Detection of signs of depression based on a multimodal approach.

PhD Dissertation Proposal

by

Karla María Valencia Segura

Doctoral Advisors:

Dr. Luis Villaseñor Pineda, INAOE

Dr. Hugo Jair Escalante Balderas, INAOE

Dr. Fernando Javier Martínez Santiago, UJA

Instituto Nacional de Astrofísica, Óptica y Electrónica

January, 2023

Santa María de Tonantzintla, Puebla, CP 72840

Detection of signs of depression based on a multimodal approach.

PhD Dissertation Proposal

by

Karla María Valencia Segura

Doctoral Advisors:

Dr. Luis Villaseñor Pineda, INAOE

Dr. Hugo Jair Escalante Balderas, INAOE

Dr. Fernando Javier Martínez Santiago, UJA

Universidad de Jaén

January, 2023

Campus Las Lagunillas s/n. 23071 - Jaén

UNIVERSIDAD DE JAÉN

Abstract

Depression is a common mental disorder affecting around 800 million people worldwide, with only a fraction receiving adequate treatment. Detection of depression is challenging due to different reasons like: patient's disposition to seek help, expert opinion, and the diversity of symptoms. Studies have shown that people with depression can present indicators of the disorder through verbal and non-verbal features. In recent years, techniques have been developed to detect and diagnose depression, including questionnaires and machine learning models. Most of the works in machine learning for depression detection use data from social networks, this data may lack the reliability and validity required for accurate depression assessment. The information shared online might be incomplete, exaggerated, or misleading, making it difficult to trust the data for diagnostic purposes. Also, most of the works are focused on using only one modality for depression detection, ignoring the different depression indicators that can be presented. Due that, integrating multimodal data, including speech, text, and non-verbal cues, can enhance diagnostic accuracy and scope.

This research focuses on detecting depression using multimodal data, including audio, video, and text from clinical interviews based on the PHQ-8 questionnaire. We propose a solution that leverages psychological evidence to extract "interest points," using both verbal and non-verbal information. The approach seeks to capture the multifaceted nature of depression, addressing the inherent complexity of the problem. The findings underscore the importance of prioritizing interest points based on psychological support. Also the preliminary results demonstrate the feasibility and novelty of their proposed solution, but the effectiveness of prioritizing interest points based on psychological support and adopting a simple approach to model interviews.

Keywords— Clinical Interviews, Depression, Detection, Multimodal

1 Introduction

According to the World Health Organization (WHO) mental illness is one of the leading causes of disability worldwide [WHO, 2019]. It is estimated that depression affects around 800 million people worldwide, making it one of the most common and recognized mental disorders in the world and a leading cause of suicide, yet only a fraction of them receive adequate treatment [Skaik and Inkpen, 2020]

One of the main challenges of depression is how difficult it is to detect it due to the fact that the diagnosis depends exclusively on the patient disposition of seek help, expert opinion and the symptoms diversity, due this often leading to insufficient treatment [WHO, 2020].Due to this, in recent years this disease has been gaining relevance for its detection and diagnosis, developing various techniques ranging from traditional ones such as the use of questionnaires to machine learning models. This is due to the fact that several studies have shown that people with depression can present indicators of this disorder through their written communication [Zimmermann et al., 2017, Bucci and Freedman, 1981, Rude et al., 2004], so analyzing the language used by patients has become an area of interest, but not only the language can give us some insights of depression symptoms, other modalities like audio and video, has features that have been related to the depression signs, like body posture, facial expression, including the voice tone can have some indicators of a depressive illness [Haque et al., 2018, Hall et al., 1995, Sobin and Sackeim, 1997], so it's important to recognize them and use them together in order to recognize and make available the necessary resources to help to identify, monitor, diagnose and deal with this disease.

While there have been recent improvements in automatically detecting depression, a persistent problem exists due to the shortage of datasets related to depression. Consequently, most works focused on depression detection have concentrated on collecting datasets from different sources, particularly those derived from social media.

However, these methods have several limitations. For example, the content of the social media posts may not be rich enough to capture depression symptoms [De Choudhury et al., 2013]. More importantly, these automated depression detection models overlook the existing methods for depression assessment endorsed by the medical and psychological community, and consequently may miss some important signals of depression. Hence, the importance of having explanatory models that allow these models to be evaluated and adopted by psychiatrists.

1.1 Motivation

One of the main difficulties of Traditional methods for depression diagnosis is that require a lot of medical resources as well economic inversion, and it not accessible for every person that could need it [Richter et al., 2021]. This frequently results in underdiagnosis and undertreatment, perpetuating the global mental health crisis. Given the unique and complementary strengths of machine learning and clinical depression assessment, a recent study showed that AI could enhance the accuracy of diagnosis and clinical decisions when combined with expert human evaluation, emphasizing the collaborative nature of AI and doctors [Guohou et al., 2020, Sezgin, 2023]. By combining both we can obtain robustness models to assess this task.

Moreover, the integration of multimodal data, which includes not only speech and text but also non-verbal cues like facial expressions, tone of voice, and body language, enhances the diagnostic accuracy and scope [Pedersen et al., 1988]. The combination between machine learning and multimodal data offers the opportunity to detect subtle indicators or symptoms that patients may present during their verbal and non-verbal communication, making it a supportive tool for detecting these symptoms and aiding in diagnostic precision.

Early and accurate detection of depression can lead to timely interventions, access to healthcare providers and reducing the burden on healthcare systems, and,

most importantly, alleviating the suffering of individuals living with depression. By exploring machine learning and multimodal information, we can develop sophisticated models capable of identifying different depression symptoms in verbal and non-verbal behaviours. This research focuses on improving the depression detection using multimodal information and provide insights to the medical staff to help in the accuracy diagnose.

For that, the motivation of this work aims to bridge the gap between the subjective nature of traditional depression diagnosis and the objective, improving the detection of depressed patients on multimodal information. Also, we plan the incorporation of explainability in our approach to provide transparency and comprehensibility in the decision-making process. By ensuring that the model's reasoning and insights can be understood and interpreted for identifying depression symptoms.

1.2 Justification

Because depression has many signs and symptoms, various studies in psychology focus on creating different ways to diagnose this mental disorder. These methods often involve interviews to understand the importance of various symptoms in relation to different factors. According to the DSM-5 [Morrison, 2015], a diagnosis of major depressive disorder (MDD) is confirmed when a person experiences at least five of the eight specified symptoms. These symptoms include persistent feelings of sadness, a significant loss of interest or pleasure in almost all daily activities, noticeable changes in weight or appetite, disruptions in sleep patterns (like trouble sleeping or sleeping too much), observable changes in physical movements (either restlessness or slowed down), feeling tired or having low energy, feelings of worthlessness or excessive guilt, difficulties in thinking, concentrating, or making decisions, recurrent thoughts of death or suicide, a suicide attempt, or having a specific suicide plan occurring daily or almost daily over the past two weeks.

It's important to note that at least one of these symptoms must be either a depressed mood or loss of interest. Additionally, the DSM-5 considers non-verbal cues as part of the diagnosis process. These include patients showing a sad facial expression, teary eyes, a furrowed brow, downturned corners of the mouth, slouched posture, limited eye contact, lack of facial expressions, minimal body movements, and changes in speech patterns (like speaking softly, lack of variation in tone, or using short words).

Consequently, the objective of the present work is to analyze and pinpoint points of interest, which can be identified across different modalities and serve as crucial indicators of emotional and psychological well-being. By focusing on these interest points, a multimodal graph neural network (GNN) can effectively capture and integrate relevant information from both verbal and non-verbal cues. Graph neural networks are particularly well-suited for modeling relationships and interactions between different data points, making them suitable for integrating information from diverse modalities.

Furthermore, the interpretability of the model can allow researchers and practitioners to understand which interest points[1] and modalities contribute most significantly to the depression detection process.

In summary, a multimodal graph neural network focusing on interest points offers a model to depression detection by leveraging the strengths of both multimodal data integration and graph-based modeling.

1.3 Problem Statement

Previous studies focused on the identification of depressive disorders have shown that this type of illness is detectable in different environments. However, most of the work

[1] An *"interest point"* is defined as a possible indicator of a relevant symptom of depression

focuses on the identification of this disorder in the social network domain, and most of these works use a single modality, which is mostly the text modality. However, these approaches set aside the diagnostic protocols established in the DSM-5 [Morrison, 2015], which guarantee the correct diagnosis based on the presence of certain symptoms over a period of time. This information is lost within the social media environment, mainly due to the fact that social networks can contain too much information unrelated to depression, making them unsuitable as they don't accurately represent real-world clinical situations [Mao et al., 2023]. They are a complex and varied data source that needs specialized techniques for effective analysis. Social media data is unstructured, lacking a set format, making it challenging to derive meaningful insights. Moreover, social media data comes in large and fast streams. Another significant aspect is its tendency to be noisy and unreliable, with users sharing false information, misleading content, or spam [2]. This creates difficulty in extracting accurate insights and is compounded by the dynamic and ever-changing nature of social media, posing challenges for learning models.

1.4 Hypothesis

We hypothesize that the combination of multimodal information, with a specific focus on interest points within each modality (e.g., facial expressions in visual data, symptoms in textual data, and prosody in auditory data), will significantly improve the detection detection. By leveraging these interest points, we anticipate that the resulting model will provide insights into the manifestation of depression symptoms, ultimately leading to a more effective and interpretable diagnostic tool.

[2] https://www.linkedin.com/pulse/maximizing-social-media-insights-data-science-pros-cons

1.5 Objectives

1.5.1 Main Objective

Develop a multimodal graph neural network model tailored to identify depression during clinical interviews and assess its severity, emphasizing the utilization of interest points from each modality. The primary goal is to create an interpretable model capable of providing insights into depression detection and its severity based on these points of interest, with the intention of surpassing the current state-of-the-art methods. By incorporating advanced techniques and leveraging the interconnected nature of multimodal data, we aim to improve the accuracy and sensitivity of depression detection compared to existing approaches.

1.5.2 Specific Objectives

To carry out the main objective, the following specific objectives are proposed:

- Identify and extract the interest points for each modality (Audio, Video and Text)
- Design a graph neural network model that learn new representation based on the combination of interest points per modality
- Evaluate the utility of the model in the depression detection task
- Design an interpretative model that help in the depression diagnose, taking in consideration the verbal and non-verbal behaviors.

1.6 Scope and limitations

The present work is limited to the English language, since the modality analyzed depend on resources and methods oriented in that language. In addition, the evaluation collections commonly used for the depression detection task are in the English language, which forces us to evaluate our proposal under this same language. In the other hand, this research does not include clinical evaluation of any type.

1.7 Expected contributions

Through this research, the following contributions are expected to be obtained:

1. A method that identify interest points in different modalities to identify the depression and severity.
2. A multimodal representation of patient's that was learned by a model combining the interest points per modality, that improves the depression detection.
3. A interpretative model that take advantage of the interest points to help in the accurate diagnose

1.8 Document Organization

Chapter 2 describes the background related to this work. Chapter 3 describe the most relevant characteristics of the works related to the research topic, found up to the date of writing this work. Chapter 4 formalizes the proposed method for this work, as well as its details. Chapter 5 describes the preliminary work, as well as the results obtained. Finally, in the last chapter the preliminaries conclusions are presented.

2 Background

In this section, we first introduce depression assessment questionnaires, which play a key role in clinical depression diagnose. We summarize studies on automatic depression detection and finally we explore the multidata fusion and the different approaches.

2.1 Depression assessment

In psychiatric diagnosis, there are different systems to carry out the diagnosis, with one of the most reliable methods being clinical interviews [Mueller and Segal, 2014].

Clinician-applied questionnaires are more effective than self-administered questionnaires. This is because they allow interviewers to personalize questions based on the respondent's previous answers, and obtain more detailed and meaningful information for diagnosis. The purpose of the clinical interviews on the mental health field is to develop a report to understand the patient symptoms are mainly characterized. This information is used to make an accurate psychiatric diagnosis, which typically guides the treatment. Interviews may vary in terms of structure, these can be either structured or unstructured.

2.1.1 Unstructured Clinical Interviews

The unstructured clinical interviews are mainly characterized by following a free-flowing conversation between the clinician and the patient, and there are no a specific topic or parameters to follow. The content and the sequence of these interviews are determinate by the clinician's [Mueller and Segal, 2014]. However the lack of structure in this approach is a serious disadvantage, since the clinicians may not gather all the information for an accurate diagnosis.

2.1.2 Structured Clinical Interviews

Unlike unstructured interviews, structured clinical interviews are characterized by following a standardized set of questions and a uniform sequence for all patients[Mueller and Segal, 2014].

The primary advantage of these interviews is that by following a structured format, clinicians increase the reliability of diagnoses. There are two types of structured interviews: *fully structured* and *semistructured*. In fully structured clinical interviews, questions are posed to the respondent exactly as written, following a predetermined sequence. The phrasing of follow-up probes is also explicitly outlined, and interviewers are instructed not to deviate from this established format. On the other hand, in a semistructured interview, while the initial questions regarding each symptom are predetermined and typically posed exactly as written to the respondent, the interviewer possesses substantial flexibility in pursuing responses. The interviewer can adapt or enhance the standard queries with personalized and context-specific probes to achieve a more precise assessment of particular psychiatric symptoms.

The deliberate development of structured and semistructured interviews enhances their methodological validity when compared to unstructured approaches. Structured and semistructured interviews are specifically designed to comprehensively and precisely evaluate clearly defined diagnostic criteria, making them generally more effective for assessing these criteria compared to unstructured interviews. Practitioners using unstructured interviews might be inclined to rush through diagnoses, prematurely limit their diagnostic considerations, and overlook the presence of comorbid conditions.

There are many semistrucutred diagnosis instruments for depression diagnosis, the most used instruments for this task are:

- *Beck Depression Inventory (BDI)* [Beck et al., 1961]: The Beck Depression Inventory (BDI) is a commonly employed tool for the assessment of depression, serving the dual purpose of screening for depressive symptoms and gauging the behavioral indicators and intensity of depression. The applicability of the BDI extends from individuals aged 13 to 80. Comprising 21 self-report items, respondents provide their responses using a multiple-choice format. Typically, the administration of the BDI takes around 10 minutes to conclude. Extensive assessments of the validity and reliability of the BDI have been conducted in diverse populations globally.

- *Montgomery-Åsberg Depression Rating Scale (MADRS)* [Montgomery and Åsberg, 1979]: The Montgomery-Åsberg Depression Rating Scale (MADRS), consisting of 10 items, is employed to gauge the level of depression severity in individuals aged 18 years and above. Respondents provide ratings for each item on a 7-point scale. This scale is derived from the Hamilton Depression Rating Scale but is more responsive to detecting changes in depression symptoms over time. It typically requires 20 to 30 minutes to complete.

- *Hamilton Depression Rating Scale (HAM-D)* [Hamilton, 1960]: The Hamilton Rating Scale for Depression, commonly referred to as HDRS, HRSD, or HAM-D, assesses the level of depression in individuals both prior to treatment, during treatment, and after treatment. This assessment tool is administered by healthcare professionals and comprises 21 items, although the scoring is derived from the first 17 items. These initial items are evaluated using either 5-point or 3-point scales. The process of administering and scoring the scale typically takes between 15 to 20 minutes to complete.

- *Patient Health Questionnaire (PHQ-8)* [Kroenke et al., 2009]: The PHQ-8 consist in the evaluation of 8 items associated with the depression. This questionnaire is half the length of many other depression instruments, has comparable

Questionnaire	Acronyms	# of items	Complete Time
Beck Depression Inventory	BDI	21	20-30m
Montgomery-Åsberg Depression Rating Scale	MARRSD	10	20-30m
Hamilton Depression Rating Scale	HAMD	21	15-20m
Patient Health Questionnaire	PHQ-8	8	5-10m

Table 1: Summary of the semistructured clinical interviews

sensitive and specificity. The PHQ-8 is dual purpose instrument that, with the same nine items, can establish provisional depressive disorder diagnoses as well as grade depressive symptom severity.

2.2 Multimodal Information

Multimodality is a representation of data using information from multiple such entities, often with multiple representations, which can be an image, a piece of audio, a piece of text or other forms [Qin et al., 2023]. One of the greatest challenges of multimodal data is to summarize the information from multiple modalities (or views) in a way that complementary information is used as a conglomerate while filtering out the redundant parts of the modalities. Due to the heterogeneity of the data, some challenges naturally spring up including different kinds of noise, alignment of modalities (or views) and, techniques to handle missing data.

Multimodal representations fall into two categories: (1) **Joint representation:** each individual modality is encoded and then placed into a mutual high dimensional space. This is the most direct way and may work well when modalities are of similar nature. And (2) **Coordinated representation:** each individual modality is encoded irrespective of one another, but their representations are then coordinated by imposing a restriction. For example, their linear projections should be maximally

correlated

2.2.1 Multimodal Data Fusion

The fusion of data is the process by which one seeks to combine various records or modalities representing the same object from different perspectives into a single representation that is coherent and clean [Du and Swamy, 2019]. The goal is to use this complementary representation to address complex tasks. There are various data fusion techniques; however, most can be classified into three main categories: early fusion, intermediate fusion, and late fusion.

- **Early fusion** or feature fusion involves extracting unimodal features from various sources of information [Snoek et al., 2005]. Subsequently, the different feature vectors are integrated into a single, large vector, which is finally used for classification. Since this vector can consist of many features, the training time can increase. The advantage of this technique is that only one training phase is needed; however, the disadvantage is the difficulty of combining features into a common representation [Snoek et al., 2005].

- **Late fusion** or decision fusion involves combining unimodal results after classification. This process predicts the final result by considering the individual labels of the involved classifiers [Ebersbach et al., 2017]. In general, these late fusion strategies are simpler to implement than early fusion, especially when different modalities vary significantly in terms of data dimensionality and sampling rates, and they often result in better performance [Ramachandram and Taylor, 2017]. However, the downside of this scheme is its learning cost, as a learning stage is required for each modality.

- Within fusion, there is a third scheme called **intermediate fusion**, which is a more flexible approach to carry out multimodal fusion. There are different

techniques that allow for intermediate fusion, but the most popular are the deep learning architectures.

In the context of deep learning architectures, once neural networks have transformed all modalities into representations, it becomes possible to merge these different representations into a single fusion layer, also known as the shared representation layer, to finally learn a joint multimodal representation [Ramachandram and Taylor, 2017]. This fusion layer or shared representation layer can be a single shared layer that merges multiple channels at a certain depth or could gradually fuse one or more modalities at a time. One of this type of architectures are the Graph Neural Networks (GNNs).

Graph Neural Networks (GNNs) exemplify one such type of architecture. As previously mentioned, GNNs constitute a class of deep learning models tailored for processing and analyzing graph-structured data. In graphs, nodes symbolize entities, while edges denote relationships between these entities. GNNs leverage these graph structures to discern intricate dependencies and patterns within the data [Wu et al., 2020]. In computer science, a graph serves as a nonlinear data structure comprising vertices and edges. Vertices are occasionally termed nodes, and edges are the lines or arcs connecting any two nodes in the graph. Formally, a graph is composed of a set of vertices V and a set of edges E, denoted by $G = (E, V)$. Graph Neural Networks are specialized neural networks adept at handling graph data structures. They draw significant inspiration from Convolutional Neural Networks (CNNs) and graph embedding. GNNs find application in predicting nodes, edges, and graph-based tasks. In section 7.1, we elaborate on the functionality of GNNs.

.

3 State of Art works

In this chapter, the most relevant related work concerning this research is presented. Various works related to the detection of depression in the clinical interviews domain are included, employing both unimodal and multimodal information. Additionally, various data fusion approaches for depression detection using multimodal information are discussed.

3.1 Automatic depression detection

In recent years, researchers have delved into the application of machine learning to identify an individual's mental health condition through their spoken or written expressions, and the integration of some non-verbal features [Benton et al., 2017, Caicedo et al., 2020]. Nevertheless, the majority of these works have primarily centered around predict if a person have or not depression empoying in most of the cases the use of a single modality, failing to encompass the full spectrum of clinical manifestations associated with depression. In our study, we have categorized these research efforts based on the number of modalities employed:

3.1.1 Unimodal features-based methods

Most of the works have focused on the usage of a single modality, generally the text modality. For example, in [Shen et al., 2013], the authors created a system for recognizing individuals who might have depression based on their written content. To address this, they suggest a two-step approach using supervised learning. In the first step, they check if users exhibit clear signs of negative emotions. Then, in the second step, they assess whether these cases indicate clinical depression or just regular sadness.

In [Shin et al., 2022], the authors conducted and recorded the Mini International Neuropsychiatric Interview (MINI) in a Korean population, a brief interview designed to diagnose mental illness [Sheehan et al., 1998].They aimed to detect depressed patients using the transcribed interview text, concluding that depression can be diagnosed using the word frequency and part-of-speech tagging, with enhanced performance when demographic variables are incorporated. In [Lopez-Otero et al., 2017], the study examines how speaker de-identification impacts a depression detection system based on speech transcriptions. The paper introduces a depression detection method using Global Vectors (GloVe) word embeddings and a word weighting strategy to reduce the impact of transcription errors. Experiments conducted within the AVEC 2016 framework demonstrate the GloVe-based approach's competitive results with manual transcriptions, though automatic transcriptions, whether de-identified or not, show a slight decrease in performance.

Another example of unimodal semi-structured clinical interviews is [Namboodiri and Venkataraman, 2019], where the authors conducted interviews with college students. They extracted facial features from recorded interviews, focusing on emotional faces, and used an SVM to determine the severity of depression (high, moderate, or low). This approach yielded lower accuracy, partly due to the limited dataset collected for this work.

Some other works have used the audio modality to identify depressed users. In [Lopez-Otero et al., 2014], the authors used a multimodal dataset but focused solely on audio features to detect depressed patients, concluding that voice indicates distinctions between depressed and non-depressed users. Another work focused on audio features is [Ma et al., 2016], where the authors proposed a deep model called *DepAudioNet* to encode different audio channel features for identifying depressed users. Finally, [Cummins et al., 2017] investigated the effects of depression using vowel-level features based on gender, concluding that depression manifests at the phoneme level of speech, and the effects of depression in speech can be captured by

features characterizing speech motor control.

3.1.2 Bimodal feature-based methods

Other works explore the integration of two modalities, such as video and text. For example, in [Joshi et al., 2013], the authors introduced a depression detection framework based on human body parts' relative and local motion patterns. They computed relative movement histograms and used a support vector-based classifier, achieving strong discriminative capabilities. To enhance the system, a bag-of-words framework was introduced to analyze dynamics within the face and local motion in various body parts. The combination of both modalities highlighted their complementary nature, demonstrating the effectiveness of a bimodal approach for depression detection.

in [Ray et al., 2019] the authors proposed a multi-level attention fusion across audio, video, and text modalities, extracting features to represent the user. They used attention layers to select the most relevant features per modality and then concatenated them. The best results were obtained using only video and text modalities. Another bimodal work is [Cohn et al., 2009], which focused on investigating the relationship between facial and vocal behaviors for clinical diagnosis, revealing the potential of nonverbal affective information in contributing to clinical research and practice.

Another recent work is [Sun et al., 2021], where the authors proposed a multi-level attention model for each modality (audio, video, image) and used attention mechanisms to focus on relevant features of each modality using late fusion. The results showed that audio and video modalities led to better performance, underscoring the need for sophisticated fusion models to capture the interplay between modalities.

3.1.3 Multimodal feature-based methods

In addition to other modalities, works have been focused on using three or more modalities. For instance, in [Zhang et al., 2020b], the authors extracted different characterizations from the three modalities and fused them using an early strategy, feeding a Multitask Deep Neural Network (DNN) with all features of each modality to detect bipolar and depressive patients. The model was found to be generalized across different mental illnesses, maintaining good performance. Finally, one of the most recent works using the E-DAIC dataset is [Saggu et al., 2022], where the authors proposed a BiLSTM and attention mechanisms to determine the severity of depression, achieving better performance than the state-of-the-art. The authors mentioned that more advanced fusion methods could be explored to improve performance.

3.2 Graph Neural Networks

In the case of depression detection has been some works that use the GNN's to model the depression indicators, such as [Burdisso et al., 2023], where the autors propose to use Graph Convolutional Network (GCN) to figure out the content of recorded conversations between a therapist and depression people in text modality. The key points of this work are: test the GCN in a new way for identifying depression in recorded interviews, and it performs better than other methods on two standard datasets; and the model proposed can help people understand its decisions, which is important when using artificial intelligence for diagnosis, and it aligns with psychology research. Another work focusing on depression detection in bimodal information is [Ghadiri et al., 2022]. Where the researchers introduced a mixed structure to forecast depression using both voice and text information. They used a deep learning model based on transformers, fine-tuned with pre-existing models, and found it had better accuracy than current approaches. Additionally, they introduced new graph-based features by converting speech signals into a complex network, leading to even

more precise results.

Because Graph Neural Networks (GNNs) are good at understanding relationships, they work well in multimodal learning. GNNs use a graph made from different types of data to connect and share information between modes, helping to bring together and enhance the understanding of multimodal data [Li et al., 2023].

As far as we know, there aren't many models that use Graph Neural Networks to combine different types of information for depression detection task. Because of this, we aim to develop an architecture that uses Graph Neural Networks to model the depression symptoms in multimodal data. For example in [Li et al., 2023] the authors introduce a new way to combine different types of information, calling it the Graph Network-based Multimodal Fusion Technique (GraphMFT) for Emotion Recognition in Conversation (ERC). They use three graphs for each conversation: one for visuals and audio, one for visuals and text, and one for audio and text. This makes it easier to mix different types of information. The experiments show that their new model can work well in understanding both similar and different types of information. When compared to the older model used as a basis, GraphMFT performs much better.

Work	Modality			Machine Learning	dataset	Performance	Features
	A	V	T				
[Burdisso et al., 2023]			X	Graph Convolutional Network	E-DAIC	F1:0.76	Word-embedding
[Shin et al., 2022]			X	NB	MINI	AC:0.83	Word frequency, part-of-speech
[Shen et al., 2013]			X	SVM	Bulletin Board System Collection	AC:0.84	TF-IDF
[Lopez-Otero et al., 2017]			X	SVM	DAIC-WOZ	AC:0.85	GloVe
[Namboodiri and Venkataraman, 2019]		X		SVM	Videos from students	F1:0.69	Emotional faces
[Joshi et al., 2013]		X	X	SVM	CI	F1:0.89	Body parts movement, BoW
[Lopez-Otero et al., 2014]	X			SVR	AVEC 2013	MAE:7.70	MFFC, RASTA-PLP,SDC,energy, spectral and prosodic
[Ma et al., 2016]	X			LSTM	AVEC 2016	F1:0.53	Mel-scale filter bank features
[Cummins et al., 2017]	X			SVM	DAIC-WOZ	F1:0.63	VL-Formants, eGeMAPS, COVAREP
[Ray et al., 2019]		X	X	Multilevel Attention	E-DAIC	RMSE:4.28	BoVW, Pose, Gaze, FAU and Embeddings
[Yin et al., 2019]	X	X	X	RNN	E-DAIC	RMSE:5.50	gaze direction, position of the head, FAUs, VGG, Semantic and emotional embeddings
[Zhang et al., 2020b]	X	X	X	Mulltitask DNN	E-DAIC	MSE:20.06	MFCC, eGeMAPS, facial landmarks, LLDs, embeddings, BoW
[Sun et al., 2021]	X	X		tranformer network	E-DAIC	RMSE:7.73	MFCC and AUposes
[Cohn et al., 2009]	X	X		SVM & LR	CI	AC:0.88	Facial actions and vocal prosody
[Ghadiri et al., 2022]	X		X	GNN & BERT	DAIC-WOZ	F1:0.82	embeddings, MFCC, Spectograms, mel-Spectograms
[Saggu et al., 2022]	X	X	X	BiLSTM	E-DAIC	RMSE:4.32	17 FAUs, gaze, voice quality, cepstral, prosodic and embeddings

Table 2: Summary of the state of art

3.3 Discussion

Table 2 provides a summary of representative studies on automatic depression detection in multiple dimensions. Reviewing the table we can conclude the following:

- Most of the works are unimodal approaches, ignoring the information that other modalities can bring to the model.
- Most of the works presented here treat depression detection as a binary classification problem.
- None of the multimodal works presented here provides an explanation of the importance of different modality inputs to depression detection.
- And finally, none of the previous studies has considered the symptoms as input feature levels. This can provide a simple explanation for the psychiatric evaluation and result in a more accurate diagnosis.

In this work, we propose a multimodal network graph, considering both binary classification and the regression task. This allows us to determine not only whether the patient has depression or not, but also the severity of the disease. We also aim to identify the rekevant points of interest in each modality to assess the importance of using these features. Finally, our model is designed to enable psychiatric professionals to understand the main symptoms or behaviors behind the model's predictions based on the identified interest points.

4 Research proposal

In this section we present the detailed methodology proposed for this works as well the work schedule that we plan to complete for the next three years.

4.1 Multimodal Graph Neural Network

In the real world, depression-relevant signs are presented in different modalities that help the medical staff determine if a person is experiencing depression. Due to this, we can conclude that depression manifests in various signs, both verbal and non-verbal, as indicated by previous psychological studies providing insights into depression signs in text, verbal, and visual modalities. Motivated by these findings, we leverage individual sign symptoms through interest points per modality.

Although multiple works have focused on addressing the task of depression detection based on multimodal information in clinical interviews, the detection performance remains constrained due to the disparity among different modalities.

For this work, beyond examining individual modalities, we propose building a structured multimodal graph, focusing on the interest points of each modality. Inspired by psychologists' investigations into signs of depression, we intend to analyze the different interest points from the following three modalities: Audio, Video, Text.

4.1.1 Problem Formulation

In this work, multimodal data consist of audio, audio transcription, and images from clinical interviews based on the PHQ-8 questionnaire, focusing on the interest points. It is expressed as follows:

$$I_p = \{I_a, I_t, I_v\}$$

Where $I_x = P_{x,1}, \ldots, P_{x,n}$ represents the interest points of each interview, and $x = a, t, v$ represents the modalities, which in our case are: audio, video, and text.

The goal of this study is to identify people with depression and determine its severity using the "interest point per modality."

To pursue this proposal using multimodal data, we suggest using a Multimodal Graph Neural Network, as illustrated in Figure 1.

Moreover, incorporating explainability into our model is essential for gaining insights into the decision-making process. This emphasis on explainability ensures that the model's decision-making is not treated as a black box but rather opens up to a better understanding of the factors contributing to the classification and severity assessments.

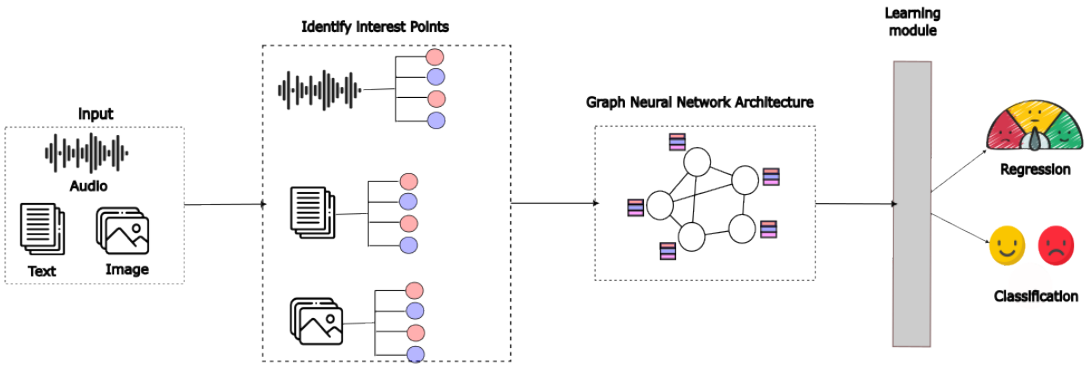

Figure 1: Multimodal Network Graph

4.1.2 Multimodal Graph Neural Network

Graph neural networks (GNNs) provide an expressive and flexible strategy to leverage interdependencies in multimodal datasets. In this work, we introduce the design of the multimodal Graph Neural Network following the methodology proposed in

[Ektefaie et al., 2023]. This methodology consists of four components:

- **Identify entities:** This involves identifying relevant entities, which can be translated into identifying the interest points from each modality and projecting them into a shared namespace.
- **Uncovering topology:** This step involves discovering the relationships and interactions among the nodes across the modalities through graph learning.
- **Propagation of information:** The third component employs convolutional or message-passing steps to learn node representations based on graph adjacencies.
- **Mixing representations:** The last component transforms learned node-level representations depending on downstream tasks. For example, a graph-level or a subgraph-level label. Popular mixing strategies include simple aggregation operators (e.g., summation or averaging) or more sophisticated functions that incorporate neural network architectures.

4.2 Methodology

In this work, our main objective is to focus on the detection of people with depression and assess the severity of their condition through semi-structured clinical interviews. The proposed work will center on creating a representation that exploits distinct points of interest for each modality and then combine them to form a multitask and interpretative model. In other words, we initiate the extraction of representative features from various modalities, such as audio, images, and text. Subsequently, using these features, we feed a model that will automatically learn how to combine them and extract the most pertinent information. This process ultimately results in a model capable of determining whether a person has depression, assessing its severity, and providing insights to aid in an accurate diagnosis.

1. **Identify and Obtain Datasets Related to Depression**: In this step, we plan to identify and obtain all datasets related to this work. The criteria for considering a dataset adequate for this work are as follows:

    - Should be a multimodal dataset,
    - The dataset should be collected under the scheme of semi-structured clinical interviews, and
    - The datasets have to be related to the depression task.

2. **Define and Develop Methods that Extract Interest Points in Different Modalities**: In this step, we propose to define the interest points per modality based on psychological evidence that can assist us in identifying depressed patients. Subsequently, extraction methods will be implemented for each of the selected interest points. This will help determine the most relevant methods and features for the task of depression detection. Some potential interest points for each modality could include:

    - **Text Modality:** Numerous studies have demonstrated that people with depression can be distinguished from others through written communication. For example, in [Zimmermann et al., 2017], the authors found that individuals with depression tend to use more first-person singular pronouns than non-depressed individuals. Additionally, people with depression use more negative emotion words [Rude et al., 2004] and tend to use more absolutist words [Bucci and Freedman, 1981].

      For the text modality there are numerous methods to extract both lexical and non-lexical characteristics however, Since we are focusing on work with semistructured questionnaires to the accurate diagnose of the depression disorder, we decide use some methods to find text similarity information of the interviews and the symptoms asked on the interviews to identify evidence of the presence of a symptom using Semantic Similar-

ity. The Semantic similarity indicates the proximity of meaning between two samples. Achieving semantic similarity involves converting samples into vectors, a process commonly referred to as embedding. This embedding can be applied to individual words or entire sentences. Various methods exist to enhance the quality of text embeddings. Options include employing TF-IDF, Word2Vec, Transformers sentence embeddings, Bag-of-Words, FastText, and other techniques. Also there are different semantic similarity measures such as:

  - Cosine similarity: It calculates the cosine angle between two n-dimensional vectors projected within a multi-dimensional space. The cosine distance consistently ranges from 0 to 1. A cosine similarity score of 1 signifies that two vectors share the same orientation. On the other hand, a value closer to 0 suggests lower similarity between the two documents.
  - Euclidean distance: Tthe Euclidian distance represents the length of a line segment between two points, which can be calculated by the Pythagorean Theorem. Therefore, in the NLP, these points are represented by words. According to the Euclidian distance, the shorter the distance between the two texts is, the more similar they are.

- **Visual Modality:** This includes body postures, gestures, facial expressions, and eye movements, among other factors [Waxer, 1974]. People with depression often avoid eye contact , generally exhibit fewer animated facial expressions [JH Balsters et al., 2012], and are more likely to hold their heads in a downward position and engage in self-touching activities (e.g., rubbing, scratching) compared to non-depressed individuals.

  Several models are used to capture different features linked to human behaviour in an video frame. These models are:

  - Facial Action Coding System (FACS) [Jacob and Stenger, 2021]: This

system describes a taxonomy of facial action units (FAU) for encoding facial expressions, based on the observed activation of muscles or muscle groups, such as Brow Lowerer or Cheek Raiser, by combining attention branch networks in a multi-task setting for focusing on the spatial regions of action units and merging branches for individual action units.

  - 3D Dense Face Alignment (3DDFA), [Zhu et al., 2017]: It serves as a comprehensive face alignment framework designed for extracting pose, depthmap, 3D model, and facial landmarks from a single facial image. 3DDFA is a robust approach capable of handling images of individuals captured from diverse angles. Additionally, it can predict the positions of facial features even when they are partially obscured due to the image's capture angle. The fundamental concept underlying 3DDFA involves optimizing model parameters—specifically, scale, rotation, translation, shape, and expression of a 3DMM face—to achieve a precise alignment with the input image.
  - ETH-XGaze [Zhang et al., 2020a]: It's a pre-trained model that capture 3D gaze vectors and head position. this model was trained over one million high-resolution images of varying gaze under extreme head poses.

- **Audio Modality:** Nonverbal communication includes how we talk beyond the actual words we use [Zhou, 2005]. This involves things like how fast or slow we speak, the pitch of our voice, how loud we are, and when we pause. It also includes sounds we make, like changes in pitch or volume, and filler words in our speech [Giri, 2009].

  When we look at people with depression compared to those without, we notice some differences in how they communicate. People with depression tend to speak more slowly and take longer pauses between questions and

answers in interviews [Nilsonne, 1988]. They also often have a speaking style that lacks variation, sounding more monotone [Pedersen et al., 1988]. Additionally, individuals with depression consistently show abnormalities in aspects like loudness, pitch patterns, and stress variations in their speech [Darby et al., 1984].

To study these nonverbal aspects, we can use two sampling methods. First, we divide the audio into segments, and then we use a pre-trained model to identify important features.

- VGGish [Hershey et al., 2017]: Derived from the well-regarded VGG network renowned for image recognition, VGGish has demonstrated its efficacy in capturing audio patterns. It transforms raw audio signals into a compact feature representation, employing its hierarchical structure to extract both: fundamental acoustic features (e.g., mel-frequency cepstral coefficients or MFCCs) and higher-level semantic features like tone and pitch.
- HuBERT [Hsu et al., 2021]: Is constructed on the established Wave2Vec framework and utilizes self-supervised learning methods to obtain resilient representations from unannotated audio data. By employing a contrastive objective, HuBERT captures significant audio features that prove useful across a range of subsequent tasks. Its notable advantage is evident in its ability to glean knowledge from extensive amounts of unlabeled audio, allowing for effective generalization across diverse domains and languages.
- Emotional Prosody: It is characterized as the combination of both segmental and supra-segmental alterations (relating to melodic elements) in our speech production when experiencing emotions. This phenomenon serves as a connection between language and emotion, as highlighted in the study by [Filippa et al., 2022]. Various emo-

tional prosody categories have been identified to align with a set of musical-like acoustic features, including rhythm, pitch, tone, amplitude, accent, pause, duration, and their progression.

For the audio and video modality, we propose using phrases from the interviews lexically related to the symptoms inquired during the interviews. This is an attempt to measure the presence of various depression indicators in both modalities and, thereby, distinguish between depressive and non-depressive patients. The extracted text fragments serve as valuable insights into the participants' psychological states. Using these fragments as anchors, we can analyze the way the phrase was pronounced and extract different audio features, as well as observe the expression of the patient at the moment they uttered the phrase. This allows for a more holistic understanding of the potential cues and manifestations associated with depression.

3. **Develop a model to create a representation that automatically combines the different interest points per modalit** This step involves integrating the multimodal interest points extracted from the previous step into a graph structure. Since Graph Neural Networks (GNNs) can be effectively used to model graph-structured data, including multimodal information where nodes or entities have multiple types of features. Here, we briefly describe how we plan to construct the graph neural network:

   - **Nodes:**
     - **Entities:** Nodes in a multimodal graph represent entities, such as objects, concepts, or individuals. The entities in our graph represent the eight symptoms from the PHQ-8 interview, which assesses depression symptoms. These symptoms are often measured on a scale of severity.

- **Multimodal Features:** Each node is associated with multiple types of features, capturing different aspects or modalities of information. For example, in this work, a node will represent the interest points from the three different modalities extracted from the previous step. We'll consider three modalities: text, audio, and video. Each symptom can have features associated with these modalities.

- **Edges:**
  - **Relationships:** Edges in the graph capture relationships between entities. These relationships can exist between entities of the same modality or between entities of different modalities. We are considering relationships like "aggravates" or "correlates with" to capture how symptoms may influence each other.
  - **Typed Edges:** Edges may be typed to indicate the type of relationship between nodes. For instance, there could be edges representing visual relationships, textual relationships, or any other relevant types.
    Depression is a complex mental health condition, and symptoms often co-occur. For example, sleep disturbance and loss of interest might mutually reinforce each other. Low energy levels can contribute to feelings of lethargy, exacerbating other symptoms, and cognitive difficulties, such as concentration problems, may impact decision-making and exacerbate feelings of guilt.

- **Adjacency Matrix:**
  - **Connectivity:** The adjacency matrix represents the connectivity of nodes in the graph. Each entry (i, j) in the matrix indicates whether there is a connection (edge) between nodes i and j.

- **Node Features:**
  - **Multimodal Feature Matrix:** Nodes are associated with a feature matrix where each row corresponds to a node, and each column cor-

responds to a feature. This matrix is usually composed of multiple submatrices, each capturing the features of a different modality.

  - **Feature Dimensionality:** The dimensionality of each feature vector may vary depending on the modality. For example, visual features could be a vector of pixel values, while textual features could be a vector of word embeddings.

- **Define Loss Function:** Since we are proposing a multitask model for classification and regression, we need to define an appropriate loss function based on the task:

  - **Regression:** Use a suitable loss function for regression, such as mean squared error (MSE).
  - **Classification:** Use cross-entropy as the loss function.

- **Graph Neural Network (GNN) Architecture:**

  - **Input Layer:** The input layer of the GNN should handle the multimodal features of nodes. There might be separate input channels or embedding layers for each modality.
  - **Aggregation and Propagation:** GNN layers aggregate information from neighboring nodes and propagate it through the graph. This process should consider both the node features and the types of edges connecting the nodes.
  - **Combining Modalities:** At some point in the architecture, features from different modalities need to be combined. This can be done through operations like concatenation, element-wise sum, or any other operation that preserves the multimodal nature of the data.

The GNN will automatically adjust these weights to capture the relationships and dependencies between different entities and modalities, allowing it to detect depression based on the multimodal information provided during the in-

terview.

4. **Design an interpretative model** In this final step, an analysis and exploration of the verbal and non-verbal behaviors per patient is proposed. It involves employing advanced models that focus on relevant multimodal interest points, previously defined by the medical community. By systematically capturing and highlighting repeated instances of specific phrases or non-verbal cues associated with a symptom, this model provide valuable insights to psychiatrists. This approach aids in a more thorough understanding of the patient's condition, enabling the psychiatrist to make more accurate diagnoses based on evidence. The emphasis here is on highlighting a patterns related to the specific symptoms recognized and defined by the medical community, thereby enhancing the diagnostic process.

4.3 Contributions

The main contribution of this work will be the efficient integration of multimodal data, such as text, images, and audio from depressive interviews. This integration will enable the model to capture the complexity and richness of information associated with depressive symptoms. Additionally, efforts will be made to develop effective graphical representations of this data, using graphs to model the relationships between different types of information. In this approach, graph nodes will represent the interest points of depressive symptoms, while edges will capture the connections and correlations between them. This method will provide an interpretable structure that reflects the interconnected nature of symptoms, thus allowing for a more comprehensible and explanatory approach to depression detection.

4.4 Work Schedule

In table 3 we can observe the work scheduled proposed for the doctoral period, this schedule includes the most relevant activities that are planned.

	2023						2024						2025						2026					
Activity	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Literature review																								
Identify and obtain datasets																								
Analyse dataset (Tex Modality)																								
Analyze and develop methods to extract information from text modality																								
Analyze preliminary results																								
Elaborate research proposal																								
Defense dissertation proposal																								
Analyze and develop methods to extract information from Audio modality																								
Research internship [3]																								
Analyze and develop methods to extract information from Video modality																								
Develop a method that automatically fuse modalities																								
Experiments and analysis of fusion approaches																								
Write Thesis																								
Revision and Thesis corrections																								
Dissertation Defense																								

	Completed activities
	Pending activities
	Pending activities to do during the intership

Table 3: Schedule of activities for the Doctoral thesis period divided by bimester.

[3] It will be an inter-ship in the University of Jaen, Spain from September 2024 to September 2025

5 Preliminary Results

In this section, we present the preliminary work and results that have been done, supporting our hypothesis and research proposal. The results presented here correspond to the first two steps of the methodology proposed in this work.

5.1 Selecting the dataset

As part of the first step of our methodology, we conduct a search for different datasets that meet all the requirements described in Step 1:"Identify and Obtain Datasets Related to Depression" of the proposed methodology. We found various datasets that fulfill these requirements, such as the MINI dataset [Sheehan et al., 1998]. This dataset was built under the Mini-International Neuropsychiatric Interview (M.I.N.I), a semi-structured diagnostic interview used to diagnose psychiatric disorders. However, this dataset only contains the transcriptions of the interviews, so we decided to discard its use since it is not a multimodal dataset.

Another identified dataset was used in [Cohn et al., 2009], where the data were collected under a semi-structured interview using the Hamilton questionnaire. Unlike the previous dataset, this one includes multimodal information as it contains both audio and transcription audio of the interviews. However, this dataset is not publicly available, so we do not have access to it.

Finally, we found the dataset used in the AVEC challenge from 2013, 2016, and 2019. We chose to work with the Extended Distress Analysis Interview Corpus (E-DAIC) dataset [DeVault et al., 2014], which was utilized in the AVEC 2019 challenge. This is the most recent dataset from the AVEC challenge, focusing on depression detection and, at the same time, being a multimodal dataset containing audio, video, and text transcriptions. The dataset comprises semi-clinical interviews designed to support the diagnosis of psychological conditions such as depression.

These interviews were collected with the goal of identifying verbal and nonverbal indicators or symptoms of this mental illness [Gratch et al., 2014].

The dataset contains three modalities: Audio, Video, and Audio transcriptions. These semi-structured interviews were conducted by a virtual agent. The dataset was partitioned into training, development, and test sets. In Table 4, we can see the statistics of this dataset. Additionally, this dataset includes detailed answers to each question on the PHQ-8 questionnaire [Kroenke et al., 2009].

Partition	# Subjects	Depressive	Non-Depressive
Training	163	37	126
Dev	56	12	44
Test	56	17	39

Table 4: Statics of E-DAIC dataset

As part of the second step of our proposed methodology, which consists of "defining and developing methods that extract interest points in different modalities," we focused on the text modality in our first experiment. To carry out this experiment, we defined the interest points for the text modality as the 8 symptoms asked on the PHQ-8 questionnaire. The PHQ-8 is recognized for its diagnostic accuracy in detecting major depressive disorders. The symptoms included in the questionnaire align with the criteria outlined in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), the authoritative guide for mental health professionals. This alignment ensures that the test effectively captures clinically relevant symptoms. Due to this, we can conclude that focusing on extracting information based on the 8 symptoms evaluated in this questionnaire provides enough indicators of an individual's mental well-being and offers valuable insights into the presence and severity of depressive symptoms.

Our first experiment consisted of evaluating our proposed approach for depression detection. This approach has two general steps: in the first step, we used a

language model resource like *Chat GPT-3* [OpenAI, 2022] and computed a series of phrases related to each symptom asked on the PHQ-8. In the second step, we used these generated phrases to represent the text using a histogram of their frequencies. This representation is named **Interest Points of Symptoms**. We can see a diagram of these two steps in Figure 2. In the next subsection, we further explain these two main steps.

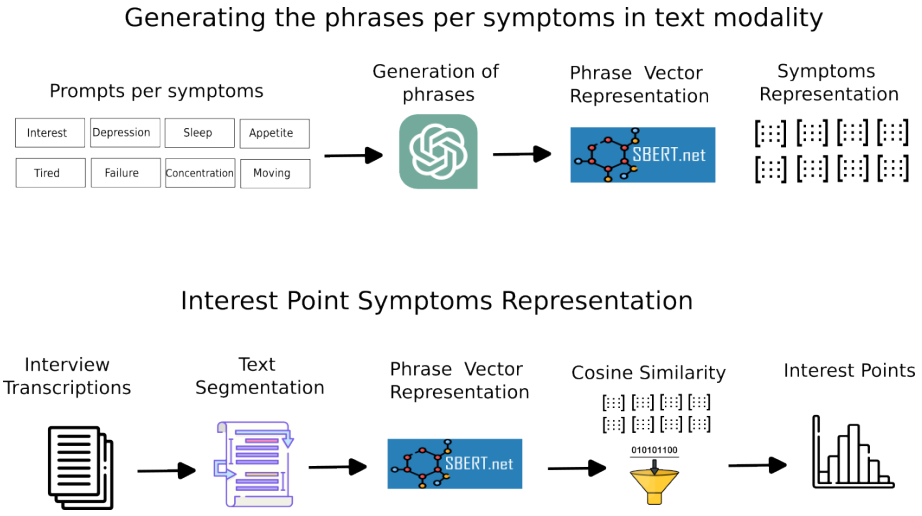

Figure 2: Diagram of the two main step for the creation of the Interest Points of Symptoms in the text modality

5.2 New representation for the interest points on text modality

5.2.1 Generating the phrases per symptoms in text modality

To generate these phrases, we use a language model as a resource. In this case, we chose *Chat GPT-3* based on the eight symptoms queried in the PHQ-8 interview: Interest, Depression, Sleep, Fatigue, Appetite, Failure, Concentration, and Movement. The prompts to generate sets of phrases were developed with the assistance of medical staff. In Table 5, you can observe the prompts that we defined to generate phrases in ChatGPT-3. The prompts were generated with the help of a medical

expert who assisted us in designing each prompt. Once these prompts were passed through ChatGPT-3, we obtained the phrases per symptom, as shown in Table 5.

Symptoms	Prompts	Phrases
Interest	Act as an individual experiencing depression, who is consulting with a psychiatrist and can only express their symptoms through written language. Your task is to replicate the typical most used 20 phrases given by individuals scoring between 10-14 on the PHQ-8 scale when answering the first question about feeling a lack of interest or pleasure in daily activities. Utilize the most frequently used phrases or their synonyms that are commonly associated with anhedonia so the synonyms are not repeated more than 2 times in all the 15 phrases that you give me	• I just can't seem to enjoy things like I used to. • Nothing really excites me anymore. • It's as if everything I used to love feels empty now. • Activities I used to find pleasurable don't bring joy anymore. • Life feels dull and uninteresting. • I can't muster any enthusiasm for anything. • I've lost my passion for the things I once loved. • It's hard to feel engaged in anything these days. • I'm struggling to find pleasure in daily activities. • Everything just seems monotonous and pointless. • Hobbies and interests no longer provide satisfaction. • I've become indifferent to the things I used to cherish. • The world seems colorless and unappealing. • Daily life feels like a never-ending chore. • I feel disconnected from what used to bring me happiness.
Depression	Act as an individual experiencing depression, who is consulting with a psychiatrist and who can only express their symptoms through written language. Your task is to replicate the typical most used 15 phrases given by individuals when answering the second question of the phq8 scale (do you feel down, depressed, or hopeless?) in a person with an average score of 10-14. Utilize the most frequently used phrases and their synonyms that are commonly associated with having dysphoria so the synonyms are not repeated more than 2 times in all the 15 phrases that you give me.	• I constantly feel down and blue. • It's like a cloud of sadness hanging over me. • I'm trapped in a never-ending feeling of sadness. • There's a deep sense of hopelessness within me. • Depression has become my constant companion. • My mood is persistently low and gloomy. • Every day is a struggle against this deep sadness. • I can't shake this overwhelming sense of despair. • It's as if I'm stuck in a pit of depression. • A heavy weight of melancholy drags me down. • My spirit is constantly weighed down by sadness. • There's this unrelenting feeling of unhappiness. • I feel an unshakable sense of despondency. • It's like I'm drowning in a sea of depression. • A pervasive sense of gloominess engulfs me daily.

Insomnia	Act as an individual experiencing depression, who is consulting with a psychiatrist and who can only express their symptoms through written language. Your task is to replicate the typical most used 20 phrases given by individuals scoring between 10-14 on the PHQ-8 scale when answering the third question about having trouble falling or staying asleep (it can be initial insomnia, middle insomnia, or terminal insomnia). Utilize the most frequently used synonyms that are commonly associated with insomnia so the synonyms are not repeated more than 2 times in all the 15 phrases that you give me.	• I find it increasingly difficult to fall asleep. • My nights are plagued by restless, sleepless hours. • Sleep eludes me, night after night. • I spend hours tossing and turning in bed. • I can't seem to get a good night's rest. • My mind races, and I can't relax enough to sleep. • I lie in bed, wide awake, unable to doze off. • It's a struggle to get any shut-eye these days. • I'm plagued by frequent awakenings during the night. • The hours tick away as I struggle to sleep. • Insomnia has become a nightly battle. • I'm tormented by chronic sleeplessness. • I wake up in the middle of the night, unable to fall back asleep. • My sleep patterns are completely disrupted. • Sleep deprivation is taking a toll on me.
Hypersonia	Act as an individual experiencing depression, who is consulting with a psychiatrist and who can only express their symptoms through written language. Your task is to replicate the typical most used 20 phrases given by individuals scoring between 10-14 on the PHQ-8 scale when answering the third question about sleeping too much. Utilize the most frequently used synonyms that are commonly associated with hypersomnia so the synonyms are not repeated more than 2 times in all the 15 phrases that you give me.	• I find myself sleeping excessively. • My days are filled with unending drowsiness. • I'm overwhelmed by constant fatigue. • It feels like I'm always in need of more sleep. • Hypersomnia has taken over my life. • I can't seem to get enough rest. • I spend most of my time in a state of drowsiness. • I struggle to stay awake during the day. • Sleeping too much has become a daily struggle. • It's as if I'm always in a haze of tiredness. • I can't break free from the cycle of oversleeping. • Fatigue clings to me like a heavy weight. • I'm constantly battling with excessive sleep. • My life is dominated by an overwhelming need to nap. • Hypersomnia has become a constant companion.
Tired	Act as an individual experiencing depression, who is consulting with a psychiatrist and who can only express their symptoms through written language. Your task is to replicate the typical most used 20 phrases given by individuals scoring between 10-14 on the PHQ-8 scale when answering the fourth question about feeling tired or having little energy to do daily activities. Utilize the most frequently used synonyms that are commonly associated with energy and fatigue so that it is not repeated more than 2 times in all the 15 phrases that you give me	• I'm constantly drained of energy. • Fatigue clings to me like a heavy weight. • It feels like I'm running on empty. • I struggle to find the energy to get through the day. • My days are marked by an overwhelming lack of vitality. • I have little motivation or enthusiasm for anything. • It's as if my energy has completely disappeared. • Daily activities leave me feeling utterly exhausted. • I can't seem to shake this persistent tiredness. • Life is an uphill battle due to my low energy. • Every task takes an immense effort to complete. • I lack the strength and vigor to do much of anything. • I'm constantly in a state of physical and mental exhaustion. • It's like there's a constant cloud of fatigue hanging over me. • My body and mind feel completely depleted.

Lost of Appetite	Act as an individual experiencing depression, who is consulting with a psychiatrist and who can only express their symptoms through written language. Your task is to replicate the typical most used 20 phrases given by individuals scoring between 10-14 on the PHQ-8 scale when answering the fifth question about having poor appetite and changes in appetite. Utilize the most frequently used synonyms that are commonly associated with hypophagia so that they are not repeated more than 2 times in all the 15 phrases that you give me.	• My appetite has almost disappeared. • I've lost interest in eating altogether. • Food no longer holds any appeal for me. • My desire to eat has significantly diminished. • It's as if I've completely lost my appetite. • I find it difficult to muster any interest in food. • Eating has become a chore I'd rather avoid. • The thought of food doesn't excite me anymore. • My appetite has dwindled to almost nothing. • I've experienced a drastic reduction in hunger. • Meals no longer bring me pleasure. • Food doesn't hold the same importance for me. • It's hard to eat when I have no appetite. • I'm constantly battling with a lack of interest in food. • My relationship with eating has changed dramatically.
Overeating	Act as an individual experiencing depression, who is consulting with a psychiatrist and who can only express their symptoms through written language. Your task is to replicate the typical most used 20 phrases given by individuals scoring between 10-14 on the PHQ-8 scale when answering the fifth question about overeating and changes in appetite. Utilize the most frequently used synonyms that are commonly associated with hyperphagia so that they are not repeated more than 2 times in all the 15 phrases that you give me.	• I can't seem to control my overeating. • My appetite has grown uncontrollable. • I find myself constantly indulging in food. • Food has become a source of comfort for me. • I'm consuming more than usual due to my emotions. • My eating habits have changed significantly. • It's like I can't stop myself from overeating. • I turn to food as a way to cope with my feelings. • I'm experiencing a significant increase in hunger. • Overeating has become a regular occurrence for me. • I'm using food as a way to numb my emotions. • My relationship with food has become unhealthy. • I constantly crave and indulge in large meals. • I'm turning to food to fill an emotional void. • My appetite is no longer under my control.
Failure	Act as an individual experiencing depression, who is consulting with a psychiatrist and who can only express their symptoms through written language. Your task is to replicate the typical most used 20 phrases given by individuals scoring between 10-14 on the PHQ-8 scale when answering the sixth question about Feeling bad about yourself or that you are a failure or have let yourself or your family down. Utilize the most frequently used synonyms that are commonly associated with failure so that they are not repeated more than 2 times in all the 15 phrases that you give me.	• I often feel like a disappointment. • There's a constant sense of failure in me. • I can't help but think I've let everyone down. • I carry an overwhelming burden of guilt. • It's as if I'm constantly falling short. • I see myself as a perpetual underachiever. • Self-doubt and worthlessness consume me. • I'm plagued by feelings of inadequacy. • Guilt and self-criticism are my constant companions. • I feel like a letdown to myself and my family. • There's a relentless inner voice telling me I'm a failure. • My self-esteem is at an all-time low. • I'm haunted by a sense of not measuring up. • The weight of my perceived failures is heavy. • I'm overwhelmed by feelings of worthlessness.

Concentrating	Act as an individual experiencing depression, who is consulting with a psychiatrist and who can only express their symptoms through written language. Your task is to replicate the typical most used 15 phrases given by individuals scoring between 10-14 on the PHQ-8 scale when answering the seventh question about having trouble concentrating on daily activities (mention some daily activities). Utilize the most frequently used synonyms that are commonly associated with concentration so that they are not repeated more than 2 times in all the 15 phrases that you give me.	• It's a constant struggle to focus on work or study. • My mind feels foggy and unfocused. • I find it hard to concentrate on simple tasks. • Daily activities like reading or cooking become overwhelming. • My attention span has significantly shortened. • Completing even basic chores feels like a challenge. • I can't keep my mind on anything for long. • Concentrating on conversations is exhausting. • I frequently lose track of what I'm doing. • My work productivity has taken a hit. • Planning and organizing have become difficult. • It's challenging to follow through with commitments. • I struggle to stay engaged in conversations. • I often make mistakes due to lack of concentration. • Even watching TV or movies feels mentally taxing.
Moving to slow	Act as an individual experiencing depression, who is consulting with a psychiatrist and who can only express their symptoms through written language. Your task is to replicate the typical most used 20 phrases given by individuals scoring between 10-14 on the PHQ-8 scale when answering the eighth question about moving or speaking so slowly that other people could have noticed. Utilize the most frequently used synonyms that are commonly associated with slow so that they are not repeated more than 2 times in all the 15 phrases that you give me.	• I've noticed that my movements have become sluggish. • My pace in daily life has significantly slowed down. • It's as if time is passing me by while I move slowly. • People around me have commented on my lack of speed. • I feel like I'm always a step behind everyone else. • My actions seem delayed and cumbersome. • My speech has become noticeably slow. • Conversations feel like they're happening in slow motion. • I'm aware that I'm not as quick as I used to be. • My lack of speed is causing me frustration. • I can see that others are more agile and responsive. • Even simple tasks take an eternity to complete. • It's like I'm in a perpetual state of slowness. • My movements are marked by a heavy sense of lethargy. • I've been told I come across as sluggish in my actions.

Table 5: Prompts designed to generate the phrases in ChatGPT-3

5.2.2 Phrase Representation

Once we obtained the phrases per symptoms, we built their *Interest Points of Symptoms* representation. The phrases obtained for ChatGPT-3 are represented in a formal way as $S = S_1, S_2, ..., S_n$, where S is the set of symptoms presented in the interview transcription, and $S_i = p_1, ..., p_n$ is the set of phrases related to each

symptom S_i generated by ChatGPT-3. Then, we computed a phrase vector using a pre-trained transformer of size 300 from *SentenceTransformer* [Reimers and Gurevych, 2019]. These vectors were trained using SNLI [Bowman et al., 2015] and MultiNLI [Williams et al., 2018] collections.

The phrase representations were obtained by computing the frequency of the phrases related to each symptom. These histograms were created by counting the number of occurrences of each interest point presented in the text using cosine similarity. This process is similar to the Bag-of-Words representation.

5.2.3 Similarity Threshold Selection

Since we use the cosine similarity, we have a range of threshold that was from $[0 to 1]$. To select the best threshold for this representation, we evaluate different range of threshold of cosine similarity. The results of this experiment can see observed in Figure 3. For this experiment we use Support Vector Machine (SVM) with linear kernel and C=1 ans class weight=balanced, since we are dealing with class imbalance.

As we observed in image below, the best results were obtained, for all metrics, was using the representation with cosine similarity of 0.2. To compare the performance of this results, we compared this representation against some other text representation.

5.2.4 Comparison with baselines and State of Art Works

To evaluate the effectiveness of the proposed approach, which focuses on "interest points" in the text modality, we compared it with other models using transformers, as well as other representations utilizing emotions in the text modality.

In Table 6, we can observe the performance of the proposed methods against some transformer models and lexical resources in the depression class. As evident,

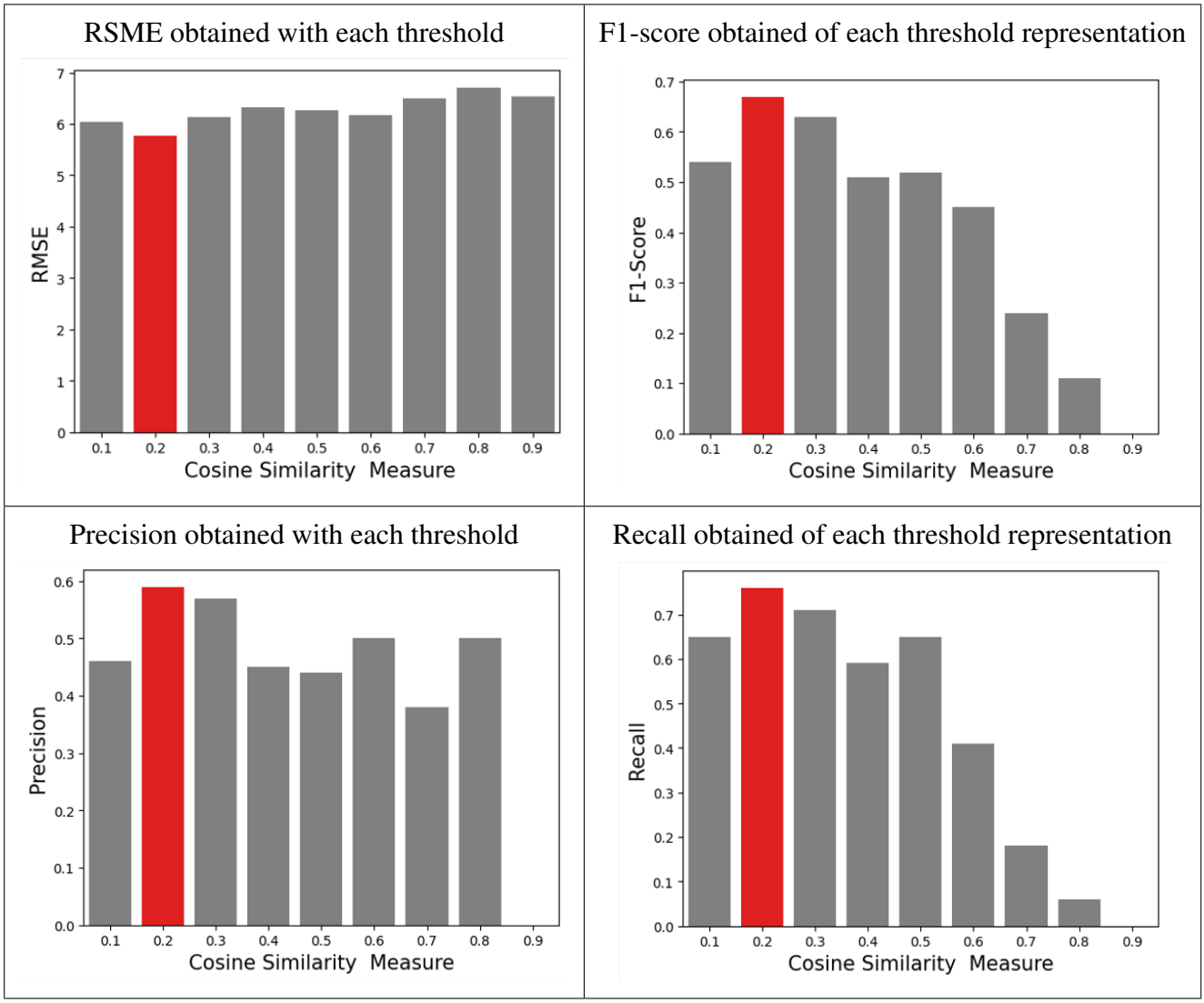

Figure 3: Metrics results obtained of each threshold representation. The red bar represent the best result achieved for each evaluation metric.

the proposed model achieves better results in F1 measure, precision, and RSME, demonstrating the effectiveness of the approach. This indicates that concentrating solely on interest points, such as the symptoms evaluated in the questionnaire, provides sufficient information to the model for identifying the depression disorder.

We also decided to compare the results of this representation against some state-of-the-art works. The comparison can be observed in Table 7.

Model	Precision	Recall	F1-score	RMSE
Interest Points Representation	**0.59**	0.76	**0.67**	**5.7**
Bert	0.45	0.59	0.51	7.8
MentalBert	0.56	0.53	0.55	6.8
Longformer	0.30	**0.88**	0.45	6.5
Depress Roberta	0.36	0.53	0.43	7.7
Vader	0.35	0.76	0.48	**5.7**
Empath	0.48	0.71	0.57	6.6

Table 6: Results over the positive class agings some popular models

Model	Precision	Recall	F1-score	RMSE
Interest Points Representation	0.59	0.76	0.67	5.7
Baseline (E-DAIC)	-	-	-	6.37
[Ray et al., 2019]	-	-	-	**4.28**
[Yin et al., 2019]	-	-	-	5.50
[Sun et al., 2021]	-	-	-	4.60
[Ghadiri et al., 2022]	-	-	-	4.32
[Burdisso et al., 2023]	-	-	**0.74**	-

Table 7: Results over the positive class agings SOA works

From the table 7 below we can highlight some points:

- We obtain results very close to the state of the art using a simpler representation, utilizing information from only one modality. This is in comparison to some works presented in this table that employ more complex models and incorporate bimodal and multimodal features.
- Focusing on the "Interest Point" results in a rich representation that provides the model with enough information to effectively distinguish between depressed and non-depressed individuals.

5.3 Results analysis

Prompt Design: During the design of prompts for generating sentences in ChatGPT-3, we experimented with two types of prompts. One involved providing detailed context and additional information to ChatGPT-3, while the other was simpler. In Table 8, you can observe the comparison of both prompts for the symptoms of interest and depression.

As we can see in the table above, the second prompt provides more accurate and detailed information to simulate a real scenario of an interview. Therefore, we generate the symptom representation as detailed in section 5.2.2 with these two kinds of prompts for each of the eight symptoms. The results can be seen in the table below:

As seen in Table 9, we achieved better results when using phrases generated with detailed prompts. Thus, we can conclude that it is necessary for the prompt to be as detailed as possible to obtain sentences that closely resemble what a patient might say in a real interview. This will help identify phrases associated with each symptom within the interviews.

Retrieved Phrases by Symptom: In Table 10, we can observe the phrases retrieved by our model for each symptom. As we can see, for some depressed individuals, there are various phrases that reflect the presence of the symptom. On the other hand, for non-depressed patients, the retrieved phrases are not as accurate as what the doctors observed during the interviews.

Symptoms	Depression Phrases	Non-Depression Phrases

Interest	• and just stay active stay healthy • outdoors in active and healthy • hiking biking rollerblading skateboarding • if I could I'm sure I would enjoy seeing all the different cultures and places and meeting all the different people and all the interesting structures • and I skateboard • I'm going to take it helps me relax • because if I don't really feel like skating hard I'll just go for a nice Sunset skate and meditate nice to our skate and just relax • not that hard you just got to get to the skate spots you know I usually like skating or some good good sweet skate spots in La streets in LA • are the beaches always good • why is sleep a lot lack of Interest	• I love it • I like the weather I like the opportunities • I took up business and administration • yeah I am here and there I'm on a break right now but I plan on going back in the next semester • probably to open up my own business • I like reading books Behringer I enjoy cooking sizing is great • I like to play sports • I enjoy I'm going out with friends and family • well I got my diploma • I finished school and I have met all the requirements • do whatever I wanted to do
Depression	• what are some things that make you really mad • the situation with my life right now • can't find a fucking job • it can be tough to find a good job these days I don't even care about it • anything that pays • I can't find a job I can't get a job so • I try I'm trying I'm trying I'm trying I'm trying I'm trying I'm trying • do you feel down • depressing • hard sucks • I just I don't know if I have what it takes to continue to do • I survived day by day • oh yeah I've always felt depressed on my life • I just missed her and I want to be with her • I'd like to give up but	• I enjoy I'm going out with friends and family • yeah I mean they've always gave me great advice
Insomnia	• how easy is it for you to get a good night sleep • desert way often • last night I couldn't sleep • just thinking about my situation • why is sleep a lot lack of Interest	

Tired	• I was always feeling down and depressed and lack of energy always wanted'	• lazy
Appetite	• my appetite was uncontrollable either lack of or I will should mean gluttonous and eating the wrong things	• I like reading books Behringer I enjoy cooking sizing is great
Failure	• the situation with my life right now • can't find a fucking job • I can't find a job I can't get a job so", " I try I'm trying I'm trying I'm trying I'm trying I'm trying I'm trying • I just I don't know if I have what it takes to continue to do • I just haven't had good luck • I didn't know I'd be here • I'm actually I don't have what it takes to stay here for two more weeks"]	• I'm not sure I've been I graduated from high school • well I got my diploma • I finished school and I have met all the requirements • do whatever I wanted to do • it's all right it could be better
Concentrating	• which one of your most memorable experiences • I don't know it's hard to say • can't really answer that one	• Even watching TV or movies feels mentally taxing.
Moving	• I was always feeling down and depressed and lack of energy always wanted	

Table 10: Example of retrieved phrases of a depressive and non-depressive patient

As we can see in the table below, the depression patients have more phrases for each symptom compared to the non-depressive phrases. Additionally, we observe that some phrases clearly indicate the presence of a symptom, such as "I've always felt depressed in my life." This is a clear indicator of depression. It's also important to note that some retrieved phrases appear in one or more symptoms. We suppose that this happens due to the similarity between some symptoms, like "problems

Symptom	Simpler Prompt	Detailed Prompt
Interest	Common phrases to express little interest or pleasure in doing things	Act as an individual experiencing depression, who is consulting with a psychiatrist and can only express their symptoms through written language. Your task is to replicate the typical most used 20 phrases given by individuals scoring between 10-14 on the PHQ-8 scale when answering the first question about feeling a lack of interest or pleasure in daily activities. Utilize the most frequently used phrases or their synonyms that are commonly associated with anhedonia so the synonyms are not repeated more than 2 times in all the 15 phrases that you give me
Depression	Common phrases to express feeling down, depressed, irritable and hopeless	Act as an individual experiencing depression, who is consulting with a psychiatrist and who can only express their symptoms through written language. Your task is to replicate the typical most used 15 phrases given by individuals when answering the second question of the phq8 scale (do you feel down, depressed, or hopeless?) in a person with an average score of 10-14. Utilize the most frequently used phrases and their synonyms that are commonly associated with having dysphoria so the synonyms are not repeated more than 2 times in all the 15 phrases that you give me.

Table 8: Comparison between different prompts

sleeping" and "feeling tired," which are not the same but share some feelings.

Finally, the segmentation used to separate the phrases is taken directly from

Prompt	Precision	Recall	F1	RSME
Simpler Prompt	**0.59**	0.65	0.59	5.9
Detailed Prompt	**0.59**	**0.76**	**0.67**	**5.77**

Table 9: Results obtained from using both representations, the simpler and the detailed.

Phrases	Symptoms
I don't enjoy traveling, because I spend long. Of time doing math and reading and not talking anybody message., I guess I play video games or it'll work out I work out I think a lot I try to at least	Interest
I'm not so pretty happy right now	Depression
horrified couldn't sleep, yeah I'm a lot more tired I sleep longer	Sleep
I'm tired and I kind of fall asleep during class and what not	Tired
' my appetite was uncontrollable either lack of or I will should mean gluttonous and eating the wrong things'	Appetite
the situation with my life right now, can't find a fucking job, I can't find a job I can't get a job so, I try I'm trying I'm trying I'm trying I'm trying I'm trying I'm trying, I just I don't know if I have what it takes to continue to do, I just haven't had good luck, I'm actually I don't have what it takes to stay here for two more weeks	Failure
don't remember, can't think of anything, I have no idea I don't know, it's hard cuz I don't always maintain focus and motivation a lot of other things distract me	Concentration
I couldn't function I couldn't drive I couldn't sleep I couldn't eat at my I couldn't do anything I was completely, shut down I felt like I felt like I was looking out a window and in life was going on outside that window and I was just stuck behind the glass III couldn't participate	Moving

Table 11: Example of the annotated interview at phrase level for each of the eight symptoms asked in the interview

the transcription provided in the dataset. We did not make any other segmentation because some phrases are longer than others. We propose, for future work, to segment the phrases in a different way to try to achieve better segmentation and more accurate phrases.

Labeled Dataset: From the interview transcriptions, we identify the phrases for every interview as p_i and classify them into one of the eight symptoms asked in the interview $s = s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8$ where s_1: 'interest', s_2: 'depression', s_3: 'sleep', s_4: 'appetite', s_5: 'tired', s_6: 'failure', s_7: 'concentration', s_8: 'moving'. The expert manually annotates each phrase p_i and classifies it in one or more symptoms S_{pi}. An example of the annotated dataset can be observed in Table 11.

5.4 Discussion

The results presented in this document correspond to the first two steps of our methodology that are 1) The search and selection of an adequate dataset for depression with multimodal information and, 2) the extraction of the interest points focused on the text modality. From the results presented here, we can conclude that the preliminary findings underscore the significance of prioritizing interest points based on psychological evidence. The results reveal that using a simpler representation for these interest points yields superior results compared to utilizing complex models, such as transformers. This suggests that, for the specific task at hand, a focused and streamlined approach to modeling interviews can be more effective than developing complex architectures.

Moreover, when we compared our work against state-of-the-art models, our results demonstrate competitive performance. However, that state-of-the-art models often extract features from multiple modalities and use more sophistical methods such as attention mechanisms. While our approach are focused only in the text modality, it's essential to recognize the nuanced trade-offs involved. The comparative of our results to the state-of-the-art reinforces the effectiveness of our strategy, particularly when considering the interpretability gained through the adoption of a simpler representation for text-based interest points.

As part of our future work, the dataset is being labeled by an expert, as mentioned before. This aims to help us gain insights into the characteristics of the data and the underlying patterns to capture with machine learning models. It also aids in evaluating our model and identifying errors, inconsistencies, or biases in our system, ensuring that the machine learning model's predictions align with the intended objectives.

In our upcoming work, we plan to initiate the extraction of interest points from both audio and video modalities, similar to the text, using psychological evidence and

evaluating the eight symptoms assessed in the interview. Additionally, we intend to model the data as a graph structure to find and define the relations between interest points and the modalities, corresponding to step three of the methodology presented in this work.

6 Conclusions

Detecting depression can be challenging for several reasons, such as the subjectivity of symptoms, as they can manifest differently among people. Additionally, inadequate access to mental health services and a shortage of mental health professionals contribute to delays in accurate diagnosis and treatment. For this reason, numerous works have focused on depression detection. However, most of these works concentrate on verbal indicators, ignoring the non-verbal symptoms that depressed people may present. Furthermore, many of these works use social networks as a data source to construct their models, disregarding diagnostic protocols established by the medical and psychological community. In this work, we focus on using multimodal data, including Audio, Video, and Text from clinical interviews based on the PHQ-8 questionnaire, to detect depression.

This work proposes a solution that leverages psychological evidence to extract depression symptoms, or what we call "interest points," encompassing both verbal and non-verbal information. By incorporating a diverse set of features, our approach seeks to capture the different indicators of depression, addressing the inherent complexity of the problem. This involves incorporating the utilization of graph structure to model the data and the relationships between different aspects of depression. This aims to obtain a model that detects the depression status and its severity in a person, providing insights into the main symptoms or behaviors behind the model's predictions based on the identified interest points.

Preliminary results presented in this document, focusing on the text modality, demonstrate the feasibility of our proposed solution. The findings underscore the effectiveness of prioritizing interest points based on psychological support and adopting a simple approach to model interviews. As we expand our research to include other modalities, we anticipate further validation of the effectiveness of this approach.

References

[WHO, 2019] (2019). Mental disorders. World Health Organization. Available at `https://www.who.int/news-room/fact-sheets/detail/mental-disorders`.

[WHO, 2020] (2020). Depressión. World Health Organization. Available at `https://www.who.int/es/news-room/fact-sheets/detail/depression`.

[Beck et al., 1961] Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., and Erbaugh, J. (1961). An inventory for measuring depression. *Archives of general psychiatry*, 4(6):561–571.

[Benton et al., 2017] Benton, A., Mitchell, M., and Hovy, D. (2017). Multi-task learning for mental health using social media text. *arXiv preprint arXiv:1712.03538*.

[Bowman et al., 2015] Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

[Bucci and Freedman, 1981] Bucci, W. and Freedman, N. (1981). The language of depression. *Bulletin of the Menninger Clinic*.

[Burdisso et al., 2023] Burdisso, S., Villatoro-Tello, E., Madikeri, S., and Motlicek, P. (2023). Node-weighted graph convolutional network for depression detection in transcribed clinical interviews. *arXiv preprint arXiv:2307.00920*.

[Caicedo et al., 2020] Caicedo, R. W. A., Soriano, J. M. G., and Sasieta, H. A. M. (2020). Assessment of supervised classifiers for the task of detecting messages with suicidal ideation. *Heliyon*, 6(8):e04412.

[Cohn et al., 2009] Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T., Zhou, F., and De la Torre, F. (2009). Detecting depression from

facial actions and vocal prosody. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7. IEEE.

[Cummins et al., 2017] Cummins, N., Vlasenko, B., Sagha, H., and Schuller, B. (2017). Enhancing speech-based depression detection through gender dependent vowel-level formant features. In *Artificial Intelligence in Medicine: 16th Conference on Artificial Intelligence in Medicine, AIME 2017, Vienna, Austria, June 21-24, 2017, Proceedings 16*, pages 209–214. Springer.

[Darby et al., 1984] Darby, J. K., Simmons, N., and Berger, P. A. (1984). Speech and voice parameters of depression: A pilot study. *Journal of Communication Disorders*, 17(2):75–85.

[De Choudhury et al., 2013] De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7.

[DeVault et al., 2014] DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., and Morency, L.-P. (2014). Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS '14, page 1061–1068, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

[Du and Swamy, 2019] Du, K.-L. and Swamy, M. (2019). Combining multiple learners: Data fusion and ensemble learning. In *Neural Networks and Statistical Learning*, pages 737–767. Springer.

[Ebersbach et al., 2017] Ebersbach, M., Herms, R., and Eibl, M. (2017). Fusion methods for icd10 code classification of death certificates in multilingual corpora. In *CLEF (Working Notes)*, page 36.

[Ektefaie et al., 2023] Ektefaie, Y., Dasoulas, G., Noori, A., Farhat, M., and Zitnik, M. (2023). Multimodal learning with graphs. *Nature Machine Intelligence*, 5(4):340–350.

[Filippa et al., 2022] Filippa, M., Lima, D., Grandjean, A., Labbé, C., Coll, S., Gentaz, E., and Grandjean, D. (2022). Emotional prosody recognition enhances and progressively complexifies from childhood to adolescence. *Scientific Reports*, 12(1):17144.

[Ghadiri et al., 2022] Ghadiri, N., Samani, R., and Shahrokh, F. (2022). Integration of text and graph-based features for detecting mental health disorders from voice. *arXiv preprint arXiv:2205.07006*.

[Giri, 2009] Giri, V. N. (2009). Nonverbal communication theories. *Encyclopedia of communication theory*, pages 690–694.

[Gratch et al., 2014] Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, S., and Morency, L.-P. (2014). The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

[Guohou et al., 2020] Guohou, S., Lina, Z., and Dongsong, Z. (2020). What reveals about depression level? the role of multimodal features at the level of interview questions. *Information & Management*, 57(7):103349.

[Hall et al., 1995] Hall, J. A., Harrigan, J. A., and Rosenthal, R. (1995). Nonverbal behavior in clinician—patient interaction. *Applied and preventive psychology*, 4(1):21–37.

[Hamilton, 1960] Hamilton, M. (1960). A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry*, 23(1):56.

[Haque et al., 2018] Haque, A., Guo, M., Miner, A. S., and Fei-Fei, L. (2018). Measuring depression symptom severity from spoken language and 3d facial expressions. *arXiv preprint arXiv:1811.08592*.

[Hershey et al., 2017] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., et al. (2017). Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

[Hsu et al., 2021] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

[Jacob and Stenger, 2021] Jacob, G. M. and Stenger, B. (2021). Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689.

[JH Balsters et al., 2012] JH Balsters, M., J Krahmer, E., GJ Swerts, M., and JJM Vingerhoets, A. (2012). Verbal and nonverbal correlates for depression: a review. *Current Psychiatry Reviews*, 8(3):227–234.

[Joshi et al., 2013] Joshi, J., Dhall, A., Goecke, R., and Cohn, J. F. (2013). Relative body parts movement for automatic depression analysis. In *2013 Humaine*

association conference on affective computing and intelligent interaction, pages 492–497. IEEE.

[Kroenke et al., 2009] Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., and Mokdad, A. H. (2009). The phq-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1):163–173.

[Li et al., 2023] Li, J., Wang, X., Lv, G., and Zeng, Z. (2023). Graphmft: A graph network based multimodal fusion technique for emotion recognition in conversation. *Neurocomputing*, page 126427.

[Lopez-Otero et al., 2014] Lopez-Otero, P., Dacia-Fernandez, L., and Garcia-Mateo, C. (2014). A study of acoustic features for depression detection. In *2nd International Workshop on Biometrics and Forensics*, pages 1–6.

[Lopez-Otero et al., 2017] Lopez-Otero, P., Fernández, L. D., Abad, A., and Garcia-Mateo, C. (2017). Depression detection using automatic transcriptions of de-identified speech. In *INTERSPEECH*, pages 3157–3161.

[Ma et al., 2016] Ma, X., Yang, H., Chen, Q., Huang, D., and Wang, Y. (2016). Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 35–42.

[Mao et al., 2023] Mao, K., Wu, Y., and Chen, J. (2023). A systematic review on automated clinical depression diagnosis. *npj Mental Health Research*, 2(1):20.

[Montgomery and Åsberg, 1979] Montgomery, S. A. and Åsberg, M. (1979). A new depression scale designed to be sensitive to change. *The British journal of psychiatry*, 134(4):382–389.

[Morrison, 2015] Morrison, J. (2015). *DSM-5® Guía para el diagnóstico clínico*. Editorial El Manual Moderno.

[Mueller and Segal, 2014] Mueller, A. E. and Segal, D. L. (2014). Structured versus semistructured versus unstructured interviews. *The encyclopedia of clinical psychology*, pages 1–7.

[Namboodiri and Venkataraman, 2019] Namboodiri, S. P. and Venkataraman, D. (2019). A computer vision based image processing system for depression detection among students for counseling. *Indonesian Journal of Electrical Engineering and Computer Science*, 14(1):503–512.

[Nilsonne, 1988] Nilsonne, A. (1988). Speech characteristics as indicators of depressive illness. *Acta Psychiatrica Scandinavica*, 77(3):253–263.

[OpenAI, 2022] OpenAI (2022). Chatgpt-3: Language model by openai. `https://www.openai.com/`.

[Pedersen et al., 1988] Pedersen, J., Schelde, J., Hannibal, E., Behnke, K., Nielsen, B., and Hertz, M. (1988). An ethological description of depression. *Acta psychiatrica scandinavica*, 78(3):320–330.

[Qin et al., 2023] Qin, Z., Zhao, P., Zhuang, T., Deng, F., Ding, Y., and Chen, D. (2023). A survey of identity recognition via data fusion and feature learning. *Information Fusion*, 91:694–712.

[Ramachandram and Taylor, 2017] Ramachandram, D. and Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108.

[Ray et al., 2019] Ray, A., Kumar, S., Reddy, R., Mukherjee, P., and Garg, R. (2019). Multi-level attention network using text, audio and video for depression prediction. In *Proceedings of the 9th international on audio/visual emotion challenge and workshop*, pages 81–88.

[Reimers and Gurevych, 2019] Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019*

Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

[Richter et al., 2021] Richter, T., Fishbain, B., Richter-Levin, G., and Okon-Singer, H. (2021). Machine learning-based behavioral diagnostic tools for depression: advances, challenges, and future directions. *Journal of Personalized Medicine*, 11(10):957.

[Rude et al., 2004] Rude, S., Gortner, E.-M., and Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*.

[Saggu et al., 2022] Saggu, G. S., Gupta, K., Arya, K., and Rodriguez, C. R. (2022). Depressnet: A multimodal hierarchical attention mechanism approach for depression detection. *Int. J. Eng. Sci.*, 15(1):24–32.

[Sezgin, 2023] Sezgin, E. (2023). Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. *Digital Health*, 9:20552076231186520.

[Sheehan et al., 1998] Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G. C., et al. (1998). The mini-international neuropsychiatric interview (mini): the development and validation of a structured diagnostic psychiatric interview for dsm-iv and icd-10. *Journal of clinical psychiatry*, 59(20):22–33.

[Shen et al., 2013] Shen, Y.-C., Kuo, T.-T., Yeh, I.-N., Chen, T.-T., and Lin, S.-D. (2013). Exploiting temporal information in a two-stage classification framework for content-based depression detection. In *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part I 17*, pages 276–288. Springer.

[Shin et al., 2022] Shin, D., Kim, K., Lee, S.-B., Lee, C., Bae, Y. S., Cho, W. I., Kim, M. J., Park, C. H. K., Chie, E. K., Kim, N. S., et al. (2022). Detection of

depression and suicide risk based on text from clinical interviews using machine learning: possibility of a new objective diagnostic marker. *Frontiers in psychiatry*, 13.

[Skaik and Inkpen, 2020] Skaik, R. and Inkpen, D. (2020). Using social media for mental health surveillance: A review. *ACM Computing Surveys (CSUR)*, 53(6):1–31.

[Snoek et al., 2005] Snoek, C. G., Worring, M., and Smeulders, A. W. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402.

[Sobin and Sackeim, 1997] Sobin, C. and Sackeim, H. A. (1997). Psychomotor symptoms of depression. *American Journal of Psychiatry*, 154(1):4–17.

[Sun et al., 2021] Sun, H., Liu, J., Chai, S., Qiu, Z., Lin, L., Huang, X., and Chen, Y. (2021). Multi-modal adaptive fusion transformer network for the estimation of depression level. *Sensors*, 21(14).

[Waxer, 1974] Waxer, P. (1974). Nonverbal cues for depression. *Journal of Abnormal Psychology*, 83(3):319.

[Williams et al., 2018] Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

[Wu et al., 2020] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24.

[Yin et al., 2019] Yin, S., Liang, C., Ding, H., and Wang, S. (2019). A multi-modal hierarchical recurrent neural network for depression detection. In *Proceedings of*

the 9th International on Audio/Visual Emotion Challenge and Workshop, pages 65–71.

[Zhang et al., 2020a] Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., and Hilliges, O. (2020a). Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 365–381. Springer.

[Zhang et al., 2020b] Zhang, Z., Lin, W., Liu, M., and Mahmoud, M. (2020b). Multimodal deep learning framework for mental disorder recognition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 344–350. IEEE.

[Zhou et al., 2020] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI open*, 1:57–81.

[Zhou, 2005] Zhou, L. (2005). An empirical investigation of deception behavior in instant messaging. *IEEE transactions on professional communication*, 48(2):147–160.

[Zhu et al., 2017] Zhu, X., Liu, X., Lei, Z., and Li, S. Z. (2017). Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*.

[Zimmermann et al., 2017] Zimmermann, J., Brockmeyer, T., Hunn, M., Schauenburg, H., and Wolf, M. (2017). First-person pronoun use in spoken language as a predictor of future depressive symptoms: Preliminary evidence from a clinical sample of depressed patients. *Clinical psychology & psychotherapy*.

7 Appendix

7.1 GNN's Architecture

The architecture of a GNN consists of multiple layers, each responsible for aggregating and updating information from neighboring nodes. The core idea behind GNNs is the "message-passing" paradigm, where information is exchanged between nodes during the training process [Zhou et al., 2020]. Here's an overview of the typical architecture components of GNNs:

- **Node Features**: Each node in the graph has associated features, which can include information about the node itself. These features serve as the input to the GNN.
- **Graph Structure**: The graph structure defines the relationships between nodes. It can be represented by an adjacency matrix or an edge list, capturing the connections in the graph.
- **Message Passing**: The core operation in GNNs is message passing. During each layer of the network, nodes exchange information (messages) with their neighbors. This allows nodes to aggregate information from their local neighborhoods.
- **Neighborhood Aggregation**: Aggregation functions (e.g., sum, mean, or attention-weighted sum) combine the messages received from neighboring nodes. This aggregation process helps each node gather information about its surroundings.
- **Update Function**: The aggregated information is then used to update the node's own features. This update function captures the node's refined representation based on both its own features and the information gathered from neighbors.

- **Multiple Layers**: GNNs typically consist of multiple layers, each performing message passing, aggregation, and feature update. Multiple layers enable the model to capture complex dependencies and patterns in the graph.
- **Graph Pooling** (Optional): In some cases, graph pooling layers may be used to downsample the graph. This operation reduces the graph size, making it computationally more efficient. Pooling is often analogous to downsampling in image-based convolutional neural networks.
- **Readout/Aggregation** (Optional): At the end of the GNN layers, a readout or aggregation function can be applied to obtain a graph-level representation. This aggregated representation can be used for various downstream tasks, such as graph classification.
- **Task-Specific Layers**: Finally, task-specific layers (e.g., fully connected layers) may be added for specific prediction tasks, such as node classification or graph classification.

By iteratively repeating the message passing and node update steps, the GNN enables information to propagate across the entire graph, allowing nodes to learn and refine their embeddings collectively.

GNNs come in various forms, each tailored to handle specific types of graph-structured data. Some common types of GNNs include:

- **Graph Convolutional Networks (GCNs)**: GCNs are one of the earliest and most widely used GNN variants. They leverage graph convolutions to aggregate and update node representations based on their local neighborhood
- **GraphSAGE**: GraphSAGE (Graph Sample and Aggregated) is another popular GNN architecture. It utilizes a sampling strategy to aggregate and update node embeddings, allowing for scalability in large graphs

- **Graph Attention Networks (GATs)**: GATs introduce attention mechanisms into GNNs, enabling nodes to selectively attend to relevant neighbors while performing message passing and aggregation
- **Graph Isomorphism Networks (GINs)**: GINs focus on capturing the structural information of graphs by applying permutation-invariant functions during the message passing and node update steps