

غرم الله القرني

Master Thesis Template-10-2-2026 %28Ghurmallah - Ahmad%29 - v 2.docx

 Midocean University

Document Details

Submission ID

trn:oid:::2945:405115253

Submission Date

Aug 6, 2026, 8:40 PM GMT+3

Download Date

Aug 9, 2026, 4:16 AM GMT+3

File Name

Master Thesis Template-10-2-2026 (Ghurmallah - Ahmad) - v 2.docx

File Size

5.2 MB

52 Pages

10,273 Words

64,929 Characters

45% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely revised by AI.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.



58 AI-generated 45%

Likely generated or likely generated and revised by AI.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely human generated as AI generated and likely AI generated as human generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.



Title

Early Mental Health Risk Detection from Social Media Posts

Name: Ghurmallah Alqarni

ID No.: 253000384

Name: Ahmad Hammad

ID No.: 261003041

Submitted in fulfillment of the requirements for the degree Master of
Science in Artificial Intelligence

At

Collage of Informatics

Midocean University

UAE

Supervisor Name: Dr. Hager Saleh

Collage of Informatics, Midocean University, Comoros

Email:

Acknowledgements

Abstract

List of Publications

Table of Contents

List of Figures.....	ix
List of Tables	ix
Chapter 1	1
Introduction.....	1
1.2 Problem Definition.....	2
1.3 Thesis Objectives.....	2
1.4 Thesis Contribution	3
1.5 Thesis Organization.....	3
2.1 Background	4
2.1.1 Artificial Intelligence (AI).....	4
2.1.2 Natural Language Processing (NLP)	5
2.1.2.1 NLP Pipeline.....	5
2.1.3 Text Preprocessing.....	6
2.1.3.1 Tokenization	6
2.1.3.2 Stop-word Removal	6
2.1.3.3 Stemming	7
2.1.4.1 Bag-of-Words (BoW).....	8
2.1.4.2 N-grams.....	8
2.1.4.3 Term Frequency–Inverse Document Frequency (TF-IDF).....	8
2.1.4.4 Word2Vec	8
2.1.4.5 GloVe.....	9
2.1.4.6 FastText	9
2.1.5 Artificial Intelligence Models.....	9
2.1.5.1 Machine Learning Models	10
2.1.5.2 Deep Learning Models	11

2.1.5.2.1 Convolutional Neural Network (CNN).....	11
2.1.5.2.2 Long Short-Term Memory (LSTM)	12
2.1.5.2.3 Bidirectional Long Short-Term Memory (BiLSTM).....	12
2.1.5.2.4 Attention Mechanism	12
2.1.5.2.5 CNN-LSTM	12
2.1.5.2.6 LSTM-Attention-CNN	13
2.1.5.3 Transformer Models.....	13
2.1.5.3.1 Bidirectional Encoder Representations from Transformers (BERT).....	13
2.1.5.3.2 DistilBERT	13
2.1.5.3.3 RoBERTa.....	14
2.1.5.3.4 DeBERTa.....	14
2.1.6 Dataset.....	17
2.2 Related Work	19
Chapter 5: Experimental Results and Discussion	41
([1] Bokolo, 2023)	43

Chapter 1: Introduction and Overview.....	1
1.2 Problem Definition	2
1.3 Thesis Objectives	3
1.4 Thesis Contribution	4
1.5 Thesis Organization.....	5
 Chapter 2: Background and Related Work.....	 31
 Chapter 3: Proposed Approach	 78
4.1 Introduction	78
 Chapter 4: Experimental Results and Discussion	 86
 Chapter 5: Conclusion and Future Work	 127
7.1 Important Features	127
7.3 Future Work	129
 References	 130

List of Abbreviations

- Artificial Intelligence.
- Area Under the Receiver Operating Characteristic Curve.
- Bag-of-Words.
- Convolutional Neural Network.
- Deep Learning.
- F1 Score.
- False Negative.
- False Positive.
- Logistic Regression.
- Long Short-Term Memory.
- Machine Learning.
- Naïve Bayes.
- Natural Language Processing.
- Random Forest.
- Receiver Operating Characteristic.
- Support Vector Machine.
- Term Frequency.
- Term Frequency–Inverse Document Frequency.
- True Negative.
- True Positive.
- World Health Organization.
- Global Vectors for Word Representation.
- Bidirectional Encoder Representations from Transformers.

List of Figures

Figure 2.1	Artificial Intelligence Hierarchy	4
Figure 2.2	Natural Language Processing Framework	5
Figure 2.3	Text Preprocessing Workflow	6
Figure 2.4	Text Representation Techniques	8
Figure 2.5	Artificial Intelligence Models Classification	10
Figure 2.6	Accuracy Formula	15
Figure 2.7	Precision Formula	15
Figure 2.8	Recall Formula	16
Figure 2.9	F1-score Formula	16
Figure 2.10	Social Media Mental Health Risk Dataset Structure	17

List of Tables

2.1 Comparison of Previous Studies

Chapter 1

Introduction

1.1 Overview

Mental health is a vital part of overall health, and it helps to shape an individual's thoughts, feelings and actions in everyday life. Mental illness, such as depression, anxiety, stress and suicidal thoughts have emerged as a significant global public health burden. The World Health Organization (WHO) estimates that millions of people worldwide suffer from mental health disorders annually, and often don't get diagnosed or timely intervention. Hence, the detection of persons at risk at an early stage is important for the researchers and healthcare professionals. Social media platforms like Twitter (X), Reddit, Facebook and Instagram have grown by a vast margin in recent years and have changed how people interact with each other and how they express their feelings, thoughts, and experiences. Users freely post their thoughts, feelings, experiences and more on the internet with some posts revealing signs of psychological well-being. As a result, social media is an excellent tool to gather textual data, which is useful to examine mental health conditions and their early signs of psychological distress. Natural Language Processing (NLP), along with Machine Learning (ML) and Deep Learning (DL), has been one of the key factors that has revolutionized the automatic textual information analysis process. NLP can help computers to interpret and understand human language, while machine learning and deep learning models can learn and categorize user with regard to their mental health status based on the information contained in the text. The evolution of transformer-based language models has also contributed to the success of more recent text classification, allowing them to capture contextual and semantic relations which are better than traditional approaches. There is still a long way to go, but it's difficult to accurately screen for mental health risks within social media. Social media texts tend to be brief, casual, ambiguous and include abbreviations, slang, emoji and spelling variations. Moreover, data sets are often skewed and the language usage of users can differ from platform to platform. The challenges bring an attention toward building and testing of robust computational models for sound and reliable early risk detection in mental health. In this thesis, methods for the early detection of mental health risks in social media posts, especially those of Machine Learning and Deep Learning are explored. It has reviewed the existing approaches and evaluated various ways of representing the text and classification models with comparison of the performance of the models using available benchmark datasets. The results are hoped to offer a detailed insight into the existing methods and will help to understand which methods are most effective in detecting early mental health risks.

1.2 Problem Definition

The rising incidence of mental health problems has led to a pressing need for effective tools that can help in the early identification of vulnerable people. Traditional mental health assessment mainly relies on clinical interview, self-report questionnaires, and practitioners' judgment. These methods offer good diagnostic accuracy but can be very time consuming, expensive, and need to be available in healthcare systems. As a result, many people with mental health issues go undiagnosed or only seek treatment once it has become far more serious, limiting the potential impact of early intervention. With the proliferation of social media, new opportunities have arisen for detecting initial symptoms of mental illnesses via humans' online written expressions. But it's still hard to accurately identify mental health risks from social media. Social media posts tend to be informal, brief, very loosely structured and include words like slang, abbreviations, emojis, misspellings, and context-dependent expressions. These features make the process of accurately detecting mental health risks more difficult, and complicate text analysis. While there have been many computational methods that have been offered for the purposes of early detection of mental health risk from social media posts, there are several issues that are yet to be addressed. Past studies have used different Machine Learning and Deep Learning models, datasets, text representation and classification algorithms, thereby making it hard to decide which one to use. Furthermore, a lot of studies only look at limited models and/or at mental health disorders, and the findings differ in the literature. Hence, a holistic analysis of the computational methods for early detection of mental health risks in social media posts, especially of Machine/Deep Learning methods is requested. The present study caters to this by reviewing and comparing the existing techniques, analyzing their strengths and limitations, and finding the most effective technique that can be used for accurate and reliable early mental health risk detection.

1.3 Thesis Objectives

This thesis aims to explore the effectiveness of computational methods to identify signs of mental health problems from text posts in social media, especially using Machine Learning and Deep Learning. To accomplish this goal, the study will:

1. Analyze current studies of tools to detect early mental health risks through data from social media and list the pros and cons of those tools.
2. Review the different text preprocessing techniques and text representation methods used to analyze social media post for mental health detection.
3. Compare various Machine Learning and Deep Learning models for social media posts classification for mental health conditions.
4. Evaluate the performance of the selected models using typical metrics such as accuracy, precision, recall, F1 score and AUC.

5. Through experiential results, compare the advantages and disadvantages of various early mental health risk identification techniques and identify the most effective ones.
6. Recommend suggestions for future studies to enhance the practicality, reliability, and accuracy of the application of AI methods in detecting mental health risks.

1.4 Thesis Contribution

This thesis makes a significant advancement in the field of early detection of mental health risks from social media posts, offering a comprehensive assessment of the computational methods, especially Machine Learning and Deep Learning approaches. This study contributes to the literature in the following ways:

- 1- Performing a full-scale state of the art review of the recent studies by focusing on the current trends, challenges, and research gaps in the field of early mental health risk detection via social media data.
2. Comparatively analyzing the widely used text pre-processing techniques, text representation methods and classification models used for mental health risk detection.
3. Assessing the performance of Machine Learning and Deep Learning models using standard and benchmark datasets and evaluation metrics such as accuracy, precision, recall, F1-score, and AUC.
4. Identifying the strengths and limitations of different approaches to identify the most effective techniques to detect mental health risks early.
5. Providing practical recommendations that might help the researchers to choose the appropriate AI technique(s) for future mental health detection studies.

1.5 Thesis Organization

This thesis consists of five chapters, each of which deals with one aspect of the research.

Chapter 1. provides an overview of the research, problem definition, objectives of research, thesis contributions, and thesis organization.

The background and related work are given in Chapter 2. It covers the basic ideas of Natural Language Processing (NLP), Machine Learning (ML), Deep Learning (DL), text preprocessing and text representation techniques, and popular evaluation metrics, and summarizes recent works for early detection of mental health risk indicators from social media data.

The research methodology is explained in chapter 3. Provides the datasets that were used in the study, data pre-processing, feature extraction techniques, selected ML models, selected DL models, experimental set-up and evaluation metrics used for testing the model performance.

The results and discussion of the experiments are shown in Chapter 4. It presents a comparison of the performance of the various models used, an analysis of the experimental results, a discussion of the merits and drawbacks of the models, and a comparison of the results obtained with the results presented in previous studies.

The final chapter of the thesis summarizes the key findings, identifies the contributions of the research, addresses limitations and offers recommendations for future research.

Chapter 2 Background and Related Work

1 Background

1.1 Artificial Intelligence (AI)

Artificial Intelligence (AI) is a branch of computer science that focuses on developing intelligent systems capable of performing tasks that normally require human intelligence, such as learning, reasoning, problem-solving, decision-making, and language understanding. AI systems analyze data, identify meaningful patterns, and make predictions or decisions using computational algorithms. The continuous advancement of AI has enabled its application in numerous domains, including healthcare, finance, education, computer vision, robotics, and natural language processing. As a result, AI has become the foundation of many modern intelligent applications and data-driven decision-making systems ([1] Bokolo, 2023)

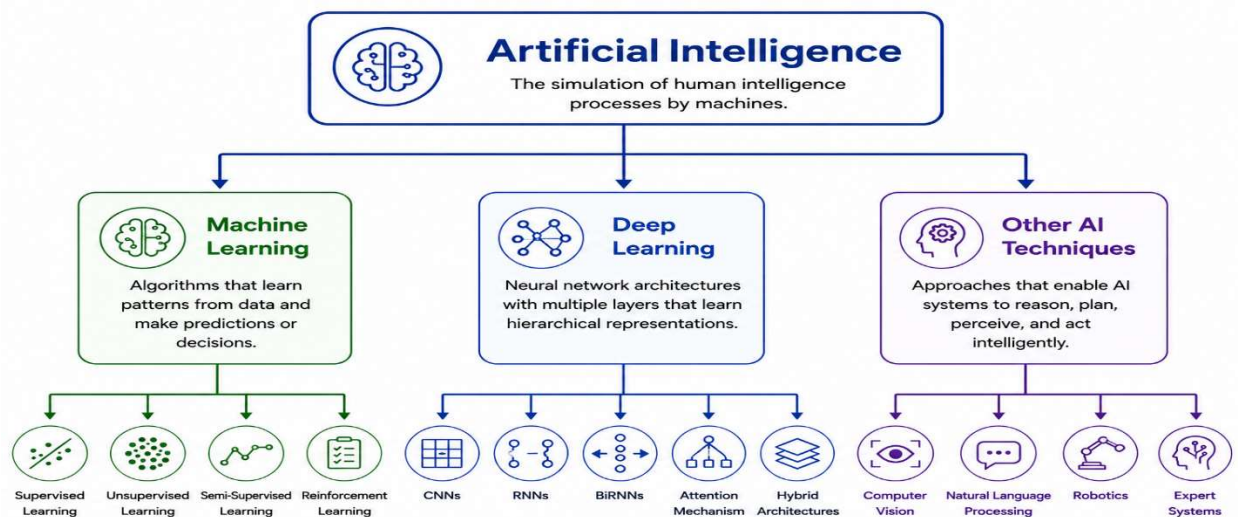


Figure 2.1 Artificial Intelligence Hierarchy

1.2 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a subfield of Artificial Intelligence that enables computers to process, understand, interpret, and generate human language. It combines computational linguistics with Machine Learning and Deep Learning techniques to transform unstructured textual data into structured representations suitable for automated analysis. NLP provides the foundation for various language-related applications, including text classification, sentiment analysis, information retrieval, machine translation, question answering, and social media analytics. In mental health research, NLP plays a fundamental role in extracting linguistic patterns that may indicate psychological conditions ([2] S. Almutairi, 2024)

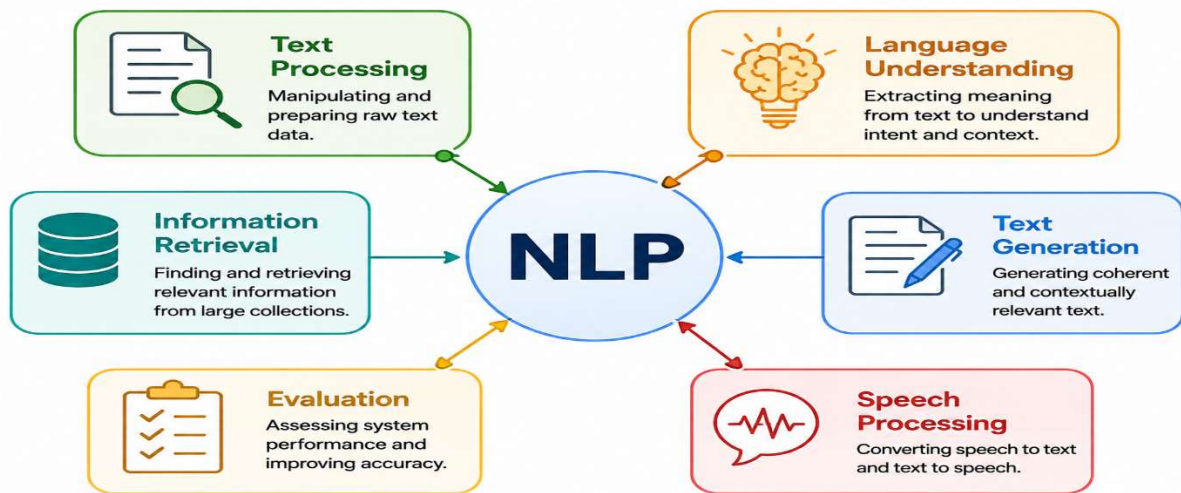


Figure 2.2 Natural Language Processing Framework, ChatGPT

1.2.1 NLP Pipeline

The Natural Language Processing (NLP) pipeline is a sequence of processing stages that converts raw textual data into structured information suitable for computational analysis. The pipeline typically consists of data collection, text preprocessing, text representation, feature extraction, model training, classification, and performance evaluation. Each stage contributes to improving data quality and enhancing the effectiveness of machine learning and deep learning models in text classification tasks ([11] Renjith, n.d.).

1.3 Text Preprocessing

Text preprocessing is the process of cleaning, normalizing, and transforming raw textual data into a structured format suitable for computational analysis. The primary objective of text preprocessing is to reduce noise, eliminate irrelevant information, and standardize textual content before feature extraction and model training. Effective preprocessing improves data quality, reduces computational complexity, and enhances the overall performance of text classification models. Common preprocessing techniques include tokenization, stop-word removal, stemming, and lemmatization ([4] Bae, 2021).

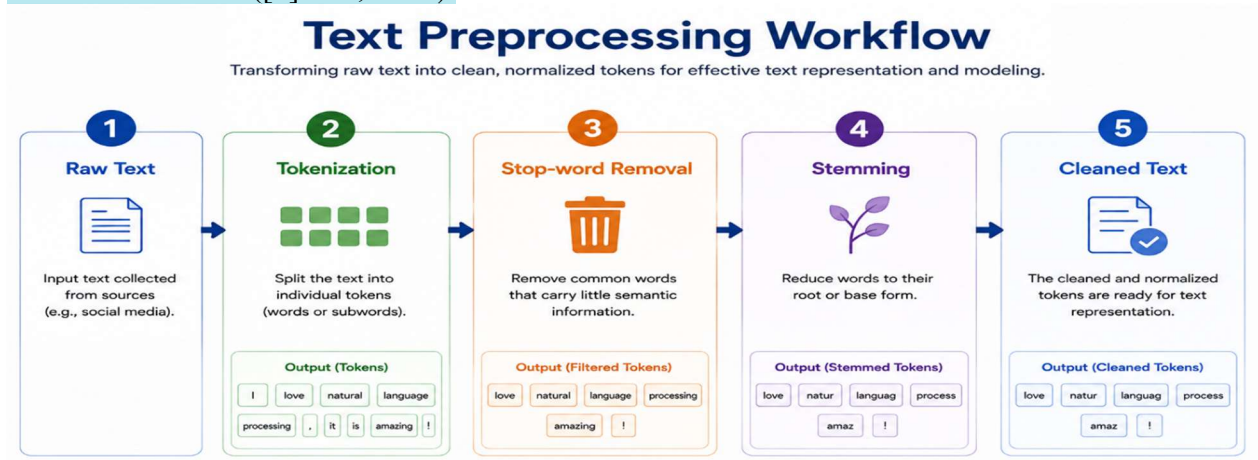


Figure 2.3 Text Preprocessing Workflow, ChatGPT

1.3.1 Tokenization

Tokenization is the process of dividing a text into smaller linguistic units called tokens, which may represent words, subwords, characters, or sentences. It is considered one of the fundamental preprocessing techniques in Natural Language Processing because it converts raw text into structured units that can be processed by machine learning and deep learning algorithms. Effective tokenization preserves the textual structure while facilitating feature extraction and subsequent language analysis ([4] Bae, 2021), ([11] Renjith, n.d.).

1.3.2 Stop-word Removal

Stop-word removal is a text preprocessing technique that eliminates frequently occurring words that contribute little semantic information to text analysis. These words, such as articles, conjunctions, and prepositions, are removed to reduce textual noise, decrease feature dimensionality, and improve computational efficiency. By retaining only informative terms, stop-

1.4.1 Bag-of-Words (BoW)

Bag-of-Words (BoW) is one of the earliest and most widely used text representation techniques in Natural Language Processing. It represents a document as a vector of word frequencies without considering grammar, syntax, or the order in which words appear. Each unique word in the corpus is treated as an independent feature, and its corresponding value is determined by its occurrence within the document. Although BoW is computationally efficient and simple to implement, it does not capture semantic relationships or contextual information between words ([1] Bokolo, 2023), ([11] Renjith, n.d.).

1.4.2 N-grams

N-grams are text representation units consisting of consecutive sequences of words extracted from textual data. Depending on the value of N , an N-gram may represent a single word (unigram), two consecutive words (bigram), or three consecutive words (trigram). Unlike the Bag-of-Words model, N-grams preserve limited word-order information, allowing local contextual relationships to be captured. This representation improves the ability of classification models to recognize common phrases and neighboring word patterns within textual data ([4] Bae, 2021), ([11] Renjith, n.d.).

1.4.3 Term Frequency–Inverse Document Frequency (TF-IDF)

Term Frequency–Inverse Document Frequency (TF-IDF) is a statistical text representation technique that measures the importance of a word within a document relative to a collection of documents. It combines term frequency, which reflects how often a word appears in a document, with inverse document frequency, which reduces the influence of words that frequently occur across the entire corpus. Consequently, TF-IDF assigns higher weights to informative words while reducing the contribution of common terms, thereby producing more discriminative feature representations for text classification tasks ([4] Bae, 2021), ([15] Qasim, 2025).

1.4.4 Word2Vec

Word2Vec is a distributed word embedding technique that learns dense vector representations by analyzing the contextual relationships among words within a text corpus. It is based on the distributional hypothesis that words appearing in similar contexts tend to have similar meanings. Consequently, semantically related words are mapped to nearby locations in the embedding space, enabling machine learning and deep learning models to capture semantic similarities more

effectively than traditional frequency-based text representation methods ([2] S. Almutairi, 2024), ([11] Renjith, n.d.).

1.4.5 GloVe

Global Vectors for Word Representation (GloVe) is a word embedding technique that learns dense vector representations by exploiting global word co-occurrence statistics collected from an entire text corpus. Unlike predictive embedding models, GloVe constructs word vectors by factorizing a word co-occurrence matrix, allowing semantic and syntactic relationships between words to be preserved. This approach produces meaningful vector representations that improve the performance of various Natural Language Processing tasks ([5] Roy, 2020), ([8] Figuerêdo, 2022).

1.4.6 FastText

FastText is a word embedding model that represents each word as a collection of character-level subword units rather than a single token. By learning embeddings for character n-grams, FastText effectively captures morphological information and generates meaningful representations for rare and previously unseen words. This capability makes FastText particularly suitable for processing noisy textual data commonly found in social media platforms ([5] Roy, 2020), ([8] Figuerêdo, 2022).

1.5 Artificial Intelligence Models

Artificial Intelligence (AI) models are computational approaches designed to learn patterns, extract knowledge, and perform prediction or decision-making tasks from data. Depending on their learning mechanisms and architectural complexity, AI models can be broadly categorized into Machine Learning (ML), Deep Learning (DL), and Transformer-based models. In Natural Language Processing, these models are widely employed to analyze textual information, identify linguistic patterns, and perform classification tasks with varying levels of contextual understanding and predictive capability ([11] Renjith, n.d.).

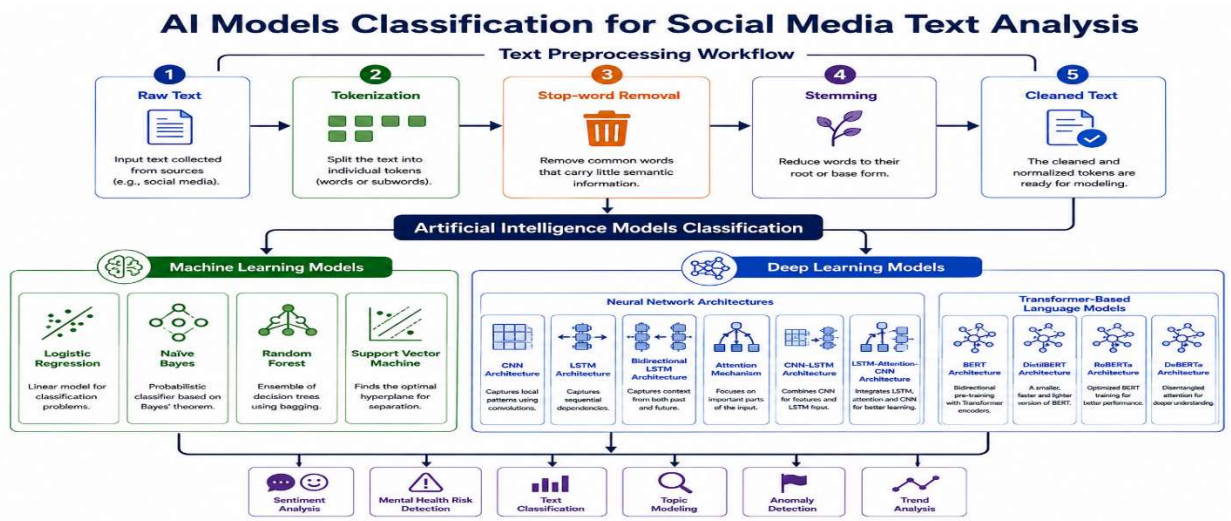


Figure 2.5 Artificial Intelligence Models Classification, ChatGPT

1.5.1 Machine Learning Models

Machine Learning (ML) is a branch of Artificial Intelligence that enables computer systems to learn patterns directly from data without being explicitly programmed. ML algorithms construct predictive models by identifying relationships between input features and target outputs through statistical learning techniques. In Natural Language Processing, Machine Learning models commonly utilize handcrafted textual features, such as TF-IDF and Bag-of-Words, to perform text classification, sentiment analysis, and mental health detection tasks ([4] Bae, 2021).

1.5.1.1 Logistic Regression

Logistic Regression is a supervised Machine Learning algorithm that estimates the probability of an input sample belonging to a specific class using the logistic (sigmoid) function. It models the relationship between input features and class labels by transforming a linear combination of features into a probability value between zero and one. Due to its computational efficiency and strong baseline performance, Logistic Regression is widely used for binary text classification tasks ([1] Bokolo, 2023).

1.5.1.2 Naïve Bayes

Naïve Bayes is a probabilistic Machine Learning algorithm derived from Bayes' theorem that predicts class membership based on conditional probability. It assumes that input features are statistically independent given the target class, allowing efficient probability estimation and

classification. Owing to its simplicity and low computational cost, Naïve Bayes is frequently applied in text mining and document classification tasks ([4] Bae, 2021).

1.5.1.3 Random Forest

Random Forest is an ensemble Machine Learning algorithm that constructs multiple decision trees using randomly selected subsets of training samples and features. The final prediction is generated by aggregating the outputs of all individual trees through a voting mechanism. This ensemble strategy improves classification accuracy, reduces overfitting, and increases the robustness of predictive models across various text classification applications ([1] Bokolo, 2023).

1.5.1.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised Machine Learning algorithm that identifies an optimal decision boundary by maximizing the margin between different classes. The model determines the hyperplane that best separates training samples while minimizing classification errors. Because of its effectiveness in handling high-dimensional feature spaces, SVM has been extensively applied in Natural Language Processing and text classification problems ([11] Renjith, n.d.).

1.5.2 Deep Learning Models

Deep Learning (DL) is a subfield of Machine Learning that employs artificial neural networks with multiple hidden layers to automatically learn hierarchical feature representations from data. Unlike traditional Machine Learning algorithms, Deep Learning models reduce the need for manual feature engineering by learning informative representations directly from raw inputs. This capability has significantly improved the performance of Natural Language Processing tasks, particularly in text classification, sentiment analysis, and mental health detection from social media posts ([9] Kumar, 2022), ([11] Renjith, n.d.).

1.5.2.1 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is a deep learning architecture designed to automatically extract discriminative features from input data through convolution operations. In Natural Language Processing, CNN applies convolutional filters over word representations to identify informative local patterns, such as phrases and neighboring word sequences. These extracted features are subsequently aggregated through pooling operations before being forwarded to fully connected layers for classification. Owing to its ability to capture local contextual information efficiently, CNN has become one of the most widely adopted architectures for text classification tasks ([2] S. Almutairi, 2024), ([9] Kumar, 2022).

1.5.2.2 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a recurrent neural network architecture specifically designed to model sequential data by capturing long-term dependencies within a sequence. Unlike conventional recurrent neural networks, LSTM incorporates memory cells together with input, forget, and output gates that regulate the flow of information during learning. These gating mechanisms enable the model to preserve relevant contextual information while discarding irrelevant content, making LSTM highly effective for Natural Language Processing applications involving sequential text analysis ([7] Sharma, 2022), ([14] Hasan, 2024).

1.5.2.3 Bidirectional Long Short-Term Memory (BiLSTM)

Bidirectional Long Short-Term Memory (BiLSTM) extends the standard LSTM architecture by simultaneously processing textual sequences in both forward and backward directions. The outputs generated from both directions are combined to construct richer contextual representations that incorporate preceding and succeeding linguistic information. This bidirectional learning strategy enables BiLSTM to capture more comprehensive contextual dependencies, thereby improving the performance of text classification and sequence modeling tasks ([9] Kumar, 2022), ([14] Hasan, 2024).

1.5.2.4 Attention Mechanism

The Attention Mechanism is a neural network technique that enables deep learning models to assign different importance weights to various elements within an input sequence. Rather than treating all words equally, the attention mechanism emphasizes the most informative textual representations while reducing the influence of less relevant information. By selectively focusing on significant contextual features, attention improves feature representation and enhances the accuracy of Natural Language Processing models ([11] Renjith, n.d.), ([14] Hasan, 2024).

1.5.2.5 CNN-LSTM

CNN-LSTM is a hybrid deep learning architecture that combines the complementary strengths of Convolutional Neural Networks and Long Short-Term Memory networks. CNN is employed to extract informative local textual features, whereas LSTM models the sequential dependencies among these extracted features. Integrating both architectures enables CNN-LSTM to learn spatial and temporal representations simultaneously, resulting in improved performance for text classification and sequence analysis tasks ([2] S. Almutairi, 2024), ([11] Renjith, n.d.).

1.5.2.6 LSTM-Attention-CNN

LSTM-Attention-CNN is a hybrid deep learning architecture that integrates Long Short-Term Memory, the Attention Mechanism, and Convolutional Neural Networks into a unified framework. LSTM first learns sequential contextual information, after which the attention mechanism identifies the most informative hidden representations. CNN subsequently extracts discriminative local features from these weighted representations before the final classification stage. This integrated architecture enhances contextual understanding while improving classification performance in Natural Language Processing applications ([11] Renjith, n.d.).

1.5.3 Transformer Models

Transformer models represent a modern family of deep learning architectures that rely primarily on the self-attention mechanism to model contextual relationships among words within a sequence. Unlike recurrent architectures, transformers process all input tokens simultaneously, allowing efficient parallel computation while capturing both short-range and long-range dependencies. Their ability to learn rich contextual representations has significantly advanced Natural Language Processing, particularly in language understanding and text classification tasks ([1] Bokolo, 2023), ([14] Hasan, 2024), ([15] Qasim, 2025).

1.5.3.1 Bidirectional Encoder Representations from Transformers (BERT)

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based pre-trained language model that learns bidirectional contextual representations by simultaneously considering both preceding and succeeding words within a sentence. After large-scale pre-training, BERT can be fine-tuned for downstream Natural Language Processing tasks by adding a task-specific output layer. Its bidirectional contextual learning capability has made BERT one of the most influential language models for text classification and language understanding ([6] Rastogi, 2022), ([15] Qasim, 2025).

1.5.3.2 DistilBERT

DistilBERT is a lightweight transformer-based language model obtained through knowledge distillation from the original BERT architecture. It significantly reduces the number of model parameters while preserving most of BERT's contextual language understanding capability. This compression strategy enables faster inference, lower computational requirements, and efficient deployment without substantial degradation in classification performance ([3] Rastogi, 2021).

1.5.3.3 RoBERTa

RoBERTa is an optimized transformer-based language model developed as an enhanced version of BERT through improvements in the pre-training strategy. By utilizing larger training corpora, dynamic masking, and optimized training procedures, RoBERTa learns richer contextual representations that improve language understanding and downstream text classification performance. Its strong contextual modeling capability has made it one of the most effective transformer models for Natural Language Processing tasks ([1] Bokolo, 2023), ([14] Hasan, 2024), ([15] Qasim, 2025).

1.5.3.4 DeBERTa

DeBERTa is an advanced transformer-based language model that enhances contextual representation by separately modeling word content and positional information through a disentangled attention mechanism. This architectural improvement enables more accurate modeling of semantic and positional relationships among tokens, resulting in stronger language understanding and improved performance across a wide range of Natural Language Processing applications, including text classification and mental health risk detection ([1] Bokolo, 2023), ([15] Qasim, 2025).

1.6 Evaluation Metrics

Evaluation metrics are quantitative measures used to assess the performance of classification models by comparing predicted labels with the corresponding ground-truth labels. These metrics provide objective criteria for evaluating the effectiveness, reliability, and generalization capability of Machine Learning, Deep Learning, and Transformer-based models. Since mental health risk detection is formulated as a text classification problem, selecting appropriate evaluation metrics is essential for accurately measuring model performance. In this study, the performance of the classification models is evaluated using the Confusion Matrix, Accuracy, Precision, Recall, and F1-score, which are among the most widely adopted evaluation metrics in Natural Language Processing and mental health detection research ([2] S. Almutairi, 2024), ([15] Qasim, 2025).

1.6.1 Confusion Matrix

A Confusion Matrix is a tabular representation used to evaluate the performance of classification models by comparing predicted class labels with the actual class labels. It summarizes classification outcomes into four categories: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These values provide the basis for calculating various evaluation metrics, including Accuracy, Precision, Recall, and F1-score, making the confusion matrix one of the most fundamental tools for assessing classification performance ([1] Bokolo, 2023), ([15] Qasim, 2025).

1.6.2 Accuracy

Accuracy measures the overall proportion of correctly classified instances among all evaluated samples. It provides a general indication of model performance by calculating the ratio of correct predictions to the total number of predictions. Although Accuracy is easy to interpret, it may produce misleading results when evaluating imbalanced datasets because it does not consider the distribution of different classes ([1] Bokolo, 2023), ([15] Qasim, 2025).

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

Figure 2.6 Accuracy Formula

1.6.3 Precision

Precision measures the proportion of correctly predicted positive instances among all instances predicted as positive. It evaluates the reliability of positive predictions by quantifying how many predicted positive samples are positive. High precision indicates that the classifier produces relatively few false positive predictions ([15] Qasim, 2025).

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

Figure 2.7 Precision Formula

1.6.4 Recall

Recall, also referred to as sensitivity or true positive rate, measures the proportion of actual positive instances that are correctly identified by the classification model. A high recall value indicates that the model successfully detects most positive cases while minimizing false negative predictions, which is particularly important in mental health risk detection applications ([1] Bokolo, 2023).

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

Figure 2.8 Recall Formula

1.6.5 F1-score

The F1-score is the harmonic mean of Precision and Recall, providing a balanced evaluation of classification performance. It is particularly useful when class distributions are imbalanced because it simultaneously considers both false positive and false negative predictions. Consequently, the F1-score is widely used for evaluating Natural Language Processing and text classification models ([15] Qasim, 2025).

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Figure 2.9 F1-score Formula

1.7 Dataset

1.7.1 Dataset Overview

The experiments presented in this thesis are based on the **social media Mental Health Risk Dataset**, a publicly available dataset hosted on Kaggle. The dataset contains **5,000** synthetic social media posts designed for mental health risk assessment. Each post is annotated with mental health categories, including **depression, stress, suicide, and normal**, together with corresponding risk labels. The dataset is intended for Natural Language Processing (NLP) applications such as text classification, mental health detection, and risk prediction, making it suitable for evaluating Artificial Intelligence models for early mental health risk detection.

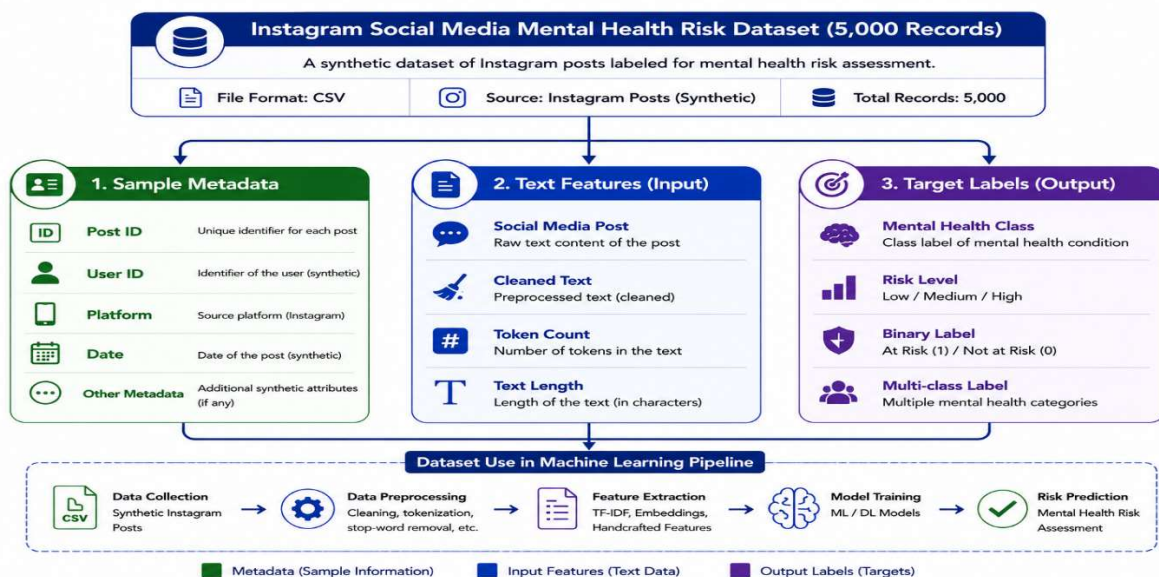


Figure 2.10 Social Media Mental Health

1.7.2 Dataset Characteristics

The dataset consists of textual social media posts collected in a structured tabular format. Each record represents a single social media post together with its corresponding mental health annotation. The dataset was designed to support supervised learning tasks by providing labeled examples that can be used for training and evaluating Machine Learning, Deep Learning, and Transformer-based classification models. The publicly available nature of the dataset also facilitates reproducibility and comparative evaluation across different Artificial Intelligence approaches.

1.7.3 Dataset Labels

The dataset categorizes each post according to predefined mental health classes representing different psychological conditions. These labels provide the ground truth required for supervised text classification and enable computational models to distinguish between different levels or categories of mental health risk. According to the dataset documentation, the primary mental health categories include:

- Depression
- Stress
- Suicide
- Normal

1.7.4 Dataset Advantages

The selected dataset offers several advantages for mental health risk detection research. First, it provides a balanced collection of labeled social media posts that can be directly used for supervised classification tasks. Second, the dataset includes multiple mental health categories, allowing comparative evaluation across different psychological conditions. Third, its public availability enables reproducibility and fair comparison with future studies. Finally, the dataset is specifically designed for Natural Language Processing applications, making it well suited for evaluating both traditional Machine Learning models and modern Transformer-based language models.

1.7.5 Dataset Limitations

Although the dataset is appropriate for benchmarking Artificial Intelligence models, it has several limitations. The dataset consists of **synthetic** social media posts rather than naturally collected user-generated content, which may limit its ability to fully represent real-world linguistic diversity. Furthermore, it primarily contains English-language posts and focuses on a predefined set of mental health categories. Consequently, the experimental results obtained from this dataset may not generalize directly to multilingual environments or real-world social media platforms without additional validation.

2 Related Work

Bokolo and Liu [([1] Bokolo, 2023)] introduced a comparative framework for detecting depression from social media tweets using the traditional Machine Learning (ML) models and transformer-based models. The experiments were performed on the publicly available Sentiment140 dataset, which was originally gathered about 1.6 million English tweets for sentiment analysis. The authors re-labelled these tweets using an algorithm that is based on polarity, since this is not the way the dataset was labelled for the purpose of depression detection. Following this process, 632,000 tweets were left for the experiments. The data was preprocessed, feature selected and label encoded before training the model. Textual features were represented as Term Frequency–Inverse Document Frequency (TF-IDF) for the traditional ML models and the transformer-based models were fine-tuned from the prepared tweet data set. The models evaluated were Logistic Regression, Bernoulli Naïve Bayes, Random Forest, DistilBERT, SqueezeBERT, DeBERTa and RoBERTa. The accuracy, precision, recall and F1-score along with 10-fold cross validation were used to evaluate the model performance. In the experiment, the model RoBERTa obtained the best performance in terms of the accuracy, which was 0.981, and the mean cross validation accuracy was around 0.97. The results of the study indicated that the transformation-based method is very appropriate for detecting depression-related language in social media posts and can be used to help detect potential mental health problems in time.

Almutairi et al. [([2] S. Almutairi, 2024)] had introduced a hybrid Deep Learning (DL) model to predict depression symptom from large scale text data. The study centered on the problems of early detection of depression based on questions posted by young users on an online platform publicly available. The experiments were performed on textual data from the Twitter, Facebook and SMILE Twitter Emotion Dataset containing 4000 depressive tweets, 7000 instances of social media and some other instances with non-depressive content. The data was preprocessed before training the model using certain steps: Data cleaning, normalization, tokenization, lemmatization, and stop-word removal. Depression-related features were extracted from the medical and psychological experts' symptom indicators and represented using Word2Vec embeddings and one-hot encoding. To enhance the extraction of features and the capturing of contextual relationships in the text, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), its improvement (Two-State LSTM) and attention mechanism were suggested to be combined in the proposed framework. The proposed model is tested against the traditional ones. The experimental results revealed that the hybrid model produced

high accuracy (97.23%), precision (98.57%), recall (97.13%), F1-score (97.84%) and specificity (97.93%) as compared to the baseline models. The study finally concluded that CNN, TS-LSTM, and attention mechanisms combined together are an effective method to predict depression symptoms from large scale social media text and can be used to contribute to early depression symptoms risk detection.

Rastogi et al. ([3] Rastogi, 2021) suggested the benchmark framework for the detection of stress from social media text utilizing traditional and transformer-based language models. Four publicly available data sets were used for the study, namely Reddit Title, Reddit Combi, Twitter Full and Twitter Non-Advert, collected from Reddit and Twitter. Stress positive and negative posts have been provided in these datasets from the official senticnet repository on GitHub and Kaggle. The data was preprocessed and automatically annotated, preparing the textual content for classification, before model training. Three categories were tested: lexicon-based, embedding-based and transformer-based. SenticNet, SO-CAL, AFINN, TensiStrength, VADER, SentiWordNet, FastText, GloVe, Word2Vec, RoBERTa, and DistilBERT were all evaluated. The accuracy and F1-score were used to evaluate the performance of the model. The experimental results indicated that the highest results were obtained by DistilBERT without the summarization of text with 98.20% accuracy and a 98.15% F1-Score on the Reddit Title dataset. Based on the results of the study, it was concluded that they were able to stress detection in social media text using transformer-based models more effectively than lexicon-based and embedding-based models.

Bae et al. ([4] Bae, 2021) developed a ML method to identify patients with schizophrenia from Reddit social media posts. The posts were pulled from the r/schizophrenia subreddit, which is a subreddit where users post about schizophrenia, and six non-mental health subreddits (fitness, jokes, meditation, parenting, relationships and teaching) served as the control group. The final data set consisted of 13,156 schizophrenia posts and 247,569 control posts between 9/2016 and 9/2020. The data were subjected to multiple preprocessing steps, such as lowercasing, tokenization, removal of punctuation and numbers, stop-word removal, lemmatization and stemming using Natural Language Toolkit (NLTK) before model training. For the study, Linguistic Inquiry and Word Count (LIWC) features, Latent Dirichlet Allocation (LDA) topic distributions, and TF-IDF were used to represent features. Four types of supervised ML algorithms were tested such as Support Vector Machine (SVM), Logistic Regression, Naïve Bayes, and Random Forest. The dataset was balanced and divided into 80/20 training and testing while hyperparameter tuning was done using 10 fold cross validation. Model performance was evaluated based on the accuracy, precision, recall, F1 score and

the Area Under the Receiver Operating Characteristic Curve (AUC). From the results of the experiments, the best performance was achieved by Random Forest with 96 Percentage, 98 Precision, 94 Recall, 96 F1 Score and AUC of 0.97. Overall, the study found that the combination of ML and Natural Language Processing (NLP) techniques can be used to effectively identify linguistic patterns associated with schizophrenia in Reddit posts and can potentially be used to help detect individuals who may be at risk for developing the condition.

Roy et al. ([5] Roy, 2020) suggested an ML approach called Suicide Artificial Intelligence Prediction Heuristic (SAIPH) to predict the potential future suicidal ideation risk from the data available in twitter. The data were gathered from public Twitter posts of users who expressed suicidal ideation and a control group. There were 512,526 tweets of 283 suicidal ideation cases and 3,518,494 tweets of 2,655 control users in the training dataset. There were 277 suicidal ideators and 3,159 control events in an independent validation dataset. The proposed framework used a neural network to estimate psychological constructs such as stress, loneliness, hopelessness, insomnia, depression, anxiety, and burden. These outputs were then served as the input features of a Random Forest (RF) Classifier for predicting suicidal ideation risk. AUC and temporal prediction analysis were the methods used for the assessment of model performance. Experimental results demonstrated an AUC of 0.88 for SAIPH, and the ability of the model to identify periods of increased suicide risk. Peak-prediction scores also suggest, roughly, a 7-fold increase in the risk of suicidal ideation in the next 10 days. The study highlighted the findings that the longitudinal study of Twitter posts using ML can be effective to identify and continuously monitor individuals at risk of suicidal ideation in an early phase.

The authors of ([6] Rastogi, 2022) have come out with a ML and DL based framework for stress detection from social media interactions by employing NLP techniques. In the study, we employed two publicly available Twitter datasets. The first data set had a total of 100,042 positive and negative tweets for binary sentiment classification. The second data set contained 7,934 tweets classified in five emotion categories: joy, sadness, neutral, anger and fear. The data was pre-processed before training the model using several techniques such as tokenization, removal of unwanted patterns, discarding stop-words and stemming. The study used Bag-of-Words (BoW) as a representation of features and LDA for topic modelling. Various models including Logistic regression, Decision Tree, Random Forest, VADER and a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model for emotion classification were evaluated. The accuracy and F1-score were used to evaluate the performance of the models. For

binary (positive and negative) sentiment classification, Random Forest had the highest performance with 97.78% accuracy and BERT had a five-class emotion classification with 94% accuracy and 83% macro F1-score. Finally, the study yielded that the combination of NLP techniques, topic modelling, ML, and emotion analysis using BERT is an effective method to identify stress from social media texts.

Proactive social-sensor system for mental health monitoring based on Twitter data was proposed by Sharma et al. ([7] Sharma, 2022). The researchers used Twitter users as social sensors and mined historical tweets for early detection of mental illness. A total of 2,138 initial tweets on six mental health conditions (depression, bipolar disorder, Alzheimer's disease, dementia, schizophrenia, and Asperger's syndrome) were collected. They then collected 979,466 historical tweets from 600 users for a six-month period. Based on the mental health content published, a total of 479 users (356 mental health users and 123 control users) were identified and included in the final dataset after filtering out inactive users and outliers. Regular expressions were used to perform detailed preprocessing of the data. As part of this process, usernames, URLs, special characters, repeated letters, acronyms and spelling mistakes were deleted. Using Text Blob, the researchers determined the sentiment polarity and created two new datasets, namely, User Summary dataset and User Weekly Negativity dataset. The evaluation of several models was considered such as SVM, Random Forest, Dense Neural Network (DNN), and multiple models using LSTM. Using stratification, the data were split with 70 % training data and 30 % testing data. From the experimental results, it can be seen that the DNN has the highest test accuracy of 84.31%. The single-layer LSTM with dropout and the SVM were found to be 82.92% and 73.43% accurate respectively. The study revealed that past sentiment analysis from Twitter can be used for proactive mental health monitoring and to identify users at risk.

Figuerêdo et al. ([8] Figuerêdo, 2022) suggested a DL approach for the early detection of depression from social media posts. The framework developed is using the CNN along with the context-independent word-embedding methods with both Early Fusion and Late Fusion to merge multiple representations of the text. The data for the study was obtained from the eRisk 2017 dataset, which consists of chronological English posts from Reddit users. A total of 887 users (135 depressive users and 752 non-depressive users) were included in the dataset. There were up to 2000 posts in total per user, including the title, publishing date and textual content. Prior to the model learning, title and text were concatenated and multiple pretrained

word representations were tested for representing features. The authors also proposed a semantic mapping method based on emoticon expression to retain the emotion information in users' posting. The experimental results demonstrated that the proposed framework obtained a precision of 0.76. Recall and precision were improved by up to 46.3% and 32.1% by semantic emoticon mapping. There was a 14.5% improvement in recall and a 40.8% improvement in precision when compared to the baseline embedding method. The authors found that combination of CNN, embedding-fusion and emoticon-based semantic information could be beneficial to achieve early detection of depression from social-media postings.

Kumar et al. ([9] Kumar, 2022) suggested a hybrid DL approach to predict depression from Twitter posts. The proposed framework consisted of two architectures, namely feature-rich CNN and Bidirectional Long Short-Term Memory (BiLSTM) network. Local semantic features were captured by CNN and contextual information was captured in forward and backward direction by BiLSTM. A benchmark twitter data set was used which was divided in three parts: D1, D2 and D3. There were 1,402 depressed users who produced a total of 292,564 tweets in D1. D2 was made up of non-depressed users whereas D3 was comprised of 36,993 depression-candidate users and over 35 million tweets. The experiments concerned only D1 and D2 with 2,558 samples from D1 and 5,304 samples from D2. The data was preprocessed by various means before the training of the model, such as removing usernames, hashtags, URLs, numbers, duplicate records, missing values, emojis, stop words and extra spaces. Stemming, tokenization, sequence padding and word embeddings were also used. The proposed CNN-BiLSTM model was compared with CNN, RNN and other baseline models. For model performance, the accuracy, precision, recall, F1-score, specificity, and AUC were used. The experimental results proved that CNN-BiLSTM model has the highest accuracy of 94.28%, precision of 96.99%, F1-Score of 94.78%, specificity of 96.35% and AUC of 95.43%. The study found that the merging of CNN-based feature extraction with Bi-directional sequence learning can lead to a better prediction of depression in Twitter text.

Kerasiotis et al. ([10] Kerasiotis, 2024) introduced a transformer-based model to classify depression from social media content. In this study, DistilBERT embeddings are fused with auxiliary information and linguistic features to enhance depression classification. The framework has extracted contextual representations from the last transformer layers and combined it with more user-level features and then classified by means of a multilayer perceptron (MLP). To enhance the class

balance and generalization, the authors also used data augmentation methods which are inspired by Easy Data Augmentation (EDA). The weighted Precision of the proposed approach was 84.26%, while the weighted Recall and F1-score were 84.18% and 84.15%, respectively, based on the experimental results. The study found that the use of transformer models combined with additional social media characteristics helps in performing depression detection when compared to text-based models.

Renjith et al. ([11] Renjith, n.d.) proposed a hybrid DL model for Identifying suicidal ideation from social media posts at an early stage. The proposed model combines LSTM, attention mechanism, and CNN to better capture the textual patterns in suicide. The University of Maryland Reddit Suicidality Dataset was used, consisting of posts from the Reddit subreddit SuicideWatch. Following data preprocessing and data annotation, 55680 posts were collected from 760 users on Reddit and classified into four levels of risk: No Risk, Low Risk, Moderate Risk, and Severe Risk. The data was pre-processed with several steps such as concatenating the post title and content, removing duplicate posts, lowercasing, removing URLs, emails, punctuation, stop words and newline symbols and then tokenization and lemmatization using tools before they were used in model training. To encode textual features, word embeddings of 300 dimensions were used by applying Word2Vec. Some basic models, such as the SVM model, CNN model, LSTM model and hybrid model LSTM-CNN were selected as basic models. Eighty percent of the data set was used for training and 20% was used for testing. The accuracy, precision, recall, and F1-score were used to assess the performance of the models. An experimental result indicated that the proposed LSTM-Attention-CNN model performed the best with accuracy, precision, recall, and F1 score 90.3%, 91.6%, 93.7% and 92.6%, respectively, as compared with all the baseline models. The study finally concluded that using the combination of LSTM, CNN, and attention mechanism is effective in capturing contextual and semantic information from the social media posts which in turn boosted the early detection of suicidal ideation.

Rabani et al. [13] came up with an Enhanced Feature Engineering Approach for Suicidal Risk Identification (EFASRI) for identification of various level of suicidality from social media post. The data used in the study was a unique data set gathered from the Twitter and Reddit platforms. To capture the informative textual features pertaining to suicidal behaviors, various Natural Language Processing (NLP) and feature engineering techniques were used. Support Vector Machine (SVM), Random Forest (RF) and the Extreme Gradient Boosting (XGB) classifiers were tested for multi-class suicidal risk classification. Performance of the proposed

EFASRI framework was found to be better than other traditional feature extraction methods. Regarding the evaluation of the models, XGBoost (XGB) model has the highest overall accuracy of 96.33% which showed that the proposed feature engineering method is effective for social media content to early identify suicidal risk.

14. Improving Understanding of Depression Detection in social media through Explainability of Transformer Models

<https://ieeexplore.ieee.org/document/10847487> (derived from social media content).

Using transformer architectures, the researchers were able to capture linguistic patterns correlated with depression, and they ensured explainability of model decisions through the implementation of explainability mechanisms. The research emphasized the need for transparency in mental health applications and revealed that transformer models can capture the psychological signals within social media texts and also offer interpretable predictions that are suitable for real world mental health monitoring systems.

A comparative study for transformer-based and LSTM-based models to detect suicidal ideation from Reddit posts was proposed by Hasan and Saquer ([14] Hasan, 2024). The goal of this study was to investigate how well perform different state of the art models based on NLP for early detection of suicidal ideation from social media posts. To ensure the quality of the data for model development, linguistic, topic modeling and statistical analyses were conducted on the data retrieved from the various subreddits of Reddit. Data pre-processing was performed and data quality was carried out before training. The study showcased the performance of the different models, such as BERT, RoBERTa, DistilBERT, ALBERT, and ELECTRA, against different types of LSTM-based models like attention-based LSTM model with BERT embedding. The accuracy and F1-score measures were used to evaluate the model performance. From the experimental results, it is found that the highest performance is achieved by RoBERTa, which got 93.22% accuracy and 93.14% F1 score, followed by Attention-LSTM with BERT embeddings with 92.65% accuracy and 92.69% F1 score. Finally, the study showed that the trained transformer-based models performed better than traditional LSTM-based models in suicidal ideation detection and that there are strong prospects for implementing such models in the creation of reliable mental health monitoring systems based on social media data. Transformer based depression severity detection framework from social media texts was proposed by.

Qasim et al. ([15] Qasim, 2025). A two-fold objective was designed to categorize the depression into Minimum, Moderate and Severe using traditional ML approaches, Sentence transformer-based approaches and End-to-End transformer architecture. The researchers took a Reddit data, containing 24 thousand posts, in which the posts were classified based on clinical psychology assessment using the Beck Depression Inventory (BDI) and DSM-V criteria. The data was preprocessed before training the model, involving several steps such as punctuation removal, numbers removal, removal of URLs, hashtags, mentions and stop words, lowering of case, tokenization, stemming and lemmatization by using Natural Language Tool Kit (NLTK). The study has tested three sets of approaches: content-based methods based on N-grams, context-based methods taking advantage of Sentence Transformers with multiple ML classifiers and transformer-based approaches represented by BERT-base-uncased, RoBERTa-base, and Microsoft/DeBERTa-base. The data set was split into a training set of 70%, a validation set of 10% and a test set of 20%. Precision, recall and F1 score were used to evaluate model performance. The experimental results revealed that RoBERTa-base model got OVEF1=0.91, while Microsoft/DeBERTa-base model showed maximum class-wise F1 score with four digits of precision and recall in Mild, Moderate and Severe classes. The study found that the transformer-based models outperformed traditional ML and sentence transformer models greatly in identifying the severity of depression in Reddit posts and serve as effective models for a scalable mental health monitoring and early intervention system.

2.1 Research Gap

Although significant progress has been achieved in applying Artificial Intelligence techniques for mental health detection from social media, several research gaps remain. First, many previous studies focus on a single mental health condition, such as depression or suicidal ideation, without providing a comprehensive comparison across different psychological disorders. Second, the reviewed studies employ different datasets, preprocessing techniques, text representation methods, and evaluation protocols, making direct comparison between reported results difficult. Furthermore, traditional Machine Learning models rely heavily on handcrafted textual features, whereas Deep Learning and Transformer-based models automatically learn contextual representations. Although Transformer models generally demonstrate superior performance, relatively few studies provide systematic comparisons between conventional Machine Learning, Deep Learning, and Transformer architectures using consistent experimental settings. Another important limitation is that most publicly available datasets are collected from English-language social media platforms, particularly Reddit and Twitter (X). Consequently, the generalizability of current approaches across different languages, cultures, and social media platforms remains insufficiently investigated. Finally, many existing studies primarily emphasize classification performance while providing limited discussion regarding model interpretability, computational complexity, and practical deployment in real-world healthcare applications. Therefore, there

remains a need for comprehensive comparative studies that evaluate different Artificial Intelligence techniques under consistent experimental settings and identify the most effective approaches for early mental health risk detection from social media posts.

2.2 Chapter Summary

The chapter presented the theoretical background required for understanding early mental health risk detection from social media posts. It introduced the fundamental concepts of Artificial Intelligence, Natural Language Processing, text preprocessing, text representation techniques, Machine Learning, Deep Learning, and Transformer-based language models. Furthermore, the chapter reviewed the most relevant studies in the field, analyzed their methodologies, highlighted current research trends, and identified the major research gaps. The knowledge presented in this chapter provides the theoretical foundation for the research methodology described in the following chapter.

1 2.1 Comparison of Previous Studies

Paper	All Models	Datasets	Dataset Link	Limitation	Best model	results
Bokolo and Liu ([1] Bokolo, 2023)	Logistic Regression, Bernoulli Naive Bayes, Random Forest, DistilBERT, SqueezeBERT, DeBERTa, RoBERTa	Sentiment140 Twitter Dataset; 632,000 tweets were repurposed and re-labeled for depression detection.	https://www.kaggle.com/datasets/kazanova/sentiment140 Sentiment140 official dataset / Kaggle mirror	The dataset was originally designed for sentiment analysis, not clinical depression diagnosis. Depression labels were inferred from negative sentiment and polarity-based relabeling, so negative	RoBERTa	Accuracy = 0.981; Mean Accuracy = 0.97 across 10-fold cross-validation

				sentiment may not always indicate depression . The study is also limited to English tweets.		
Almutairi et al ([2] S. Almutairi, 2024)	CNN, LSTM, Two-State LSTM, CNN-LSTM with Attention Mechanism	Large-scale textual dataset from a publicly accessible online platform for young people	https://www.kaggle.com/datasets/infamouscoder/mental-health-social-media	Dataset source is limited to text questions from young users ; generalizability to other platforms such as Twitter or Reddit may require further validation.	Proposed Hybrid CNN + Attention-based TS-LSTM	Accuracy = 97.23 %; Precision = 98.57 %; Recall = 97.13 %; F1-score = 97.84 %; Specificity = 97.93 %
Rastogi ([3] Rastogi, 2021)	SenticNet, SO-CAL, AFINN, TensiStrength, VADER, SentiWordNet, FastText, GloVe, Word2Vec	Stress Detection from Social Media Articles; Reddit and Twitter datasets for stress-positive and stress-	https://www.kaggle.com/datasets/mexwell/stress-detection-from-social-media-articles	The task focuses on stress detection rather than direc	DistilBERT	Accuracy = 98.20 %; F1-score = 98.

	c, RoBERTa , DistilBERT	negative classification	GitHub official dataset / Kaggle mirror	t clinic al diagn osis of depre ssion or ment al illnes s. Twitt er posts are short er and noisi er than Redd it posts .		15 %
Bae et al. ([4] Bae, 2021)	SVM, Logistic Regression, Naive Bayes, Random Forest	Reddit posts: r/schizophrenia + control subreddits: fitness, jokes, meditation, parenting, relationships, teaching. Final dataset: 13,156 schizophrenia posts and 247,569 control posts	Dataset available from the correspo nding author upon reasonab le request. Code: GitHub repositor y mention ed in the paper	No clinic al diagn osis confi rmati on for Redd it users ; limit ed to Redd it; tradit ional ML only, not deep learn ing	Rando m Forest	Acc uracy 96 %, Pre cisi on 98 %, Rec all 94 %, F1- scor e 96 %, AU C 0.9 7
Roy et al. ([5]	Neural Networks,	Twitter data: 512,526 tweets from	Not publicly available	Limit ed to Twitt	SAIPH: Neural Networ	AU C = 0.8

Roy, 2020)	SVM, Random Forest, Sentiment Analysis	283 SI cases and 3,518,494 tweets from 2,655 controls; validation on 277 suicidal ideators and 3,159 control events	due to Twitter data-sharing policies	er; code not public; outcome based on expression of suicidal ideation on social media; limited demographic data	ks + Random Forest	8; temporal prediction showed ~7-fold increase in risk within 10 days
Rastogi et al. ([6] Rastogi, 2022)	Logistic Regression, Decision Tree, Random Forest, Vader, LDA, BERT	Sentiment dataset: 100,042 tweets; Emotion dataset: 7,934 tweets with joy, sadness, neutral, anger, fear labels	Sentiment Analysis dataset: IEEE Data Port. Emotion dataset: data. World "Detecting Emotions in Text"	Does not directly diagnose clinical stress; stress is inferred from sentiment/emotion; limited to tweet text; future work suggests addin	Random Forest for sentiment; BERT for emotion classification	Random Forest accuracy = 97.78%; BERT accuracy = 94%; BERT macro F1-score = 83%

				g healt h para mete rs		
Sharm a et al. ([7] Sharm a, 2022)	SVM, Random Forest, Single Dense Neural Network, Single- Layer LSTM, Single- Layer LSTM with Dropout, Double- Layer LSTM, Double- Layer LSTM with Dropout	Twitter dataset. Initial set: 2,138 mental- health-related tweets. Historical set: 979,466 tweets from 600 users. Final analysis: 479 users, including 356 positive cases and 123 controls.	No public dataset link was provided . The authors created their own dataset using Twitter Search API and a custom crawler.	Small user- level samp le; label s were base d on self- repor ted posts and majo rity votin g rathe r than clinic al diagn osis; datas et was imba lance d; limit ed to Twitt er; senti ment analy sis may be affec ted by sarca sm, slang	Single Dense Neural Networ k	Acc urac y = 84. 31 %. Sin gle- Lay er LS TM wit h Dro pou t = 82. 92 %. SV M = 73. 43 %

				, and regional language.		
Figueredo et al. ([8] Figueredo, 2022)	CNN, individual pre-trained word embeddings, Early Fusion, Late Fusion, semantic emoticon mapping	eRisk 2017 Reddit dataset: 887 users, including 135 depressive and 752 non-depressive users; chronologically ordered English posts, with up to 2,000 posts per user.	eRisk research collection; access is generally controlled through the eRisk organizers and may require a data-use agreement. No unrestricted direct download link was provided in the paper.	Labels are based on social media evidence rather than direct clinical assessments; data are limited to Reddit and English; the dataset is imbalanced; results may depend on emoticon usage; ethical and privacy concerns	CNN with fusion-based word embeddings and semantic emoticon mapping	Precision = 0.76; recall improvement up to 46.3%; precision improvement up to 32.1%; compared with baseline embeddings, recall gain = 14.5% and precision gain =

				limit data access.		40.8%
Kumar et al. ([9] Kumar, 2022)	CNN, RNN, BiLSTM, CNN-BiLSTM, baseline ML/DL approaches	Twitter benchmark dataset derived from Shen et al. [54]. D1: 292,564 tweets from 1,402 depressed users. D2: non-depressed Twitter users. D3: 36,993 depression-candidate users and more than 35 million tweets. Experiments used D1 and D2 with 2,558 and 5,304 samples.	The paper cites the benchmark dataset source in Shen et al. [54], but does not provide a direct public download link in the supplied text.	Limited to English Twitter text; labels depend on self-reported depression patterns rather than confirmed clinical diagnosis; emojis and non-English content were removed; dataset description is partly inconsistent because very	CNN-BiLSTM	Accuracy = 94.28%, Precision = 96.99%, F1-score = 94.78%, Specificity = 96.35%, AUC = 95.43%

				large raw collections were reduced to smaller experimental samples.		
Kerasiotis et al ([10] Kerasiotis, 2024)	DistilBERT + Auxiliary Features + MLP	Social Media Depression Dataset	https://link.springer.com/article/10.1007/s13278-024-01360-4			Precision = 84.26%, Recall = 84.18%, F1-score = 84.15%
Renjith et al. ([11] Renjith, n.d.)	SVM, CNN, LSTM, LSTM-CNN, LSTM-Attention-CNN	University of Maryland Reddit Suicidality Dataset; 55,680 Reddit posts from 760 users; four classes: No Risk, Low Risk, Moderate Risk, Severe Risk	No direct public dataset link was provided in the paper. The dataset is described as the University of Maryland Reddit Suicidality Dataset.	The dataset is limited to Reddit posts and lacks clinical records, validated diagnoses, and self-reports.	LSTM-Attention-CNN	Accuracy = 90.3%, Precision = 91.6%, Recall = 93.7

				t quest ionna ires. The post- level label s were also deriv ed from user- level label s, whic h may intro duce labeli ng uncer taint y and limit gener alizat ion to other platf orms and langu ages.		%, F1 - sc ore = 92. 6 %
Rabani ([12] Rabani , 2023)	SVM, Random Forest (RF), XGBoost (XGB)	Novel dataset collected from Twitter and Reddit	Dataset not publicly released		XGBoo st (XGB)	Acc urac y = 96. 33 %
Explai nable Depres sion	Transform er-based Models with Explainabi lity Framewor k DistilBER T,	Social Media Depression Dataset	https://w ww.kagg le.com/d atasets/k azanova/ sentimen t140		Explain able Transfo rmer Model	Imp rov ed inte rpre tabi lity whi le mai

	SqueezeBERT, DeBERTa, RoBERTa					maintaining competitive depression detection performance Recall = 99% , F1-score = 98% , Mean 10-fold Cross-Validation Accuracy = 97%
Hasan and Saquer ([14] Hasan, 2024)	BERT, RoBERTa, DistilBERT, ALBERT, ELECTRA, Multiple LSTM models, Attention-LSTM with	Curated Reddit dataset collected from multiple subreddits	Not specified in the paper	The study relied only on Reddit data, which may limit gener	RoBERTa	Accuracy = 93.22% , F1-score = 93.14% ; Attention

	BERT embeddings			alization to other social media platforms and languages.		n-LS TM with BERT embeddings achieved Accuracy = 92.65% , F1-score = 92.69%
Qasim et al. ([15] Qasim, 2025)	N-grams + RF, DT, NB, GNB, GBC, AB, MLP, KNN, SVM, LR; Sentence Transformers (multi-qa-mpnet-base-dot-v1, paraphrase-mpnet-base-v2, all-MiniLM, all-distilroberta, distiluse, etc.); BERT-base-uncased, RoBERTa-base, Microsoft/	Reddit dataset containing 24,000 clinically annotated posts labeled as Minimum, Moderate, and Severe depression	Not publicly provided in the paper	The study relies only on Reddit data and English text, which may limit generalization to other social media platforms, languages,	Microsoft/DeBERTa-base	F1-score = 0.90 ; Class-wise F1-scores: Mild = 0.96 , Moderate = 0.95 , Severe = 0.96 . RoBERTa

	DeBERTa -base			and clinic al popu latio ns. Tran sfor mer mode ls also requi re subst antial comp utati onal resou rces for traini ng and deplo ymen t.		- bas e achi eve d the hig hest ove rall F1- sco re = 0.9 1 amo ng the eval uate d tran sfor mer mo dels .
--	------------------	--	--	---	--	--

Chapter 4 Proposed Approach (Methodology)

Chapter 5: Experimental Results and Discussion

Chapter 6 Conclusion and Future Work

References

([1] Bokolo, 2023)

Bokolo, B. G., & Liu, Q. "Deep Learning-Based Depression Detection from Social Media: Comparative Evaluation of ML and Transformer Techniques." *Electronics*, vol. 12, no. 21, 4396, 2023.

([2] S. Almutairi, 2024)

S. Almutairi, M. Abohashrh, H. H. Razzaq, M. Zulqarnain, A. Namoun and F. Khan, "A Hybrid Deep Learning Model for Predicting Depression Symptoms from Large-Scale Textual Dataset," in *IEEE Access*, vol. 12, pp. 168477-168499, 2024.

([3] Rastogi, 2021)

Rastogi, A., Liu, Q., & Cambria, E. "Stress Detection from Social Media Articles: New Dataset Benchmark and Analytical Study." 2022 International Joint Conference on Neural Networks, IJCNN, 2022.

([4] Bae, 2021)

Bae, Yi Ji, Midan Shim, and Won Hee Lee. "Schizophrenia detection using machine learning approach from social media content." *Sensors* 21.17 (2021): 5924.

([5] Roy, 2020)

Roy, Arunima, et al. "A machine learning approach predicts future risk to suicidal ideation from social media data." *NPJ digital medicine* 3.1 (2020): 78.

([6] Rastogi, 2022)

Rastogi, K., Rani, R., Prakash, K. B., & Gupta, D. (2022). Stress detection using natural language processing and machine learning over social interactions. *Journal of Big Data*, 9, 33.

([7] Sharma, 2022)

Sharma, S. K., Gaur, A., Gupta, D., & Wang, L. (2022). Machine learning-based proactive social-sensor service for mental health monitoring using Twitter data. *Journal of Innovation & Knowledge*, 7(4)

([8] Figuerêdo, 2022)

Figuerêdo, J. S. L., Maia, A. L. L. M., & Calumby, R. T. (2022). Early depression detection in social media based on deep learning and underlying emotions. *Online Social Networks and Media*, 31, 100225.

<https://doi.org/10.1016/j.osnem.2022.100225>

([9] Kumar, 2022)

Kumar, A., Sharma, A., & Arora, A. (2022). A hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM. *Multimedia Tools and Applications*, 81, 23649–23685.

([10] Kerasiotis, 2024)

Kerasiotis, M., Ilias, L., & Askounis, D. (2024). *Depression Detection in Social Media Posts Using Transformer-Based Models and Auxiliary Features*. *Social Network Analysis and Mining*, 14, 196.

([11] Renjith, n.d.)

Renjith, Shini, Annie Abraham, Surya B. Jyothi, Lekshmi Chandran, and Jincy Thomson. “An Ensemble Deep Learning Technique for Detecting Suicidal Ideation from Posts in Social Media Platforms.” *Journal of King*

([12] Rabani, 2023)

Rabani, Syed Tanzeel, Akib Mohi Ud Din Khanday, Qamar Rayees Khan, Umar Ayoub Hajam, Ali Shariq Imran, and Zenun Kastrati. Detecting Suicidality on Social Media: Machine Learning at Rescue.” *Egyptian Informatics Journal*, vol. 24, no. 2, 2023, pp. 291–302.

<https://doi.org/10.1016/j.eij.2023.04.003>

([14] Hasan, 2024)

Hasan, K., & Saquer, J. (2024). A Comparative Analysis of Transformer and LSTM Models for Detecting Suicidal Ideation on Reddit. *Proceedings of ICMLA 2024*.

([15] Qasim, 2025)

Qasim, A., Mehak, G., Hussain, N., Gelbukh, A., & Sidorov, G. (2025). Detection of Depression Severity in Social Media Text Using Transformer-Based Models. *Information*, 16(2), 114.