Multimodal Machine Learning Models for Depression Detection from Clinical Interviews

Thesis

submitted in fulfilment of the requirements for the degree of

Masters of Technology

in

Computer Science and Engineering

by

Srijanak De

(19CS30047)

Under the supervision of

Prof. Partha Pratim Chakraborty

Department of Computer Science and Engineering

Indian Institute of Technology Kharagpur

Academic Year, 2023-24

April, 2024

DECLARATION

I certify that

(a) The work contained in this report has been done by me under the guidance of my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Date: April, 2024 (Srijanak De)
Place: Kharagpur (19CS30047)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

KHARAGPUR - 721302, INDIA

CERTIFICATE

This is to certify that the project report entitled "**Multimodal Machine Learning Models for Depression Detection from Clinical Interviews**" submitted by **Srijanak De** (Roll No. 19CS30047) to the Indian Institute of Technology Kharagpur towards fulfilment of requirements for the award of the degree of Masters of Technology in Computer Science and Engineering is a record of bonafide work carried out by him under my supervision and guidance during Academic Year, 2023-24.

Date: April, 2024
Place: Kharagpur

Prof. Partha Pratim Chakraborty
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Kharagpur - 721302, India

Acknowledgements

I would like to thank my Thesis supervisor *Prof. Partha Pratim Chakraborty* for their exceptional guidance and support, without which this project would not have been possible. They have always motivated me to explore as much as I can, look into multiple papers and try as many ideas as I can. They have always supported me through whatever problems I faced during the project.

I recognize that this project would not have been possible without the support from the Department of Computer Science and Engineering, IIT Kharagpur. Many thanks to all those who made this project possible.

I am grateful to my friends, who were my constant support in every situation, and last but not least, my parents for their invaluable trust and support in all my choices.

Srijanak De

Abstract

Name of the student: **Srijanak De** Roll No: **19CS30047**
Degree for which submitted: **Masters of Technology**
Department: **Department of Computer Science and Engineering**
Thesis title: **Multimodal Machine Learning Models for Depression Detection from Clinical Interviews**
Thesis supervisor: **Prof. Partha Pratim Chakraborty**
Month and year of thesis submission: **April, 2024**

This thesis explores a novel automated depression detection system that utilizes a multi-modal machine learning approach to detect depression in patients. Employing the DAIC dataset, which includes textual interviews and audio-visual recordings from 100 patients, the project began with text-based exploratory analysis using techniques like PCA, UMAP, and t-SNE, followed by initial depression classification using traditional and neural network-based models, achieving F1 scores between 0.52 and 0.61. BERT achieved the maximum average F1 score of 0.67. For audio analysis, Random Forest classifiers were applied to features extracted from .wav files using mel-frequency cepstral coefficients (MFCCs), achieving an F1 score of 0.52. HuBERT and VGGish models for feature extraction from raw audio also used Random Forest, yielding F1 scores of 0.59 and 0.65 respectively. To process the video files and extract features, SVM was used, resulting in an F1 score of 0.51. A comprehensive multi-modal machine learning model was then created, by passing embeddings from BERT for transcripts, VGGish from audio, and SVM from video through an MLP Classifier, yielding an F1 score of 0.72.

Contents

Declaration i

Certificate ii

Acknowledgements iii

Abstract iv

Contents v

List of Figures viii

1 Introduction 1
- 1.1 Dataset Overview . . . 1
- 1.2 Motivation . . . 2
- 1.3 Aims and Research Focus . . . 2
- 1.4 Contribution . . . 2
- 1.5 Structure of the Report . . . 3

2 Literature Review 4
- 2.1 Related Work . . . 4
- 2.2 Literature Gaps . . . 4
- 2.3 Dataset . . . 5
  - 2.3.1 The Distress Analysis Interview Corpus (DAIC) [5] . . . 5
- 2.4 Detection of Depression . . . 6
  - 2.4.1 The Verbal and Non-Verbal Signals of Depression – Combining Acoustics, Text and Visuals for Estimating Depression Level [8] 6
  - 2.4.2 Text-based depression detection on sparse data [4] . . . 6
  - 2.4.3 Affective Conditioning on Hierarchical Attention Networks applied to Depression Detection from Transcribed Clinical Interviews [10] . . . 7

3 Methodology 8
- 3.1 Major Depressive Disorder (MDD) . . . 8

3.1.1 Diagnostic Criteria for MDD 8
3.1.1.1 Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) 8
3.2 Creating a baseline questionnaire structure 10
3.3 Making the baseline adaptive to newer contexts and rules 11
3.4 Dataset Overview 14
3.4.1 Data Description 14
3.4.2 Structure of the Data 14
3.4.3 Supplementary Data Files 15
3.4.4 Preprocessing 15
Pseudocode 17
3.5 Exploratory Data Analysis 17
3.5.1 Uniform Manifold Approximation and Projection (UMAP) . . 17
3.5.2 Principal Component Analysis (PCA) 19
3.5.3 t-Distributed Stochastic Neighbor Embedding (t-SNE) 19
3.5.4 Word Cloud 20
3.6 Feature Extraction 22
3.6.1 TF-IDF 22
3.7 Word2Vec 23
3.7.1 Mathematical Formulation 23
3.7.2 Training of the Model 24
3.8 Model Training and Results 24
3.8.1 Logistic Regression 24
3.8.2 Multinomial Naive Bayes 25
3.8.3 Support Vector Machine (Linear RBF Kernel) 26
3.9 Advanced Modeling with BERT and DistilBERT 27
3.9.1 BERT 27
3.9.2 DistilBERT 27
3.10 Audio Analysis with Random Forest, HuBERT, and VGGish 28
3.10.1 Random Forest on MFCCs 28
3.10.2 HuBERT and VGGish for Feature Extraction 29
3.11 Video Analysis with SVM 29
3.11.1 SVM for Feature Extraction 29
3.12 Multi-Modality Analysis: Integrating Audio, Video, and Transcript Data 30
3.12.1 Ensemble Approaches with MLP Classifier 30
3.13 Model Evaluation and Comparison 31
3.13.1 Performance Metrics 31
3.13.2 Analysis of Results Across Modalities 32
3.14 Iterative Approach 32
3.14.1 Sequential Model Improvement Across Modalities 32

3.14.2 Lessons Learned 32
3.15 Challenges and Limitations 33
3.15.1 Data-Related Challenges Across Modalities 33
3.15.2 Model-Specific Limitations 33
3.16 Conclusion of Methodology 34
3.16.1 Summary 34
3.16.2 Implications 34

4 Conclusion and Future Work **35**
4.1 Conclusion 35
4.2 Future Work 36

Bibliography **37**

List of Figures

3.1 Example SCID-5-CV Baseline Question-Answer Pathway 10
3.2 Sample Question Asked in Webapp 11
3.3 A Part of the Questionnaire Graph Based on SCID-5-CV 13
3.4 Generalization Tool with SCID-5-CV as Default 13
3.5 Demonstration of the Working Generalization Tool 13
3.6 Text Analysis Pipeline 16
3.7 UMAP 18
3.8 PCA 20
3.9 tSNE 21
3.10 Word Cloud 21
3.11 Multi-modal Model Pipeline 30

Chapter 1

Introduction

The prevalence of mental health disorders globally necessitates the development of efficient and scalable diagnostic tools. Depression, being one of the most pervasive of such disorders, presents unique challenges and opportunities for the field of computational psychiatry. This research utilizes a novel approach combining natural language processing (NLP) and machine learning (ML) techniques to address these challenges. Below, we present a structured overview of the work conducted, delineating the dataset, the objectives, the contributions of the study, and the layout of this report.

1.1 Dataset Overview

The dataset employed in this study originates from structured clinical interviews consisting of 100 patient transcripts, audio recordings, and video recordings, collected as part of the Distress Analysis Interview Corpus (DAIC) project. Each component of the dataset, transcript, audio, and video—features provides a rich, multi-modal representation of depression. The transcripts offer a language-based insight, while the audio files provide vocal characteristics and the video files capture visual behaviors, all contributing to a comprehensive dataset.

1.2 Motivation

The motivation behind this research is twofold. On the clinical front, it aims to advance the diagnostic capabilities for depression, enabling faster and more accurate identification of the disorder through automated means. From a technological perspective, it seeks to push the boundaries of NLP and ML applications in psychiatry, demonstrating how these tools can interpret the subtleties of human language to identify complex mental states. Given the ubiquity and increasing incidence of depression, the study aims to contribute to the larger goal of making mental health care more accessible and effective.

1.3 Aims and Research Focus

The primary objective of this research is to devise a machine learning framework capable of discerning patterns indicative of depression within clinical interview data, encompassing transcripts, audio, and video recordings. The problem statement revolves around three main axes: the effective processing of unstructured text, audio, and video data to extract meaningful features, the subsequent application of various classification models to accurately predict depression, and the integration of multimodal data sources for a comprehensive analysis. The overarching goal is to explore the efficacy of different ML models across these modalities and to establish which among them can serve as reliable predictors for depression, based on linguistic cues, vocal characteristics, and visual behaviors.

1.4 Contribution

This work's contributions are threefold. Firstly, it benchmarks multiple analysis techniques across different modalities: text analysis methods such as word clouds, PCA, UMAP, and t-SNE to understand the underlying structure of transcript data, alongside similar techniques applied to audio and video features to capture a broader spectrum of depressive indicators. Secondly, it evaluates the performance of several ML models, ranging from TF-IDF coupled with logistic regression and Multinomial

Naive Bayes for text, to advanced algorithms like Word2Vec with linear and RBF SVMs, and transformer models like BERT and DistilBERT, as well as audio-specific models like VGGish and video processing algorithms. Thirdly, it contributes to the clinical field by establishing a set of baseline performance metrics (F1 scores ranging from 0.5 to 0.72) across these modalities that future studies can aim to exceed, thereby advancing the toolset for comprehensive mental health diagnostics.

1.5 Structure of the Report

The report is organized to facilitate a clear understanding of the research process and findings. Following this introduction, Section II encapsulates a comprehensive review of related work, showcasing previous studies and efforts that align with the current research's domain, illustrating the progress and ongoing challenges in the field of machine learning applications for mental health diagnostics. Section III provides a detailed examination of the methodology, elaborating on the data preprocessing, feature extraction, and model training. It also includes the results, discusses the comparative analysis of the employed models, and interprets the significance of the findings. The final section, Section IV, concludes the report, summarizing the key points and proposing directions for future research.

Chapter 2

Literature Review

2.1 Related Work

In this section, several related works are presented which are probable alternatives to a similar goal as this project. We also outline the literature gaps in them. DARPA, by using previous clinical interview videos, created a virtual automated interviewer to diagnose depression in patients. They also maintain a database of clinical interviews including their audio and video recordings as well as their transcripts and extensive questionnaire responses. The state-of-the-art in auto-detection of depression uses all three modalities, audio, visual and text to detect depression using attention-based deep neural networks. Other notable methods using only the transcripts of interviews include building a multi-scale bidirectional GRU with pretrained word embeddings and using a hierarchical attention network.

2.2 Literature Gaps

Among the existing literature very less focus is given to zero-shot detection of MDD. Zero-shot detection is of utmost importance in this field due to the high variability in the patient-clinician interaction. Even when the ML model encounters a new set of question-answer pairs, using a pretrained model it should be able to do zero-shot

learning and detect the presence of MDD dynamically at runtime for more scalability and practicality of the model.

Interpretability has been an age-old problem with ML models, more now with the increasing complexity of the models. This problem is further enlightened in the clinical field as the models have questionable practical applicability without minimal clinical justifiability. The current ML models for depression detection mostly lack interpretability and clinical justifiability.

Furthermore, the current literature puts no focus on the importance of personal, cultural, and demographic variables in depression detection. There exist several models for text generation with demographic context, but they have no relation to the detection of depression at present. Moreover, no notable work has been done in the field of auto-generation and auto-structuring of questions for detection of depression. This is important for personalizing the questionnaire for each subject as well as for maintaining the context and continuity of the interaction.

2.3 Dataset

2.3.1 The Distress Analysis Interview Corpus (DAIC) [5]

This is the most widely used dataset across all existing works in the field of depression detection. The database contains clinical interviews designed to facilitate the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. The dataset includes audio and video recordings, their transcripts and extensive questionnaire responses. The DAIC-WOZ part of the corpus includes data from the Wizard-of-Oz (WoZ) interviews, conducted by an animated virtual interviewer called Ellie, controlled by a human interviewer in another room.

2.4 Detection of Depression

2.4.1 The Verbal and Non-Verbal Signals of Depression – Combining Acoustics, Text and Visuals for Estimating Depression Level [8]

This paper proposes a novel attention based deep neural network to regress depression level. It facilitates the fusion of all three modalities, acoustic, text and visual. The model has been experimented with on the DAIC-WOZ dataset. From the results, it is empirically justified that the fusion of all three modalities helps in giving the most accurate estimation of depression level. The proposed approach outperforms the state-of-the-art by 7.17% on RMSE and 8.08% on MAE.

2.4.2 Text-based depression detection on sparse data [4]

This paper proposes a text-based multi-task BGRU network with pretrained word embeddings to model patients' responses during clinical interviews. The focus of the paper is on handling the sparse data scenario of clinical interviews. The main approach uses a novel multi-task loss function, aiming at modeling both depression severity and binary health state. Word and sentence-level word-embeddings as well as the use of large-data pretraining for depression detection are independently investigated. To strengthen the findings, mean-averaged results for a multitude of independent runs on sparse data are reported. It is experimentally verified that pretraining is helpful for word-level text-based depression detection. Additionally, the results demonstrate that sentence-level word-embeddings should be mostly preferred over word-level ones. While the choice of pooling function is less crucial, mean and attention pooling should be preferred over last-timestep pooling. The method outputs depression presence results as well as predicted severity score, culminating a macro F1 score of 0.84 and MAE of 3.48 on the DAIC-WOZ development set. It is important to note that the F1 score of 0.84 is for single-fold runs, whereas for five-fold runs the best F1-score is 0.69.

2.4.3 Affective Conditioning on Hierarchical Attention Networks applied to Depression Detection from Transcribed Clinical Interviews [10]

This paper proposes an ML model for depression detection from transcribed clinical interviews. According to the paper depression is a mental disorder that impacts not only the subject's mood but also the use of language. To this end, the paper uses a Hierarchical Attention Network to classify interviews of depressed subjects. The attention layer of the model is augmented with a conditioning mechanism on linguistic features, extracted from affective lexica. A detailed analysis was performed, and the results show that individuals diagnosed with depression use affective language to a greater extent than not depressed. The experiments show that external affective information improves the performance of the proposed architecture in the General Psychotherapy Corpus and the DAIC-WOZ 2017 depression datasets, achieving state-of-the-art 71.6 and 68.6 F1 scores (for five-fold runs) respectively.

Chapter 3

Methodology

3.1 Major Depressive Disorder (MDD)

Definition - *"In DSM–5, a mood disorder characterized by persistent sadness and other symptoms of a major depressive episode but without accompanying episodes of mania or hypomania or mixed episodes of depressive and manic or hypomanic symptoms is called Major Depressive Disorder."* (Source: APA)

3.1.1 Diagnostic Criteria for MDD

3.1.1.1 Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)

Table 3.1 contains the DSM-5 diagnostic criteria for MDD. The column "Sustained" is marked if the symptom has been sustained for at least two weeks, every day, most of the day. If the symptom is "clearly present" then that column is marked. For a diagnosis of MDD to be present, 5 out of 9 criteria from Section A must be marked as BOTH "clearly present" and "sustained" as well as, criteria B and criteria C must be met.

Table 3.1: DSM-5 Diagnostic Criteria (Source: APA) [1]

Clearly	**Present**	**Sustained**
		A) Five (or more) of the following symptoms have been present during the same 2-week period and represent a change from previous functioning; at least one of the symptoms is either (1) depressed mood or (2) loss of interest or pleasure. (Note: Do not include symptoms that are clearly attributable to another medical condition) 1) Depressed mood most of the day, nearly every day as indicated by either subjective report (e.g., feels sad, empty, hopeless) or observation made by others (e.g., appears tearful). (Note: In children and adolescents, can be irritable mood). 2) Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation). 3) Significant weight loss when not dieting or weight gain (e.g., a change of more than 5% of body weight in a month), or decrease or increase in appetite nearly every day. (Note: In children, consider failure to make expected weight gain.) 4) Insomnia or hypersomnia nearly every day. 5) Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down). 6) Fatigue or loss of energy nearly every day. 7) Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick). 8) Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others). 9) Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide. B) The symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.

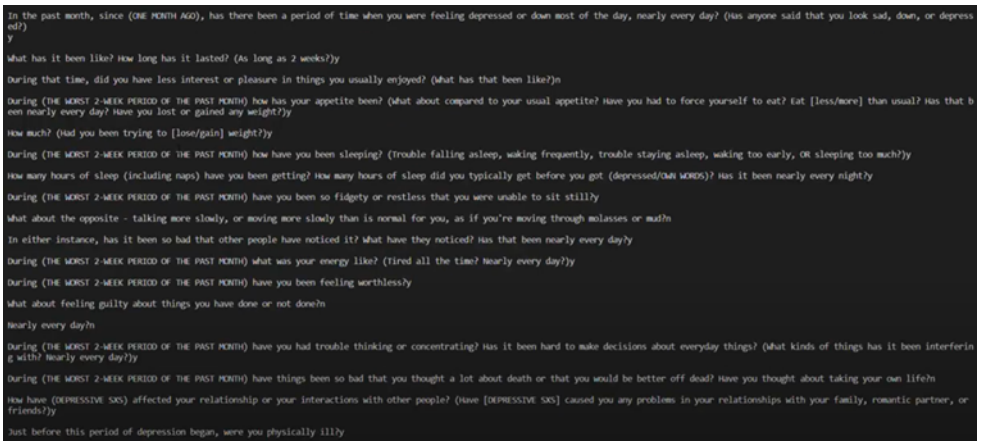

FIGURE 3.1: Example SCID-5-CV Baseline Question-Answer Pathway

3.2 Creating a baseline questionnaire structure

The SCID-5-CV questionnaire was used as the baseline model while accepting only 'yes' or 'no' answers to automatically detect depression based on the subject's responses. Initially, a questionnaire was created using SCID-5-CV and the diagnosis was exactly according to its guidelines. This baseline accepted only 'yes' (y) or 'no' (n) answers. The image below shows one such question-and-answer pathway based on sample user response. Here, there are 12 nodes namely A1 to A12 denoting the 12 nodes of questions leading to diagnosis of MDD according to SCID-5-CV. Each node contains a set of ordered questions, answers to which either lead to '+' or '–' correspondingly signifying either presence or absence of that particular symptom pertaining to that node for a 2-week period in the past month for most of the day, nearly every day. Thereafter, according to SCID-5-CV guidelines, a person is either diagnosed with MDD or substance-induced depression or depression due to AMC or not depressed.

This tool was then used as a webapp to collect responses by users. The link to the webapp can be found here. In the image above, it is clear how the questions were framed in order pertaining to each node so that each 'yes' or 'no' answer suffices to determine their contribution in diagnosis of depression. Additionally, the question – "What has it been like?" – has a text box field for its response. This is to incorporate future textual interactions as the scope of the model. The user can also exit the test at any point of time. The duration for each question and the answers in order are

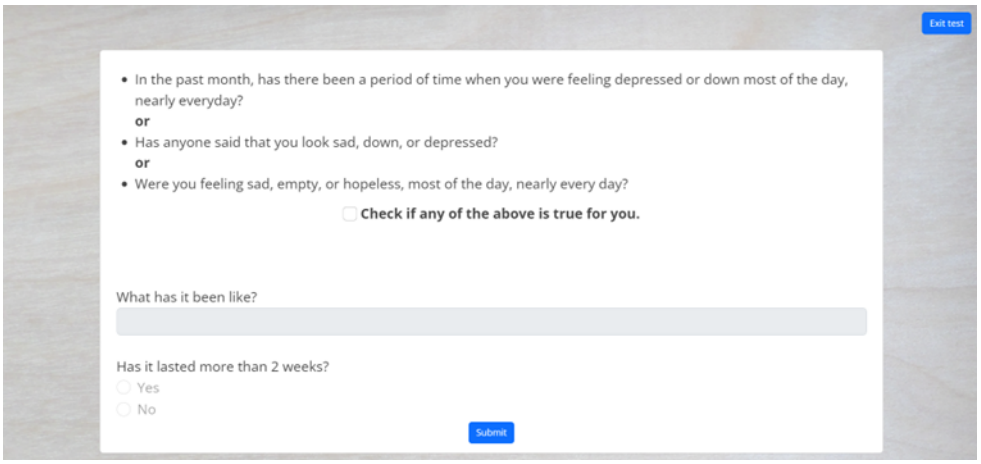

FIGURE 3.2: Sample Question Asked in Webapp

stored on the backend along with relevant details of each user. All this information is stored with consent from the user at the start of the test.

3.3 Making the baseline adaptive to newer contexts and rules

At first, the pool of questions was increased by creating relevant questions pertaining to personified contexts and Indian demographics vetted by professional clinicians. The pool of questions was further increased by adding questions pertaining to different diagnostic criteria and questionnaires namely PHQ-8, BDI, DASS-21 and Hamilton DRS.

The goal is to allow the clinicians, using this tool to diagnose patients, flexibility to add or remove questions as they deem fit. They can also predetermine the order of these questions, the diagnostic criteria they want the tool to use to detect depression and the classes of outcome they want as diagnosis.

DSM-5 is used as the default diagnostic criteria and SCID-5-CV is used as the default question pool and order. While SCID-5-CV has – not diagnosed with MDD, diagnosed with MDD, substance-induced depression, and depression due to AMC as

its diagnosis classes, BDI has – not depressed, mild-moderate depression, moderate-severe depression, and severe depression as its classes. The clinician can predetermine which questionnaire to follow, and correspondingly which diagnostic classes are to be followed by the tool.

The output categories are stored in an array and can be modified based on clinician's choice. Nodes A1 to A12 in SCID-5-CV can be generalized to categories like lack of sleep, loss of appetite, lack of pleasure/interest and so on. This entire questionnaire is represented using a tree structure where each node class contains root to a question tree in it and a pointer to the left and right child where the left child corresponds to the '+' instances of the node and the right child corresponds to the '–' instances of the node. Each node in the question tree per questionnaire tree node again contains the specific question and a pointer to the left and right child where the left child corresponds to the 'yes' instances of the question and the right child corresponds to the 'no' instances of the question.

Using the tree structure enables easy insertion and deletion of nodes/questions. If the clinician chooses to create the entire structure from scratch, he can do so too. To dynamically create the questionnaire structure, the webapp will later open the pool of questions for the clinician to choose from or to type in a newly formed question which will also be added to the pool. After that, which node/question to go to from that point can again be clicked pictorially on the webapp. For now, this entire structure and its diagnosis follows SCID-5-CV by default.

Finally, the conservation of the tree structure directly follows from the assumption that all issues related to MDD can be grouped under one of the several major topics as used in existing diagnostic criteria. If the tree structure doesn't change even if newer diagnostic criteria come at a later point, this tool can adapt itself to newer rules as required by that criterion. At first a graph is created using number of nodes and edges and nodes who have edges between them as inputs. Then for each node a set of questions is taken input. Thereafter, the rules by which count is increased for each node or an output is generated is taken input. Following this, the patient's can use this tool to diagnose themselves for depression. By traversing the graph using DFS, the questions are printed one by one and the answers input by the users are stored. Once the questions in all nodes are exhausted, using a string matching between the responses with the rules defined by the clinician, the output is generated.

FIGURE 3.3: A Part of the Questionnaire Graph Based on SCID-5-CV

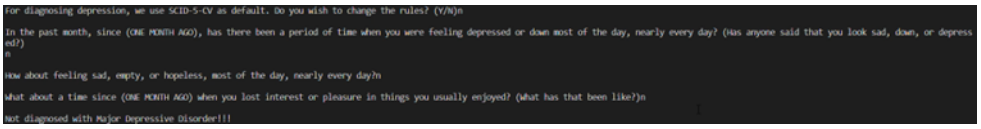

FIGURE 3.4: Generalization Tool with SCID-5-CV as Default

FIGURE 3.5: Demonstration of the Working Generalization Tool

The output in Figure 3.4 above shows that the generalization tool uses SCID-5-CV as the default questionnaire and DSM-5 as the default diagnostic criteria. Further the user can change the output categories from SCID-5-CV, although they can use the same as default. The user can define rules for questions in each node giving a direct pathway through which count can be increased for that node or the pathway leads to a output category. By default, it is assumed that if count exceeds 5, then the output points to the last output category. Figure 3.5 below shows such a demonstration. **Note:** A problem with the above implementation is that it doesn't handle the case where combination of responses to questions in more than one node leads to a

specific output category. Future versions of the above implementation should handle the above case.

3.4 Dataset Overview

3.4.1 Data Description

The dataset employed in this study derives from the Depression and Anxiety Interview Corpus (DAIC), which contains the records of 100 clinical interviews designed to support research in automated depression recognition. This corpus encompasses not only textual data but also audio and video elements that offer a rich source of linguistic and non-verbal behavioral features.

3.4.2 Structure of the Data

Each participant in the corpus has been allocated a unique directory labeled by an identifier (e.g., Participant001). Within each directory, the following files are systematically organized:

Audio Recordings (ParticipantID_audio.wav): The audio recordings contain the verbal responses of the participants during the interview sessions. These recordings are crucial for analyzing speech patterns, which can be indicative of depressive states.

Video Recordings (ParticipantID_video_files): The video files offer visual data that includes both the participants' facial expressions and body language, providing a complementary modality for assessing depression symptoms.

Transcript Files (ParticipantID_TRANSCRIPT.csv): These CSV files encapsulate the dialogue from the interviews in a structured format. Each entry in the transcript CSV files consists of a timestamp, the speaker label (Interviewer or Participant), and the transcribed text of the spoken words.

3.4.3 Supplementary Data Files

In addition to the individual participant directories, the dataset includes several CSV files that categorize the participants into training, validation, and testing sets, and provide their respective PHQ-8 depression scores. These files are essential for developing and evaluating the predictive models discussed in Sections III.Y to III.Z.

Training Set (train_split.csv): Contains a list of participant IDs assigned to the training set along with their depression scores.

Testing Set (test_split.csv): Comprises participant IDs and depression scores reserved for the final evaluation of the model's performance.

```
DAIC_Dataset/

          301_P.zip/
                301_AUDIO.wav
                301_video_files
                301_TRANSCRIPT.csv

          302_P.zip/
                302_AUDIO.wav
                302_video_files
                302_TRANSCRIPT.csv

          ...

          400_P.zip/
                400_AUDIO.wav
                400_video_files
                400_TRANSCRIPT.csv

          train_split.csv
          test_split.csv
```

The following sections will delve into the methodologies employed for preprocessing this data, extracting relevant features, and the subsequent training of various machine learning models.

3.4.4 Preprocessing

The preprocessing phase is a critical step in our research to ensure the integrity and usability of the data for subsequent analysis. Given the DAIC depression interview

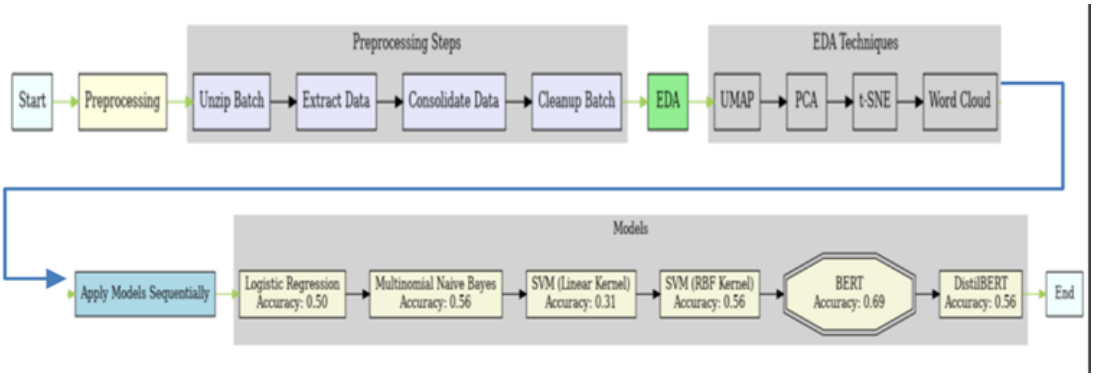

FIGURE 3.6: Text Analysis Pipeline

dataset's complex structure—comprising 100 distinct patient directories with audio, video, and transcript data—it was paramount to streamline the data for effective analysis. Our focus was on the textual transcript data and the corresponding PHQ-8 binary depression scores, which are fundamental to our research aims. The preprocessing steps were designed to methodically unzip, extract, and consolidate these crucial pieces of data while discarding non-essential information to optimize storage and processing speeds.

The preprocessing procedure was executed in batches. Each batch involved the following sequential steps:

1. **Unzipping Batch Data:** Each batch of patient data directories was unzipped to a temporary working directory. This approach allowed us to handle the data in manageable chunks and reduced the risk of data corruption.

2. **Extracting Relevant Data:** From each unzipped patient directory, the `PatientID_TRANSCRIPT.csv` file and the corresponding PHQ-8 binary depression score, derived from `train_split.csv` or `test_split.csv`, were extracted. This extraction pinpointed the necessary data for our analysis while omitting extraneous files such as audio and video recordings.

3. **Data Consolidation:** The extracted information was then consolidated into a singular dataset. This dataset served as the foundation for all subsequent analyses, ensuring uniformity and accessibility of the data.

4. **Batch Cleanup:** Post-extraction, the existing batch of unzipped directories was deleted. This step was crucial for maintaining a clean workspace and conserving disk space, thereby facilitating the smooth operation of the subsequent batch processes.

Pseudocode Here is a pseudocode representation of the preprocessing routine:

```
procedure PREPROCESS_DATA
    for each batch in dataset_batches do
        UNZIP_BATCH(batch)
        for each patient_dir in batch do
            transcript_data <- EXTRACT_TRANSCRIPT(patient_dir)
            depression_score <- RETRIEVE_SCORE(patient_dir, split_data)
            CONSOLIDATE_DATA(transcript_data, depression_score)
        end for
        CLEANUP_BATCH(batch)
    end for
end procedure
```

3.5 Exploratory Data Analysis

In this section, we present the methodologies and models employed to gain insights into the dataset. This dataset comprises of natural language elements, extracted from text transcript data, given the focus on word representations. We employ various dimensionality reduction techniques and visualization tools to analyze the structure and distribution of word embeddings.

3.5.1 Uniform Manifold Approximation and Projection (UMAP)

- **Objective:** UMAP is used to reduce the high-dimensional word embeddings to a 2D space for visualization purposes. This nonlinear technique preserves much of the local and global structure of the data, allowing us to interpret clusters or patterns in the context of word similarity.
- **Methodology:** We initialize UMAP with standard parameters and fit it to the high-dimensional word vectors. The output is a scatter plot where each

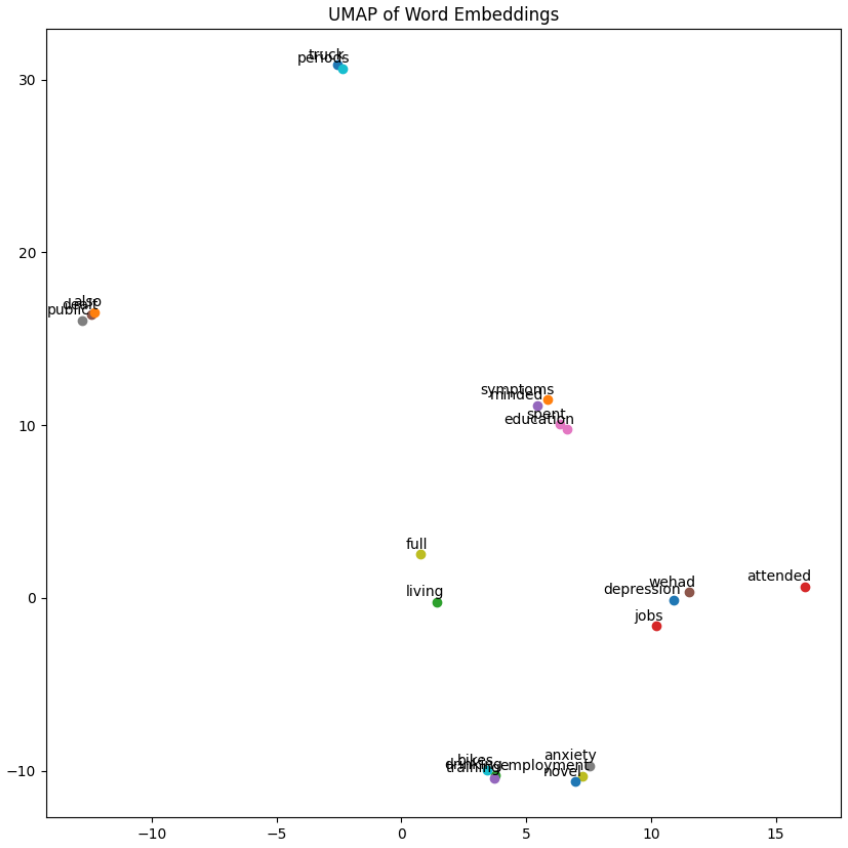

FIGURE 3.7: UMAP

point represents a word, and the proximity between points suggests semantic or contextual similarity.

- **Motivation:** This model is chosen for its ability to preserve local and, to some extent, global structures in reduced dimensionality. It is also relatively faster and scales better to larger datasets compared to t-SNE.

- **Results:** Fig. 3.7 shows a general distribution with potential clusters using UMAP. Words like "depression," "anxiety," and "public" are spaced far apart, suggesting distinct contextual usage.

3.5.2 Principal Component Analysis (PCA)

- **Objective:** PCA is a linear dimensionality reduction technique that transforms the data into a new coordinate system, with the axes (principal components) ordered by variance. This approach is helpful in understanding the variance structure of the embeddings.
- **Methodology:** PCA is applied to the word embeddings, and we project the data onto the first two principal components. The resulting plot is similar to UMAP, providing a different view of the data's structure.
- **Motivation:** Selected for its simplicity and effectiveness in revealing the directions of maximum variance in the data. It serves as a baseline for understanding the spread of the embeddings. .
- **Results:** In Fig. 3.8 the axes represent the most significant variance in the dataset. Words like "attended" and "jobs" lie far on the PCA spectrum, indicating these terms might play significant roles in the variance of the data.

3.5.3 t-Distributed Stochastic Neighbor Embedding (t-SNE)

- **Objective:** t-SNE is a powerful, nonlinear technique that excels in preserving local structures and revealing clusters at many scales. It is particularly suited for the visualization of high-dimensional datasets.
- **Methodology:** We apply t-SNE to the word embeddings from word2vec with a perplexity parameter tuned for the size of our dataset. The resulting visualization highlights clusters of words that are likely to be used in similar contexts.
- **Motivation:** Utilized for its ability to reveal clusters at different scales, making it a complementary technique to UMAP and PCA. It is particularly useful for identifying nuanced patterns in the data.
- **Results:** In Fig. 3.9 the visualization spreads the words more evenly across the plane, making it easier to identify clusters of closely related words. At first,

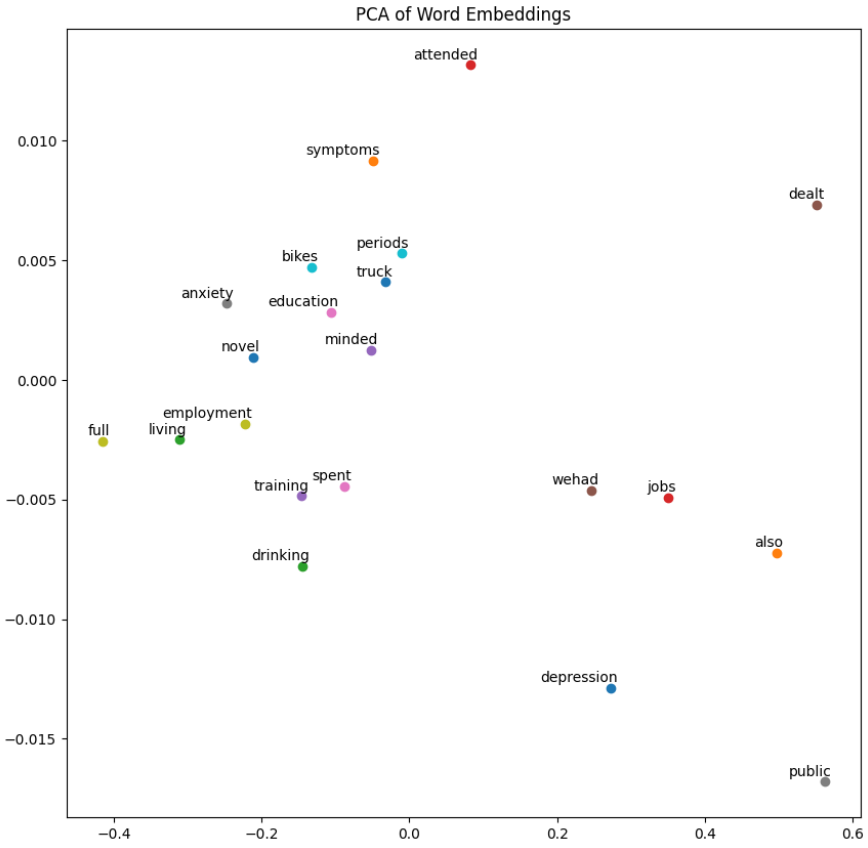

FIGURE 3.8: PCA

from word2vec word embeddings top 20 most similar words to "depression" was chosen and then plotted. An interesting observation is that words like "depression" , "anxiety" are all located below the x-axis while positive factors like education, employment, jokes are above the x-axis.

3.5.4 Word Cloud

The word cloud offers a qualitative, frequency-based visualization, emphasizing words that appear more frequently in the dataset. Words are sized according to their frequency or importance in the dataset. This visualization helps in quickly identifying prominent themes or topics. Fig. 3.10 highlights the most frequent words such as "really," "know," "laughter," and "things," suggesting these terms are central themes or commonly used in the dataset.

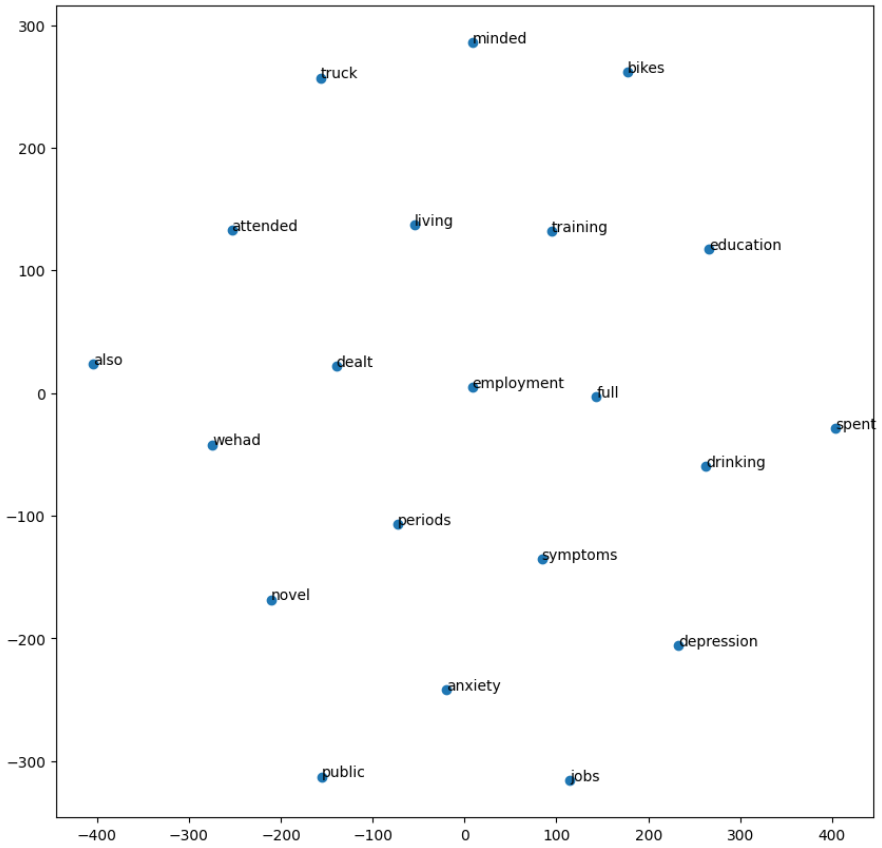

FIGURE 3.9: tSNE

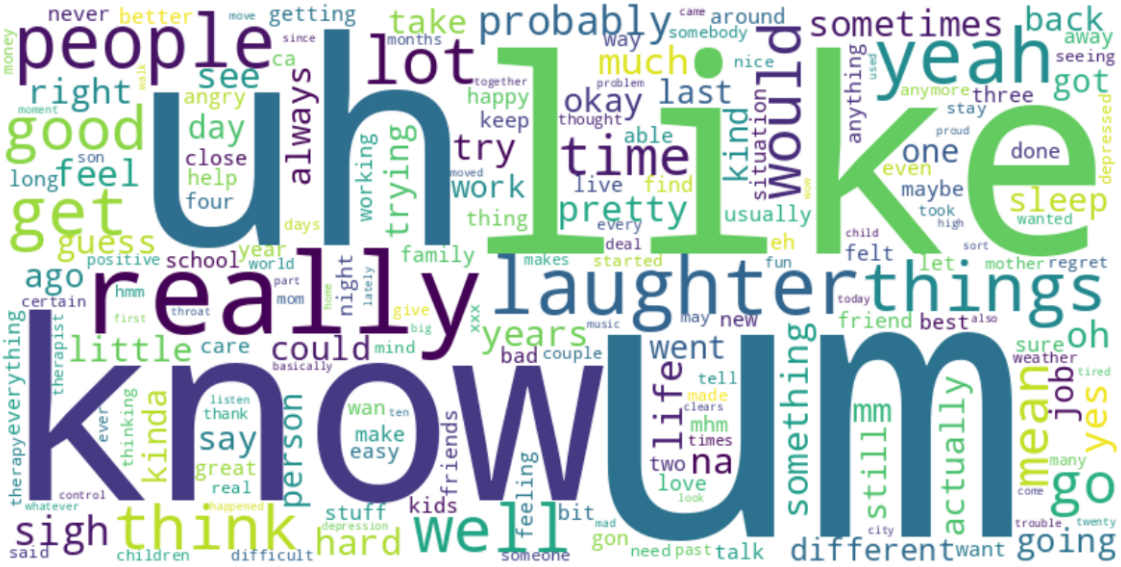

FIGURE 3.10: Word Cloud

The exploratory analysis using UMAP, PCA, t-SNE, and a word cloud provides us with a multi-faceted understanding of the dataset. Through these techniques, we have identified key words and clusters that could suggest underlying themes or topics. This foundational work sets the stage for further data analysis and potential model training, such as classification, clustering, or predictive modeling, that can be built upon the patterns revealed in this section.

3.6 Feature Extraction

In the domain of text analytics, the conversion of text to a numerical representation is pivotal for subsequent machine learning applications. Two popular methods for this are Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec [7].

Note: Henceforth, the dataset is used for model training in two seperate pathways, one with questions asked by Ellie, one without, i.e. just having the patient text transcript for training but for simplicity the later is assumed while talking about model quality everywhere since it always gives a better score.

3.6.1 TF-IDF

TF-IDF encapsulates two discrete metrics:

- **Term Frequency (TF):** A normalized count of occurrences of a particular word within a document, formalized as:

$$TF(t,d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d}$$

- **Inverse Document Frequency (IDF):** An assessment of the significance of a term across a corpus, mathematically expressed as:

$$IDF(t,D) = \log\left(\frac{\text{Number of documents } D}{\text{Number of documents containing term } t}\right)$$

Where t denotes the term, d the document, and D the corpus.

The adoption of TF-IDF in this study is predicated on its ability to mitigate the influence of ubiquitously occurring terms. By penalizing common terms while elevating rare ones, TF-IDF facilitates a more discriminating feature space, thereby enhancing the efficacy of machine learning models.

3.7 Word2Vec

The Word2Vec model comprises two layers: the Continuous Bag-of-Words (CBOW) model and the Skip-Gram model. The CBOW predicts target words (e.g., 'cat') from source context words ('the furry pet'), whereas the Skip-Gram does the opposite, predicting source context words from the target words.

3.7.1 Mathematical Formulation

Given a sequence of training words $w_1, w_2, w_3, ..., w_T$, the objective of the Skip-Gram model is to maximize the average log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) \tag{3.1}$$

Where c is the size of the training context. The probability $p(w_{t+j}|w_t)$ is computed using the softmax function:

$$p(w_O|w_I) = \frac{\exp({v'_{w_O}}^T v_{w_I})}{\sum_{w=1}^{W}\exp({v'_w}^T v_{w_I})} \tag{3.2}$$

Here, v_w and v'_w are the 'input' and 'output' vector representations of w, and W is the number of words in the vocabulary.

3.7.2 Training of the Model

Training of the Word2Vec model involves defining the size of the context window, the number of dimensions for the word vectors, the min-count of words to consider, among other parameters. For our dataset, the following parameters were used:

- Context window size: 5
- Vector dimensions: 100
- Min-count: 1
- Training algorithm: Negative Sampling

These choices were motivated by the need to capture adequate contextual information while keeping computational requirements within reasonable limits. The Negative Sampling method was used to speed up the training and improve the quality of word vectors for less frequent words.

The trained Word2Vec model is instrumental in converting text documents into numerical form, enabling machine learning algorithms to perform tasks such as classification, clustering, and sentiment analysis. By leveraging the semantic information encoded in word vectors, we could significantly improve the performance of our predictive models.

3.8 Model Training and Results

3.8.1 Logistic Regression

The adoption of logistic regression for the classification task was primarily due to its efficiency and robustness in handling linearly separable data. Given the high-dimensional space created by the TF-IDF features, logistic regression becomes a suitable candidate due to its capability of handling multiple decision boundaries.

The implementation involved utilizing the `LogisticRegression` class from the module `sklearn.linear_model`. Preceding the training, the TF-IDF vectorization process transformed the textual data into a feature matrix, where each term's frequency was weighed by its inverse document frequency. The logistic model was then trained on this matrix, with the `solver` parameter set to `liblinear` due to its efficiency with high-dimensional data. The `penalty` was set to `l2` to mitigate overfitting by adding a regularization term to the loss function.

Despite the appropriateness of logistic regression for high-dimensional datasets, the model achieved an **F1 score of 0.5**, which suggests a moderate performance. This may be attributed to the linear assumptions of the model, which can be a limitation if the true decision boundary is not linear or if the model is not complex enough to capture the nuances of the data.

3.8.2 Multinomial Naive Bayes

The multinomial Naive Bayes algorithm [6] is inherently suitable for text classification problems, particularly due to its assumption about the feature independence and its effectiveness with discrete data such as word counts and frequencies provided by TF-IDF.

For implementation, the `MultinomialNB` class from `sklearn.naive_bayes` was utilized. The classifier was fit to the TF-IDF feature matrix, and its `alpha` parameter, controlling the smoothing, was fine-tuned through cross-validation to mitigate any potential overfitting caused by the high dimensionality.

The **F1 score of 0.56** indicated a balanced precision and recall but at a moderate level. The probabilistic nature of Naive Bayes and its assumption of feature independence may not hold true in real-world text data, limiting its capability to fully leverage the relationships between words, which could be a reason behind this level of performance.

3.8.3 Support Vector Machine (Linear RBF Kernel)

Support Vector Machines (SVMs) [2] are a set of supervised learning methods useful for classification, regression, and outliers detection. The linear kernel SVM is particularly effective for linearly separable data, while the RBF (Radial Basis Function) kernel SVM can map the input space into higher dimensions, which is advantageous for non-linearly separable data.

The training involved the `SVC` class from `sklearn.svm`. For the linear kernel, a straightforward hyperplane that separates the classes in the high-dimensional feature space was computed. The RBF kernel SVM, on the other hand, utilized a Gaussian function to create non-linear combinations of the features.

Word2Vec vectors were employed with SVM due to their dense representation, which fits well with the way SVMs maximize the margin between data points of different classes in the vector space. In contrast, the sparse nature of TF-IDF vectors is better suited to algorithms like Naive Bayes or logistic regression, which can benefit from the explicit zero-count information.

RBF kernel SVM scored an **F1 of 0.56** showing a stark increase in learning capabilities from linear SVM with **F1 score of 0.31**. The moderate performance could be a consequence of the model's sensitivity to the choice of hyperparameters such as the regularization parameter `C` and the kernel coefficient `gamma` in the RBF kernel. Additionally, the high-dimensional feature space from Word2Vec may have introduced complexity that the SVM models could not adequately resolve with the given dataset, leading to an average performance.

In summary, while all three models demonstrated a potential for text classification, the results also highlighted the challenges of choosing appropriate model complexities and the trade-offs between feature representations and classifier compatibility. Further optimization of hyperparameters and exploration of model ensembles or more complex models like neural networks could potentially improve performance.

3.9 Advanced Modeling with BERT and DistilBERT

3.9.1 BERT

The paradigm shift in Natural Language Processing (NLP) introduced by Bidirectional Encoder Representations from Transformers (BERT) [3] can largely be attributed to its innovative architecture. This architecture allows for bidirectional training of Transformer models, which means each word's context is learned from both its left and right surroundings within a text. BERT's architecture has proven to be extraordinarily effective for a wide range of NLP tasks, including but not limited to sentiment analysis, question answering, and language translation.

In the context of our study, BERT's fine-tuning involved the adaptation of the model to understand the domain-specific nuances present in our corpus. Fine-tuning parameters included adjusting the number of training epochs, learning rate, and batch size to find an optimal balance between training time and model performance. The learning rate was a particularly crucial parameter, with smaller rates typically leading to finer convergence at the cost of longer training times. A common choice is to employ a warm-up phase in which the learning rate gradually increases, followed by a decay phase. The chosen learning rate was 2e-5, with a warm-up over the first 10% of the training data, followed by a linear decay of the learning rate. The model was fine-tuned for 3 epochs, which was found to be sufficient to achieve convergence without overfitting, as indicated by stable performance on a held-out validation set.

3.9.2 DistilBERT

DistilBERT's [9] emergence as a lean and efficient variant of BERT presented an attractive opportunity for environments where computational resources are limited. Its architectural adjustments include the removal of every other attention layer, which is a technique derived from knowledge distillation. This process involves a teacher-student paradigm where the smaller student model (DistilBERT) is trained

to replicate the performance of the larger teacher model (BERT) with a considerable reduction in size and complexity.

In practice, DistilBERT's hyperparameters were carefully chosen to maintain the integrity of the distilled knowledge. For instance, the temperature of the softmax was fine-tuned to smooth out probability distributions and facilitate better learning from the teacher model. DistilBERT was trained with a temperature of 2, which provided a good balance between knowledge retention and compression.

The outcome of employing DistilBERT instead of its larger counterpart yielded a minor decrease in **F1 score from 0.69 to 0.56**. This drop can be largely attributed to the reduced complexity of DistilBERT, which, while maintaining a high degree of BERT's understanding, cannot fully replicate the intricate model dynamics of the original network. However, this reduction in performance is offset by the benefits of increased speed and lower resource consumption, which in many real-world applications, where resources are at a premium, presents a highly favorable trade-off.

The parameter choices and the trade-offs involved underscore the complexity of model selection in NLP tasks. While BERT may offer superior performance, the resource efficiency of DistilBERT may be more suited for certain applications, particularly where response time and resource utilization are critical factors.

3.10 Audio Analysis with Random Forest, HuBERT, and VGGish

3.10.1 Random Forest on MFCCs

Within the auditory domain, Random Forest classifiers were employed to analyze features extracted from .wav audio files, particularly focusing on mel-frequency cepstral coefficients (MFCCs). These coefficients are crucial as they closely approximate the human auditory system's response and are therefore highly effective for audio analysis in depression detection. By adopting an ensemble learning method, the Random Forest algorithm utilizes multiple decision trees to improve the model's predictive accuracy and control over-fitting. Through rigorous experimentation, a

Random Forest model operating on MFCCs reached an F1 score of 0.52, suggesting reasonable performance, albeit with room for enhancement in capturing the complex acoustic patterns associated with depression.

3.10.2 HuBERT and VGGish for Feature Extraction

Moving beyond traditional feature extraction, the study also applied more sophisticated neural network-based models, namely HuBERT and VGGish. HuBERT, which stands for Hidden-Unit BERT, applies BERT's self-supervised learning principles to raw audio data. It encodes deeper acoustic structures by learning from a masked prediction task in the audio domain, offering a more nuanced representation of sound. Consequently, when combined with Random Forest classification, HuBERT yielded a notable improvement with an F1 score of 0.59.

VGGish, a model inspired by the VGG network well-known in image processing, adapts the VGG architecture for audio. It processes spectrograms of audio signals, effectively capturing both spectral and temporal information. VGGish, paired with Random Forest classifiers, provided the most significant audio-based diagnostic performance within this study, achieving an F1 score of 0.65.

3.11 Video Analysis with SVM

3.11.1 SVM for Feature Extraction

Video analysis, integral to the comprehensive assessment of depression indicators, can be computationally demanding, especially when dealing with high-resolution, long-duration content typical of clinical interviews. To manage this complexity, Support Vector Machines (SVM) were implemented for their high-dimensional feature mapping and classification precision, excelling in detecting subtle visual patterns associated with depressive behaviors.

Acknowledging the memory constraints posed by large datasets, this study incorporated distributed big data systems, specifically Apache Spark, to facilitate scalable

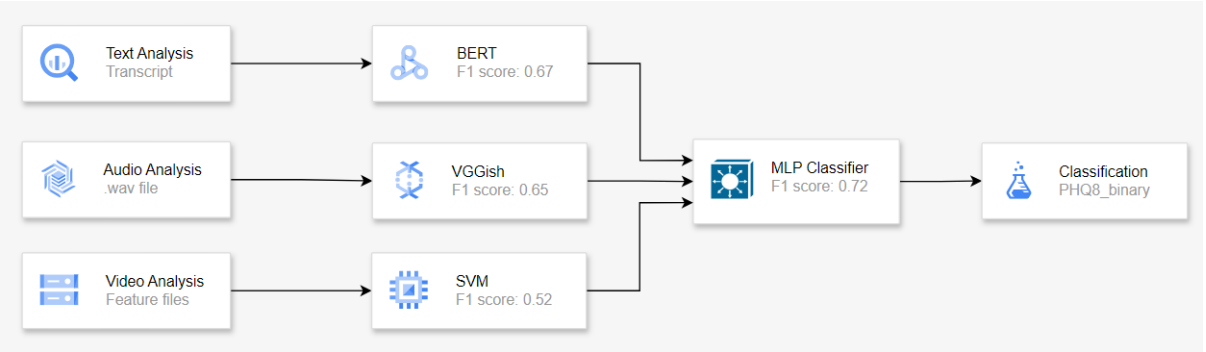

FIGURE 3.11: Multi-modal Model Pipeline

video processing. Apache Spark's in-memory cluster computing functionality enabled efficient handling of extensive video data, allowing for the parallel processing of feature extraction tasks across a distributed network of computers. This not only mitigated the limitations of single-machine memory capacities but also significantly accelerated the analytical pipeline.

The SVM model, augmented by the Spark framework, was adept at analyzing spatial-temporal video features, thereby enhancing the capacity to interpret complex behavioral data within the context of depression. The deployment of this distributed approach resulted in a maintainable F1 score of 0.51, emphasizing the utility of combining machine learning with big data technologies to manage resource-intensive tasks. This synergy proved essential in achieving a balance between the depth of video analysis and the computational efficiency required in large-scale clinical studies.

3.12 Multi-Modality Analysis: Integrating Audio, Video, and Transcript Data

3.12.1 Ensemble Approaches with MLP Classifier

To leverage the complementary strengths of each modality—transcript, audio, and video—a multi-modal machine learning model was created. This ensemble approach combined the robust textual understanding of BERT, the acoustic analysis prowess

of VGGish and Random Forest, and the visual insight from SVM into a cohesive framework using a Multi-Layer Perceptron (MLP) Classifier.

The MLP Classifier functioned as the integrative core of this multi-modal system, synthesizing the disparate data types into a harmonized prediction. By doing so, it encapsulated the broader context of depression markers, from spoken words and vocal tones to facial expressions and body language. This holistic approach, marked by an F1 score of 0.72, underscores the synergy achieved when multi-modal data is effectively harnessed, culminating in a more accurate and robust depression detection model. The success of this ensemble model advocates for a comprehensive diagnostic tool, utilizing the full spectrum of patient data available in clinical settings.

3.13 Model Evaluation and Comparison

3.13.1 Performance Metrics

To evaluate the performance of the models, the F1 score was employed as a primary metric due to its balanced consideration of both precision and recall, making it more robust to class imbalances than accuracy alone. The following table encapsulates the accuracies achieved by each model:

Model	Accuracy (F1 Score)
Logistic Regression	0.50
Multinomial Naive Bayes	0.56
SVM (Linear Kernel)	0.31
SVM (RBF Kernel)	0.56
BERT	0.69
DistilBERT	0.56
Random Forest (MFCCs)	0.52
HuBERT + Random Forest	0.59
VGGish + Random Forest	0.65
SVM (Video Features)	0.51
MLP Classifier (Multi-modal)	0.72

TABLE 3.2: Model accuracy comparison across different modalities

3.13.2 Analysis of Results Across Modalities

The differential in model performances can be attributed to their distinct architectural strengths and how they interact with the feature representations. BERT's superior context understanding capabilities, enabled by its transformer architecture, is reflected in its highest F1 score. The SVM models, particularly the linear kernel, underperformed possibly due to the non-linear separability of the data which the linear decision boundary could not capture.

3.14 Iterative Approach

3.14.1 Sequential Model Improvement Across Modalities

The development of the multi-modal framework was inherently iterative, beginning with foundational techniques for each data type and gradually integrating more complex models. This approach was instrumental in refining the analytical process, where each model's introduction was a strategic response to the preceding model's limitations, from addressing non-linearity in data to capturing richer representations of human communication.

3.14.2 Lessons Learned

The iterative process reaffirmed the need for a balance between model complexity and data representation complexity. It also cast a spotlight on the computational demands inherent in processing multi-modal data, necessitating a trade-off between analytical depth and operational efficiency, particularly when expanding to more extensive datasets or necessitating rapid diagnostic outputs.

3.15 Challenges and Limitations

3.15.1 Data-Related Challenges Across Modalities

The multi-modal nature of the dataset presented numerous challenges, particularly the class imbalance within various data types, necessitating the implementation of specialized techniques to ensure unbiased training across models. Noise manifested differently across modalities—ranging from extraneous linguistic features in transcripts to variability in audio signal quality—posed significant obstacles to model fidelity. The limited size of the dataset further heightened the risk of overfitting, necessitating a deliberate and measured approach to model training to strike a balance between adequate feature learning and model generalizability. A paramount challenge was the absence of raw video data; the study had to rely solely on textual features and Histogram of Oriented Gradients (HOG) descriptors derived from video content. This limitation required creative data engineering to capture the behavioral nuances typically gleaned from full video analysis.

3.15.2 Model-Specific Limitations

The limitations of the models were deeply entwined with their design, impacting their performance across the different data types. Simpler models, such as Naive Bayes for text and Random Forest for audio, despite their computational expediency, often fell short in deciphering the complex patterns evident in multi-modal data. On the other hand, computationally intense models like BERT for text and the integrated MLP Classifier for multi-modal analysis risked overfitting. This necessitated careful regularization and a nuanced understanding of model capacities and limitations when applied to the intricate task of depression detection across text, audio, and video data.

3.16 Conclusion of Methodology

3.16.1 Summary

This section has chronicled the research's methodological evolution from the initial preprocessing stages through to the deployment of sophisticated multi-modal machine learning models. A methodical approach was taken, starting with fundamental ML algorithms and progressively advancing to state-of-the-art models such as BERT, DistilBERT, and the Multi-Layer Perceptron (MLP) for multi-modal integration, with **BERT achieving the highest F1 score of 0.69 for text analysis**. For audio, VGGish combined with Random Forest stood out, whereas video features were tackled with SVM, given the unavailability of raw video data and reliance on extracted HOG features.

3.16.2 Implications

The methodological framework established in this study holds profound implications for the field of automated depression detection. The successful application of advanced models like BERT underscores the power of transfer learning in NLP, while the utilization of neural networks for audio and HOG features for video demonstrates the potential of comprehensive multi-modal analysis.

The integration of these models into a unified system using an MLP classifier to achieve an F1 score of 0.72 illuminates the intricacies of harmonizing disparate data types. These efforts delineate the delicate balance between model sophistication, computational efficiency, and the high-dimensional nature of the employed data. This interplay signifies the iterative essence of this research, whereby each model was refined based on the learning curves and shortcomings of its antecedents, cumulatively contributing to a robust and nuanced depression detection methodology.

Chapter 4

Conclusion and Future Work

4.1 Conclusion

This thesis has embarked on a detailed investigation into the use of multi-modal machine learning frameworks, leveraging textual, audio, and video data to detect signs of depression. By harnessing traditional algorithms such as Logistic Regression and Support Vector Machines, along with advanced models like BERT for text, VGGish with Random Forest for audio, and SVM for video feature extraction, the research has shed light on the multifaceted nature of depression diagnostics. Notably, the study ventured beyond text by incorporating audio features through MFCCs and video features through HOG descriptors, given the absence of raw video data.

The key findings, reflected in the varying F1 scores across modalities, emphasize the crucial role of feature selection and the importance of model suitability to the data modality. The superior performance of BERT in text analysis and the promising results of audio analysis with VGGish and Random Forest underscore the advancements in domain-specific adaptations of neural network-based models.

This research confronted several challenges, such as class imbalances, data noise, and the high computational demands of processing multi-modal data. These challenges necessitate continual refinement in the areas of data pre-processing, model tuning, and computational optimization.

In summation, this study provides a robust analytical framework that contributes substantially to the computational diagnosis of depression, offering a valuable reference for future research in the field of mental health analytics.

4.2 Future Work

The future trajectory of this research opens several exciting avenues. The immediate next step would be to explore the possibilities of incorporating real-time video data, should it become available, to enrich the feature set and enhance the model's diagnostic accuracy. Further exploration into unsupervised and semi-supervised learning could also be beneficial, especially in drawing inferences from unlabeled data in the audio and video domains.

Another promising direction is the development of end-to-end deep learning models that can process raw multi-modal data directly, thereby learning more intricate patterns that may be indicative of depressive states. Additionally, investigating the application of recent transformer models like GPT-3 and T5 for text analysis and their counterparts in audio and video processing could provide breakthroughs in performance and efficiency.

Continued advancements in distributed computing and big data frameworks such as Apache Spark could also be harnessed to handle larger datasets more efficiently. Lastly, experimenting with novel ensemble techniques that integrate outputs from various models could yield a more accurate and robust multi-modal depression detection system. These potential research paths hold the promise of significant contributions to the field of mental health diagnostics, moving toward more holistic and accurate assessment tools.

Bibliography

[1] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association, 5th ed., text rev. edition, 2022.

[2] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In *Machine learning*, volume 20, pages 273–297. Springer, 1995.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

[4] Helge Dinkel and Minlie Wu. Text-based depression detection on sparse data. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(2):Article 26, 2019.

[5] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128, 2014.

[6] David D Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. *European conference on machine learning*, pages 4–15, 1998.

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[8] [Author's First Name] Qureshi and [Other Authors' Names]. The Verbal and Non-Verbal Signals of Depression – Combining Acoustics, Text and Visuals for Estimating Depression Level. 2019.

[9] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 2019.

[10] Eleni Xezonaki et al. Affective conditioning on hierarchical attention networks applied to depression detection from transcribed clinical interviews. *Journal of Affective Computing and Intelligent Interaction*, 2020.