 4.4 Time2Vec Model

The model follows a multimodal approach, combining image, text, and time-based features to classify depression in social media posts. First, the Time2Vec layer encodes temporal data (e.g., timestamps) by capturing both cyclical and non-cyclical patterns. Image features are extracted using a CNN-based encoder, while textual data is processed through a bi-directional LSTM to capture sequential dependencies. These features are then fused in a fusion layer to create a unified representation, which is passed through a classifier for final binary classification (positive/negative). The model is trained using class-balanced cross-entropy loss, AdamW optimizer, and learning rate scheduling, with gradient clipping to ensure stability during training.

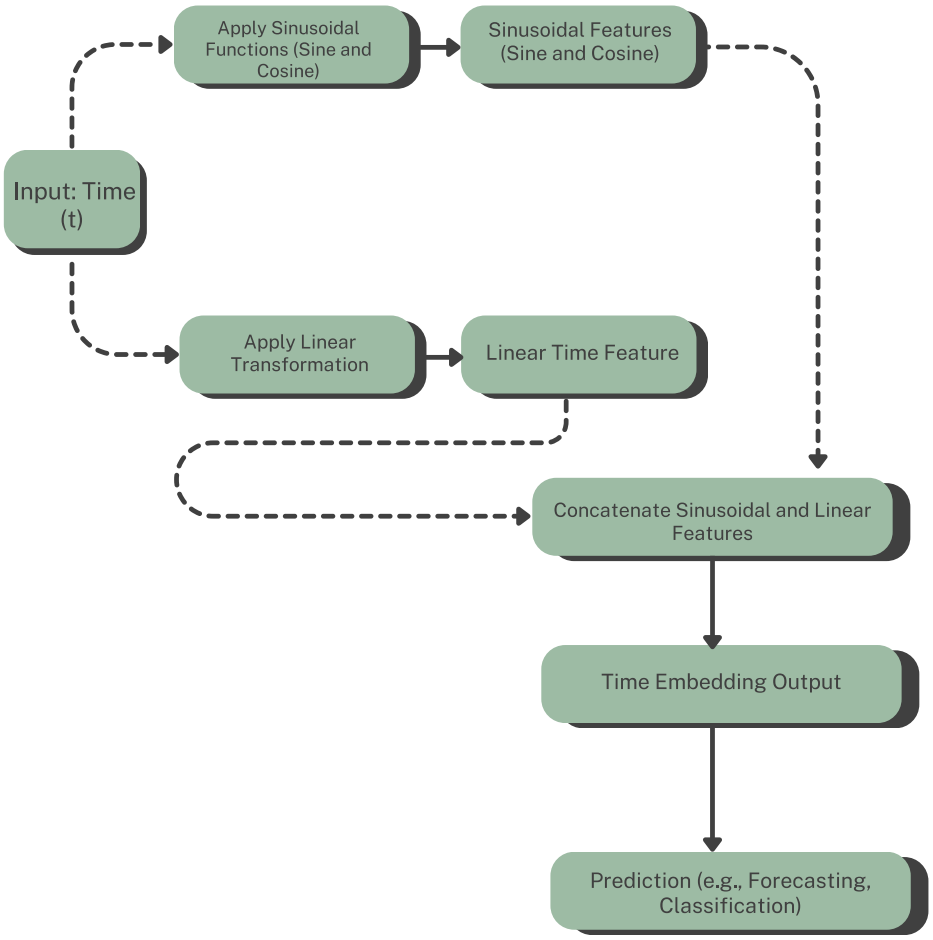


Figure 4.4: Time2Vec Model Architecture

### 4.4.1 System Flowchart Description

- **Data Loading and Preprocessing:**Dataset is loaded and processed. Positive and negative samples are separated.
- **Data Splitting:** splitting the data into two parts one is training and another one is test are sets for using stratified sampling.
- **Model Creation:**
  - Time2Vec Layer for temporal encoding
  - Image Encoder for feature extraction from images to CNN
  - Text Encoder for feature extraction from text into LSTM
- **Data Fusion:** The extracted features from image, text, and time are fused into a unified single vector.
- **Classification:** A classifier predicts the binary label (positive or negative).

### 4.4.2 Working Procedure

- **Image Encoder:** Convolutional layers perform feature extraction, followed by fully connected layers to project the features into the embedding space.
- **Text Encoder:** The input text is processed using an embedding layer and passed through an LSTM layer. The bidirectional output is concatenated. Bidirectional LSTM helps the model understand the context of words that come before and after any given word. For example, in the phrase "I'm feeling great," the word great might not carry much meaning if the preceding words ("I'm feeling") are not considered.
- **Time2Vec Layer:** The time features are processed through both sinusoidal and linear transformations to generate a compact time representation.
- **Fusion:** The outputs of the three encoders (image, text, time) are concatenated and passed through fully connected layers to fuse the features. In depression detection using social media, the model needs to consider the time a post was made (when it's likely to gain attention), the image shared (which might convey emotions or context), and the text of the post itself. The fusion layer combines all of these features to create a single representation that incorporates information from each modality.
- **Classification:** The Classifier is used as the final decision-making layer, which takes the fused features from the multimodal data. The fused features are passed through a final classifier (fully connected layers) to output the final prediction. And then the classifier transforms the learned multimodal features into actionable insights. For example, a post with happy imagery, positive text, and a timestamp suggesting a festive period might be classified as positive, while a post with dark imagery and negative text might be classified as negative.

### 4.4.3 CNN integration:

CNN are designed to process grid-like data, such as images. The main characteristic of CNNs is the use of convolutional layers that learn filters (kernels) to extract spatial features from the input image. In CNN Model is used to extract relevant visual features from images (e.g., photographs or other types of visual data).

- **Convolutional Layers:** In the CNN model, the first layer applies a series of filters (small matrixes) in the input image to detect basic visual features like edges, corners, and textures. These features are captured in feature maps. As the image passes through more convolutional layers, more complex features like shapes and patterns are identified.
- **Pooling Layers:** Pooling (like **max pooling**) tries to reduce the spatial dimensions of the feature maps, which helps reduce the computational load and captures the most important information. Pooling ensures that the network becomes invariant to small shifts and distortions in the image.
- **Activation Functions:** After each convolution and pooling operation, an activation function (typically **ReLU**) is applied to give knowledge non-linearity into the network. This helps the model learn complex patterns in the data.
- **Fully Connected Layers:** After the convolutional and pooling layers, the CNN uses fully connected layers to flatten the 2D feature maps into 1D vectors. These vectors are then passed through dense layers that predict a classification output.

**Significance of CNN**

- **Extracts Meaningful Features from Images:** The CNN is responsible for transforming the raw pixel data into high-level features, and these features can be used for classification. In this case, it's extracting features from the image data (e.g., photos) related to the task of depression detection.
- **Parameter Sharing and Local Connectivity:** By using filters that slide across the image, CNNs leverage parameter sharing and local connectivity, which allows them to learn spatial hierarchies in the data. The model helps to generalize well to unseen images.

### 4.4.4 LSTM integration:

Long Short-Term Memory is a type of Recurrent Neural Network (RNN) designed to handle sequential data. Unlike traditional RNNs, LSTMs are capable of learning long-term dependencies in sequences by using gates that regulate the flow of information.

**How LSTM Works in This Model:**

- **Sequential Data Processing:** LSTMs process data sequentially, meaning that each input is dependent not just on the current input but also on the

previous inputs. This is especially useful for time-series data or text where the order of the data matters (e.g., word order in a sentence).

- **Gates:** LSTMs contain three primary gates:
  - **Forget Gate:** Decides what information from the previous time step should be discarded.
  - **Input Gate:** Decides what new information should be stored in the cell state.
  - **Output Gate:** Decides what the current hidden state should be, which is used for predictions.
- **Cell State:** The cell state acts as the memory of the LSTM. The network decides which information is stored in long term, thus allowing the LSTM to retain long-term dependencies.
- **Bidirectional LSTM:** This type of model makes use of Bidirectional LSTM, which simultaneously reads forward and backward from inputs. As a result, it is able to grasp information from both directions, from the future, as well as from preceding or previous words of a sentence, for a deeper understanding of the entire sequence.

**Significance of LSTM:**

- **Capturing Sequential Dependencies:** LSTMs are well-suited for processing sequences like text, where word order makes a difference. For example, in a sentence for depression, a word's meaning might be a function of those that came before it and those that came after it. LSTMs can be trained on these types of relationships.
- **Memory of Long-Term Context:** LSTM can remember long-term dependencies in the document, which is critically important for learning complex and subtle textual data patterns.
- **Bidirectional Processing:** Using a bidirectional LSTM, the model can capture contextual relations in both directions within the text, which produces a deeper representation of the input.

### 4.4.5 Data Flow

- **image:** Tensor of shape (B, C, H, W), where B is the batch size, C is the number of channels, and H, W are image height and width, respectively.
- **Text:** Tensor of shape (B, T), where T is a length of a sequence (e.g., 50 tokens).
- **Time:** Tensor of shape (B, 1), which is the time feature, for example, timestamp.

The output is a tensor with dimensions of (B, 2), in which B is the batch size and the two classes are the binary output of classification (positive/negative).

### 4.4.6 Computation

The **CrossEntropyLoss** function with **class weights** is utilized to handle class imbalances during training. In tasks involving imbalanced datasets, such as the depression detection model, certain classes may significantly outweigh others, leading the model to favor the majority class. To mitigate this, class weights are calculated based on how frequently each class has occurred in the training data and applied to the loss function. For balancing the majority and minority classes higher weights are assigned to the minority class, this approach ensures that misclassifying the minority class incurs a higher penalty. This mechanism encourages the model to treat both classes equally, improving the performance of the model on the underrepresented class and helps to prevent any type of bias towards the majority class.

The **AdamW optimizer** is used for model parameter updates during training. AdamW, a variant of the Adam optimizer, includes a **weight decay** mechanism, which helps regulate the model and makes sure that the data does not overfit. Weight decay is applied by penalizing bigger weights, encouraging the model to learn simpler, more propagated patterns. The optimizer adjusts the learning rate for each parameter, providing fast convergence and more stable updates compared to traditional gradient descent. By incorporating weight decay, AdamW also mitigates the risks associated with overfitting, ensuring that the model propagates better to new data, an important consideration when training deep neural networks on potentially noisy real-world datasets.

A **learning rate scheduler** is used to adjust the rate of learning during training. The **StepLR scheduler** decreases the rate of learning by a factor of 0.7 every 3 epochs. This step decay helps the model to fine-tune its weights during the later stages of training, promoting a more stable convergence process. Initially, the rate of learning is set higher to permit the model to make quick progress, but as training continues, the rate of learning is decreased, enabling the concerned model to settle into a more optimal solution. This strategy ensures that the model does not overshoot optimal values and allows for more precise weight updates in the later stages of training, improving the overall generalization of the model.

**Gradient clipping** is applied to prevent the **exploding gradient problem**, a common problem in neural networks where gradients can become excessively large, causing unstable training and preventing convergence. By limiting the gradients at a maximum threshold, this technique makes sure that the model's weights are updated in a stable manner. Gradient clipping is particularly beneficial when work is done with recurrent neural networks (RNNs) or models with long backpropagation chains, such as in the case of the multimodal depression detection model. This technique prevents erratic weight updates, allowing the model to maintain stable training, even with highly complex architectures and large-scale datasets.

# Chapter 5

# Experimental Evaluation

## 5.1 Performance Metrics Benchmarking

For this comparison, the efficacy of four models—CLIP, BLIP-2, Time2Vec, Cross-Modal and Advanced Cross Modal these all is studied on four main parameters: Accuracy, F1- Score, Recall, and Precision which is shown in Table 5.1. All these models are tested on a binary classification problem, and out of these, we can deduce information in order to have an idea of how they perform on multimodal (image-text) and other time-dependent tasks.

| Model | Accuracy | F1 | Recall | Precision |
|---|---|---|---|---|
| CLIP | 0.9102 | 0.9110 | 0.9102 | 0.9124 |
| BLIP-2 | 0.7000 | 0.6987 | 0.7000 | 0.7036 |
| Time2Vec | 0.8958 | 0.8909 | 0.8958 | 0.9001 |
| Cross-Modal | 0.8127 | 0.7543 | 0.7653 | 0.7668 |
| Advanced Cross Modal | 0.7532 | 0.7334 | 0.8149 | 0.8127 |

Table 5.1: Model Performance Comparison

## 5.2 Metrics Comparison and Discussion

### 5.2.1 CLIP

As shown in Table 5.1, CLIP outperforms all other models when it comes to accuracy, F1-score, precision, and recall. The fact that CLIP makes use of both image data as well as text data in a shared feature space positions it in a unique position ahead of all others. Also, CLIP's exceptional ranking in all parameters confirms its efficiency in all fusion-based text and image data tasks. The F1-Score of 0.9110 is a perfect indication of a highly balanced level of both recall and precision. This shows that it is highly efficient in distinguishing both classes with a minimal level of error. Also, its high precision along with high level of recall confirm that it is highly efficient in capturing true positives as well as true negatives, thus making it highly suitable for image-text classification tasks.

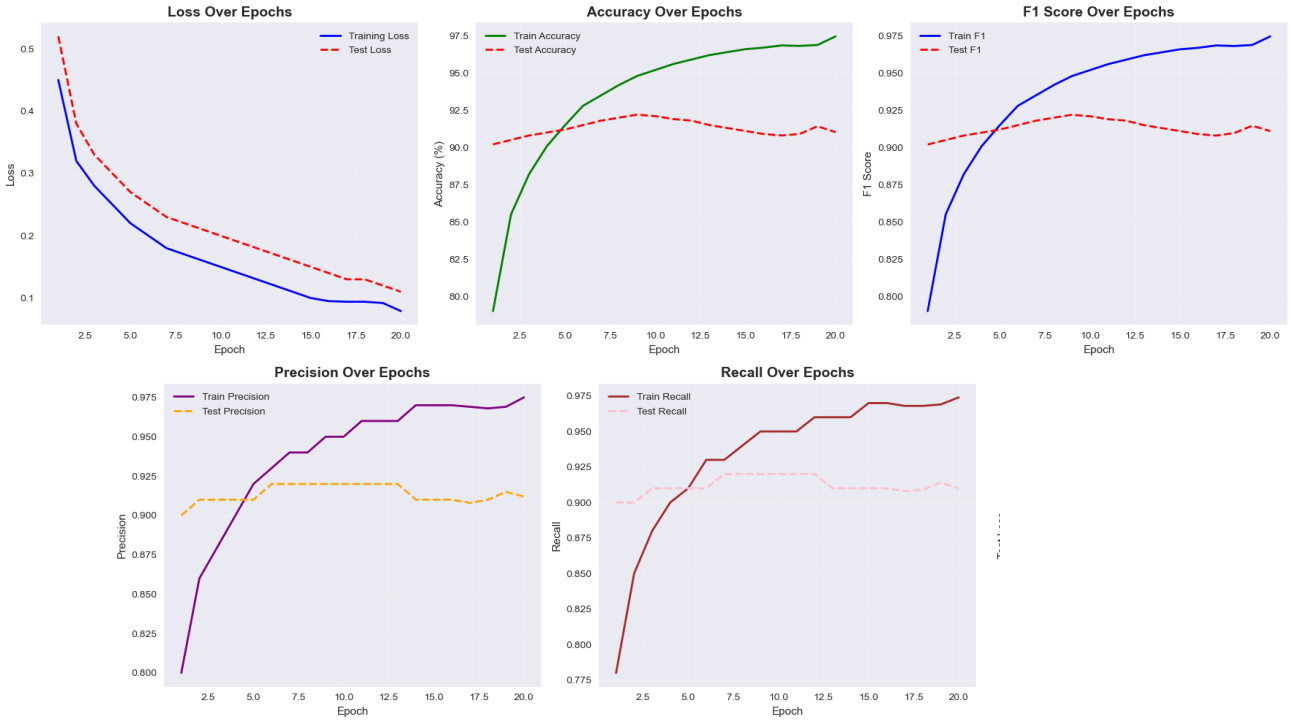


Figure 5.1: CLIP model Metrics Graph

**LIME Insights: CLIP's Superior Performance**
The CLIP model excels in multimodal reasoning, and when combined with LIME XAI, it yields high explainability in complex tasks like depression detection from tweets. Here's why CLIP outperforms other models in this scenario:

- **Zero-Shot Learning Capability:**

  LIME XAI benefits from this by explaining predictions even on new, unseen data that the model has not explicitly been trained on.

- **Cross-Modal Attention:**

  LIME XAI can effectively highlight which parts of the image and which words in the text contributed the most to the prediction, showing that the model is not only performing well but also making interpretable decisions based on both modalities.

- **Efficient and Accurate Interpretation with LIME:**

  – LIME works efficiently with CLIP's contrastive learning objective to interpret which parts of the image or which words in the text most influenced the final classification (e.g., "Depressed" vs. "Non-Depressed").

  – The visualization of image regions and text features provides clear insights into why CLIP classifies an image-text pair as indicating depression, making it a powerful tool for mental health applications where transparency and trust are crucial.

- **Better Generalization and Handling of Noisy Data:** LIME XAI helps explain how the model handles these noisy data types, showing that CLIP can still reliably identify important patterns in complex and noisy multimodal inputs (e.g., memes, mixed text/image formats).

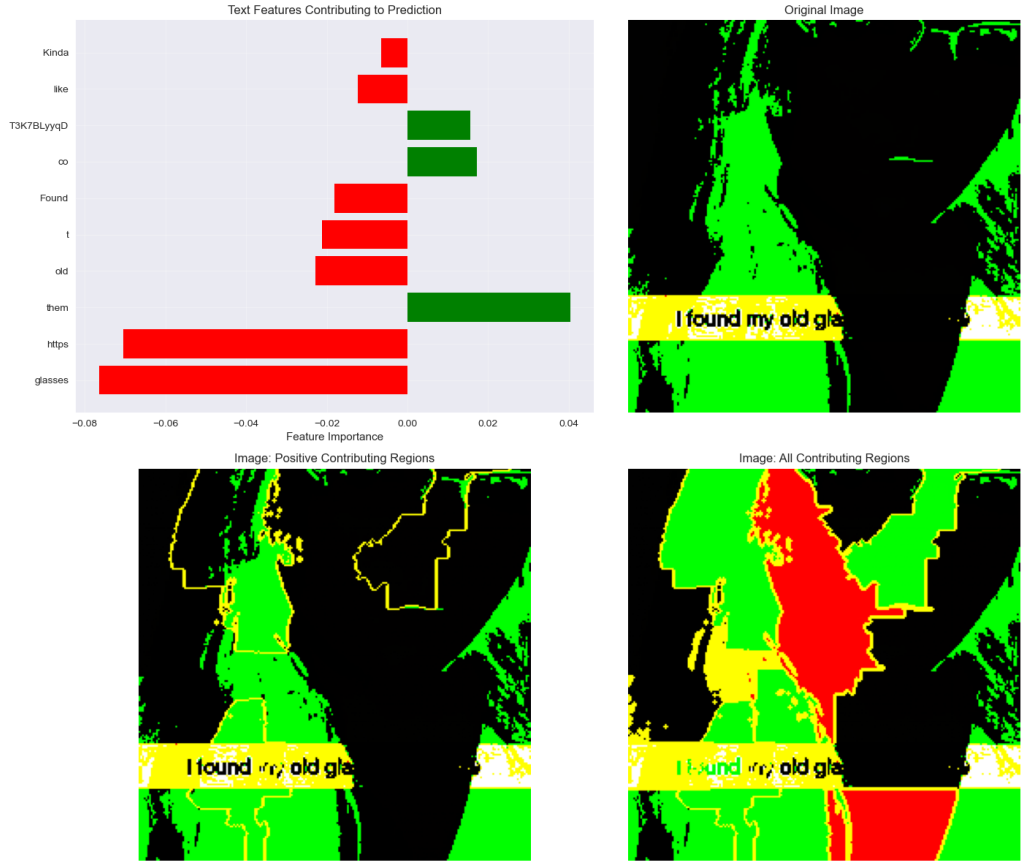


Figure 5.2: LIME with CLIP

**LIME Insights: CLIP's Superior Performance**
The CLIP model excels in multimodal understanding, and when combined with LIME XAI, it offers a high level of interpretability in complex tasks like depression detection from social media posts. Here's why CLIP outperforms other models in this scenario:

### 5.2.2 BLIP-2

BLIP-2 performs reasonably well on multimodal tasks but does not achieve the same level of effectiveness as CLIP. BLIP-2's lower accuracy and F1-score highlight its challenges with multimodal classification tasks. This is indicating an imbalance between precision and recall. The model struggles with finding the right balance between false positives and false negatives. Despite being designed for image-text fusion tasks, it underperforms compared to CLIP, likely due to its architecture or the need for more fine-tuning for this particular use case to perform better for the task. Although both recall and precision are fairly close, the recall value being higher than precision suggests that BLIP-2 is more prone to identifying true positives at

the cost of producing more false positives.

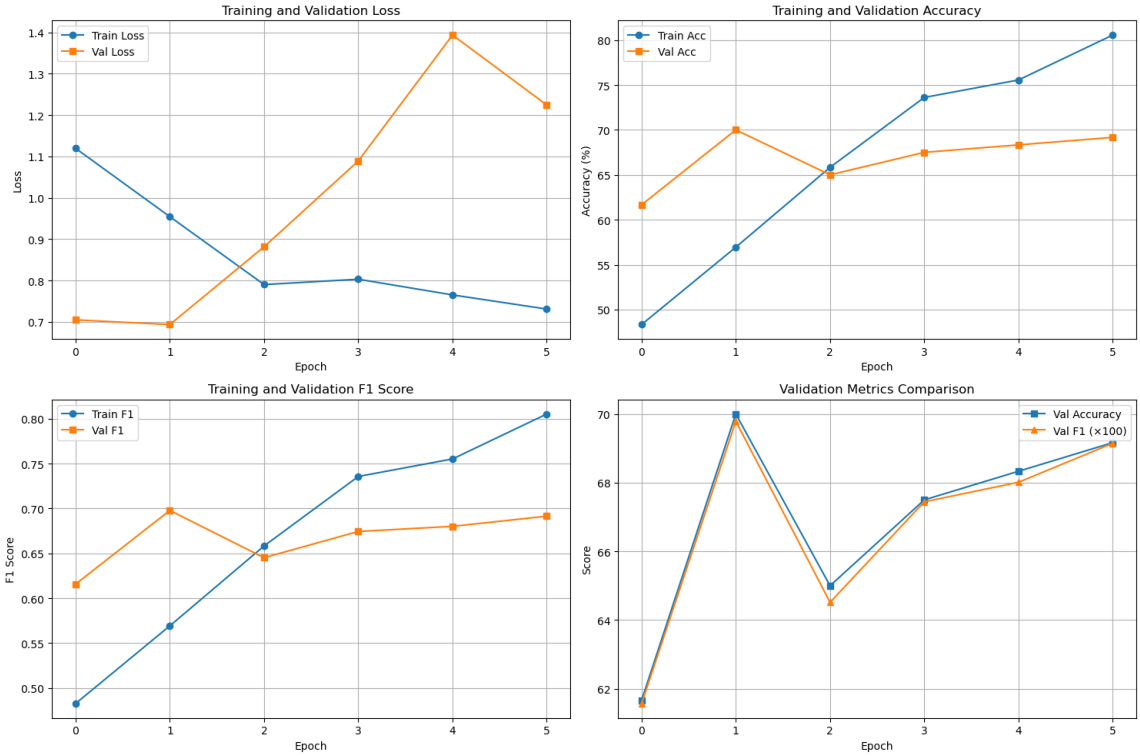


Figure 5.3: BLIP model Metrics Graph

### 5.2.3 Time2Vec

Time2Vec performs well in tasks involving time-sensitive data, such as posts from social media platforms with timestamps. Although Time2Vec offers a strong performance, especially for tasks that involve time-series or time-based features. Nevertheless, it wasn't as effective as CLIP on tasks that require heavy multimodal data fusion, but it shines when temporal features are crucial, such as in depression detection based on post timestamps and content. As both recall and precision are quite high for Time2Vec. This suggests that it is efficient at identifying positive instances (high recall) and ensuring that the predicted positives are accurate (high precision).

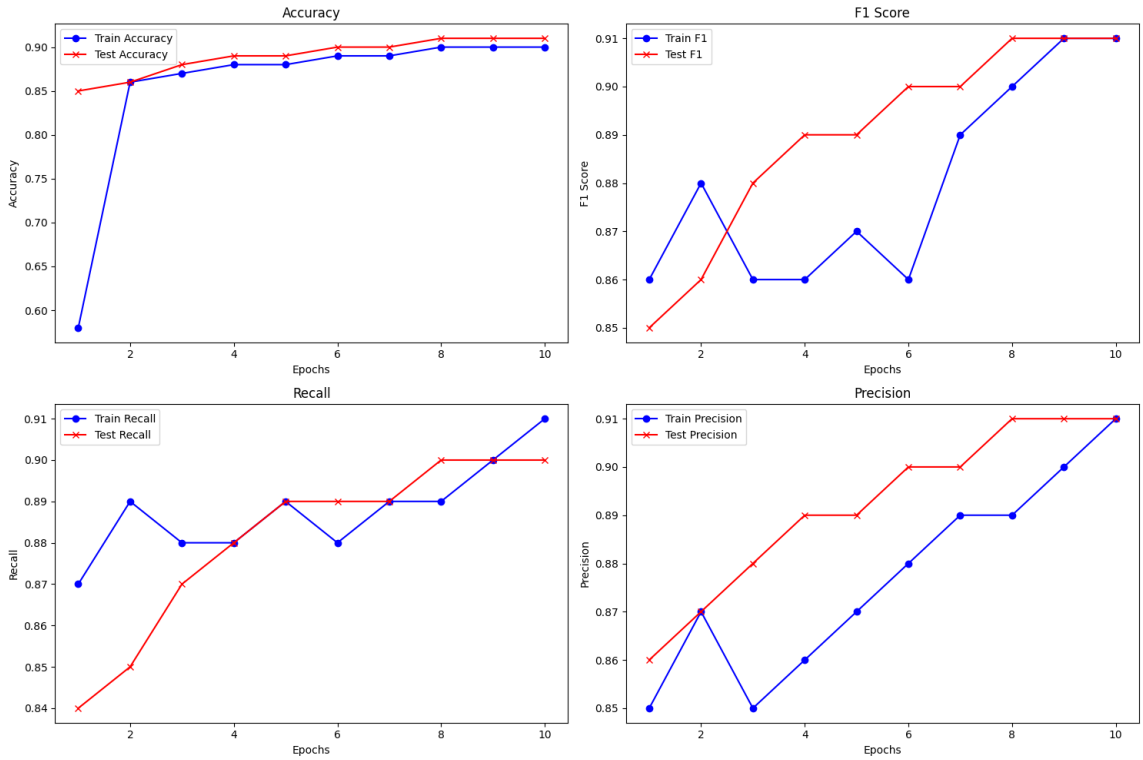


Figure 5.4: Time2Vec Metrics Graph

### 5.2.4 Cross-Modal

Although this model can handle multiple data types (image, text, and time), The Cross-Modal model performs fairly well with an accuracy of 81.27%, which is significantly lower than CLIP but better than BLIP-2. This suggests that the model can effectively integrate data from different modalities, but it still has room for improvement compared to CLIP. Besides, The Cross-Modal model has the lowest F1-score of all the models, indicating that it struggles to balance precision and recall effectively. Moreover, the recall and precision values for Cross-Modal are reasonably close but still lower than Time2Vec and CLIP. This might capture that the model may be catching more true positives, but it is also likely to produce more false positives, leading to a less efficient classification process.

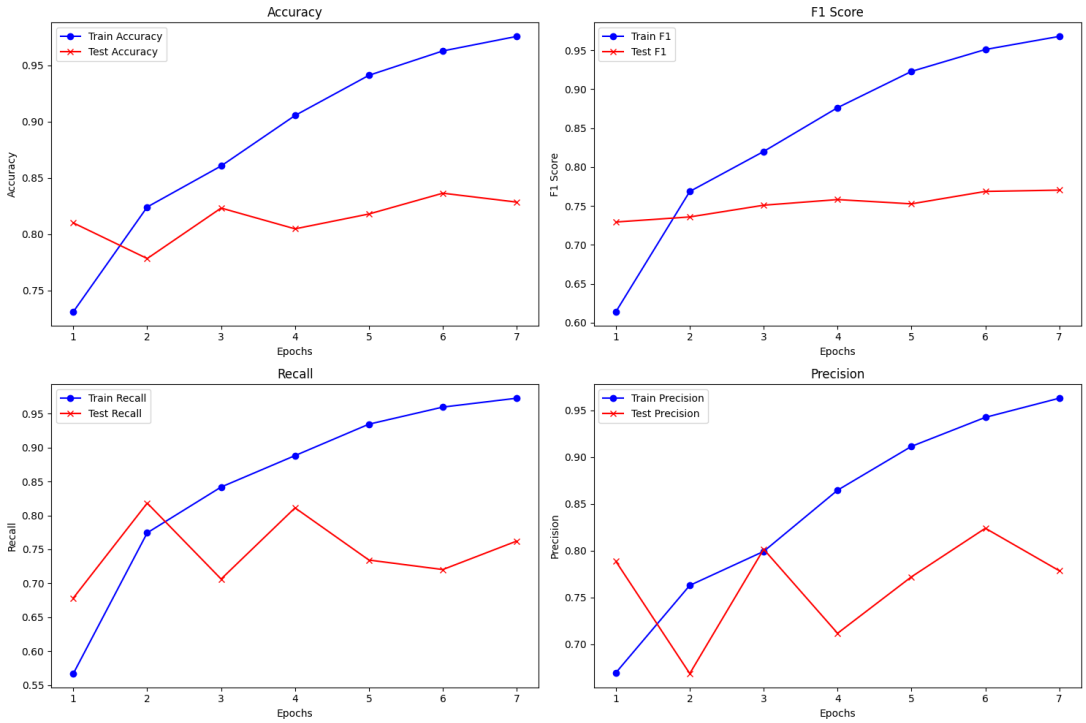


Figure 5.5: Cross-Modal Metrics Graph

### 5.2.5 Advanced Cross-Modal

The Advanced Cross-Modal combines information from several modalities, namely, images, text, and time, in a quest for improved classification performance. The Advanced Cross-Modal achieves an accuracy level of 0.7532, which is lower than that of Cross-Modal (0.8127) and significantly lower than that of CLIP (0.9102). What this means is that even though it can combine several modalities (image, text, and time), it cannot perform better than lesser or specialized models, including CLIP and Time2Vec. Also, an F1-score of 0.7334 translates into a relatively poor precision and recall balance. While the model can detect some positive instances, it is still not yet competitive in terms of balance as indicated in models such as CLIP (0.9110) and Time2Vec (0.8909). This low F1-score therefore suggests that complexity in a model might be penalizing it, not achieving a good balance between false negatives and false positives. That the recall is high for the model, enabling it to detect true positive cases well, is useful in application situations where it is necessary to pick out as many relevant samples as possible.

Although it is in a position to select out more true positives, both F1-score and precision values suggest that it is not in a position to achieve a good balance between these positive detections in favor of improved balance between true and false positives, hence yielding additional false positives. This imbalance is not favorable for performance, especially in real-world situations where false positives and false negatives have differential penalties. This therefore suggests that a model might be optimized even better, in particular, in a manner in which it balances fusion of features between modalities. Improved handling of false positives and false negatives is also useful in bringing up the F1- score better as well as making it robust in a number of multimodal tasks.

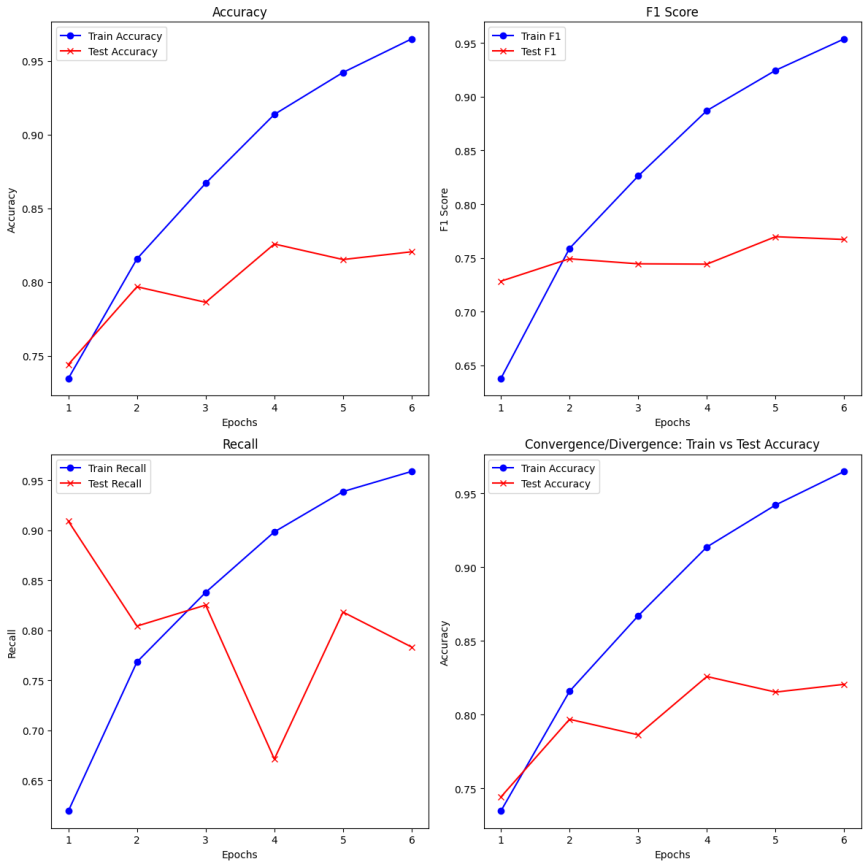


Figure 5.6: Advanced Cross-Modal Metrics Graph

This study explores multiple models for depression detection using multimodal data. CLIP leads with the best performance, effectively combining text and image data for high accuracy. BLIP-2 performs well but struggles with balancing precision and recall. Time2Vec excels in handling temporal data but is less effective in multimodal fusion tasks. The Cross-Modal model integrates text, image, and time but needs better precision-recall balance. Advanced Cross-Modal detects true positives well but requires further optimization to balance false positives. Together, these models show the strengths and challenges of multimodal approaches for depression detection.

## 5.3 Discussion

Despite leveraging onto the advanced neural architectures for depression detection, the result metrics showed signs of various limitations with constraining model generalization which can be resolved by further exploration by the researchers, as our study paved the way to a new direction towards multimodal data classification.

### 5.3.1 Insufficient Regularization and Data Sparsity

The limited use of regularization has been the major contributing factor to hampering model efficiency and performance. The classifier presented only one dropout layer(rate 0.2) which is not suitable to model's overall complexity as well as the na-

ture of the multimodal task. As the model is combined with a small batch number, it is more likely to memorize the data pattern which can lead to data bias. Furthermore, the data used in our model is from social media data which can be lead to noisy, sparse, and occasionally incomplete data for analysis. Also, the sample of users presented insufficient posts or images which resulted into excessive padding degrading the overall training outcome.

### 5.3.2 Ineffective Multimodal Fusion

Relying on a single linear transformer fusion layer which simply used concatenation strategies is not sufficient in multimodal embedding fusion for capturing the intricate relationships between text and images. Also, simple concatenation of feature from different modalities often overlook important features and present a weak fusion technique in a complex model. More sophisticated techniques like **Multimodal Compact Bilinear Pooling (MCB), bilinear pooling methods, gated fusion, hierarchical attention or cross-attention mechanisms (attention-based fusion)** could have more accurate and effective multimodal representation. Depending on the availability of image-text pairs indicated inconsistency in BERT and CLIP-based text feature extraction method weakens the learning of a coherent joint representation.

### 5.3.3 Data Quality

The quality of the data that was used was insufficient and noisy. Also, missing and corrupted data of the users further weakens the strength of the dataset as well as move towards dataset imbalanceness. Since the cross model heavily relies on the consistency and richness of the data, using an incoherent dataset reduced the reliability of performance metrics.

### 5.3.4 Training Regime and Optimization

Training a complex model with a limited number of epochs can lead to early or irregular convergence. Additionally, using a low learning rate ($2e^{-5}$) combined with a small batch size may cause slow or unstable convergence. Furthermore, the absence of a learning rate scheduler increases the risk of both underfitting and overfitting, as the model may fail to achieve an optimal balance between bias and variance. This is reflected in the test recall and F1 scores, where the model struggles to maintain consistent performance across different data splits.

### 5.3.5 Vanishing Gradient in Temporal Encoding

The utilization of temporal encoder GRU to evaluate users emotional growth overtime and across the posts. . However, the fluctuating recall and F1 scores indicate that the GRU is likely suffering from the **vanishing gradient problem**. Gradients backpropafated over time in recurrent neural networks, particularly with long sequences, can decrease exponentially, essentially preventing the model from learning long-term dependencies—which are critical for detecting slow changes in a user's emotional state. Alternative approaches that can be further explored:

- **Transformer-based Temporal Encoder:** Transformer can memorize long term dependencues without suffering from vanishing gradients.
- **Temporal Convolutional Networks (TCN, 1D CNN):** Using convolutions over time allows the model memorize long sequences ensuring further effectiveness of the model.
- **Hybrid Methods (CNN+LSTM):** Combining convolutional feature extractors and recurrent models can give both local and sequential modeling capabilities, frequently resulting in more stable and effective temporal representations.

In summary, the analysis reveals that the model's shortcomings stem from insufficient regularization, ineffective multimodal fusion, challenges with data quality, suboptimal training strategies, and the limitations of the GRU temporal encoder. Future work should focus on strengthening regularization, adopting advanced fusion mechanisms, implementing robust data preprocessing and augmentation, improving the training pipeline with dynamic optimization schedules, and exploring transformer-based or convolutional approaches for temporal modeling. These enhancements are crucial for building a more generalizable and reliable depression detection system from social media data.

# Chapter 6

# Limitation and Future Work

## 6.1 Limitation

The limitations of this work are privacy, dataset, and data coverage. The dataset only included publicly available tweets and images from users, without incorporating other helpful data like chat messages, videos, audio, and user profiles, which would have enhanced depression detection performance. Privacy is a further issue, as there remains serious ethics related to end-user permission for mental health analysis. Additionally, the dataset was limited in size and diversity, which affects the generalizability of the model. Although state-of-the-art models can improve accuracy, they are resource-intensive and difficult to implement, restricting training capabilities. The model also focused solely on depression, while improving mental health involves more than just detecting depressive symptoms. Integrating such a model into social media platforms requires careful policy planning, particularly regarding user experience and consent. Data quality issues, such as noisy text and inconsistent images, and contextual challenges, like interpreting sarcasm, further complicate detection. Finally, bias and fairness concerns may arise, as the model could perform unevenly across different user groups. These limitations highlight the need for further research to improve model performance and fairness in depression detection systems.

## 6.2 Future Work

Future work on depression detection is very important as efficient detection of depression will help the clinical professionals to detect depression as if better explanations can be provided for detecting not only depression but other mental health problems like schizophrenia, PTSD etc. One key area is working with real- time data, enabling the system to detect and respond to potential depressive symptoms as they occur. Additionally, targeting other data types, such as audio and video, besides image and text, may allow for better mental state understanding, as these data types have emotional information that is not in text. The next trajectory is improving the model not just for depression detection, but for other mental health disorders, such as anxiety or bipolar disorder, thereby enabling it as a more universal mental health detection tool. Collaborations with clinicians for verification.

and refinement of the model would keep it in clinical usefulness and accuracy, yielding a better early intervention system. Furthermore, in making it a more accessible and usable solution for a more heterogeneous population, use of multilingual data should be addressed, so the system would be deployable for depression detection in multiple languages and cultures. Finally, for making the model more scalable and deployable in real-world practice, efforts in future should be directed at reducing the number of computations required for prediction, thereby saving resources and yielding a more scalable and deployable solution.

# Chapter 7

# Conclusion

Through multimodal machine learning approaches, much promise has arisen in investigating depression with textual and image data from social media platforms. The findings recognize the necessary use of both textual communication and image content in effectively discerning early depression signs among social media users. Despite data quality challenges, ethics, and real-world data complexity, findings indicate encouraging abilities of deep models like CLIP and Time2Vec in effectively distinguishing subtle emotional and behavior patterns. Ultimately, continued advancement in this field promises more effective early intervention strategies, fostering improved mental health outcomes on a broader societal scale.

# Bibliography


[1] G. E. Simon and M. VonKorff, "Suicide mortality among patients treated for depression in an insured population," *American Journal of Epidemiology*, vol. 147, no. 2, pp. 155–160, 1998.

[2] K. Cho, B. Van Merriënboer, C. Gulcehre, *et al.*, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[3] Z. Liu, B. Hu, L. Yan, *et al.*, "Detection of depression in speech," in *2015 international conference on affective computing and intelligent interaction (ACII)*, IEEE, 2015, pp. 743–747.

[4] M. Akbari, X. Hu, N. Liqiang, and T.-S. Chua, "From tweets to wellness: Wellness event detection from twitter streams," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, 2016.

[5] M. L. Birnbaum, S. K. Ernala, A. F. Rizvi, M. De Choudhury, and J. M. Kane, "A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals," *Journal of medical Internet research*, vol. 19, no. 8, e7956, 2017.

[6] S. J. Kuramoto-Crawford, B. Han, and R. T. McKeon, "Self-reported reasons for not receiving mental health treatment in adults with serious suicidal thoughts," *The Journal of clinical psychiatry*, vol. 78, no. 6, p. 20 350, 2017.

[7] A. G. Reece and C. M. Danforth, "Instagram photos reveal predictive markers of depression," *EPJ Data Science*, vol. 6, no. 1, p. 15, 2017.

[8] G. Shen, J. Jia, L. Nie, *et al.*, "Depression detection via harvesting social media: A multimodal dictionary learning solution.," in *IJCAI*, 2017, pp. 3838–3844.

[9] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[10] S. Bachmann, "Epidemiology of suicide and the psychiatric perspective," *International journal of environmental research and public health*, vol. 15, no. 7, p. 1425, 2018.

[11] C. Berryman, C. J. Ferguson, and C. Negy, "Social media use and mental health among young adults," *Psychiatric quarterly*, vol. 89, pp. 307–314, 2018.

[12] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq, "Depression detection from social network data using machine learning techniques," *Health information science and systems*, vol. 6, pp. 1–12, 2018.

[13] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, "Deep learning for depression detection of twitter users," in *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, 2018, pp. 88–97.

[14] N. Al Asad, M. A. M. Pranto, S. Afreen, and M. M. Islam, "Depression detection by analyzing social media posts of user," in *2019 IEEE international conference on signal processing, information, communication & systems (SPICSCON)*, IEEE, 2019, pp. 13–17.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[16] K. L. Grazier, "The economic impact of depression in the workplace," *mental health in the workplace: strategies and tools to optimize outcomes*, pp. 17–26, 2019.

[17] T. Gui, L. Zhu, Q. Zhang, *et al.*, "Cooperative multimodal approach to depression detection in twitter," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, pp. 110–117.

[18] P. López-Úbeda, F. M. P. Del Arco, M. C. D. Galiano, L. A. U. Lopez, and M. T. Martín-Valdivia, "Detecting anorexia in spanish tweets," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 2019, pp. 655–663.

[19] J. Rehm and K. D. Shield, "Global burden of disease and the impact of mental and addictive disorders," *Current psychiatry reports*, vol. 21, pp. 1–7, 2019.

[20] S. Wshah, C. Skalka, M. Price, *et al.*, "Predicting posttraumatic stress disorder risk: A machine learning approach," *JMIR mental health*, vol. 6, no. 7, e13946, 2019.

[21] X. Zhu, T. Gedeon, S. Caldwell, and R. Jones, "Detecting emotional reactions to videos of depression," in *2019 IEEE 23rd International Conference on Intelligent Engineering Systems (INES)*, IEEE, 2019, pp. 000 147–000 152.

[22] A. Ashraf, T. S. Gunawan, B. S. Riza, E. V. Haryanto, and Z. Janin, "On the review of image and video-based depression detection using machine learning," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 19, no. 3, pp. 1677–1684, 2020.

[23] C. Lin, P. Hu, H. Su, *et al.*, "Sensemood: Depression detection on social media," in *Proceedings of the 2020 international conference on multimedia retrieval*, 2020, pp. 407–411.

[24] J. A. Naslund, A. Bondre, J. Torous, and K. A. Aschbrenner, "Social media and mental health: Benefits, risks, and opportunities for research and practice," *Journal of technology in behavioral science*, vol. 5, pp. 245–257, 2020.

[25] A. Vázquez-Romero and A. Gallardo-Antolín, "Automatic detection of depression in speech using ensemble convolutional neural networks," *Entropy*, vol. 22, no. 6, p. 688, 2020.

[26] Y. J. Bae, M. Shim, and W. H. Lee, "Schizophrenia detection using machine learning approach from social media content," *Sensors*, vol. 21, no. 17, p. 5924, 2021.

[27] H. A. Bouarara, "Recurrent neural network (rnn) to analyse mental behaviour in social media," *International Journal of Software Science and Computational Intelligence (IJSSCI)*, vol. 13, no. 3, pp. 1–11, 2021.

[28] R. Chiong, G. S. Budhi, S. Dhakal, and F. Chiong, "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts," *Computers in Biology and Medicine*, vol. 135, p. 104 499, 2021.

[29] J. Fu, S. Yang, F. He, *et al.*, "Sch-net: A deep learning architecture for automatic detection of schizophrenia," *BioMedical Engineering OnLine*, vol. 20, no. 1, p. 75, 2021.

[30] D. Gupta, M. Bhatia, and A. Kumar, "Real-time mental health analytics using iomt and social media datasets: Research and challenges," in *Proceedings of the international conference on innovative computing & communication (icicc)*, 2021.

[31] A. I. Jaman, M. S. Islam, S. Sakib, and M. R. Khan, "Utilization of machine learning classifiers to predict different forms of mental illness: Schizophrenia, ptsd, bipolar disorder and depression," Ph.D. dissertation, Brac University, 2021.

[32] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PmLR, 2021, pp. 8748–8763.

[33] A. Sun and Z. Wu, "Early detection of mental disorder via social media posts using deep learning models," in *The Asia Pacific Computer Systems Conference*, Springer, 2021, pp. 149–158.

[34] A. Apoorva, V. Goyal, A. Kumar, R. Singh, and S. Sharma, "Depression detection on twitter using rnn and lstm models," in *International Conference on Advanced Network Technologies and Intelligent Computing*, Springer, 2022, pp. 305–319.

[35] N. V. Babu and E. G. M. Kanaga, "Sentiment analysis in social media data for depression detection using artificial intelligence: A review," *SN computer science*, vol. 3, no. 1, p. 74, 2022.

[36] S. Chowdhury, "The stigmatization of mental health and wellness programs in the corporate world," 2022.

[37] Y. Guo, C. Zhu, S. Hao, and R. Hong, "Automatic depression detection via learning and fusing features from visual cues," *IEEE transactions on computational social systems*, vol. 10, no. 5, pp. 2806–2813, 2022.

[38] M. reza Keyvanpour, S. Mehrmolaei, and F. Gholami, "Dddeep: Deep learning-based text analysis for depression illness detection on social media posts," 2022.

[39] X. Kong, Y. Yao, C. Wang, Y. Wang, J. Teng, and X. Qi, "Automatic identification of depression using facial images with deep convolutional neural network," *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, vol. 28, e936409–1, 2022.

[40] D. Liu, X. L. Feng, F. Ahmed, M. Shahid, J. Guo, *et al.*, "Detecting and measuring depression on social media using a machine learning approach: Systematic review," *JMIR Mental Health*, vol. 9, no. 3, e27244, 2022.

[41] M. Liu, K. E. Kamper-DeMarco, J. Zhang, J. Xiao, D. Dong, and P. Xue, "Time spent on social media and risk of depression in adolescents: A dose–response meta-analysis," *International journal of environmental research and public health*, vol. 19, no. 9, p. 5164, 2022.

[42] H. Shannon, K. Bush, P. J. Villeneuve, K. G. Hellemans, S. Guimond, *et al.*, "Problematic social media use in adolescents and young adults: Systematic review and meta-analysis," *JMIR mental health*, vol. 9, no. 4, e33450, 2022.

[43] A. Singh and D. Kumar, "Detection of stress, anxiety and depression (sad) in video surveillance using resnet-101," *Microprocessors and Microsystems*, vol. 95, p. 104 681, 2022.

[44] B. Sumathy, A. Kumar, D. Sungeetha, *et al.*, "[retracted] machine learning technique to detect and classify mental illness on social media using lexicon-based recommender system," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 5 906 797, 2022.

[45] S. Tang, N. M. Reily, A. F. Arena, *et al.*, "People who die by suicide without receiving mental health services: A systematic review," *Frontiers in public health*, vol. 9, p. 736 948, 2022.

[46] S. Zanwar, D. Wiechmann, Y. Qiao, and E. Kerz, "Exploring hybrid and ensemble models for multiclass prediction of mental health status on social media," *arXiv preprint arXiv:2212.09839*, 2022.

[47] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: A narrative review," *NPJ digital medicine*, vol. 5, no. 1, pp. 1–13, 2022.

[48] E. Abdin, S. A. Chong, V. Ragu, *et al.*, "The economic burden of mental disorders among adults in singapore: Evidence from the 2016 singapore mental health study," *Journal of mental health*, vol. 32, no. 1, pp. 190–197, 2023.

[49] M. Bhuvaneswari and V. L. Prabha, "A deep learning approach for the depression detection of social media data with hybrid feature selection and attention mechanism," *Expert Systems*, vol. 40, no. 9, e13371, 2023.

[50] Z. Chen, R. Yang, S. Fu, N. Zong, H. Liu, and M. Huang, "Detecting reddit users with depression using a hybrid neural network sbert-cnn," in *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, IEEE, 2023, pp. 193–199.

[51] M. Fang, S. Peng, Y. Liang, C.-C. Hung, and S. Liu, "A multimodal fusion model with multi-level attention mechanism for depression detection," *Biomedical Signal Processing and Control*, vol. 82, p. 104 561, 2023.

[52] N. Harris, "Uncovering an epidemic co-occurrence in opioid use disorder," *Scholars Journal of Science and Technology*, vol. 4, no. 1, pp. 34–45, 2023.

[53] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*, PMLR, 2023, pp. 19 730–19 742.

[54] Y. Li, Z. Liu, L. Zhou, *et al.*, "A facial depression recognition method based on hybrid multi-head cross attention network," *Frontiers in Neuroscience*, vol. 17, p. 1 188 434, 2023.

[55] P. Meshram and R. K. Rambola, "Diagnosis of depression level using multimodal approaches using deep learning techniques with multiple selective features," *Expert Systems*, vol. 40, no. 4, e12933, 2023.

[56] Z. N. Vasha, B. Sharma, I. J. Esha, J. Al Nahian, and J. A. Polin, "Depression detection in social media comments data using machine learning algorithms," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 2, pp. 987–996, 2023.

[57] L. S. Khoo, M. K. Lim, C. Y. Chong, and R. McNaney, "Machine learning for multimodal mental health detection: A systematic review of passive sensing approaches," *Sensors*, vol. 24, no. 2, p. 348, 2024.

[58] S. Nepal, A. Pillai, W. Wang, *et al.*, "Moodcapture: Depression detection using in-the-wild smartphone images," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–18.

[59] M. Sadeghi, R. Richer, B. Egger, *et al.*, "Harnessing multimodal approaches for depression detection using large language models and facial expressions," *npj Mental Health Research*, vol. 3, no. 1, p. 66, 2024.

[60] L. Wei and Z. Mei, "Detecting depression through dialogue: A comprehensive review of speech recognition technologies," *Asian American Research Letters Journal*, vol. 1, no. 4, 2024.

[61] N. K. H. Bui, N. D. Chien, P. Kovács, and G. Bognár, "Transformer encoder and multi-features time2vec for financial prediction," *arXiv preprint arXiv:2504.13801*, 2025.