ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crowned our effort with success. We are grateful to our institution, **National Institute of technology Srinagar** with its ideals and inspirations for having provided us with the facilities, which has made this project a success.

We wish to extend our profound thanks to **Dr. Mohammad Ahsaan Chishti, Head Department of Computer Science & Engineering at NIT SRI**, for giving us the consent to carry out this project.

We owe our sincere thanks to our Project Supervisor , **Dr. Ranjeet Kumar Rout Assistant professor, Department of Computer Science & Engineering, NIT SRINAGAR** for his immense help during the project and also for his valuable suggestions on the project.

We would like to thank all the teaching and non-teaching staff for their valuable advice and support.

We would like to express our sincere thanks to our parents and friends for their support.

ABSTRACT

Mental health has become a major issue in modern life, and depression is a common and distressing condition that affects millions of people worldwide. We explore the intersection of cognition and technology, focusing on using deep learning tools to understand trauma through multiple modalities (audio, text, and video).

We begin by delving into the challenges of depression and mental health in contemporary society The high prevalence of depression places a significant burden on individuals, families, and societies, affecting overall health, employment, and quality of life. Mental health stigma exacerbates the problem and hinders timely diagnosis and treatment. Recognizing the need for innovative solutions, we recognize the potential of deep learning techniques to transform the diagnosis of dementia. Deep learning, has demonstrated an incredible ability to analyse and interpret complex data. By harnessing the power of deep learning programs, we can find new ways to identify and intervene with dementia early.

We focus on the multimodal approach to depression detection, emphasizing the analysis of audio, text, and video recordings/ features. The utilization of audio data involves examining speech patterns, tone, and vocal characteristics, which may reveal subtle indicators of depressive symptoms. Textual data, can be analysed for linguistic markers and sentiment to aid in early detection. Video recordings provide an additional layer by capturing non-verbal cues, facial expressions, and body language that may offer valuable insights into an individual's mental state. This exploration at the intersection of mental health and deep learning underscores the promise of technology in advancing our understanding and management of mental health challenges in contemporary society.

Contents

INTRODUCTION

Mental health has emerged as a defining societal challenge in the 21st century, with depression standing out as a pervasive and debilitating condition affecting millions globally. The complex interplay of genetic, biological, environmental, and psychological factors renders depression a multifaceted phenomenon that transcends individual experiences, permeating communities, and straining healthcare systems. The impact of mental health conditions on the quality of life, productivity, and social dynamics necessitates a paradigm shift in our approach to identification, intervention, and support.

In recent years, technology has emerged as a transformative force with the potential to revolutionize the landscape of mental health care. The integration of artificial intelligence, specifically deep learning, presents an innovative avenue for addressing the challenges associated with the early detection of depression. As societal attitudes toward mental health evolve and the stigma surrounding these conditions gradually erodes, the time is ripe to explore comprehensive solutions that harness the capabilities of advanced technologies.

The Burden of Depression in Society

Depression's pervasive influence extends far beyond the individual, infiltrating families, workplaces, and communities. According to the World Health Organization (WHO), more than 264 million people worldwide suffer from depression, and the numbers are steadily rising. The consequences of untreated depression are far-reaching, impacting not only the mental and emotional well-being of individuals but also their physical health. Increased risk of chronic conditions, impaired cognitive function, and heightened vulnerability to other mental health disorders contribute to the intricate web of challenges posed by depression.

Moreover, the societal ramifications of depression are profound. The economic burden, stemming from decreased productivity and increased healthcare expenditures, places a strain on both developed and developing nations. The social fabric is also affected, as individuals grappling with depression often face stigma, discrimination, and barriers to seeking help. Understanding and addressing the multifaceted nature of depression requires a nuanced and comprehensive approach that goes beyond traditional diagnostic methods.

The Evolving Landscape of Mental Health Technology:

Against this backdrop, technology emerges as a catalyst for change, offering innovative solutions to long-standing challenges in mental health care. Artificial intelligence (AI), and more specifically, deep learning, has gained prominence for its ability to process vast amounts of data and discern intricate patterns that may elude human analysis. This technological prowess positions deep learning as a promising tool for enhancing our understanding of mental health conditions and improving the accuracy of diagnostic processes.

Deep Learning in Mental Health

Deep learning, a subset of machine learning inspired by the structure and function of the human brain's neural networks, has demonstrated exceptional capabilities in diverse fields, ranging from image and speech recognition to natural language processing. In the context of mental health, the application of deep learning holds the potential to unlock new insights into the early signs and patterns associated with depression.

This exploration is particularly salient in the realm of depression detection, where traditional methods often rely on subjective assessments and self-reporting. By leveraging the power of deep learning algorithms, we can move beyond the limitations of current diagnostic approaches, introducing a data-driven and objective dimension to the identification of depressive symptoms. The integration of AI in mental health care aligns with the broader trend of precision medicine, tailoring interventions to individual characteristics and needs.

The Multimodal Approach to Depression Detection

Recognizing the heterogeneity of depression and the myriad ways it manifests, a multimodal approach to detection becomes imperative. This approach involves the simultaneous analysis of various data sources, including audio, text, and video recordings, to capture a more comprehensive picture of an individual's mental state.

Analysing audio data involves examining speech patterns, tone, and vocal characteristics, as alterations in these aspects may serve as subtle indicators of depressive symptoms. Textual data, derived from sources such as social media, electronic health records, or personal writings, can be mined for linguistic markers and sentiment analysis to gain deeper insights into an individual's emotional well-being. The inclusion of video recordings introduces a visual component, capturing non-verbal cues, facial expressions, and body language that convey nuanced information about an individual's mental state.

The multimodal approach recognizes that depression is a complex, dynamic condition with diverse manifestations, and no single modality can fully encapsulate its complexity. By combining insights from different data sources, a more holistic understanding of an individual's mental health can be achieved, enhancing the accuracy and reliability of depression detection.

This exploration at the intersection of mental health and deep learning aims to navigate the evolving landscape of technology-enabled mental health care. As we delve deeper into the potential of deep learning for depression detection through audio, text, and video recordings, we embark on a journey that holds promise for transforming the way we understand, diagnose, and address mental health challenges in contemporary society. The subsequent sections of this study will delve into the specific applications of deep learning in each modality, examining the opportunities and challenges associated with this groundbreaking approach to mental health care.

Challenges in depression detection:

1. Heterogeneity of Depression Expression

- *Audio:* Depressed individuals may exhibit variations in vocal characteristics, and the heterogeneity in speech patterns among individuals with depression poses a challenge. Some may display hyperactivity, while others may present lethargy, making it difficult to establish universal audio features.
- *Text:* Expressions of depression in text can be highly subjective, context-dependent, and influenced by cultural nuances. Sarcasm, humor, or metaphorical language may mask underlying depressive sentiments.
- *Visual:* Facial expressions and body language associated with depression can vary widely. Individuals may adopt different coping mechanisms, impacting the consistency of visual cues.

2. Lack of Ground Truth Annotations

- *Audio:* Establishing ground truth for audio features often involves relying on self-reported depression scales. However, individuals may underreport their symptoms due to social desirability bias, impacting the accuracy of annotations.
- *Text:* Ground truth in text-based depression detection often relies on clinical diagnoses or self-reporting, introducing subjective elements. Mislabelling and misinterpretation of language nuances further complicate accurate annotation.
- *Visual:* Depression severity may not be uniformly evident in facial expressions or body language, leading to challenges in accurately annotating visual features. Objective ground truth labels are challenging to establish.

3. Data Privacy and Ethical Concerns

- *Audio:* Collecting audio data for depression detection raises privacy concerns. Capturing and storing individuals' voice recordings may be perceived as intrusive, requiring strict adherence to ethical guidelines and obtaining informed consent.
- *Text:* Analysing personal text data, especially from social media, raises ethical issues related to privacy and consent. Preserving user anonymity while ensuring data integrity is a constant challenge.
- *Visual:* Processing facial images for depression detection necessitates caution to prevent potential misuse or unauthorized access. Consent, anonymity, and secure storage are paramount in handling visual data.

4. Cross-Cultural and Multilingual Variability

- *Audio:* Vocal patterns and expressions associated with depression can vary across cultures and languages. Developing universal models that are sensitive to cultural nuances is a significant challenge.
- *Text:* Language-specific idioms, expressions, and linguistic nuances impact the generalization of depression detection models. Models trained on one language may not perform optimally across diverse linguistic landscapes.
- *Visual:* Facial expressions and body language may differ culturally, influencing the universality of visual features. Addressing cross-cultural variations is crucial for robust depression detection.

5. Interplay of Comorbidities

- *Audio:* Depression often coexists with other mental health conditions. Distinguishing depression-specific audio features from those associated with comorbidities is a complex task.
- *Text:* Individuals with depression may also have anxiety or other mental health issues, leading to overlapping linguistic patterns. Isolating depression-specific text features becomes challenging in such cases.

- *Visual:* Identifying visual cues solely linked to depression amidst the presence of comorbidities like anxiety or stress adds complexity to feature extraction.

6. Subjectivity in Ground Truth Annotations

- *Audio:* The interpretation of audio features linked to depression is subjective, involving judgments about tone, pitch, and pace. Variations in annotator interpretations may introduce inconsistency.
- *Text:* Clinical diagnoses or self-reports used for ground truth in text-based detection rely on subjective assessments. Lack of standardized criteria for labelling poses challenges in creating reliable datasets.
- *Visual:* Interpreting facial expressions and body language is inherently subjective. The absence of clear, objective criteria for annotating visual features introduces variability in labelled datasets.

7. Limited Availability of Diverse Datasets

- *Audio:* Access to diverse and representative audio datasets with sufficient examples of depression across demographics is limited. Imbalances in dataset composition may result in biased models.
- *Text:* Obtaining diverse and culturally representative text datasets for depression detection is challenging. Datasets often reflect the dominant language and cultural background, limiting model generalization.
- *Visual:* Datasets for visual depression detection may lack diversity in terms of age, ethnicity, and cultural backgrounds, impacting the model's ability to generalize across populations.

8. Complexity of Multimodal Fusion

- Combining audio, text, and visual features for comprehensive depression detection introduces challenges in fusion techniques. Determining the optimal way to integrate diverse modalities while avoiding information redundancy is non-trivial.

- Balancing the weights assigned to each modality and addressing potential biases in fusion models requires careful consideration. Inconsistencies in modalities may affect the overall robustness of multimodal depression detection systems.

9. Scarcity of Explainability

- *Audio:* Interpreting the significance of specific audio features in the context of depression may lack transparency. Models based on complex audio patterns may lack explainability.
- *Text:* Understanding the basis for text-based depression predictions, especially in deep learning models, can be challenging. Explainability is crucial for gaining insights into model decisions.
- *Visual:* The interpretability of visual features linked to depression may be challenging, particularly in deep learning models. Transparent models are essential for building trust in the reliability of visual depression detection.

Depression detection using audio, text, and visual features is a complex interdisciplinary endeavour, demanding a nuanced approach that addresses the intricacies within each modality while considering the multifaceted nature of depression. Overcoming these challenges requires collaborative efforts from researchers, clinicians, ethicists, and technologists to ensure the development of reliable, ethical, and culturally sensitive depression detection models.

LITERATURE REVIEW

A Systematic Review on the Application of Machine Learning Methods in Mental Health Detection in Online Social Networks (OSNs)

Examines the landscape of mental health detection within Online Social Networks (OSNs). Focusing on data sources, machine learning techniques, and feature extraction methods, the study meticulously analyses 22 articles published between 2007 and 2018, employing a rigorous screening process based on titles, abstracts, and full texts. The review underscores the significance of OSNs as viable data sources for mental health detection and advocates for the synergistic integration of traditional methods with OSNs analysis to enhance research outcomes. The primary data sources, including Twitter, Facebook, Sina Weibo, and microblogs, were scrutinized using machine learning and deep learning techniques, with Support Vector Machines emerging as the predominant choice. Feature extraction methods, such as N-Gram, Term Frequency-Inverse Document Frequency, and Linguistic Inquiry and Word Count, among others, are thoroughly explored for their relevance in mental health detection. Notably, the review identifies Support Vector Machines as the most frequently employed machine learning technique in the analysed studies.

Machine Learning-Based Depression Detection from Social Network Data: Insights from Islam et al. (2019)

The research conducted by Islam et al., as published in the journal Health Information Science and Systems (2019), offers a comprehensive exploration of the potential of machine learning techniques in detecting depression from social network data, specifically focusing on Facebook comments. Recognizing the growing prominence of social media as a means of communication, the authors

delve into the opportunity it presents for analysing user-generated content to infer mental health status. Addressing key research questions, the investigation seeks to define depression, identify indicative factors within social media content, extract these factors, and ascertain the most influential times for engaging with potentially depressed individuals online. Methodologically, the study involves the collection of public Facebook data, preprocessing through the Linguistic Inquire and Word Count (LIWC) software, and the application of supervised machine learning approaches, including Decision Tree, k-Nearest Neighbour (kNN), Support Vector Machine (SVM), and Ensemble classifiers.

Depression Detection using Emotion Artificial Intelligence

This paper delves into the realm of sentiment analysis, specifically focusing on the identification of tweets indicative of depression. Utilizing a dataset comprising 10,000 tweets obtained through the Twitter API, the study adopts an 80:20 split for training and testing, respectively. The training set incorporates a curated list of words suggesting depression tendencies, while the test set encompasses a mixture of neutral and negative tweets. The investigation employs two classification algorithms: the Naive Bayes Classifier and the Support Vector Machine (SVM). The Naive Bayes Classifier, based on Bayes' theorem and recognized for its speed and accuracy, specifically utilizes the Multinomial Naive Bayes for text classification. On the other hand, SVM, a versatile supervised learning algorithm capable of both linear and non-linear classification through kernel trick, is employed to categorize tweets into predefined classes.

Sense Mood: Depression Detection on Social Media

This paper introduces Sense Mood, an innovative system designed to efficiently detect and analyse potential users with depression on social media, with a specific focus on Twitter. Sense Mood leverages a deep visual-textual multimodal learning approach, integrating features extracted from both textual (tweets) and visual

(images) data posted by users. This involves the use of a Convolutional Neural Network (CNN)-based classifier for images and BERT for text, extracting deep features that are then combined to capture users' emotional expressions. The system employs a neural network for the classification of users into those with depression and normal users, generating an automatic analysis report. Addressing limitations in existing work, particularly the incomplete utilization of text and visual signals, and the challenge of integrating these representations, Sense Mood aims to provide an effective solution for detecting depression risk based on users' textual and visual tweets. Experimental results demonstrate that Sense Mood exhibits the potential to detect depression accurately and promptly in users on social media. This capability holds promise for identifying users at risk of depression and prompting them to take active preventive measures.

.

Multi-modal Depression Detection Based on Emotional Audio and Evaluation Text

This paper offers a comprehensive exploration into depression detection using a multi-modal approach, specifically analysing emotional audio features and evaluation text. The research focuses on identifying signs of depression through an extensive experimental process, incorporating feature analysis, and utilizing a carefully selected dataset. The study engaged native Chinese-speaking participants aged between 18 and 65, employing diagnostic criteria for mental disorders. A control group of non-depressed patients was included, and participants underwent conversations with physicians, along with assessments using the Hamilton Depression Rating Scale (HAMD). The experimental setup included the use of Kinect 2.0 and a Canon camera for recording, alongside a standalone sound card for audio recording in a controlled environment. Audio feature analysis delved into low-level features such as RMS energy, zero-crossing rate, voice probability, and MFCC. Correlation analysis with HAMD scores identified patterns associated with depression. Text feature analysis involved face-to-face conversations, with responses extracted as text data. Baidu voice recognition and manual correction

facilitated text feature extraction, employing Word2vec for vectorization and retaining word order information. The study explored a multi-modal fusion model to enhance depression recognition, revealing positive effects of fusing emotional changes. Results provided insights into differences across case and control groups, as well as various age groups and genders.

Deep Learning for Depression Detection of Twitter Users

This paper presents a noteworthy study focusing on the application of deep learning techniques for the detection of depression in social media posts, with a particular emphasis on Twitter. Addressing the challenges inherent in identifying mental illnesses through social media platforms, the research grapples with the scarcity of annotated training data required for supervised machine learning approaches. The primary objective of the study is to identify the most effective deep neural network architecture for this challenging task and to assess the performance of various models and word embeddings. The paper delineates the overall design of the approach, introducing an efficient neural network architecture that optimizes word embeddings. It elucidates the evaluation process, comparing the optimized embeddings produced by the proposed architecture with commonly used word embeddings. The methodology extends to a comparative evaluation of several deep learning architectures prevalent in natural language processing tasks, specifically tailored for detecting mental disorders. The study reports its findings based on the performance evaluation conducted on two publicly available datasets, CLPsych2015 and Bell Lets Talk. The comparison encompasses the performance of Convolutional Neural Network (CNN)-based models against Recurrent Neural Network (RNN)-based models.

Gender Bias in Depression Detection Using Audio Features

This paper delves into the critical issue of gender bias within the commonly utilized DAIC-WOZ dataset for depression detection research, emphasizing its potential adverse impact on machine learning model accuracy. The authors not only bring attention to the challenges of accurately detecting depression but also highlight the existence of biases within datasets, specifically focusing on gender bias, and propose methods to alleviate such biases. The paper commences by underscoring the complexities associated with precise depression detection and acknowledges potential biases in datasets that may skew classification performance. The authors draw attention to the notion of Fair Machine Learning, advocating for fair, unbiased classification to contribute towards a more equitable society through the application of machine learning.

The examination of the DAIC-WOZ dataset reveals previously unacknowledged gender bias, with a significant frequency difference in individuals with depression between genders. This revelation is expressed through the inequality $p(D \mid g = f) > p(D \mid g = m)$, where g represents gender, f denotes female, and m stands for male. Additionally, the authors identify a comparable class imbalance and gender bias within the validation set.

Investigation of Speech Landmark Patterns for Depression Detection

The authors extract speech landmarks using the SpeechMark toolbox and then calculate counts and durations of sequential landmark groups or n-grams. They find that durations of consecutive bigrams and onset-offset pairs are statistically significant in characterizing depression across two datasets. The authors also investigate the effectiveness of different feature sets for depression detection and find that the proposed count-based and duration-based features achieve state-of-the-art performance on both datasets. They further show that the landmark-based systems are complementary to the acoustic systems and that the count and duration

systems are complementary to each other. The paper also explores the task-specific analysis of landmarks and shows that tailored statistics for each elicitation task can improve depression detection. The authors conclude that duration-based features, especially for large percentiles, are useful and effective in detecting depression.

Hybrid CNN-SVM classifier for efficient depression detection system

The paper introduces a novel approach for the automatic detection of depression using a hybrid CNN-SVM model. The proposed system aims to address the increasing prevalence of depression by providing an alternative solution based on machine learning techniques. The study utilizes the DAIC-WOZ dataset, which includes audio recordings, video recordings, and psychiatric questionnaire responses in text format from 189 participants. The dataset was split into training, validation, and test sets, with a total of 122 subjects (38 depressed and 84 non-depressed) used for the experiments. The hybrid model combines the feature extraction capabilities of a Convolutional Neural Network (CNN) with the classification abilities of a Support Vector Machine (SVM). The CNN serves as a trainable feature extractor, while the SVM acts as the classifier, receiving the extracted features from the CNN as a feature vector for classification.

Design and Implementation of Attention Depression Detection Model Based on Multimodal Analysis

the authors propose a novel multimodal analysis-based attention depression detection model that utilizes both voice and text data for improved accuracy in depression detection. The model addresses the limitations of existing depression detection methods that rely solely on single data sources by integrating a fusion of text and voice data. The proposed model leverages the BERT-CNN model for natural language analysis and the CNN-BiLSTM model for voice signal processing, incorporating an attention mechanism to enhance the accuracy of depression detection. The authors present experimental results demonstrating the effectiveness

of the proposed model. They compare the classification results and accuracy of the validation data set for each model, highlighting the stability and improved performance achieved by the multimodal analysis-based attention depression detection model. The experimental results indicate that the proposed model exhibits enhanced accuracy and stabilized graphs, overcoming the instability and rapid loss increase associated with single data usage.

Furthermore, the paper discusses the significance of multimodal classification in depression detection, emphasizing the complementary nature of different data types such as text and voice. The authors cite previous research that supports the effectiveness of multimodal analysis in various applications, including personal recommendation systems, emotion recognition, and sentiment analysis.

Automatic Detection of Depression in Speech Using Ensemble Convolutional Neural Networks

The author proposes a speech-based method for automatic depression classification using ensemble learning for Convolutional Neural Networks (CNNs). The system is evaluated using the data and experimental protocol provided in the Depression Classification Sub-Challenge at the 2016 Audio–Visual Emotion Challenge. The proposed system consists of three 1D-CNNs, each with a different kernel size, followed by a fully connected layer. The outputs of the three CNNs are concatenated and fed into a final fully connected layer for classification. The system is trained using a 5-fold cross-validation approach, and the final metrics are computed by concatenating the partial results obtained over the test set with the model trained in each one of the 5-fold iterations with the corresponding training + validation configuration.

The proposed ensemble system is compared to two reference systems: the baseline system provided by the AVEC-2016 challenge, which is based on an SVM-based classifier and uses hand-crafted features, and the DepAudionet depression detection system, which uses a 1D-CNN, LSTM, and fully connected layers. The proposed system outperforms both reference systems in terms of classification accuracy, achieving an accuracy of 75.5%.

Detection of Depression Using Multimodal Models Based on Text and Voice Features by Solieman et al.

The authors proposed a novel approach for the automatic diagnosis of depression using multimodal models based on text and voice quality features. The study utilized the DAIC-WOZ database as a primary data source, which contains audio recordings and transcripts of interviews with individuals. The authors employed a sequential multimodal model that can predict depression using two types of data: text and audio features. The text analysis model processed the transcripts of the participants' interviews using natural language processing (NLP) techniques, specifically focusing on sentiment analysis. On the other hand, the voice quality analysis model utilized glottal flow voice features extracted from the participants' voices during the interviews. The performance of the developed models was evaluated, with the text analysis model achieving an F1-score of 0.7 on the non-depressed group of the test set, and the voice quality model achieving an F1-score of 0.65. The authors suggested that with further improvements, these models could potentially serve as self-diagnostic tools for depression without the need for clinical help. The study also emphasized the potential for automated detection of other mental illnesses based on the developed models.

DATASETS

1.DAIC-WOZ Dataset: A Multimodal Perspective on Depression Assessment

The Dialogue and Interaction with Computers - Western Ontario Dataset (DAIC-WOZ) is a seminal resource in the field of affective computing, specifically designed for the analysis of depression and related mental health conditions. Developed collaboratively by the University of Southern California's Signal Analysis and Interpretation Laboratory (SAIL) and Western University in Ontario, Canada, this dataset serves as a cornerstone for researchers aiming to advance the understanding and detection of depression through the lens of multimodal data.

The DAIC-WOZ dataset was created for the Audio/Visual Emotion Challenge (AVEC) in 2016 and 2017. Data collected include audio and video recordings and extensive questionnaire responses; this part of the corpus includes the Wizard-of Oz interviews, conducted by an animated virtual interviewer called Ellie, controlled by a human interviewer in another room. Data has been transcribed and annotated for a variety of verbal and nonverbal features. Its primary objective is to facilitate the development and evaluation of machine learning models for the detection of depression and related affective states. The dataset consists of 189 sessions, each lasting between five to ten minutes, involving participants with varying mental health conditions, including depression.

Modality Components:

1. Audio Recordings:
The audio component captures participants' speech patterns, intonation, and other acoustic features during interactions with a computerized virtual agent. This information provides insights into the vocal characteristics associated with different affective states, aiding in the development of algorithms for depression detection.

2. Text Transcriptions:

- Transcripts of the verbal interactions are included, offering textual data for linguistic analysis. The text component allows researchers to explore linguistic markers, sentiment, and the content of participants' responses, providing a rich source for natural language processing and textual analysis.

3. Video features:

Video features that are exacted by open face library are given. These include gaze direction, facial features etc.

Virtual Agent Interactions:

Participants engage with a computerized virtual agent during the sessions. The virtual agent is programmed to follow specific scripts designed to elicit a range of emotional responses. These interactions aim to simulate real-life conversations, allowing for the collection of data in a controlled yet ecologically valid setting.

To facilitate the development and evaluation of machine learning models, each session in the DAIC-WOZ dataset is accompanied by ground truth annotations. Trained human raters assessed the participants' depression levels using standardized clinical instruments, including the Patient Health Questionnaire-9 (PHQ-9). These annotations serve as reference labels for training and testing machine learning algorithms. The dataset includes participants with varying degrees of depression, ensuring a diverse representation of affective states. This diversity is crucial for training models that can generalize well to different manifestations of depression, contributing to the robustness and applicability of the developed algorithms. Given the sensitive nature of mental health data, the creators of the DAIC-WOZ dataset have taken measures to prioritize participant privacy. All data are anonymized, and stringent ethical guidelines are followed to ensure the responsible use of the collected information.

The multimodal nature of the DAIC-WOZ dataset makes it an invaluable resource for researchers exploring the potential of artificial intelligence, machine learning, and deep learning in depression detection. The dataset's richness and complexity

pave the way for innovative approaches to understanding, diagnosing, and addressing mental health conditions.

Problems related to DAIC-Woz Dataset

In the context of working with small datasets, several challenges arise, including the risk of overfitting and dimensionality problems. Overfitting occurs when a model learns to perform well on the training data but fails to generalize to new, unseen data.

To mitigate overfitting, the use of dropout layers and early stopping during the training process is preferred. Dropout layers are a technique commonly used in deep learning models to randomly drop units (along with their connections) from the neural network during training, which helps prevent overfitting by reducing the reliance on any individual unit . Early stopping involves monitoring the performance of the model during training and stopping the training process when the model's performance on a validation dataset starts to degrade, thus preventing the model from overfitting to the training data.

Another challenge associated with small datasets is the dimensionality problem, which refers to the high number of features relative to the number of samples. This can lead to difficulties in training models effectively.

Furthermore, class-imbalance is a common issue in many datasets, where one class may be significantly more prevalent than others. This imbalance can negatively impact model performance, leading to increased error rates. To address class-imbalance, the authors propose solutions at both the model-level and data-level.

Twitter Tweets Dataset

Detecting depression from social media platforms comes with numerous challenges. First, identifying
depression-oriented actions on social media can be difficult, since users engage in a diverse range of actions, such as liking, commenting, and posting. However, only a small percentage of these actions exhibit depressive symptoms, such as negative self-talk, reduced social activity, and expressions of hopelessness. Second, the raw data obtained from social media platforms require advanced preprocessing to derive valuable insights. Third, a significant number of users employ various means, including emojis, images, and links to other posts or blogs, to express their emotions. Drawing conclusions from these types of posts presents its own set of challenges. However, social media posts, including tweets, are generally short in length and may not provide enough context
about the user. Moreover, these methods did not consider other modalities, such as images and user profiles, which can provide valuable information about the user's mental health. Pre-processing the dataset

METHODOLOGY

Twitter Tweets Dataset

First, we check the dataset for missing or repeating values. If there is such data then we drop the row. Then we proceed to check the labels of the data. If the data is highly imbalanced, then we may proceed with various sampling techniques. After this we proceed with text pre-processing and feature extraction so that we can use various Machine Learning and Deep Learning models on the dataset.

Text Pre-processing

Text pre-processing is the process of cleaning the raw text data by removing the noise such as punctuations, emojis and common words to make it ready for our model to train. It is very important to remove unhelpful data or parts from our text. Text pre-processing improves the performance of an NLP system. There are several types of text pre-processing techniques, one must use the ones suitable for their use case in an order that makes sense. Some common text pre-processing techniques include tokenization, stemming, lemmatization, stop word removal etc.

Tokenization: This is a common task in NLP. It is a fundamental step in both traditional NLP methods like Count Vectorizer and Advanced Deep Learning-based architectures like Transformers. Tokens are the building blocks of Natural Language. Tokenization is a way of separating a piece of text into smaller units called tokens. By breaking text into tokens, it becomes easier to analyse and process the text programmatically.

Stemming: Stemming is the process of reducing a word to its base or root form, by removing any suffixes or prefixes that may be attached to it. The resulting word is called the stem, and it may not necessarily be a valid word in the language. Stemming is often used as a pre-processing step in natural language processing

(NLP) tasks, such as information retrieval and text classification. By reducing words to their base form, it becomes easier to group related words together and perform analysis on them. For example, consider the following words: "walking", "walked", and "walks". These words all have the same root word "walk". By stemming these words, we can reduce them to their base form.

Lemmatization: Lemmatization is a text normalization technique used for Natural Language Processing (NLP). It can convert any word's inflections to the base root form. This allows us to categorize words with similar meanings, no matter how they are spelled. Lemmatization brings context to the words. By reducing words to their base form, it becomes easier to perform analysis on them and to identify the meaning of the text. For example, consider the following words: "better", "best", and "good". These words have different forms, but they all have the same lemma "good". Lemmatization is typically more accurate than stemming, but it is also computationally more expensive. There are various algorithms used for lemmatization in NLP, including the WordNet lemmatizer and the spaCy lemmatizer.

Stop words removal: Stop words are words that are commonly used in a language and are usually removed from texts because they do not carry important meaning. Examples of stop words in English include "the", "a", "an", "and", "or", "of", "in", "is", "that", "to", and "with". Stop words removal is a common technique used in natural language processing (NLP) to improve the efficiency and accuracy of text analysis. By removing stop words from a text, we can reduce the size of the data and focus on the more important words that carry meaning and contribute to the analysis.

Feature Extraction

Embeddings are generated for the text using various models such as Count vectorizer and TF/IDF, Bert, Roberta etc

Count vectorizer is a simple technique that counts the frequency of each word in a document and represents the document as a vector of word counts.
TF-IDF (Term Frequency-Inverse Document Frequency) is a more sophisticated technique that takes into account the frequency of each word in the entire corpus of documents, as well as the frequency of the word in the specific document being represented.

BERT (Bidirectional Encoder Representations from Transformers). This is a pre-trained language model. The Transformer model consists of an encoder and a decoder. The BERT model is a variant of the Transformer encoder architecture that is trained on a large amount of text data using a language modelling objective. The BERT encoder is a stack of transformer blocks, where each block consists of a multi-head self-attention mechanism and a feed-forward neural network. The self-attention mechanism allows the model to attend to different parts of the input sequence, while the feed-forward network provides non-linear transformations of the input. The encoder takes as input a sequence of tokens (usually words or sub-words), which are first converted into vector representations through an embedding layer. The resulting embeddings are then passed through a series of transformer blocks, where each block updates the embeddings by attending to the other tokens in the sequence. The bert embedding gives a pooler output vector of size 768.

Comparison Models

We generate the textual features using various methods including the state-of-the-art transformers for text including Bert. Then these features are passed on the models such as logistic regression and Artificial Neural Networks for classification.

DAIC WOZ dataset:

Audio Feature Extraction

We extracted a comprehensive set of acoustic features from the audio recordings to capture various aspects of phonation and voice quality. These features provide valuable insights into the spectral, temporal, and source characteristics of the audio signals, contributing to our understanding of emotional expression and vocal behaviour. The extracted features include Mel-Frequency Cepstral Coefficients (MFCCs) which contain information for differentiating different sounds and insights into pitch and higher-frequency components of the voice. These features were extracted using covarep app and python audio library librosa.

Video feature extraction

We extracted set of video features using open face library which includes facial landmarks which provide spatial information about the facial geometry such eye corner, nose tips etc and head pose which estimates the head pose in terms of rotation angles which provides insights into the orientation of the face in the video frame and eye gaze direction which estimates gaze direction of a subject by analysing position of eyes relative to facial landmarks. Facial Expressions features are the result of analysis on facial landmarks to infer the presence and intensity of facial expressions, such as happiness, sadness, or surprise.

RESULTS

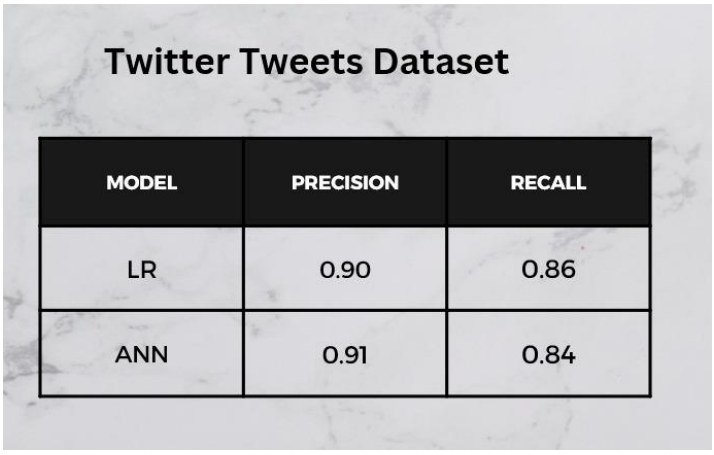
Twitter Tweets Dataset

Logistic Regression Model

After generating the features using bert transformer we use logistic regression model to classify the tweets.

Artificial Neural Networks Model

Using the generated features from transformer models, these are passed through a neural networks model consisting of 5 layers.
First is the input dense layer which consists of 12 nodes that takes an input size of 768 and has relu as activation function. Second layer is the Dropout layer with .1 dropout rate. Third layer is another dense layer with 6 nodes with relu as activation function. Fourth layer is another dense layer with 3 nodes with relu as activation function. Final layer is another dense layer with single node and sigmoid activation function.
We use the adam optimiser for loss minimization.



CONCLUSION

In conclusion, the integration of deep learning techniques across various modalities such as audio, text, and video recordings present a promising avenue for comprehensive depression detection. By leveraging these advanced technologies, we can potentially revolutionize mental health assessment, offering more timely and accurate diagnoses.

The multi-modal approach allows for a more nuanced understanding of individuals' mental states, capturing subtle cues and patterns that may not be evident through any single modality alone. This holistic approach enhances the reliability and effectiveness of depression detection, enabling earlier interventions and personalized treatment strategies.

However, it's crucial to acknowledge the ethical considerations and potential biases inherent in the development and deployment of such systems. Careful attention must be paid to issues of privacy, consent, and fairness to ensure that these technologies benefit individuals without causing harm or perpetuating discrimination.

Moving forward, further research and collaboration between experts in both mental health and machine learning will be essential to refine and validate these methods. Ultimately, the goal is to harness the power of deep learning to improve the lives of those affected by depression, fostering a future where mental health support is more accessible, proactive, and tailored to individual needs.

REFERENCES

1. A. Saidi, S. B. Othman and S. B. Saoud, "Hybrid CNN-SVM classifier for efficient depression detection system," 2020 4th International Conference on Advanced Systems and Emergent Technologies (IC_ASET), Hammamet, Tunisia, 2020, pp. 229-234, doi: 10.1109/IC_ASET49463.2020.9318302.

2. A. Bailey and M. D. Plumbley, "Gender Bias in Depression Detection Using Audio Features," 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 2021, pp. 596-600, doi: 10.23919/EUSIPCO54536.2021.9615933

3. Esaú Villatoro-Tello, Gabriela Ramírez-de-la-Rosa, Daniel Gática-Pérez, Mathew Magimai.-Doss, and Héctor Jiménez-Salazar. 2021. Approximating the Mental Lexicon from Clinical Interviews as a Support Tool for Depression Detection. In Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21). Association for Computing Machinery, New York, NY, USA, 557–566. https://doi.org/10.1145/3462244.3479896

4. Z. Huang, J. Epps and D. Joachim, "Investigation of Speech Landmark Patterns for Depression Detection," in IEEE Transactions on Affective Computing, vol. 13, no. 2, pp. 666-679, 1 April-June 2022, doi: 10.1109/TAFFC.2019.2944380.

5. Sara Sardari, Bahareh Nakisa, Mohammed Naim Rastgoo, Peter Eklund, Audio based depression detection using Convolutional Autoencoder, Expert Systems with Applications, Volume 189, 2022. https://doi.org/10.1016/j.eswa.2021.116076.

6. Park, Junhee, and Nammee Moon. "Design and implementation of attention depression detection model based on multimodal analysis." Sustainability 14.6 (2022): 3569.

7. Zhao, Y., Liang, Z., Du, J., Zhang, L., Liu, C. and Zhao, L., 2021. Multi-head attention-based long short-term memory for depression detection from speech. Frontiers in Neurorobotics, 15, p.684037.

8. Solieman, H. and Pustozerov, E.A., 2021, January. The detection of depression using multimodal models based on text and voice quality features. In 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus) (pp. 1843-1848). IEEE.

9. Dinkel, H., Wu, M. and Yu, K., 2019. Text-based depression detection on sparse data. arXiv preprint arXiv:1904.05154.

10. Fang, M., Peng, S., Liang, Y., Hung, C.C. and Liu, S., 2023. A multimodal fusion model with multi-level attention mechanism for depression detection. Biomedical Signal Processing and Control, 82, p.104561.

11. Vázquez-Romero, A. and Gallardo-Antolín, A., 2020. Automatic detection of depression in speech using ensemble convolutional neural networks. Entropy, 22(6), p.688.

12. Ye, J., Yu, Y., Wang, Q., Li, W., Liang, H., Zheng, Y. and Fu, G., 2021. Multi-modal depression detection based on emotional audio and evaluation text. Journal of Affective Disorders, 295, pp.904-913.

13. Yang, L., 2019, September. Multi-modal depression detection and estimation. In 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) (pp. 26-30). IEEE.