BABEŞ BOLYAI UNIVERSITY, CLUJ NAPOCA, ROMÂNIA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Early Discovery of Anxiety/Depression in Teenagers Using Digital Tools

– MIRPR report –

Toth Alexandra-Melania
Vlădău Andra-Ioana
Mihalcea Ioan-Alexandu
Emails: alexandra.melania.toth@stud.ubbcluj.ro,
andra.vladau@stud.ubbcluj.ro, ioan.mihalcea@stud.ubbcluj.ro
Specialisation: Computer Science Romanian

2024-2025

Contents

List of Figures

Chapter 1

Introduction

1.1 Objective:

Anxiety/Depression for teenagers - To enhance mental health support for teenagers by developing digital tools that can proactively identify signs of anxiety and depression. These tools aim to engage adolescents in their digital environments, whether through chatbots, social media, video games, or other innovative platforms, to provide early intervention and emotional support.

1.2 Main Idea:

The Early Discovery of Anxiety/Depression in Teenagers solution uses digital platforms to identify and monitor mental health challenges among adolescents. The system combines multiple approaches, including AI-powered chatbots capable of conversational analysis, sentiment evaluation through social media interactions, and mental health assessments embedded in video game experiences. By engaging teenagers where they spend most of their time, be it online or in games, the system aims to provide real-time insights and early warnings of anxiety or depression. This integrated solution offers a holistic approach, blending digital engagement with predictive analytics and personalized intervention strategies.

1.3 Motivation:

Teen anxiety and depression are pressing issues, often going unnoticed due to social stigma, lack of awareness, or teens' reluctance to seek help. Traditional detection methods are limited in reach and timing, missing critical opportunities for early intervention. By embedding intelligent algorithms in digital spaces where teens naturally engage, like social media, chat apps, and video games, we can monitor signs of mental distress in real time.

AI-powered tools, especially those using natural language processing and machine learning, can analyze behavioural patterns, detect early signs of anxiety or depression, and respond immediately, offering instant, personalized support. These algorithms not only provide proactive intervention, but also reduce barriers to help seek by creating a safe, accessible way for adolescents to explore mental health support in familiar digital environments. This intelligent approach can help prevent more serious issues, offering timely and compassionate care to teens who need it most.

Chapter 2

The Scientific Problem Addressed

The prevalence of anxiety and depression among teenagers has reached alarming levels in recent years, with studies indicating a sharp increase in mental health issues within this demographic. According to the World Health Organization (WHO), nearly 10 - 20% of adolescents worldwide experience mental health conditions, with many cases undiagnosed and untreated. These issues, if unaddressed, can lead to long-term psychological, academic, and social consequences.

A significant challenge in addressing teenage mental health lies in early detection of symptoms. Adolescents often hesitate to seek help due to stigma, lack of awareness, or limited access to traditional healthcare services. Additionally, current diagnostic methods rely heavily on self-reporting or clinical observations, which may not capture the nuanced and early signs of anxiety and depression.

In the digital age, teenagers increasingly interact with technology, spending substantial amounts of time on social media platforms, messaging apps, and video games. These environments often capture behavioral patterns and emotional cues that can serve as indicators of mental health challenges. However, traditional mental health interventions have largely overlooked the potential of these digital spaces to provide proactive and non-intrusive support.

The scientific problem addressed by this project centers on bridging this gap by leveraging advanced digital tools to identify, monitor, and provide early interventions for anxiety and depression. Specifically, the project seeks to answer the following key questions:

- How can artificial intelligence and sentiment analysis be utilized to identify early signs of anxiety and depression in teenagers through their digital interactions?
- What methods can effectively integrate mental health assessments into engaging and non-invasive digital environments such as video games and chatbots?
- How can predictive analytics be used to ensure timely and personalized interventions for at-risk adolescents?

By addressing these questions, this project aims to create an innovative, evidence-based approach to teenage mental health care. It leverages the ubiquity of digital platforms to provide scalable, accessible, and proactive solutions to a growing global health challenge.

Chapter 3

Existing Problem-Solving Methods (Related Work)

3.1 Description of the Datasets

This study leverages two publicly available datasets sourced from Kaggle to develop and evaluate the proposed digital tools for the early detection of anxiety and depression in teenagers. These datasets include the *Reddit Mental Health Dataset* [1] and the *Emotion Recognition Dataset* [2]. Below, we provide a detailed description of each dataset and the preprocessing steps undertaken.

3.1.1 The Datasets

3.1.1.1 Reddit Mental Health Dataset

The *Reddit Mental Health Dataset* contains posts from Reddit users, labeled based on various mental health conditions. For this project, we focused on two relevant categories: *Anxiety* and *Depression*. The dataset originally includes the following categories:

- 0: Stress
- 1: Depression
- 2: Bipolar Disorder
- 3: Personality Disorder
- 4: Anxiety

For our analysis, only posts labeled as *Depression* and *Anxiety* were retained. The *title* column was removed, and the data was concatenated with the *Emotion Recognition Dataset*. To ensure data consistency, null values were removed, and the dataset was balanced to prevent bias towards the larger class.

3.1.1.2 Emotion Recognition Dataset

The *Emotion Recognition Dataset* includes text samples labeled with six emotion categories:

- 0: Anger
- 1: Fear
- 2: Joy
- 3: Love

- 4: Sadness
- 5: Surprise

The labels were transformed into three broader categories to align with our focus on mental health:

- **Anxiety:** Includes posts labeled as *Fear* and *Surprise*.
- **Depression:** Includes posts labeled as *Sadness* and *Anger*.
- **Normal:** Includes posts labeled as *Joy* and *Love*.

The resulting dataset combines information from both sources, providing a diverse set of features for model training and evaluation. Oversampling was applied to the *Normal* class to balance its representation with *Anxiety* and *Depression*.

3.1.2 Quantitative Analysis of the Datasets

Small Dataset: For initial experiments and prototyping, we created a smaller subset by selecting the first 500 rows from each category (*Anxiety* and *Depression*), resulting in a balanced dataset of 1,000 rows.

Large Dataset: The larger dataset used for final model training and testing contains:

- **Anxiety:** 20,290 rows
- **Depression:** 21,796 rows

This dataset provides a robust foundation for training predictive models and evaluating their performance.

3.1.3 Related Work and Usage of Datasets

Both datasets have been used in prior research to analyze mental health and emotional patterns through text. For instance:

- The *Reddit Mental Health Dataset* has been utilized to study language patterns associated with mental health conditions and develop classification models for mental health prediction. Studies have achieved notable accuracy in identifying conditions like anxiety and depression using natural language processing (NLP) techniques.
- The *Emotion Recognition Dataset* has been widely used for emotion classification tasks, leveraging machine learning models to distinguish between different emotional states.

In this project, we extend these works by combining both datasets, re-labeling their categories to focus specifically on anxiety, depression, and normal states, and balancing the data to enhance model performance. By addressing class imbalances and tailoring the datasets to the problem, we aim to improve the detection of mental health conditions in digital interactions.

3.1.4 Preprocessing and Data Transformation

- Extracted relevant labels and mapped them to *Depression*, *Anxiety*, and *Normal*.
- Removed unnecessary columns such as *title* from the Reddit dataset.
- Checked and ensured no null values in the combined dataset.
- Performed oversampling on the *Normal* class to address class imbalance.

The prepared datasets provide a strong foundation for training advanced AI models capable of identifying early signs of mental health challenges in adolescents.

3.2 First Approach - Small Data

3.2.1 Anxiety Data

We used a data set of 500 rows (small data).

- **Hyperparameters:** ID, Text, Target
- **Target:** Normal (246), Anxiety (254)
- **Models:** Support Vector Classification (Support Vector Machine) with encoding Word2Vec
- **Results:**
  - **Accuracy:** 0.64
  - **Precision (average):** 0.6481
  - **Precision for Anxiety Class:** 0.6481
  - **Precision for Normal Class:** 0.6304
  - **Recall:** 0.6731
  - **Confusion Matrix:**

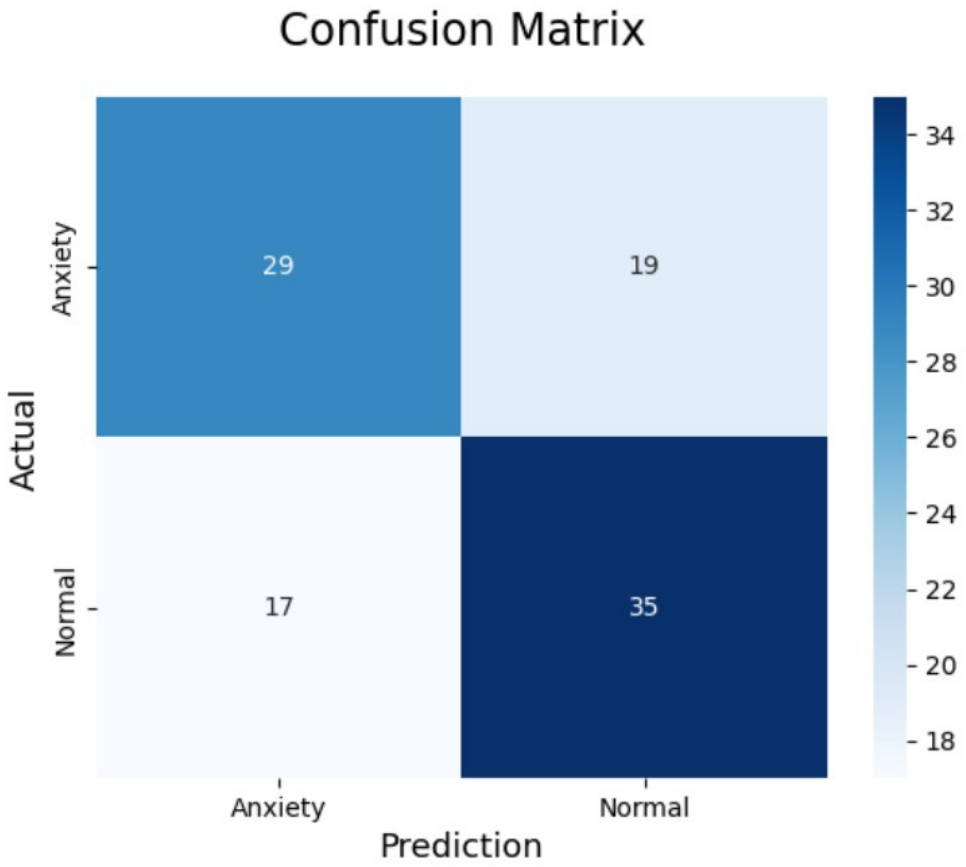

Figure 3.1: Confusion matrix for anxiety data (small dataset). This matrix shows the classification performance: 29 true positives and 17 false positives for normal cases, and 35 true positives and 19 false negatives for anxiety cases.

3.2.2 Depression Data

We used a data set of 500 rows (small data).

- **Hyperparameters:** ID, Text, Target
- **Target:** Normal (249), Depression (251)
- **Models:** Support Vector Classification (Support Vector Machine) with encoding Word2Vec
- **Results:**
  - **Accuracy:** 0.62
  - **Precision (average):** 0.6182
  - **Precision for Depression Class:** 0.6182
  - **Precision for Normal Class:** 0.6222
  - **Recall:** 0.6667
  - **Confusion Matrix:**

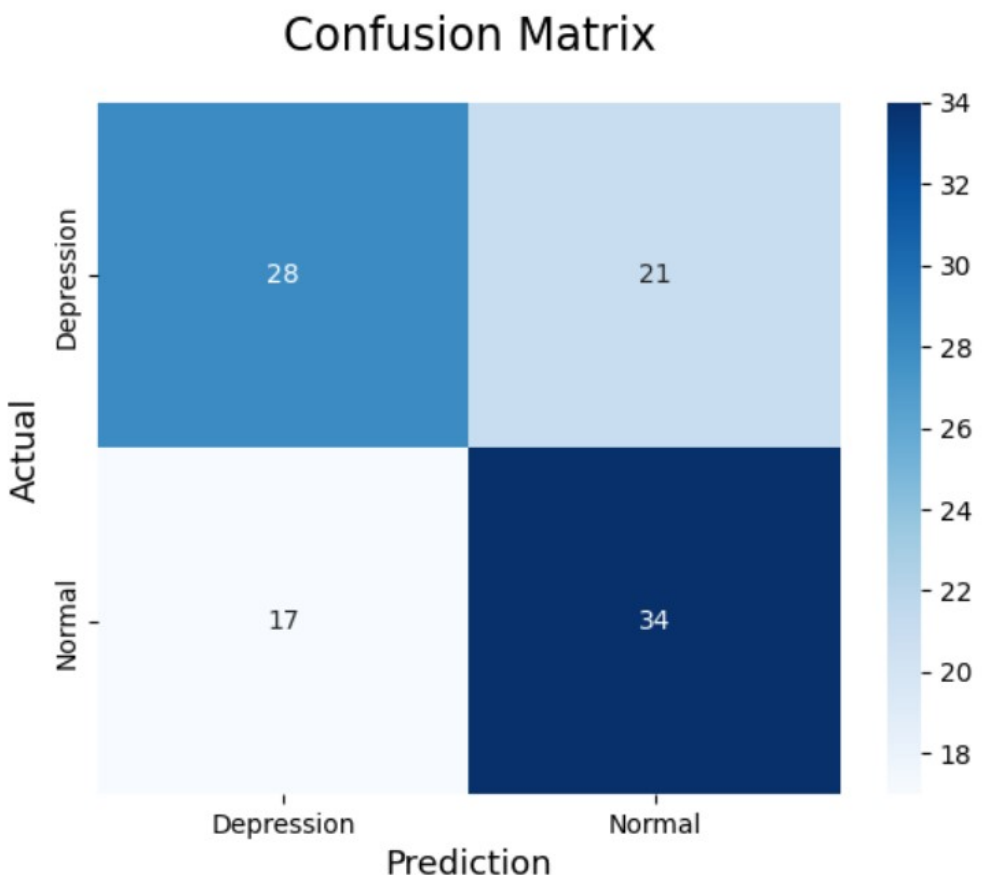

Figure 3.2: Confusion matrix for depression data (small dataset). This matrix illustrates the results with 28 true positives and 17 false positives for normal cases, and 34 true positives and 21 false negatives for depression cases.

3.3 Second Approach - Big Data

3.3.1 Anxiety Data

3.3.1.1 Linear Regression

- **Hyperparameters:** ID, Text, Target
- **Target:** Normal (246), Anxiety (254)
- **Models:** Linear Regression with transformer Sentence Transformer with the data set encoding 'all-MiniLM-L6-v2'
- **Results:**
  - **Accuracy:** 0.82
  - **Average Precision:** 0.82
  - **Precision for Anxiety Class:** 0.82
  - **Precision for Normal Class:** 0.82
  - **Recall for Anxiety Class:** 0.81
  - **Recall for Normal Class:** 0.83
  - **Confusion Matrix:**

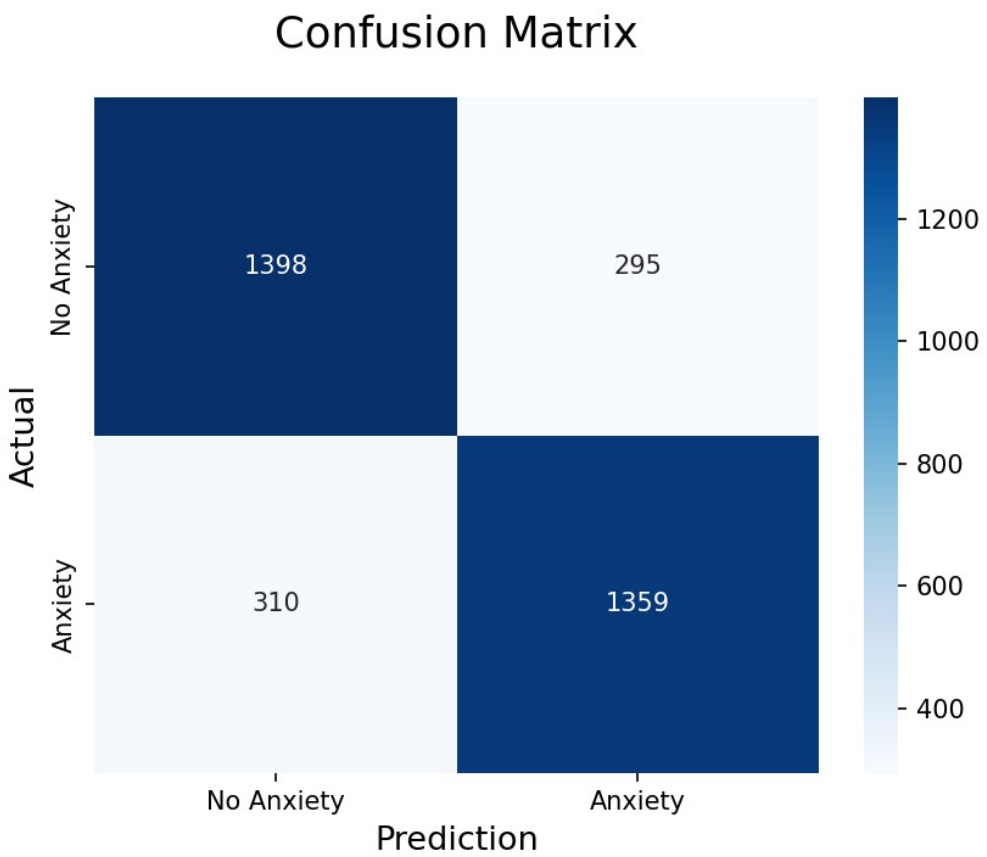

Figure 3.3: Confusion matrix for anxiety data using Linear Regression (large dataset). It details 1398 true positives and 310 false positives for normal cases, along with 1359 true positives and 295 false negatives for anxiety cases.

3.3.1.2 Decision Tree

- **Hyperparameters:** ID, Text, Target
- **Target:** Normal (246), Anxiety (254)
- **Models:** Decision Tree with transformer Sentence Transformer with the data set encoding 'all-MiniLM-L6-v2'
- **Results:**
  - **Accuracy:** 0.83
  - **Average Precision:** 0.83
  - **Precision for Anxiety Class:** 0.88
  - **Precision for Normal Class:** 0.79
  - **Recall for Anxiety Class:** 0.75
  - **Recall for Normal Class:** 0.90
  - **Confusion Matrix:**

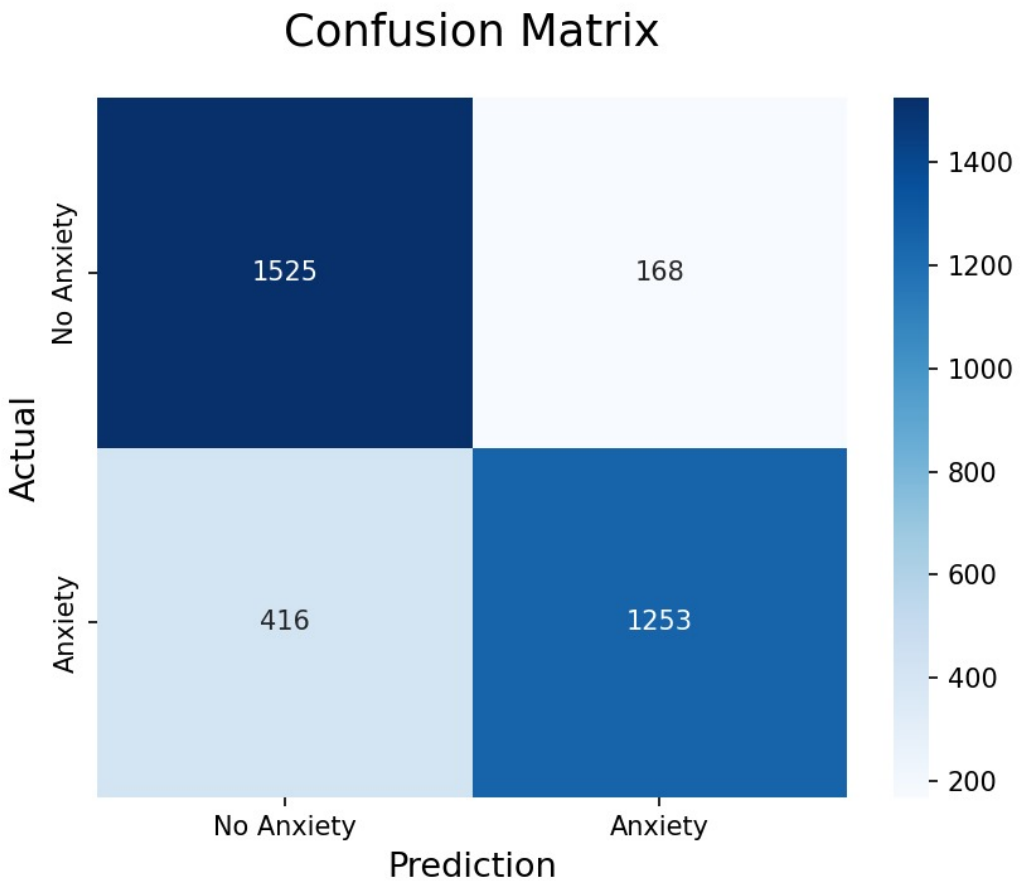

Figure 3.4: Confusion matrix for anxiety data using a Decision Tree (large dataset). The matrix indicates 1525 true positives and 416 false positives for normal cases, with 1253 true positives and 168 false negatives for anxiety cases.

3.3.1.3 Multi-layer Perceptron Classifier

- **Hyperparameters:** ID, Text, Target
- **Target:** Normal (246), Anxiety (254)
- **Models:** Multi-layer Perceptron Classifier with transformer Sentence Transformer with the data set encoding 'all-MiniLM-L6-v2'
- **Results:**
  - **Accuracy:** 0.93
  - **Average Precision:** 0.93
  - **Precision for Anxiety Class:** 0.95
  - **Precision for Normal Class:** 0.90
  - **Recall for Anxiety Class:** 0.90
  - **Recall for Normal Class:** 0.95
  - **Confusion Matrix:**

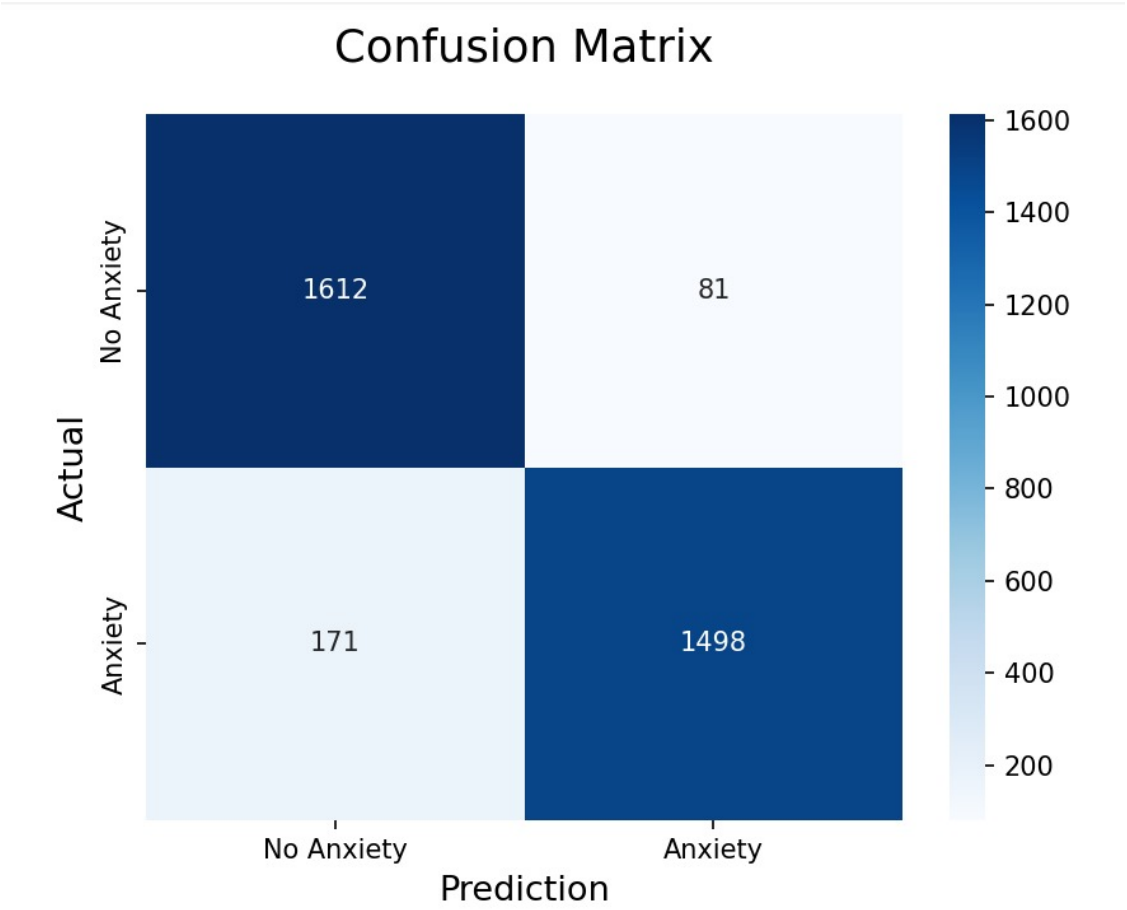

Figure 3.5: Confusion matrix for anxiety data using an Multi-layer Perceptron Classifier (large dataset). It shows 1612 true positives and 171 false positives for normal cases, and 1498 true positives with 81 false negatives for anxiety cases.

3.3.2 Depression Data

3.3.2.1 Linear Regression

- **Hyperparameters:** ID, Text, Target
- **Target:** Normal (249), Depression (251)
- **Models:** Linear Regression with transformer Sentence Transformer with the data set encoding 'all-MiniLM-L6-v2'
- **Results:**
  - **Accuracy:** 0.80
  - **Average Precision:** 0.81
  - **Precision for Depression Class:** 0.81
  - **Precision for Normal Class:** 0.80
  - **Recall for Depression Class:** 0.80
  - **Recall for Normal Class:** 0.81
  - **Confusion Matrix:**

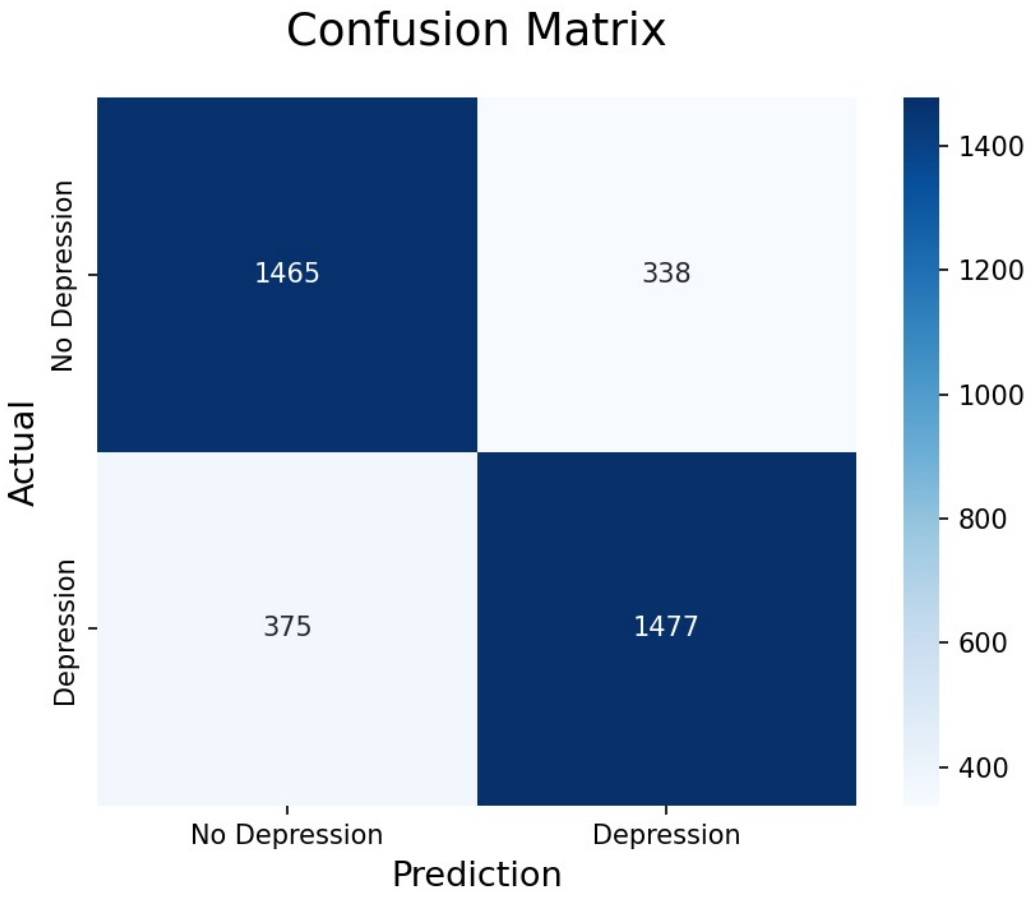

Figure 3.6: Confusion matrix for depression data using Linear Regression (large dataset). The results include 1465 true positives and 375 false positives for normal cases, as well as 1477 true positives and 338 false negatives for depression cases.

3.3.2.2 Decision Tree

- **Hyperparameters:** ID, Text, Target
- **Target:** Normal (249), Depression (251)
- **Models:** Decision Tree with transformer Sentence Transformer with the data set encoding 'all-MiniLM-L6-v2'
- **Results:**
  - **Accuracy:** 0.83
  - **Average Precision:** 0.84
  - **Precision for Depression Class:** 0.89
  - **Precision for Normal Class:** 0.78
  - **Recall for Depression Class:** 0.75
  - **Recall for Normal Class:** 0.91
  - **Confusion Matrix:**

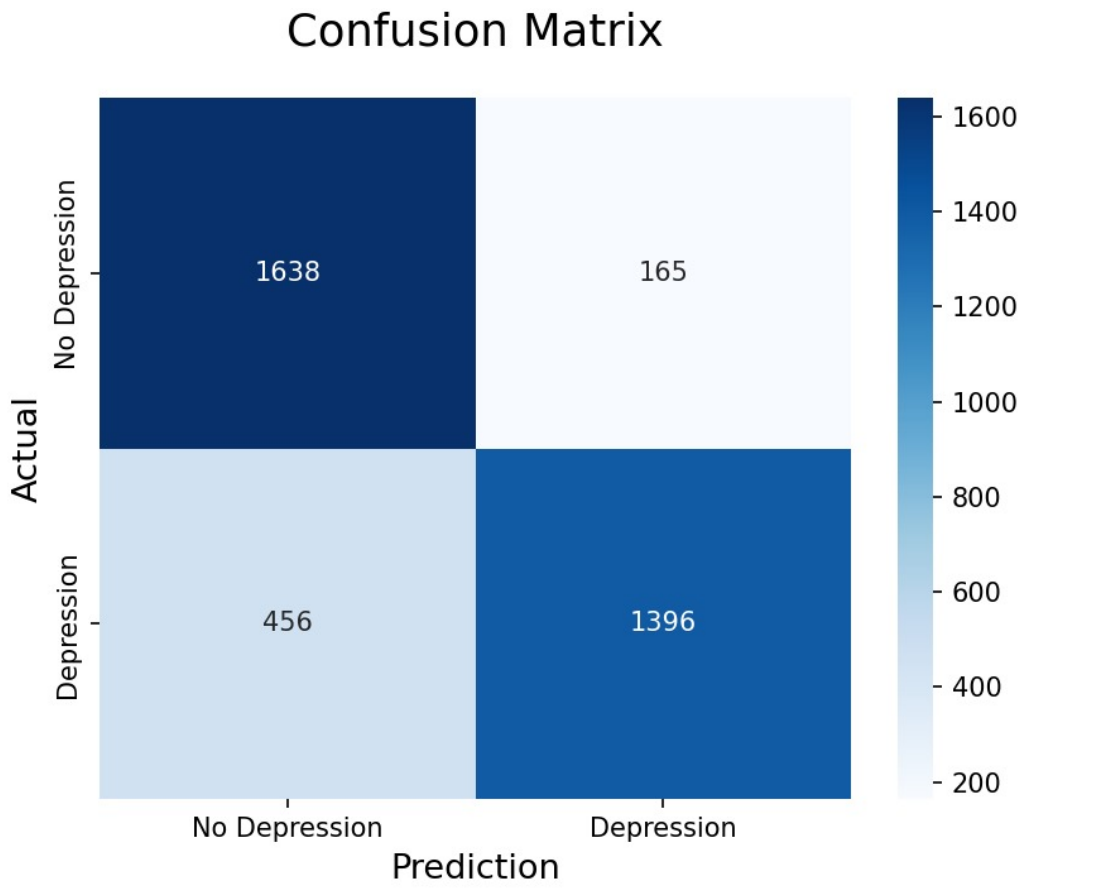

Figure 3.7: Confusion matrix for depression data using a Decision Tree (large dataset). It highlights 1638 true positives and 456 false positives for normal cases, with 1396 true positives and 165 false negatives for depression cases.

3.3.2.3 Multi-layer Perceptron Classifier

- **Hyperparameters:** ID, Text, Target
- **Target:** Normal (249), Depression (251)
- **Models:** Decision Tree with transformer Sentence Transformer with the data set encoding 'all-MiniLM-L6-v2'
- **Results:**
  - **Accuracy:** 0.93
  - **Average Precision:** 0.93
  - **Precision for Depression Class:** 0.96
  - **Precision for Normal Class:** 0.90
  - **Recall for Depression Class:** 0.89
  - **Recall for Normal Class:** 0.97
  - **Confusion Matrix:**

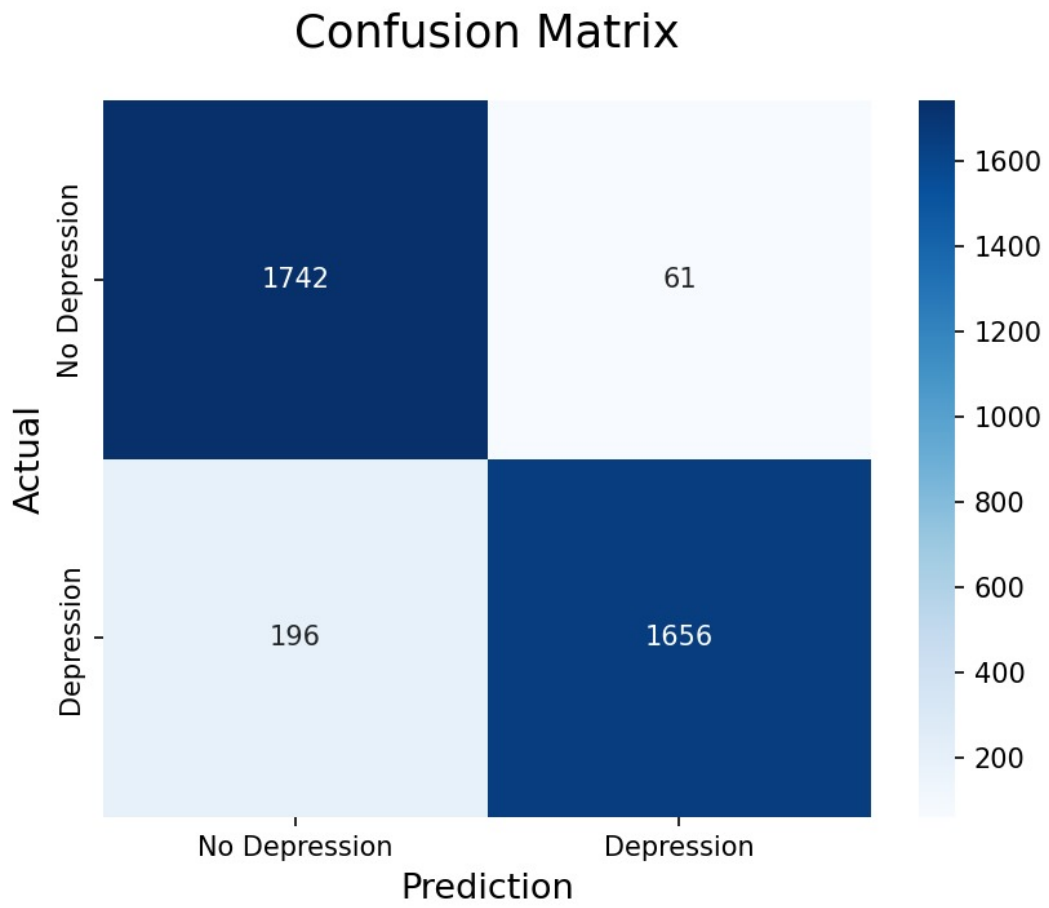

Figure 3.8: Confusion matrix for depression data using an Multi-layer Perceptron Classifier (large dataset). The matrix reports 1742 true positives and 196 false positives for normal cases, alongside 1656 true positives and 61 false negatives for depression cases.

Bibliography

[1] *Reddit Mental Health Dataset*. Neel Ghoshal (2020).
Available at: https://www.kaggle.com/datasets/neelghoshal/reddit-mental-health-data.

[2] *Emotion Dataset*. Parul Pandey (2020).
Available at: https://www.kaggle.com/datasets/parulpandey/emotion-dataset/data?select=training.csv.