Normandie Université

UNIVERSITÉ
CAEN
NORMANDIE

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité **INFORMATIQUE**

Préparée au sein de l'**Université de Caen Normandie**

En cotutelle internationale avec l'**Université de Tartu - Estonie, ESTONIE**

Estimation of Depression Level from Text: Symptom-Based Approach, External Knowledge, Dataset Validity

Présentée et soutenue par

KIRILL MILINTSEVICH

Thèse soutenue le 18/10/2024
devant le jury composé de :

M. GAEL DIAS	Professeur des universités - Université de Caen Normandie	Directeur de thèse
MME KAIRIT SIRTS	Maître de conférences - Université de Tartu - Estonie	Co-directeur de thèse
M. EDUARD BARBU	Chercheur - Université de Tartu - Estonie	Membre du jury
MME BEATRICE DAILLE	Professeur des universités - Nantes Université	Membre du jury
M. FABRICE MAUREL	Maître de conférences - Université de Caen Normandie	Membre du jury
M. XAVIER TANNIER	Professeur des universités - Sorbonne Université	Membre du jury
MME NATALIA GRABAR	Chargé de recherche au CNRS - Universite de Lille	Rapporteur du jury
M. ROMAN KLINGER	Professeur - Université Otto-Friedrich de Bamberg	Rapporteur du jury

Thèse dirigée par **GAEL DIAS** (Groupe de recherche en informatique, image et instrumentation de Caen) et **KAIRIT SIRTS** (Université de Tartu - Estonie)

GREYC
Laboratoire de recherche
en sciences du numérique

KIRILL MILINTSEVICH

Estimation of Depression Level from Text: Symptom-Based Approach, External Knowledge, Dataset Validity

CAEN 2024

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia.
Electronics and Computer Science Laboratory (GREYC), CNRS UMR 6072, University of Caen Normandy, France.

Supervisors

Assoc. Prof. Kairit Sirts
Institute of Computer Science
University of Tartu, Estonia

Prof. Gaël Dias
GREYC Laboratory - CNRS UMR 6072
University of Caen Normandy, France

Opponents

Prof. Roman Klinger
Faculty Information Systems and Applied Computer Sciences
University of Bamberg, Germany

Dr. Natalia Grabar
Savoirs, Textes, Langage (STL) - CNRS UMR 8163
University of Lille, France

The public defense will take place on October 18, 2024 at 09:30 in Salle des thèses, UFR Sciences 3, Campus 2, 6 boulevard Maréchal Juin, 14032 Caen.

The publication of this dissertation was financed by the University of Caen Normandy and University of Tartu.

UNIVERSITY OF TARTU
Institute of Computer Science

ABSTRACT

Major Depressive Disorder (MDD) is one of the most prevalent psychiatric disorders globally, often resulting in disability and an increased risk of suicide. The recent COVID-19 pandemic has further exacerbated depression rates in countries such as France and Estonia, and worldwide. However, the stigma surrounding mental illnesses and the limited availability of psychiatric treatment prevents many individuals from receiving proper diagnosis and care.

Natural Language Processing (NLP) research community has long been interested in automatic depression detection through text. Initial linguistic studies identified differences in vocabulary usage between depressed and non-depressed individuals. Advances in machine and deep learning have since enabled the detection of depression through social media texts and clinical interview transcriptions. However, most of the researchers approach depression detection as a binary classification task, which overlooks crucial symptomatic details. Moreover, the scarcity of high-quality data for depression detection poses another significant challenge, as clinical datasets are often restricted by regulations. Social media provides abundant data, but the lack of professional oversight in labeling raises questions about the validity of this data.

The primary aim of this thesis was to develop symptom-based models for automated depression estimation from text and explore ways to integrate existing domain knowledge into neural models. This led to the following research questions: (RQ1) How does predicting depression as a collection of symptoms compare with predicting depression as a binary diagnosis? (RQ2) Does including external knowledge into current state-of-the-art neural architectures improve automatic depression estimation? While working on RQ2, we noticed that the social media dataset failed to show any improvement, particularly for the lack of interest symptom, prompting us to study whether the annotations in this dataset align with the definition of this symptom (RQ3).

First, we explored **symptom-based depression prediction** for automatic depression estimation through text. Instead of approaching automatic depression estimation through text as a binary problem, we built a multi-target regression neural model to predict the frequency of each depression symptom individually. This model achieved state-of-the-art results in symptom-based depression estimation, producing symptom scores that can be easily converted into a binary label yet provide more information. Second, for **external knowledge integration**, we used a simplistic input marking approach to incorporate the information from the sentiment and emotion lexicons and psychiatrists' expertise into pre-trained language models (PLM). Finally, for **annotation validity**, we advocated for rigorous and standardized mental health dataset annotation, emphasizing the need for greater involvement of domain experts. A higher-quality social-media text dataset for anhedonia detection was built and made publicly accessible.

We also put forward several paths for future work. The rising popularity of Large

Language Models (LLMs) presents new opportunities for depression estimation, though their biases and hallucination tendencies require careful consideration. Further exploration of external knowledge integration into models presents another direction for future research. Additionally, annotating more texts with various symptoms and collecting data for languages other than English is necessary for advancing the field.

CONTENTS

LIST OF ABBREVIATIONS

Acronyms

BDI Beck Depression Inventory. 23

BERT Bidirectional Encoder Representations from Transformers. 24, 27, 39–47, 49, 52

DAIC-WOZ Distress Analysis Interview Corpus Wizard-of-Oz. 14, 15, 19–22, 24, 25, 27–31, 33, 35, 39–44, 46, 52

DSM-5 Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. 14, 17, 47, 52

E-DAIC Extended Distress Analysis Interview Corpus. 21, 22

EULA End-User Licence Agreement. 21

GRU Gated Recurrent Units. 29

LSTM Long Short-Term Memory. 29, 32, 38

MADRS Montgomery-Åsberg Depression Rating Scale. 47–49

MAE Mean Absolute Error. 27, 28, 30, 33–35, 40

MDD Major Depressive Disorder. 6, 14, 17, 18, 21, 31, 51

MHP Mental Health Professional. 23, 40, 47–49

MLP Multilayer Perceptron. 29

NLP Natural Language Processing. 6, 14, 17, 24, 31, 46, 51

PHQ-8 Patient Health Questionnaire. 14, 28, 30, 31, 33, 40, 42, 45

PHQ-9 Patient Health Questionnaire. 20, 21, 23, 30

PLM Pre-trained Language Model. 6, 37, 42, 45, 46, 50, 52, 53

PTSD Post-Traumatic Stress Disorder. 21

RMSE Root Mean Square Error. 27, 28

RNN Recurrent Neural Network. 25

RRMSE Relative Root Mean Square Error. 28, 33–35

WOZ Wizard-of-Oz. 21

Depression Symptoms

CON Diminished ability to think or **con**centrate, or indecisiveness. 17, 34, 35, 40–42, 47

DEP **Dep**ressed mood. 17, 34, 35, 40, 42, 47, 48

EAT **Eat**ing problems: significant weight loss when not dieting or weight gain, or decrease or increase in appetite nearly every day. 17, 34, 35, 40–42, 47

ENE Fatigue or loss of **ene**rgy. 17, 34, 40–42, 46–48

LOI **L**ack **o**f **i**nterest in doing things, markedly diminished interest or pleasure in all, or almost all, activities (anhedonia). 17, 34, 35, 40, 42, 46–50

LSE **L**ow **s**elf-**e**steem, feelings of worthlessness or excessive or inappropriate guilt. 17, 34, 35, 40, 42, 46, 47

MOV Psychomotor agitation or retardation, **mov**ing too fast or too slow so that the others might have noticed. 17, 34, 35, 40–42, 46, 47

SLE Problems with **sle**ep: insomnia or hypersomnia. 17, 34, 35, 40–42, 47

SUI Recurrent thoughts of death or recurrent **sui**cidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide. 17, 42, 47

Nomenclature

Cls Classification head. 25, 31

$\mathbf{Enc}^{\mathbf{int}}$ Turn-level interview encoder. 25, 31

$\mathbf{Enc}^{\mathbf{turn}}$ Token-level turn encoder. 25, 31

$h^{\mathbf{int}}$ Interview hidden representation. 25, 31

$h_i^{\mathbf{s}}$ *i*-th turn hidden representation. 25

LIST OF ORIGINAL PUBLICATIONS

Publications included in the thesis

I. **Milintsevich, K.**, Sirts, K., & Dias, G. (2023). Towards Automatic Text-Based Estimation of Depression through Symptom Prediction. *Brain Informatics, 10*, 4. doi:10.1186/s40708-023-00185-9
Author's contributions: Performed the experiments and analyses, wrote the code, and had a major role in writing the paper.

II. **Milintsevich, K.**, Dias, G., & Sirts, K. (2024). Evaluating Lexicon Incorporation for Depression Symptom Estimation. In *Proceedings of the 6th Clinical Natural Language Processing Workshop (Clinical NLP 2024)* (pp. 322–328). Association for Computational Linguistics.
doi:10.18653/v1/2024.clinicalnlp-1.28
Author's contributions: Performed the experiments and analyses, wrote the code, and had a major role in writing the paper.

III. Agarwal, N.*, **Milintsevich, K.***, Métivier, L., Rothärmel, M., Dias, G., & Dollfus, S. (2024). Analyzing Symptom-based Depression Level Estimation through the Prism of Psychiatric Expertise. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp.974–983). ELRA and ICCL. doi:10.18653/v1/2024.lrec-main.87
Author's contributions: Developed the neural model used in all the experiments, was involved in establishing the data annotation procedure, and participated in writing and reviewing the text.

IV. **Milintsevich, K.**, Sirts, K., & Dias, G. (2024). Your Model Is Not Predicting Depression Well And That Is Why: A Case Study of PRIMATE Dataset. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)* (pp. 166–171). Association for Computational Linguistics. doi:10.18653/v1/2024.clpsych-1.13
Author's contributions: Performed the experiments and analyses, wrote the code, and had a major role in writing the paper.
* – authors contributed equally.

Publications not included in the thesis

V. **Milintsevich, K.**, & Agarwal, N. (2023). Calvados at MEDIQA-Chat 2023: Improving Clinical Note Generation with Multi-Task Instruction Finetuning. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, (pp. 529—535). Association for Computational Linguistics. doi:10.18653/v1/2023.clinicalnlp-1.56

LIST OF ORIGINAL RESOURCES

Datasets

I. **Milintsevich, K.** & Sirts, K. (2024). Reddit anhedonia [Data set]. https://huggingface.co/datasets/tartuNLP/reddit-anhedonia

Software

I **Milintsevich, K.** (2022). Hierarchical Depression Symptom Classifier (v1.0). Université de Caen Normandie and University of Tartu. doi:10.5281/zenodo.12657260

II **Milintsevich, K.** (2024). Dialogue Classifier (v1.0.1). Université de Caen Normandie and University of Tartu. doi:10.5281/zenodo.12657477

1. INTRODUCTION

Major Depressive Disorder (MDD) is one of the most common psychiatric disorders worldwide that often causes disability and increases the risk of suicide (World Health Organization et al., 2017). Moreover, after the recent COVID-19 pandemic, depression levels are increasing in France (Léon et al., 2023), Estonia,[1] and worldwide.[2] However, mental illnesses are frequently stigmatized, and psychiatric treatment might not be available to many. Because of that, many people cannot receive an appropriate diagnosis followed by treatment. Hence, developing methods for the automated early detection of potentially depressed individuals is necessary to mitigate these challenges.

Automatic depression detection from text has been the interest of the Natural Language Processing (NLP) and linguistic communities for many years. First, linguistic studies have shown differences in the choice of vocabulary between the depressed and non-depressed populations (e.g., Coppersmith et al. (2014b), De Choudhury et al. (2013), Rude et al. (2004), and Yazdavar et al. (2017)). Later, various machine and deep learning solutions were adapted to detect depression through social media texts (e.g., Ji et al. (2022) and Yadav et al. (2020)) or transcriptions of clinical interviews (e.g., Mallol-Ragolta et al. (2019), Villatoro-Tello et al. (2021), and Xezonaki et al. (2020)).

It is noteworthy that most of the previous works have approached automatic depression detection from text as a binary classification task. However, potentially, the most widely used definition of MDD comes from the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (American Psychiatric Association, 2022). According to the DSM-5, depression diagnosis is defined as a co-occurrence pattern of specific symptoms. Thus, there are numerous different symptom profiles behind the same diagnostic label. Consequently, adopting the symptom-based approach for automatic depression detection from text would provide more information and transparency than binarized diagnosis prediction.

The lack of high-quality data is another challenge to automatic depression estimation. Clinical datasets, such as recordings of patient-therapist conversations, are collected in hospitals, which are usually bound by strict regulations that prohibit any data sharing. One of the rare exceptions is the DAIC-WOZ dataset (Gratch et al., 2014), which is publicly available under the end-user license agreement. In this dataset, before the conversation, each interviewee filled in the PHQ-8 (Kroenke et al., 2001), a questionnaire that measures the severity of depression based on the frequency of symptoms from the DSM-5 criteria. Hence, this dataset has become the foundation of many research initiatives, including this thesis.

On the other hand, social media is a goldmine of publicly available data. Nu-

[1] https://inimareng.ee/en/1-4-mental-health-problems-among-estonias-adult-population/

[2] World Health Organization et al., 2022.

merous works leverage data collected from social media platforms like Reddit[3] and X[4] (former Twitter) for automatic depression detection. However, most of this data is labeled either automatically (Pirina & Çöltekin, 2018; Syarif et al., 2019) or with the help of layperson crowd workers who have little to no training in clinical psychology (Gupta et al., 2022; Yates et al., 2017). Undoubtedly, involving mental health professionals in the annotation process is challenging. Nevertheless, their absence from or little participation in this loop puts the validity of such data to the question.

However, dataset validity is an important concern. Based on the data from Harrigian et al. (2021), out of 20 social-media-based depression datasets,[5] only three include manual annotation, and only one dataset involved a clinical professional during the annotation procedure. Furthermore, Pérez et al. (2023) tasked one mental health professional and two computer scientists to annotate the Reddit-based data with the first three BDI-II (Beck et al., 1996) symptoms and reported low inter-annotator agreement (median Cohen's Kappa of 0.38).

Another type of data that can be used for the automatic depression detection from text is in the form of various lexicons. Several studies have shown differences in language usage between depressed and non-depressed individuals (Pennebaker et al., 2003). This is reflected, among other things, in the increased use of negatively valenced terms and first-person pronouns (Coppersmith et al., 2014b; Rude et al., 2004) or emotional words (De Choudhury et al., 2013) by depression-prone people. At the same time, several lexicons encoding the emotion (Mohammad & Turney, 2013), sentiment (Nielsen, 2011), or depression-specific (Yazdavar et al., 2017) vocabulary have been created over time. Given that the lexicons alone have been previously used to detect depression from text (e.g., Chung and Pennebaker (2011) and Losada and Gamallo (2020)), the models for automatic depression detection from text can potentially benefit from this external knowledge.

Research Questions. The main goal of this thesis was to develop symptom-based models for automated depression estimation from text. We also explored the ways of introducing the existing linguistic knowledge into the neural models.

Thus, we establish the **Research Questions (RQ)** of this thesis:

RQ1 How does predicting depression as a collection of symptoms compare with predicting depression as a binary diagnosis?

RQ2 Does including external knowledge into current state-of-the-art neural architectures improve automatic depression estimation?

Finally, while working on the RQ2, the social-media-based dataset, PRIMATE (Gupta et al., 2022), behaved differently from the DAIC-WOZ dataset by failing to benefit neither from the choice of a base model nor external knowledge. This led us

[3] https://www.reddit.com/

[4] https://x.com/

[5] Only considering the datasets that could be accessed either directly or through signing a user agreement.

to pursue the case study concerning the validity of the annotations on this dataset. After benchmarking the dataset with a larger variety of base models, we still failed to see any improvement, specifically for the lack of interest symptom, also known as anhedonia. Thus, we decided to study whether the annotations for anhedonia in the PRIMATE dataset are actually in line with the definition of anhedonia (RQ3).

Outline. This dissertation is structured as an integrated collection of publications. In Chapter 2, we outline the common background of the thesis, which ties together all the included publications. In this chapter, we give a brief psychological background on depression; then, we discuss how it affects language production and which linguistic resources have captured the linguistic differences. Finally, we present the recent datasets and approaches for automated depression estimation from text. The next chapters summarize each included publication and aim to answer the research questions. Hence, Chapter 3 tackles **RQ1** and presents a state-of-the-art approach for the symptom-based depression estimation from text. Chapter 4 investigates the incorporation of external resources into pre-trained language models for depression estimation and additionally presents an iterative improvement on the symptom prediction model (**RQ2**). Chapter 5 describes a case study of a social-media-based dataset, outlines the shortcomings of layperson annotators, and presents the pathway for a better annotation of social-media-based data (**RQ3**). Finally, all the publications are presented in their original form at the end of this manuscript.

2. BACKGROUND

Major Depressive Disorder (MDD) is one of the most common psychiatric disorders (World Health Organization et al., 2017). Unsurprisingly, it has attracted the interest of the scientific community, particularly the NLP community, to propose solutions for automatic depression detection. However, most of these approaches have treated the prediction of depression as a binary classification task without considering the psychiatric diagnostic criteria that define the diagnosis based on symptoms.

Symptom-Based Approach in Depression. In the Diagnostic And Statistical Manual Of Mental Disorders, Fifth Edition (DSM-5) (American Psychiatric Association, 2022), MDD is defined by nine symptoms:

1. Depressed mood (DEP);
2. Markedly diminished interest or pleasure in all, or almost all, activities (anhedonia) (LOI);
3. Significant weight loss when not dieting or weight gain, or decrease or increase in appetite nearly every day (EAT);
4. Insomnia or hypersomnia (SLE);
5. Psychomotor agitation or retardation (MOV);
6. Fatigue or loss of energy (ENE);
7. Feelings of worthlessness or excessive or inappropriate guilt (LSE);
8. Diminished ability to think or concentrate, or indecisiveness (CON);
9. Recurrent thoughts of death or recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide (SUI).

To assign the MDD diagnosis, an individual must have five or more symptoms, one of which must be either (1) depressed mood (DEP) or (2) anhedonia (LOI). In addition, the symptoms must be present nearly every day during the same 2-week period and cause clinically significant distress or impairment in important areas of functioning. Taking into account the fact that all symptoms except "depressed mood" have sub-symptoms, almost 1,000 unique combinations of symptoms can be classified as MDD (Fried & Nesse, 2015a). This heterogeneity also leads to a poor agreement between human experts in assigning an MDD diagnosis following DSM-5 guidelines (Regier et al., 2013). Hence, by viewing automatic depression estimation as a binary classification task, all of the symptomatic information is neglected.

In clinical practice, MDD is routinely assessed by rating scales, such as the Patient Health Questionnaire (Kroenke et al., 2001), a self-assessment questionnaire of nine questions, each of which is mapped to a DSM-5 symptom. Each symptom question is rated on a scale from 0 to 3, where the score increases together with the frequency of the symptom. In most of the datasets for automatic depressed estimation from text based on the PHQ, the final score is obtained by summing all the item scores and then is usually binarized using a cut-off point.

Outline. In this chapter, we have so far introduced the motivation of predicting MDD based on symptoms in contrast to a binary class. In Section 2.1, we review the studies that observed the differences in language production between depressed and non-depressed people. Later, in Section 2.2, we describe existing lexicons and datasets relevant to the automatic depression estimation from text. In particular, we present lexical resources that have been commonly used to assess the language of depressed individuals in Section 2.2.1 followed by an overview of clinical and social-media-based datasets in Section 2.2.2. Section 2.3 concludes this chapter by presenting the main deep learning approaches, evaluation metrics, and previously published results for automatic depression estimation from text.

2.1. Language of Depression

Depression is related, among other things, to one's language production. This is explained by the change in the cognitive process of a depressed or depression-prone person.

Beck (1979) formulated a cognitive theory according to which individuals who are vulnerable to depression possess deep-level knowledge structures or depressive schemata. These schemata lead them to view themselves and their environment in systematically negative terms. Beck (1979) further proposed that the interaction of these cognitive processing biases with a negative life event or stressor predisposes individuals to experience a pattern of negative automatic thoughts concerning themselves, the world, and the future (referred to as the 'cognitive triad'), along with accompanying negative mood. This is typically expressed by an increased use of negatively valenced terms by depression-prone individuals (Al-Mosaiwi & Johnstone, 2018; Coppersmith et al., 2014b; Rude et al., 2004). Additionally, Pennebaker et al. (2003) have also shown that language reflects the psychological state of a person.

Another characteristic of a depressed mind is self-focused attention. Pyszczynski and Greenberg (1987) have proposed that individuals suffering from depression tend to excessively ruminate about themselves. According to Pyszczynski and Greenberg (1987), following the loss of a significant source of self-worth, individuals may become trapped in a self-regulatory cycle focused on attempting to regain what has been lost. This engenders heightened self-focus, which is believed to amplify negative emotions and self-blame while hindering effective control efforts by diverting attentional resources. In line with this observation, numerous studies showed a high correlation between the increased use of first-person pronouns (Coppersmith et al., 2014b; De Choudhury et al., 2013; Mehl, 2004; Rude et al., 2004; Tadesse et al., 2019; Yazdavar et al., 2020) or other self-focused cognitive distortions (Bathina et al., 2021) and depression.

Various studies show other differences in linguistic arsenals among the depressed population. For example, Al-Mosaiwi and Johnstone (2018) observed increased usage of absolutist terms in people with anxiety, depression, and suicidal ideations. In their research, absolutist and nonabsolutist terms serve to express magnitudes or probabilities. Absolute words convey such notions without nuance, using terms like “always,” “totally,” or “entire.” In contrast, nonabsolute words introduce a degree of nuance, employing terms such as “rather,” “somewhat,” or “likely.” Yazdavar et al. (2020) found that depressed people are more likely to use more authentic, less confident and certain language, as well as an increasing number of informal and swear words. Similar findings have also been reported by Coppersmith et al. (2014b). Yazdavar et al. (2017) have also shown the difference in language between the different age groups; the difference in authenticity, informal, and sexual lexicons is higher among adolescents than among adults. De Choudhury et al. (2013) have reported the increased use of emotional words. Habermas et al. (2008) and Trifu et al. (2017) have observed that the depressed population used past tense more when speaking about their experiences. In summary, the discussed studies have demonstrated that systematic differences can be found in language usage between depressed and non-depressed people.

2.2. Language Resources

This section touches upon the data since it is arguably the most important aspect of depression estimation. With mental health being an extremely sensitive topic, publicly available clinical data is practically non-existent. We start by describing the relevant work on lexicons that have been used to find differences in the texts between depressed and non-depressed individuals. After that, we present the DAIC-WOZ dataset, the only publicly available dataset of clinical conversations. Finally, we finish this section with a compilation of datasets collected from social media platforms, another important source of depression-related data.

2.2.1. Lexicons

Based on previous research that established the differences in language production between depressed and non-depressed individuals, researchers have used different heuristic methods to construct lexicons containing specific depression-related terms. Neuman et al. (2012) used a search engine to find web pages containing the expression “depression is like *”, where * is a wildcard and extracted metaphoric descriptions of depression. Then, they used the corpus of contemporary American English to retrieve first- and second-order synonyms for each extracted term. This resulted in a lexicon that includes 1723 phrases associated with depression. De Choudhury et al. (2013) created a depression lexicon based on the corpus collected from the “Mental Health” category of Yahoo! Answers. The researchers compiled 900,000 question-answer pairs by extracting all questions and their corresponding best answers. Following tokenization of the question-answer texts, they proceeded to calculate, for each word within the corpus, its association with the regular expression "depress*" using both pointwise mutual information (PMI) and log-likelihood ratio (LLR). The final lexicon was defined as the union of the top 1% of terms in terms of LLR and PMI. Yazdavar et al. (2017) built a lexicon of depression-related terms based on the PHQ-9 questionnaire. Using techniques similar to the previous researchers, they collected a list of depression-related words and their synonyms, which were later validated and revised with the help of mental health professionals.

Several recent works on evaluating and enriching the depression lexicons with computational methods have been carried out. Losada and Gamallo (2020) evaluated two aforementioned lexicons (De Choudhury et al., 2013; Neuman et al., 2012) on eRisk 2017 test collections (Losada et al., 2017) and used automatic methods to expand and re-build the lexicons. The authors used corpus-based and thesaurus-based approaches to extend the lexicons. In the corpus-based strategy, new terms were extracted from Wikipedia using distributional similarity. In the case of the thesaurus-based approach, the lexicons were enhanced with the associations from the Wordnet.[1]

Other types of language resources used in depression detection from text are sentiment and emotion lexicons. One such resource is Linguistic Inquiry and Word Count (LIWC),[2] (Boyd et al., 2022) a text analysis software manually constructed by psychologists, which includes a set of dictionaries covering various categories, like personal pronouns, positive/negative emotion words, terms related to time orientation (past, present or future), etc. NRC Word-Emotion Association Lexicon[3] (aka EmoLex) (Mohammad & Turney, 2013) is a list of 14,182 English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive), which was annotated with the help of crowdsource workers. Finally, AFINN lexicon[4] (Nielsen, 2011) is a publicly available wordlist of 2,477 English terms manually rated by Nielsen for valence with an integer between minus five (negative) and plus five (positive). All aforementioned language resources have been used partially, individually, or in combination to detect depression from text (Chung & Pennebaker, 2011; Coppersmith et al., 2014b; Coppersmith, Dredze, Harman, & Hollingshead, 2015; De Choudhury et al., 2013; Gkotsis et al., 2016; Losada & Gamallo, 2020; Park et al., 2012; Rude et al., 2004; Safa et al., 2022; Xezonaki et al., 2020).

2.2.2. Depression Datasets

DAIC-WOZ dataset. Distress Analysis Interview Corpus (Gratch et al., 2014) constitutes a multimodal compilation of semi-structured clinical interviews. It was crafted to emulate conventional protocols aimed at identifying individuals

[1] A lexical database of English: https://wordnet.princeton.edu/

[2] https://www.liwc.app/

[3] https://www.saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

[4] http://www2.imm.dtu.dk/pubdb/pubs/6010-full.html

susceptible to post-traumatic stress disorder (PTSD) and major depressive disorder (MDD). These interviews were gathered within a broader initiative aimed at developing a computer agent capable of conducting interviews and discerning verbal and nonverbal cues indicative of mental health issues (DeVault et al., 2014). Participants in the study were sourced from two separate demographics residing in the Greater Los Angeles metropolitan area: veterans of the U.S. armed forces and members of the general public. They were categorized for depression, PTSD, and anxiety utilizing established psychiatric questionnaires. The corpus contains four interview formats:

- **Face-to-face** interviews: These involved direct interactions between participants and a human interviewer.
- **Teleconference** interviews: Conducted remotely via a teleconferencing system by a human interviewer.
- **Wizard-of-Oz** interviews: In this format, an animated virtual interviewer named Ellie conducted the interview. However, Ellie was controlled by a human interviewer who was situated in a separate room.
- **Automated** interviews: Participants engaged in interviews where Ellie operated autonomously as an agent in a fully automated capacity.

The collection process commenced with interpersonal interviews, encompassing both face-to-face interactions and teleconferencing sessions. Subsequently, Wizard-of-Oz interviews and automated interviews were conducted. Face-to-face and teleconference interviews typically spanned 30 to 60 minutes, whereas Wizard-of-Oz interviews lasted approximately 5 to 20 minutes, and automated interviews ranged from 15 to 25 minutes. The interviews followed a semi-structured format, starting with neutral questions to foster rapport and ensure participant comfort. They then transitioned to more targeted inquiries regarding symptoms and experiences associated with depression and PTSD. Finally, a “cool-down” phase was incorporated after the interview to mitigate the risk of participants departing in a distressed state of mind.

Before each interview, the participants completed different questionnaires to establish basic demographic variables and measure psychological distress and current mood. The Positive and Negative Affect Scale (PANAS) was used to assess mood (Watson & Clark, 1994), the PTSD Checklist – Civilian Version, the Patient Health Questionnaire (Kroenke et al., 2001), and the State-Trait Anxiety Inventory (Spielberger et al., 1971) were used to assess psychological condition. Only the scores of the Patient Health Questionnaire are available in the dataset version that is shared with the end-users.

The dataset is distributed upon signing the End-User Licence Agreement[5] and is available in two versions: the Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WOZ) and the Extended Distress Analysis Interview Corpus (E-DAIC) (Ringeval et al., 2019). The datasets are pre-split into training, validation, and

[5] https://dcapswoz.ict.usc.edu/

Depression severity	**PHQ-8 Score**	**DAIC-WOZ**			**E-DAIC**		
		Train	**Dev**	**Test**	**Train**	**Dev**	**Test**
No symptoms	[0..4]	47	17	22	77	26	19
Mild	[5..9]	29	6	11	36	15	16
Moderate	[10..14]	20	6	11	26	8	10
Moderately severe	[15..19]	7	6	7	17	6	9
Severe	[20..24]	4	1	2	7	1	2
Total		107	35	47	163	56	56

Table 1: Number of interviews for each depressive symptom severity category (as per Kroenke and Spitzer, 2002) in DAIC-WOZ and E-DAIC databases.

test sets, which are shown in Table 1. Both datasets contain the audio of the conversations with their text transcriptions and facial features from the video. The E-DAIC database extends the DAIC-WOZ database by adding the interviews with the fully automated agent. Furthermore, E-DAIC contains text transcriptions produced with the Google Cloud's speech recognition service (Ringeval et al., 2019) while the conversations in the DAIC-WOZ were transcribed manually (Gratch et al., 2014).

Below is an excerpt from the DAIC-WOZ dataset (the spelling is kept as is):

ELLIE: *do you have roommates*

PATIENT: *yes i do*

ELLIE: *tell me more about that*

PATIENT: *um they're they're friendly it's just that they're very quiet*

PATIENT: *'cause i'm not used to that environment*

ELLIE: *oh*

ELLIE: *what's it like for you living with them*

...

Social-media-based datasets. While multiple depression-related datasets exist based on social media texts, most of them only present binary annotation, i.e., whether the user is depressed or not. Table 2 presents an overview of several datasets. We aimed to review commonly used datasets as well as recent ones[6]. The most common sources of data are Reddit (Gupta et al., 2022; Losada & Crestani, 2016; Naseem, Dunn, et al., 2022; Pirina & Çöltekin, 2018; Sampath & Durairaj, 2022; Yates et al., 2017; Zhang et al., 2022) and X (former Twitter)[7] (Coppersmith, Dredze, Harman, Hollingshead, & Mitchell, 2015; Kabir et al., 2023; Syarif et al., 2019; Yadav et al., 2020). Most of the studies use automatic methods of

[6] Harrigian et al. (2021) have compiled an exhaustive list of mental health-related social media datasets. However, it is limited to the period between January 2012 and December 2019.

[7] Since February 2023, X (former Twitter) revoked free access to its API (application programming interface) for academics. This change rendered the use of existing datasets and the collection of new data extremely challenging.

Dataset	Manual review	Labels
From Reddit		
Losada and Crestani (2016)	Authors	Binary
Yates et al. (2017)	Layperson	Binary
Pirina and Çöltekin (2018)	None	Binary
Losada et al. (2019, 2020) and Parapar et al. (2021)	Self-assessment	BDI
Sampath and Durairaj (2022)	MHP	3 severity levels
Naseem, Dunn, et al. (2022)	Yes	4 severity levels
Gupta et al. (2022)	Layperson	PHQ-9
Zhang et al. (2022)	MHP	38 symptom classes
From X (former Twitter)		
Coppersmith, Dredze, Harman, Hollingshead, and Mitchell (2015)	Authors	Binary
Syarif et al. (2019)	None	4 severity classes
Yadav et al. (2020)	MHP	PHQ-9 + FL
Kabir et al. (2023)	MHP	4 severity classes

Table 2: Overview of social-media-based datasets.

annotations, such as regular expression matching of self-reported terms, like "I have been diagnosed with depression". Some of them perform manual verification and annotation either via layman crowd workers (Yates et al., 2017) or by the authors themselves (Coppersmith, Dredze, Harman, Hollingshead, & Mitchell, 2015; Losada & Crestani, 2016).

Recently, an interest in more fine-grained depression annotation has emerged. In particular, the two recent datasets, D2S (Yadav et al., 2020) and PRIMATE (Gupta et al., 2022), identify depressed social media posts from X and Reddit, respectively, and annotate them with PHQ-9 symptoms (Kroenke & Spitzer, 2002). Both datasets have been annotated with the help of crowd workers and later verified by Mental Health Professionals (MHP). However, the verification process was different. For D2S, conflicting annotations were resolved with the majority voting, and a psychiatrist resolved the ties. Afterward, 100 random samples were selected for quality control and verified by a psychiatrist. Additionally, Zirikly and Dredze (2022) annotated a random sample of D2S with the explanations for each symptom with the help of two MHPs, increasing the validity of the data. In the case of PRIMATE, no information is given on the quality control procedure. Another symptom-based annotation dataset was collected for the eRisk initiative (Losada et al., 2017, 2019, 2020; Parapar et al., 2021). This dataset is based on the Reddit posts (Losada & Crestani, 2016) supplied with the results from the self-assessment from 90 users who evaluated their mental state with the BDI questionnaire. Over the years, more data has been validated with the help of the eRisk shared task,[8]

[8] https://erisk.irlab.org/

expanding the dataset.

2.3. Automatic Depression Estimation from Text

This section describes the recent advances in automatic depression estimation from text. Here, we discuss neural network approaches for text-based automatic depression prediction. First, we start with the neural approaches used for processing dyadic texts, which is the format of the DAIC-WOZ dataset. We then also briefly describe the methods used for automatic depression estimation from the social-media-based datasets. We finish this section with a description of the main evaluation metrics that will be used in this work. We also present the recent results in the field of automatic depression estimation from text.

2.3.1. Approaches for Automatic Depression Estimation

DAIC-WOZ dataset. The DAIC-WOZ dataset is frequently used for testing automatic depression detection systems. In the DAIC-WOZ, each data sample is a conversation between a participant and a virtual assistant, Ellie. Considering this, some researchers use only participants' part as input (Burdisso et al., 2023; Mallol-Ragolta et al., 2019; Villatoro-Tello et al., 2021; Xezonaki et al., 2020), and others use both participant's and Ellie's speech (Agarwal, Dias, et al., 2024a; Shen et al., 2022; Toto et al., 2021; Williamson et al., 2016). While, in general, using the whole conversation produces better results than using only the participant's speech, Burdisso et al. (2024) suggest that Ellie's speech contains biases that allow models to distinguish between depressed and control participants more easily.

Another challenge is the length of the textual transcriptions of the conversation in the DAIC-WOZ. Since the appearance of pre-trained transformer-based (Vaswani et al., 2017) models, like BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021; He et al., 2020), they have rapidly become state-of-the-art for many NLP tasks[9]. However, most of the state-of-the-art pre-trained transformer-based models are limited in their effective input length, which most often is equal to 512 tokens. At the same time, the average input length of an interview in the DAIC-WOZ is $\approx 2,000$ tokens. While transformer-based models like Longformer (Beltagy et al., 2020) support input sequences up to 4,096 tokens, they have not gotten much traction for depression estimation. In fact, some researchers report that Longformer-based models underperform on the DAIC-WOZ compared to classical bag-of-words machine learning approaches (Chua et al., 2022) or graph neural networks (Agarwal, Dias, et al., 2024b; Burdisso et al., 2024).

One solution is to use a variation of the hierarchical neural classifier (Z. Yang et al., 2016), where an interview is encoded on two levels: the token and sentence

[9] GLUE leaderboard: https://gluebenchmark.com/leaderboard and SuperGLUE leaderboard: https://super.gluebenchmark.com/leaderboard.

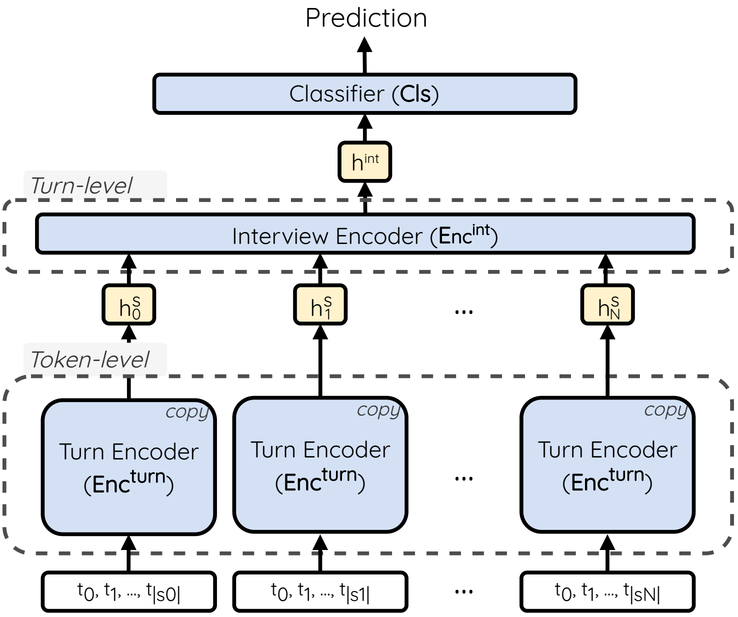


Figure 1: General architecture of a hierarchical classifier model.

level. This model has been successfully adopted for the DAIC-WOZ and showed good performance compared to other methods (Lau et al., 2023; C. Li et al., 2022; Mallol-Ragolta et al., 2019; Xezonaki et al., 2020). Figure 1 shows the hierarchical classifier in its general form. It is formulated as follows: given N turns s each containing $|\mathrm{s}_i|$ tokens t, the model first encodes each turn token-by-token with a token-level turn encoder $\mathbf{Enc}^{\mathbf{turn}}$ to get the i-th turn representation h_i^{s} (2.1), which are later encoded with a turn-level interview encoder $\mathbf{Enc}^{\mathbf{int}}$ to get an interview representation h^{int} (2.2). Finally, the prediction is made with a classification head **Cls**.

$$h_i^{\mathrm{s}} = \mathbf{Enc}^{\mathbf{turn}}(\langle t_0^i, t_1^i, \ldots, t_{|\mathrm{s}_i|}^i \rangle) \tag{2.1}$$

$$h^{\mathrm{int}} = \mathbf{Enc}^{\mathbf{int}}(\langle h_0^{\mathrm{s}}, h_1^{\mathrm{s}}, \ldots, h_N^{\mathrm{s}} \rangle) \tag{2.2}$$

In this model, $\mathbf{Enc}^{\mathbf{turn}}$ and $\mathbf{Enc}^{\mathbf{int}}$ can be any neural network that can produce an encoding from a sequence, for example, a recurrent neural network (RNN) as in Mallol-Ragolta et al. (2019) and Xezonaki et al. (2020) or a Transformer-based encoder as in Lau et al. (2023). A classification head **Cls** is usually represented with one or several fully connected layers, also called a linear layer, which consists of a learnable weight matrix W_o together with a bias vector b_o, and it applies the linear transformation:

$$\hat{y} = h^d W_o^\top + b_o \tag{2.3}$$

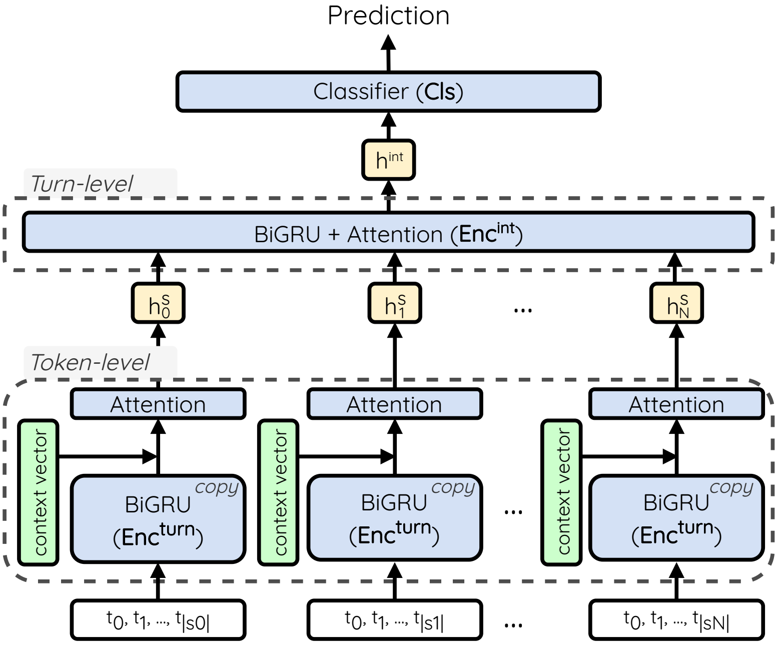


Figure 2: Hierarchical model with attention conditioning proposed by Xezonaki et al. (2020).

where the prediction $\hat{y}$ can be a real number in case of binary classification or regression or a vector of real numbers in case of multi-class classification, multi-target classification, or multi-target regression.

As discussed in Section 2.1, a large body of lexical resources on the language of depression have been collected in the past years. Furthermore, the connection between depression and change in sentiment and emotional expression has been found (De Choudhury et al., 2013). This has found a place in the domain of automatic depression estimation: several works have presented different ways of incorporating this external knowledge into the neural models to improve depression estimation. For example, Xezonaki et al. (2020) encoded external knowledge for various affective lexicons as a feature context vector for each input token. They later concatenated the context vector with each token representation in the hierarchical neural classifier. Figure 2 shows an overview of their hierarchical model with attentional conditioning.

Another research direction on incorporating external knowledge into automatic depression estimation is via multi-task learning. In multi-task learning, a model is trained on two or more different tasks at the same time, in contrast with single-task learning, which we have seen so far. These different tasks can have equal or different importance. For example, Qureshi et al. (2020) trained a classifier on the depression level and emotion intensity simultaneously. Another work by C. Li et al. (2022) incorporates depression, topic, dialog act, and emotion tasks into a single multi-task hierarchical model.

Social-media-based datasets. So far, we have discussed the neural approaches for the DAIC-WOZ dataset, which has an interview format and a longer input length. As previously discussed in Section 2.2.2, social-media-based datasets are most commonly sourced either from Reddit or X. Due to the nature of these platforms, the input text is much shorter (especially in the case of X). Thus, fine-tuning transformer-based pre-trained language models is much more prevalent in the works that use such data (Gupta et al., 2022; Yadav et al., 2020; Zhang et al., 2022).

Those language models are, however, pre-trained on general domain texts. Hence, an initiative to pre-train a domain-specific language model has emerged, resulting in MentalBERT and MentalRoBERTa (Ji et al., 2022). These models are based on general-domain BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) models, which were later adapted to the mental health domain using domain-adaptive pre-training (Gururangan et al., 2020). Ji et al. (2022) collected a corpus of texts from mental-health-related subreddits[10] and continued pre-training BERT and RoBERTa on this corpus. According to Ji et al. (2022), fine-tuning MentalBERT and MentalRoBERTa for mental health tasks, such as depression estimation, gives higher performance than fine-tuning the general-domain models. Other works using these models have also shown high performance for depression estimation on social-media-based datasets (Naseem, Lee, et al., 2022; Xu et al., 2024; K. Yang et al., 2022; K. Yang et al., 2024).

2.3.2. Evaluation Metrics

Most of the works treat depression estimation as a binary task, for which the performance is often measured with a macro-averaged F_1-score. F_1-score (also known as micro-averaged F_1-score or miF_1) is defined as:

$$miF_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{2.4}$$

where precision is the fraction of relevant instances among the retrieved instances, and recall is the fraction of relevant instances that were retrieved. For macro-averaged F_1-score (maF_1), first a class-specific F_1^c-score is computed for each class separately, and then the F_1^c-scores are averaged:

$$maF_1 = \frac{\sum_{c \in C} miF_1^c}{|C|} \tag{2.5}$$

For the regression, common measures are micro- and macro-averaged mean absolute error (*mi*MAE and *ma*MAE) and root mean square error (RMSE), defined in Equations 2.6, 2.7 and 2.8 respectively, where y_i is the true score and $\hat{y}_i$ is the predicted score. Additionally, for *ma*MAE, C is the set of classes, *mi*MAE^c

[10] A thematic community on Reddit.

denotes the *mi*MAE for the class c. MAE[11] is commonly used when the total score of the depression scale is predicted as a regression task (e.g., Lin et al. (2020) and Qureshi et al. (2020)) to preserve the scale of the PHQ score.

$$miMAE = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{2.6}$$

$$ma\mathrm{MAE} = \frac{\sum_{c \in C} mi\mathrm{MAE}^c}{|C|} \tag{2.7}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\hat{y}_i - y_i)^2} \tag{2.8}$$

While MAE is generally an effective and easily interpretable metric for evaluating regression tasks, it can give artificially low error scores when the data set is imbalanced, and the model tends to predict scores close to the mean value. A more complex version of RMSE, the Relative Root Mean Square Error (RRMSE) can give a better view of the performance in those cases, as it penalizes more the model that tends to predict scores close to the mean value of the training set (Borchani et al., 2015). RRMSE is defined in Equation 2.9, where $\bar{y}$ is the mean score of the training set. RRMSE values are positive; the RRMSE of 1 indicates the performance equal to the mean score, with smaller values showing the improvement over the mean.

$$RRMSE = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}} \tag{2.9}$$

2.3.3. Published Results

DAIC-WOZ dataset. Table 3 shows an overview of the previously published results on the DAIC-WOZ. Surprisingly, none of the works predict individual symptoms but rather a binary diagnosis (Table 3a) or a total PHQ-8 score (Table 3b). Binary diagnosis is obtained by a cut-off of a total PHQ-8 score, where PHQ-8 < 10 is classified as non-depressed and PHQ-8 ≥ 10 as depressed.

Modern neural architectures, such as Graph Convolutional Networks (GCN) and Transformer-based models, outperform other methods, even without introducing external knowledge. We would like to note, however, that DAIC-WOZ validation and test sets are small (as previously shown in Table 1), which increases the variance of the results among different runs. Only three works (Agarwal, Dias, et al., 2024a, 2024b; **Milintsevich, Kirill** et al., 2023) accounted for this by reporting average metrics over several runs. Another issue is that not all the authors (Mallol-Ragolta et al., 2019; Niu et al., 2021) explicitly stated which version of F_1-score

[11] Henceforth, MAE refers to both micro-averaged MAE (*mi*MAE) and macro-averaged MAE (*ma*MAE).

Model	Architecture	EK	Results	
			Dev F_1	Test F_1
†Mallol-Ragolta et al. (2019)	H-BiGRU	✗	0.51	0.63
Xezonaki et al. (2020)	H-BiGRU	✓	0.69	-
Villatoro-Tello et al. (2021)	MLP	✗	0.64	-
†Niu et al. (2021)	H-BiGRU+GAT	✗	0.77	-
C. Li et al. (2022)	H-BiLSTM	✓	-	0.71
‡**Milintsevich, Kirill** et al. (2023)	H-BiLSTM	✗	0.72	0.74
Burdisso et al. (2023)	GCN	✗	0.84	(0.61)
Burdisso et al. (2024)	Longformer	✗	0.79	-
Burdisso et al. (2024)	GCN	✗	0.90	-
‡Agarwal, Dias, et al. (2024a)	Transformers	✗	0.77	0.80
‡Agarwal, Dias, et al. (2024b)	GCN	✗	0.76	0.81

(a) Depression as a binary classification task.

Model	Architecture	EK	Results	
			Dev MAE	Test MAE
Qureshi et al. (2020)	LSTM	✓	-	3.69
Lin et al. (2020)	BiLSTM	✗	3.88	-
Niu et al. (2021)	H-BiGRU+GAT	✗	3.73	-
Hong et al. (2021)	GNN	✗	3.76	-
‡**Milintsevich, Kirill** et al. (2023)	H-BiLSTM	✗	3.61	3.78
‡**Milintsevich, Kirill**, Dias, et al. (2024)	H-Transformers	✓	-	3.59

(b) Depression as a regression task.

Table 3: Main previously published results on DAIC-WOZ. **EK** stands for **External Knowledge**. The architectures are the following: BiGRU – Bi-directional Gated Recurrent Unit; BiLSTM – Bi-directional Long Short-Term Memory; MLP – Multilayer Perceptron; GCN – Graph Convolutional Network; GNN – Graph Neural Network; GAT – Graph Attention Network. Prefix H- stands for Hierarchical. A dagger (†) signals that the authors did not specify whether they used a micro- or macro-averaged F_1-score. A double dagger (‡) indicates that the results are reported as an average over several runs. The score in parentheses comes from replicating the experiments locally.

they used.[12] Finally, Burdisso et al. (2023) and Burdisso et al. (2024) chose their best models based on the F_1-score of the validation set, which is coincidentally the only metric they reported. However, the high variance of the results increases the risk of overfitting the model selection, which makes the results biased (Cawley & Talbot, 2010). We investigated it further by replicating the experiments of Burdisso et al. (2023) on the DAIC-WOZ test set;[13] the model showed 0.61 F_1-score, in

[12] Micro- and macro-averaged versions of F_1-score can give drastically different results when the classes are unbalanced.

[13] The code from Burdisso et al. (2024) was not available at the moment of writing this text.

contrast to the high 0.84 F_1-score on the validation set. Finally, the cutpoint of 10 to convert the PHQ-8 score into a binary label is somewhat arbitrary. According to Kroenke and Spitzer (2002) and Kroenke et al. (2001), there is a "gray zone" in the range of [10..14] points. Furthermore, the difference between the symptom severity of a person with 9 and 10 points is most likely to be marginal. However, they would be assigned different binary labels. Thus, all comparisons should be considered with due care.

Because of the reasons mentioned above, predicting the total PHQ-8 score as a regression task instead of the binary classification would be preferable since it takes into account the whole range of the PHQ-8 score, thus alleviating the issues introduced by the strict cutpoint. Table 3b shows that only a few works regard the DAIC-WOZ dataset as a regression task. Overall, the MAE in the range of [3.59..3.78] points can be considered state-of-the-art for the automatic depression estimation from text.

Social-media-based datasets. Comparing the results of depression estimation on the social-media-based datasets is exceptionally challenging due to their extreme heterogeneity. In 2021, Harrigian et al. conducted a study of 102 datasets, 42 of which were aimed at depression detection. Most of these datasets are either inaccessible or unique to one study only. Furthermore, the annotation scheme varies greatly from one dataset to another, e.g., some works use PHQ-8 or PHQ-9 as a guideline, while others use the Center for Epidemiologic Studies Depression (CES-D) scale, and other works do not specify their definition of depression. Considering all these differences in social-media-based datasets, we cannot present a comparative table summarizing the results.

3. SYMPTOM-BASED AUTOMATIC DEPRESSION ESTIMATION (PUBLICATION I)

As shown in Chapter 2, representing a mental disorder, specifically an Major Depressive Disorder (MDD), as a profile of individual symptoms provides a more detailed mental picture of a person. However, this approach has not yet been fully explored by the NLP community, which is reflected in the lack of work on automatic symptom-based depression estimation. This chapter answers our first research question (RQ1): *"How does predicting depression as a collection of symptoms compare with predicting depression as a binary diagnosis?"* To investigate this question, we present a multi-target hierarchical regression model for symptom-based depression estimation on the DAIC-WOZ dataset. Our model achieves results that are on par with state-of-the-art models on both binary diagnostic classification and depression severity prediction while providing a more fine-grained overview of individual symptoms for each person.

3.1. Methodology

To efficiently encode the interviews, we employed a hierarchical architecture (Z. Yang et al., 2016), described in Section 2.3.1. Since we aim at predicting scores for individual symptoms, we adopted a prediction head that produces eight regression outputs, effectively making it a multi-target regression model.

Figure 3 shows an overview of the model. The classification head **Cls** is a feed-forward network that maps the interview representation h^{int} to a label vector $\hat{l} = [\hat{l}_1, \hat{l}_2, \ldots, \hat{l}_7, \hat{l}_8]$ (3.1, 3.2, 3.3), where each predicted label $\hat{l}_k \in [0,3]$ represents a symptom score for a corresponding question in PHQ-8. The feed-forward classifier consists of two linear layers (W_1, W_2) with biases (b_1, b_2), with a LeakyReLU activation function and a LayerNorm layer (Ba et al., 2016) in-between.

$$z' = \text{LeakyReLU}(h^{\text{int}} W_1^{\top} + b_1) \tag{3.1}$$

$$z = \text{LayerNorm}(z') \tag{3.2}$$

$$\hat{l} = z W_2^{\top} + b_2 \tag{3.3}$$

The token-level turn encoder $\textbf{Enc}^{\textbf{turn}}$ uses a distilled RoBERTa-based model from the SentenceTransformers (S-RoBERTa).[1] Distilled models keep most of the capabilities of their full-sized counterparts while being almost twice as small and fast (Sanh et al., 2019). Decreasing the computational complexity of our model is crucial due to the fact that all turns of the interviews have to be processed in parallel, i.e., several copies of $\textbf{Enc}^{\textbf{turn}}$ are created, and their respective computational graphs are stored during training. The turn-level interview encoder $\textbf{Enc}^{\textbf{int}}$ deploys a single

[1] https://huggingface.co/sentence-transformers/all-distilroberta-v1

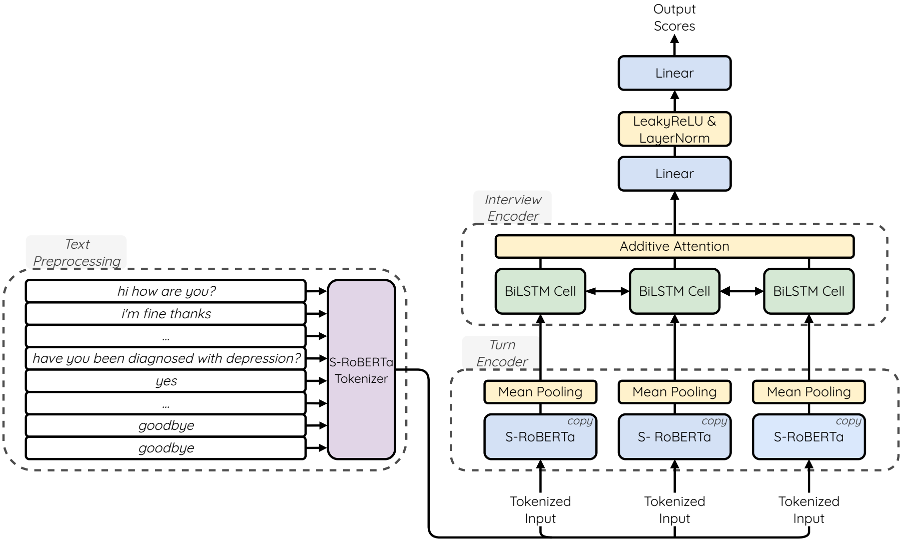


Figure 3: Overview of the model. On the turn level, the same instance of S-RoBERTa is used to encode each turn. Mean Pooling is the operation that averages all the token representations output by S-RoBERTa.

layer BiLSTM with a hidden dimension of 300 and an additive attention layer on top of it.

As a training objective for the symptom prediction task, the Smooth L_1 loss was used. Smooth L_1 loss is less sensitive to outliers than, for example, MSE loss and, in some cases, prevents exploding gradients (Girshick, 2015). Smooth L_1 loss is defined as in (3.4) for multi-target regression:

$$\text{Smooth}_{L_1}(\hat{l}, l) = \frac{1}{K} \sum_{k=1}^{K} \text{Smooth}_{L_1}(\hat{l}_k, l_k) \tag{3.4}$$

where $\hat{l}_k$ and l_k are the predicted and true scores for the k-th symptom respectively, $K = 8$ is the number of symptoms, and with

$$\text{Smooth}_{L_1}(\hat{l}_k, l_k) = \begin{cases} 0.5(\hat{l}_k - l_k)^2, & \text{if } |\hat{l}_k - l_k| < 1 \\ |\hat{l}_k - l_k| - 0.5, & \text{otherwise} \end{cases} \tag{3.5}$$

Since distinct random seeds can lead to substantially different results (Dodge et al., 2020), each model was trained five times using different random seeds, and the average of the five runs is reported. Each model was trained for 200 epochs using AdamW optimizer with the learning rate of $3e^{-5}$ and a linear warm-up scheduler. A model checkpoint was saved after each epoch, and the checkpoint with the highest micro-averaged F1-score on the development set was chosen as the final model.

3.2. Data and Experimental Setup

Data. All the experiments were carried out on the DAIC-WOZ dataset, described in Section 2.2.2.

Models. To provide some validity to the symptom prediction approach, we compared the results of our model to three baseline tasks adopted in previous works: 1) Binary Diagnostic classification, where a patient is said to be depressed if their PHQ-8 score is at least 10, and non-depressed otherwise, 2) multi-class classification into five classes with differing severity as depicted in Table 1, i.e., no symptoms, mild, moderate, moderately severe and severe depression, and 3) depression severity prediction modeled as the PHQ-8 total score regression ranging from 0 to 24.

The outputs of the multi-target regression model predicting symptom scores could be recast to a suitable format for these three tasks. For the depression severity prediction task (**regression**), the symptom scores were summed up to give the estimate of the final PHQ-8 value. For the **binary** and **multi-class** classification tasks, the summed total score could be converted either into a binary label at a cut-off of 10 for the binary diagnostic classification or converted into five classes for the multi-class classification, such that [0..5) stands for no symptoms, [5..10) mild, [10..15) moderate, [15..20) moderately severe and [20..24] severe depression estimate.

For comparison, we trained three baseline models that predict the three tasks directly, i.e., the model predicts one of the two classes for the binary diagnostic prediction (**BINARY DIAGNOSTIC**), one class out of five for the multi-class severity prediction (**5-CLASS SEVERITY**), and a continuous score for the total depression severity regression (**PHQ-8 SEVERITY**). All baseline models use the same hierarchical architecture shown in Figure 3; only the output layer of the feed-forward classifier network is different. Whereas the output layer for the **SYMPTOM PREDICTION** model has multiple regression heads, the PHQ-8 Severity model has a single regression head, and the Binary Diagnostic and the 5-Class Severity models have a classification head that predicts one of the two or five classes, respectively.

Metrics. We evaluated the Binary Diagnosis Eval task with micro- and macro-averaged F1-scores (Equations 2.4 and 2.5). For the PHQ-8 Score Severity Eval, mean absolute error (*mi*MAE) was used (Equation 2.6) alongside its macro-averaged version (*ma*MAE), defined in Equation 2.7. For the symptom-based evaluation, relative root mean squared error (RRMSE) was used (Equation 2.9) along with the previously mentioned metrics.

3.3. Results and Discussion

Table 4 compares our SYMPTOM PREDICTION model to three baselines: BINARY DIAGNOSTIC, 5-CLASS SEVERITY, and PHQ-8 SEVERITY models. Our model

Model	Binary Classification		Regression	
	miF$_1$ $_{\pm\sigma}$	*maF*$_1$ $_{\pm\sigma}$	*mi*MAE$_{\pm\sigma}$	*ma*MAE$_{\pm\sigma}$
BINARY DIAGNOSTIC	0.719$_{\pm 0.016}$	0.701$_{\pm 0.010}$	-	-
5-CLASS SEVERITY	0.711$_{\pm 0.026}$	0.683$_{\pm 0.024}$	-	-
PHQ-8 SEVERITY	0.681$_{\pm 0.019}$	0.584$_{\pm 0.024}$	5.03$_{\pm 0.09}$	5.69$_{\pm 0.12}$
SYMPTOM PREDICTION	**0.766**$_{\pm 0.023}$	**0.739**$_{\pm 0.025}$	**3.78**$_{\pm 0.13}$	**4.19**$_{\pm 0.13}$

Table 4: Experimental results on the test set of the DAIC-WOZ dataset. All models were run five times with different seed values, and the average values with standard deviation are presented.

Symptom	MAE $\pm\sigma$	RRMSE $\pm\sigma$	*miF*1 $\pm\sigma$	*maF*1 $\pm\sigma$
LOI	0.529 ± 0.047	0.877 ± 0.067	0.800 ± 0.024	0.669 ± 0.043
DEP	0.550 ± 0.027	0.733 ± 0.022	0.821 ± 0.019	0.729 ± 0.024
SLE	0.753 ± 0.073	0.805 ± 0.060	0.774 ± 0.055	0.757 ± 0.047
ENE	0.638 ± 0.031	0.816 ± 0.030	0.745 ± 0.030	0.709 ± 0.035
EAT	0.811 ± 0.049	0.972 ± 0.064	0.762 ± 0.035	0.685 ± 0.026
LSE	0.620 ± 0.018	0.796 ± 0.012	0.817 ± 0.024	0.779 ± 0.021
CON	0.830 ± 0.040	0.878 ± 0.012	0.681 ± 0.034	0.557 ± 0.029
MOV	0.438 ± 0.022	0.976 ± 0.035	0.936 ± 0.000	0.484 ± 0.000

Table 5: Test scores for each symptom. All models were run five times with different seed values, and the average values with standard deviation are presented. For computing the F1-scores, the predicted scores were binarized, such that the scores < 1.5 were treated as negative class instances, and the scores ≥ 1.5 were treated as positive class instances.

generally outperformed or matched the baselines across all tasks, particularly excelling in binary classification and regression tasks. For the multi-class classification task, which is not included in the table, the 5-CLASS SEVERITY model performed better on the micro-F1 score, while both models performed similarly on the macro-F1 score. The PHQ-8 SEVERITY model performed poorly on both classification tasks. Compared to previous works on DAIC-WOZ data, which also used only text input, our SYMPTOM PREDICTION model achieved comparable results, except for the multi-class classification task where the model by Qureshi et al. (2020) significantly outperformed it.

We then evaluated the SYMPTOM PREDICTION model for each symptom using MAE and micro- and macro-averaged F1-scores. Since each symptom score ranges from 0 to 3, binary labels for F1-scores were determined with a cutoff of 1.5 points. MAE can be misleading with imbalanced datasets, so we used Relative Root Mean Square Error (RRMSE) (Equation 2.9) for better evaluation. RRMSE (Borchani et al., 2015) can give a better view of the performance in those cases, as it penalizes more the model that tends to predict scores close to the mean value of the training set.

Table 5 shows that the core depression symptoms like depressed mood (DEP) and lack of interest (LOI) are well-predicted. Symptoms related to sleep (SLE) and feelings of failure (LSE) are also accurately predicted. Movement-related symptom (MOV) appears to be the most accurately predicted one judging from the MAE and $miF1$-score, but this is misleading due to dataset bias. In our sample, the moving symptom (MOV) has a relatively low score for most participants, biasing the model towards always predicting low scores. The RRMSE reveals predictions close to the mean, and a high micro-F1 combined with low macro-F1 indicates the model often predicts scores that fall into the negative class.

The results reflect the nature of the DAIC-WOZ data since the topics related to the most accurately predicted symptoms are discussed the most during each interview. Some of the well-predicted symptoms are addressed in the interview, even though less directly, e.g., assessing the feeling of being a failure (LSE) by asking what the interviewee's friends and family think about them. The sleep-related symptom (SLE) is also predicted relatively accurately; there are indeed questions about the person's sleep problems, but they are not present in every interview. Finally, the symptoms related to eating (EAT), problems with concentration (CON), and slowed down or overly agitated movement (MOV) are not detected accurately by the model. Interestingly, the results in Table 5 show a RRMSE score close to 1 for these symptoms, which can indicate that there is little textual evidence of these symptoms in the data and thus, the model just learns an average score for these symptoms across the training dataset.

Every interview also includes the question, "Have you been diagnosed with depression?". Thus, it is plausible that the model can extract information relevant to predictions only from the answer to this question, thus using it as a shortcut. We investigated more thoroughly whether this question strongly correlates with the model's predictions. First, we classified the answers to this question into three categories: "yes", "no", and "other". "Yes" and "no" categories were assigned to the answers that can be clearly interpreted as positive or negative. If a participant tried to avoid the question or started to give extra information about their condition, the answer was classified as "other". Fisher's exact test at the p-value < 0.05 was used to decide whether the depressed and non-depressed participant groups were different in their "yes" and "no" answers to this question. Similar analyses were conducted for every symptom with the groups formed by the symptom scores. Based on these analyses, we can conclude that the answers to the question "Have you been diagnosed with depression?" differ significantly between the groups formed based on different symptom scores. Thus, the model is suspect in utilizing these differences when making predictions. To estimate how dependent the model is on these answers, we replaced all the "yes" answers with a random answer variation from the "no" answer set and vice versa. Additionally, we replaced each "other" answer with another random answer from the "other" answer set as well. The same model was run on this perturbed test set, showing no drop in the miF_1 score (-0.00%) and an insignificant minor drop in the maF_1 score (-0.52%). Similar pattern was observed for *mi*MAE (+0.06) and *ma*MAE (+0.11). Thus, we can conclude that the model did not use this question with its explicit answers as a shortcut for making complex predictions.

3.4. Conclusions and Future Work

The publication on which this chapter is based is the first and the most substantial contribution to this thesis. Here, we established a neural architecture that produced state-of-the-art results for symptom-based depression estimation. This architecture was also fundamental for the experiments in the next chapter. We also showed that the predicted scores of each individual symptom, when summed and converted to the binary label, produced better results than training the model directly on the binarized labels. At the same time, these multi-target predictions provided more information about the symptomatic profile (**RQ1**). In the next chapter, we continued this work by improving the architecture and introducing depression and sentiment lexicons into the model to find out whether this external knowledge helps to improve the prediction of symptoms.

4. EXTERNAL KNOWLEDGE INCORPORATION FOR DEPRESSION SYMPTOM ESTIMATION (PUBLICATIONS II AND III)

In the previous chapter, we showed that treating depression as a system of symptoms rather than a binary diagnosis is better for automated depression prediction. We demonstrated it by using a multi-target hierarchical regression model, which achieved state-of-the-art results in depression symptom level prediction. However, this approach relied on the information encoded by a pre-trained language model (PLM), which was trained on a general domain text. At the same time, the vast amount of carefully collected depression-related lexical resources, described in Section 2.1, stays unvisited. Also, incorporating psychiatrists' expertise into the neural models is underexplored. In this chapter, we aim to answer the second research question (RQ2) *"Does including external knowledge into current state-of-the-art neural architectures improve automatic depression estimation?"* For this purpose, we used a simplistic approach of input marking to highlight the words from the sentiment and emotion lexicons described in Section 2.2.1, as well as psychiatrists' annotations collected as part of Publication III. This method allowed us to incorporate the external knowledge from these lexicons into PLMs without changing the architecture. In addition, we modified the hierarchical neural classifier proposed in the previous chapter to make the training more efficient. Our experiments showed that incorporating the lexical resources into the domain-specific PLM (MentalBERT in our case) improved automated depression symptom estimation.

4.1. External Knowledge Incorporation via Input Marking

To incorporate external knowledge into the model, we use three lexicons described in Section 2.2.1: AFINN (Nielsen, 2011), NRC (Mohammad & Turney, 2013), and SDD (Yazdavar et al., 2017). To provide the reader with a quick reminder, AFINN is a sentiment valence lexicon, NRC is an emotion and sentiment lexicon, and SDD is a lexicon of depression-related words and phrases.

Another source of external knowledge is the psychiatrists' annotations (PA). Three psychiatrists from public hospitals were employed to undertake span-based annotation of the transcripts. The task given to the psychiatrists consisted of highlighting information within transcripts that might have influenced a psychiatrist's decision during an interview. Since it is a subjective task that lacks a definitive right or wrong answer, a common consensus on the importance of various utterances within the transcripts might not exist. Even within the field of medicine, professionals do not universally agree on the significance of various pieces of information, and subtle differences in opinion exist between psychiatrists based on their individual knowledge and experience (Reed et al., 2018). As such, after various meetings and discussions with the psychiatrists, it was agreed that the medical annotators should have complete freedom to annotate the transcripts without any constraints in order to capture their true judgment. As a consequence, we forwent defining detailed annotation protocols and relied on the annotator's judgment as experts in the field for the reliability of their annotations. However, they were encouraged not only to identify information that suggests the presence of depression but also to pinpoint clues that indicate its absence. Furthermore, the expected lack of consensus within the task renders inter-annotator agreements less informative. In case multiple annotators are assigned per transcript, a simple union of annotated spans would be used to capture knowledge from all assigned annotators. Unfortunately, at this stage of our research, only one annotator per transcript could be assigned due to the workload experienced by the annotators, particularly due to the radical increase of mental care demand after the COVID pandemic coupled with the shortage of mental health professionals. The current annotation process had lasted nearly 5 months, and we anticipated this time frame would scale linearly with the increase in the number of annotators per transcript.

Following Zhou and Chen (2022), we annotated the lexicon words and psychiatrists annotations in the input text by marking them with the "@" token on either side (see Table 6 for an example). This way, the pre-trained model's architecture remains unchanged.

Illustration of the lexicon-based input marking
a) i'm pretty much good because see by me being a bus operator you run into circumstances and situations you gotta remain calm and still remain professional at the same time
b) i'm @ pretty @ much @ good @ because see by me being a bus operator you run into circumstances and situations you gotta remain @ calm @ and still remain professional at the same time
c) i'm @ pretty @ much @ good @ because see by me being a bus operator you run into circumstances and situations you gotta remain @ calm @ and still remain @ professional @ at the same @ time @

Table 6: Example of the input marking. Text a) is the original text without markings, b) and c) show text with terms from AFINN and NRC lexicons marked.

4.2. Model Modifications

While the model presented in the previous chapter already shows state-of-the-art results for symptom-based depression estimation, it suffers from high memory consumption during training because its input processing is not optimal. To improve it, we propose two modifications. First, the BiLSTM utterance-level encoder is replaced with a randomly initialized 4-layer 12-head transformer encoder. Second, we change the way the input data is represented. In the original model,

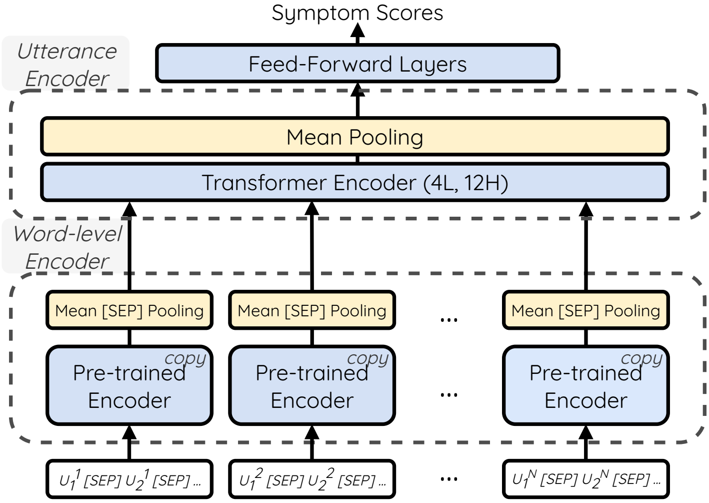


Figure 4: Overview of the model architecture. U_i^N stands for *i*-th utterance of *N*-th input. *Symptom Scores* are $|L|$ real numbers, where $|L|$ is the number of symptoms to predict.

each utterance of the interview is encoded separately by a word-level encoder. This is far from optimal since most of the utterances are short (<10 tokens); thus, a lot of computation is wasted on padding tokens. Instead, the utterances are concatenated into one input text separated by the `[SEP]` special token. This way, the number of passes through the encoder is reduced by ~40 times for each input. After, we perform the *Mean `[SEP]` pooling* on the tokens representing each utterance to get the final utterance representation. The overview of the model architecture is presented in Figure 4.

4.3. Results and Discussion

Experimental setup. We used two pre-trained models in the word-level encoder of our architecture: BERT-Base model (Devlin et al., 2018) and MentalBERT (Ji et al., 2022). Due to the time difference between the experiments, psychiatrists annotations were originally tested using `all-mpnet-base` model[1] as a pre-trained model in the word-level encoder. To make the comparison smoother, we additionally used BERT-Base and MentalBERT as pre-trained models for the psychiatrists' annotations[2]. As for the data, we tested our approach on the two datasets: DAIC-WOZ for lexicon and psychiatrists annotations and PRIMATE for lexicon annotations (see Section 2.2.2 for more details). Since the train, validation, and test splits are not provided with the PRIMATE dataset, we randomly split the data using an 80/10/10 ratio.

[1] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[2] This explains slight differences between the results reported in this chapter and in Publication III; however, the findings stay the same.

Model	LOI	DEP	SLE	ENE	EAT	LSE	CON	MOV	PHQ-8
BERT	0.56	**0.63**	0.77	0.87	0.81	0.78	0.74	0.34	4.38
+SDD	0.70	0.88	0.94	0.94	1.00	0.97	0.87	0.34	5.60
+AFINN	**0.50**	0.70	0.79	0.81	0.85	0.72	0.77	0.34	4.56
+NRC	**0.50**	0.66	**0.73**	0.77	0.81	0.71	**0.73**	0.34	**4.31**
+ALL-LEX	**0.50**	0.69	0.81	**0.74**	0.81	**0.69**	0.74	0.34	4.56
+PA	0.52	0.68	0.80	0.83	**0.79**	0.75	0.77	0.34	4.65
+RAND	0.59	0.69	0.77	0.81	0.82	0.74	0.77	0.34	4.59
MEBERT	0.59	0.64	0.91	0.92	0.89	0.71	0.71	0.35	4.71
+SDD	0.69	0.72	0.89	0.92	0.93	0.85	0.78	0.34	5.07
+AFINN	0.48	0.62	0.71	0.78	0.79	0.70	0.74	0.34	4.27
+NRC	0.60	0.68	0.71	0.78	0.80	0.74	0.71	0.34	4.35
+ALL-LEX	**0.44**	**0.55**	**0.63**	**0.72**	**0.69**	0.67	**0.67**	0.34	**3.59**
+PA	0.51	0.58	0.81	0.84	0.83	**0.64**	0.70	0.34	4.26
+RAND	0.58	0.69	0.70	0.78	0.83	0.72	0.72	0.34	4.50
SOTA	0.53	**0.55**	0.75	**0.64**	0.81	**0.62**	0.83	0.44	3.78
HUMAN	**0.44**	0.66	**0.56**	0.70	–	0.88	–	–	–

Table 7: Results for the DAIC-WOZ test set. The mean MAE is reported for five runs. For symptom scores, the standard deviation is $0.00 \leq \sigma \leq 0.12$; for the PHQ-8 score, the standard deviation is $0.13 \leq \sigma \leq 0.42$. MEBERT is short for MentalBERT. The best MAE for each symptom is **in bold**. SOTA means current state-of-the-art results in the literature (**Milintsevich, Kirill** et al., 2023).

Results. Table 7 shows the results for the DAIC-WOZ dataset. Additionally, we finetuned the +RAND version of both BERT and MEBERT to verify if the improvement comes only from the input marking by randomly marking 8% of the words in each interview. The results showed slight overall improvement when the NRC lexicon was introduced to the BERT model. The combination of all lexicons is marginally beneficial only for some symptoms, and results have deteriorated with the exclusive introduction of the SDD lexicon. On the other hand, for the MEBERT model, the combination of all lexicons (+ALL-LEX) produces the best results overall, both symptom-wise and for the global PHQ-8 score.

Psychiatrists' annotations showed behavior similar to that of the lexicons on the BERT model, i.e., without clear improvement. For the MEBERT model, psychiatrists' annotations showed consistent improvement for all symptoms, although to a lesser extent than the combination of all the lexicons. Additionally, +RAND models performed on the same level as the baseline models, suggesting that the content of the marking is the key part influencing the performance of the model and not the input markings themselves.

We also compared neural models to the human annotators. For this, we have tasked our MHPs with completing the self-assessment PHQ-8 questionnaire on behalf of each patient only based on their interview transcripts. Missing values in

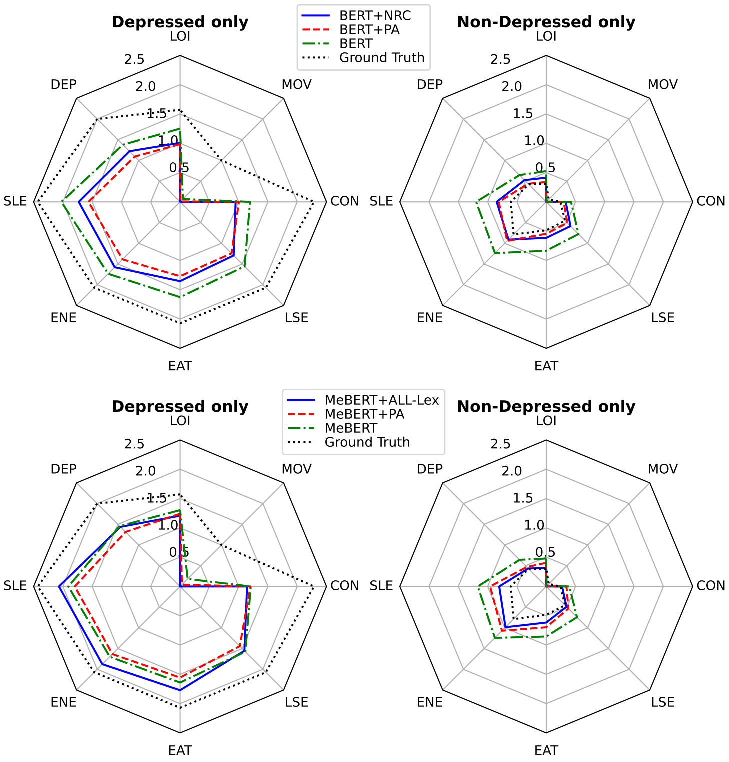


Figure 5: Average predicted values for depressed and non-depressed patients of the DAIC-WOZ test set.

Table 7 for eating, (EAT), concentration (CON), and movement (MOV) problems are due to a low number of annotated transcripts, i.e., human annotators did not find any sufficient evidence in the texts of most transcripts to assign a score to a symptom. The results showed that the best-performing model, MEBERT+ALL-LEX, performed on par or better than the human annotators on all symptoms except sleeping problems (SLE) and lack of energy (ENE).

Figure 5 depicts a more detailed overview of the best-performing lexicon-based models: BERT+NRC and MeBERT+ALL-Lex, as well as the models using psychiatrists' annotations: BERT+PA and MeBERT+PA. The results show that the improvement for the BERT+NRC model comes from the non-depressed population, while it loses to the baseline model for the depressed population. The MeBERT+All-Lex model, however, improves for both depressed and non-depressed populations. BERT+PA falls behind the lexicon-infused model in both depressed and non-depressed populations; the same is true for MeBERT+PA.

Table 8 shows the results for the PRIMATE dataset. Contrary to the results on the DAIC-WOZ, introducing external knowledge failed to improve performances

Model	LOI	DEP	SLE	ENE	EAT	LSE	CON	MOV	SUI
BERT	**0.59**	**0.65**	0.81	0.62	0.75	0.60	**0.65**	0.81	0.82
+SDD	0.58	0.62	0.81	**0.64**	0.74	**0.63**	0.63	**0.82**	0.82
+AFINN	0.57	0.60	0.80	0.62	0.76	0.59	0.64	0.81	**0.83**
+NRC	0.55	0.62	**0.82**	0.60	0.79	0.59	0.61	0.80	0.82
+ALL-Lex	0.56	0.63	0.79	0.61	**0.80**	0.58	0.61	**0.82**	0.82
+Rand	0.56	0.63	0.80	0.61	0.77	0.59	0.62	0.80	**0.83**
MeBERT	0.58	0.58	0.82	0.62	0.78	0.60	0.62	**0.82**	0.84
+SDD	0.53	**0.60**	**0.83**	0.62	0.79	0.60	0.61	0.81	**0.86**
+AFINN	0.57	0.55	**0.83**	0.62	0.79	**0.63**	0.58	0.81	0.85
+NRC	0.57	0.58	0.82	**0.63**	0.79	**0.63**	0.61	0.80	0.85
+ALL-Lex	0.56	0.59	0.80	0.62	**0.80**	0.61	**0.63**	**0.82**	0.84
+Rand	**0.60**	0.59	0.78	0.62	0.75	0.62	0.61	0.81	0.83

Table 8: Results for the PRIMATE test set. The mean macro-F1 score is reported for five runs. The best macro-F1 for each symptom is **in bold**. As standard splits are not provided, we cannot present Sota results.

for PRIMATE. The models that used the lexicon input marking showed signs of improvement for some symptoms yet were largely inconsistent.

Discussion. The results from the DAIC-WOZ show that PLMs can indeed benefit from the introduction of external knowledge about the sentiment and emotional value of the words. Surprisingly, the introduction of the depression-specific lexicon had the opposite effect. We hypothesize that two reasons could cause it. First, SDD covers less than 0.5% of words in the interview, almost 15 times less than AFINN and NRC. Thus, the introduced signal might be too weak for the model to learn. Second, the SDD lexicon was based on Twitter data, while DAIC-WOZ contains transcripts of real conversations. From our observations, the people describe their problems more explicitly in their social media posts. At the same time, DAIC-WOZ conversations are more generally themed, and the PHQ-8 scores are based on the person's self-assessment test rather than the conversations themselves. This brings us back to the conceptual difference between the DAIC-WOZ and PRIMATE datasets. While the first one aims at establishing the link between the underlying person's mental condition and their speech, the latter one sets a goal of detecting whether a particular symptom is mentioned in the text. This difference might explain the greater impact of the AFINN and NRC lexicons on modeling the DAIC-WOZ dataset.

4.4. Exploring Model's Attention

By analyzing the models' attention mechanism, we investigated how much the models already know about the lexicon content by itself and whether the models learn to use the marked content. In particular, we wanted to see how much the models already pay attention to the words in our lexicon without any marking and whether marking the lexicon words will make the models pay more attention to these words. For that purpose, we defined the relative lexicon attention score S_h^l for each attention head h of each layer l, which was calculated as shown in Equation 4.1 where T refers to all input tokens in the dataset, Lex is a set of lexicon tokens, and $A_h^l(t_i)$ is the attention score of token t_i. A higher relative lexicon attention score shows that the attention scores that the model assigns to the tokens from the lexicon are higher than the attention scores for the other tokens.

$$S_h^l = \frac{1}{|T|} \frac{\sum_{i=1}^{|T|} A_h^l(t_i) \cdot \mathbb{1}_{t_i \in Lex}}{\sum_{i=1}^{|T|} A_h^l(t_i)} \tag{4.1}$$

Figure 6 presents the relative lexicon attention scores S_h^l for three models: the pre-trained model (MentalBERT) without any fine-tuning, fine-tuned on DAIC-WOZ MEBERT, and MEBERT+ALL-LEX, which were tested on the DAIC-WOZ interviews with and without input markings. Results show that models have more uniform lexicon attention scores when no input markings are used `[A-C]`. Input marking makes the attention scores higher for the marked tokens, even for the models that did not have marked data during training, which is shown by a larger light-colored area in `[D, E]`. Fine-tuning on marked data has an even greater effect on attention scores `[F]`. This evidence suggests that input marking is an effective strategy to guide model attention. Additionally, even when the input text has no markings, the fine-tuned MEBERT model has higher attention scores for words from the ALL-LEX lexicon `[B]` compared to the model that was not fine-tuned on the DAIC-WOZ `[A]`. In conclusion, this attention score analysis shows that although the models learn to use the markings by paying more attention to the marked words, fine-tuning the model on the DAIC-WOZ data already induces the importance of the sentimental and emotional words[3].

We concluded a similar experiment for the psychiatrists' annotations. Unlike lexicons, the psychiatrists' annotations are not limited to individual words or phrases. Hence, we investigated the attention scores in the utterance encoder. For each turn u_t, we computed an average attention score $\mathbb{S}_t$ which is defined as:

$$\mathbb{S}_t = \frac{1}{l \cdot h} \sum_{i=1}^{l} \sum_{j=1}^{h} A_j^i(u_t) \tag{4.2}$$

where l is a layer, h is an attention head, and $A_h^l(u_t)$ is the attention score of turn u_t at layer l and attention head h. Figure 7 shows the distribution of average attention scores over the turns. Interestingly, MEBERT+PA and MEBERT+ALL-LEX models show clear attention clusters, dividing each interview into four parts. This partitioning follows the structure of the interviews in the DAIC-WOZ dataset, where each conversation starts with a general discussion to make the patient feel

[3] Models based on BERT show similar results.

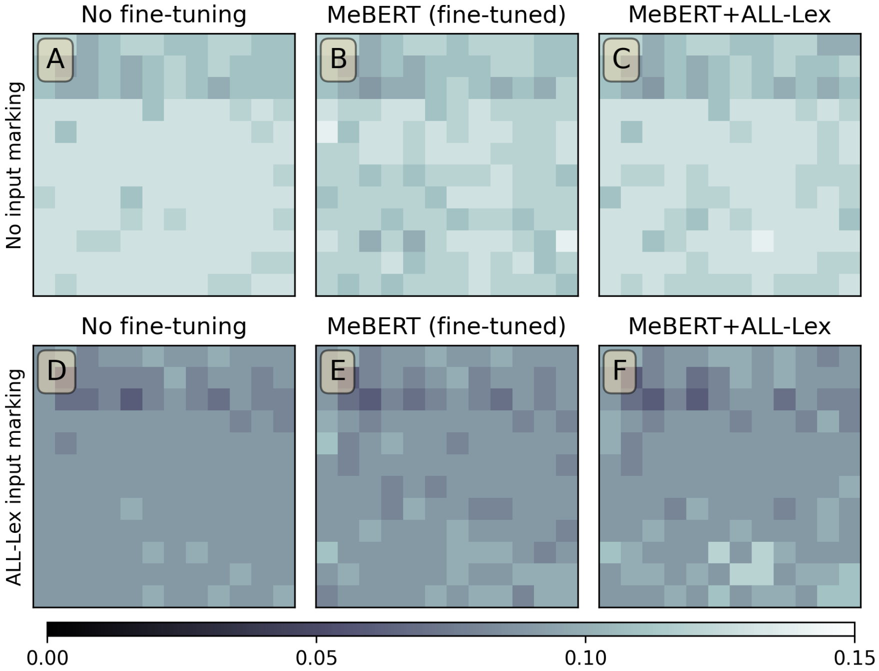
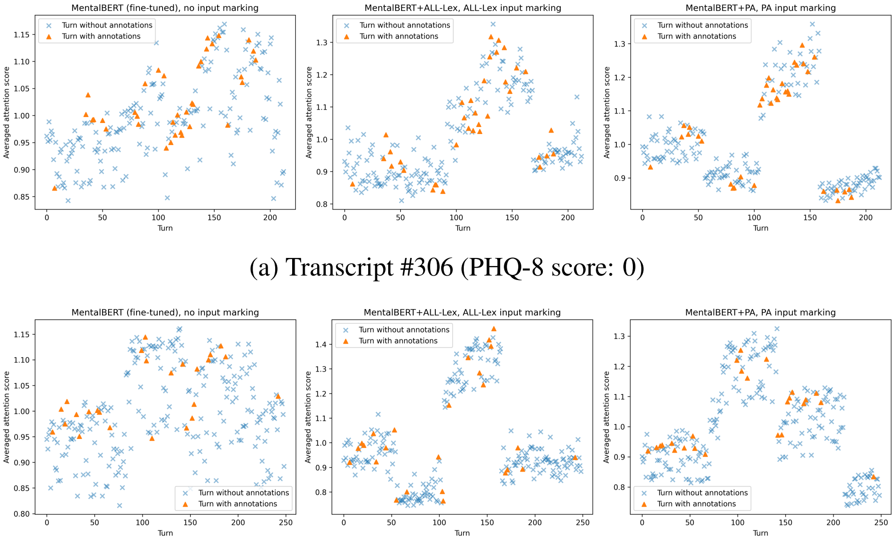


Figure 6: Relative lexicon attention scores. For each heatmap, the rows and columns correspond to layers and attention heads, respectively. The top row `[A-C]` shows the relative attention scores for the models tested on the inputs without any markings, and the bottom row `[D-F]` shows the scores tested on the inputs with ALL-LEX markings. The results are obtained on the test split of the DAIC-WOZ.

more comfortable, followed by more depression-targeted questions, and finishes with a cool-down phase to make the patient feel at ease again (Gratch et al., 2014). While both MEBERT fined-tuned without input markings, MEBERT+PA and MEBERT+ALL-LEX generally assign higher attention scores to the middle parts of the interview, MEBERT+PA and MEBERT+ALL-LEX assign attention scores in a more targeted way[4].

These results bring us to an interesting conclusion. Input marking seems to serve as an attention-guiding mechanism for all the models that we used in the experiments. However, not all the models benefit from this in the same way: MEBERT showed the highest performance boost when external knowledge was introduced via the input marking, while BERT and `all-mpnet-base` demonstrated only slight improvement or even slight decrease in the performance.

[4] Models based on `BERT` and `all-mpnet-base` show similar results.



(a) Transcript #306 (PHQ-8 score: 0)

(b) Transcript #332 (PHQ-8 score: 18)

Figure 7: Average turn attention scores.

4.5. Conclusions and Future Work

The work presented in this chapter was a logical continuation of the experiments presented in the previous chapter. We showed that pre-trained language models could still benefit from existing lexical resources for symptom-based depression estimation (**RQ2**). In particular, we discovered that a domain-specific PLM, like MentalBERT, benefits from the lexicon-based external knowledge and, though to a lesser extent, from the psychiatrists' expertise, more than a general-domain PLM like BERT. Further analysis of the attention scores suggested that the input marking played an attention-guiding role during fine-tuning, redirecting the model's attention toward the marked areas in the input on the word level and toward the depression-related interview parts on the utterance level. Moreover, we presented an incremental improvement of the neural architecture to model text in dialog format. The improved model uses a transformer-based utterance-level encoder and requires less computation power for training and inference by virtue of optimized input representation. Finally, conflicting results on the PRIMATE dataset raised suspicions about the annotation quality, which we will continue to study in the next chapter. In future work, we plan on experimenting with other methods of external knowledge introduction to the transformer-based models, for example, by modifying the attention mechanism or loss function. Furthermore, to better understand the model's behavior, we can use more faithful and sophisticated methods of constructing saliency maps, like ALTI (Ferrando et al., 2022) instead of simple attention weights exploration.

5. SOCIAL-MEDIA-BASED DEPRESSION DATASETS VALIDITY (PUBLICATION IV AND DATASET I)

So far, we have predominantly discussed the methods for symptom-based depression estimation (Chapter 3) and investigated whether the incorporation of depression and sentiment lexicons can help to improve the symptom detection from text (Chapter 4). While the results in the previous chapter showed that lexicons did help for the DAIC-WOZ dataset, they did nothing substantial for the PRIMATE (Gupta et al., 2022) dataset. At first, we experimented with the more performant pre-trained language models (PLM), expecting better performance after fine-tuning. However, the other models still failed to show any improvements for the PRIMATE dataset, leading us to investigate the annotations in more detail. A practicing clinical psychology intern[1] reannotated a subset of PRIMATE data for the lack of interest in doing things (anhedonia) symptom (LOI) with more fine-grained labels and span-based explanations. As a result, the new annotations showed extremely low agreement with the original labels, which raised concerns about the validity of this dataset.

5.1. Benchmarking Pre-Trained Models on PRIMATE

In the previous chapter, we saw that, unlike for DAIC-WOZ, predictions for PRIMATE benefited neither from the MentalBERT pre-trained model nor from lexicon information. Moreover, the previous chapter showed that the choice of the base model could significantly affect the performance. Thus, the first goal was to experiment with different base models of various sizes to see if any of those make a difference for PRIMATE.

Experimental setup. We fine-tuned multiple state-of-the-art transformer-based pre-trained language models (PLMs) on the PRIMATE dataset, ranging from 66 to 345 million parameters. We first chose DistilBERT (Sanh et al., 2019) as a baseline and BERT-Base (Devlin et al., 2018), RoBERTa-Base, RoBERTa-Large (Liu et al., 2019), DeBERTa-Base, and DeBERTa-Large (He et al., 2020) as higher-performing models. In particular, DeBERTa has shown constant improvements in various NLP tasks and replaced BERT and RoBERTa as the state-of-the-art model for many of them.[2] We used the same splits as in Chapter 4.

Results. The results presented in Table 9 showed that larger models, such as RoBERTa-Large and DeBERTa-Large, performed better on average than other models. However, the improvement is marginal, specifically for the DeBERTa-Large model, which is very close to the DistilBERT baseline. Concerning the symptoms, RoBERTa-Large and DeBERTa-Large performed better for predicting lack of energy (ENE), low self-esteem (LSE), hyper or lower activity (MOV), and

[1] Dr. Kairit Sirts—one of the supervisors of this thesis.

[2] https://gluebenchmark.com/leaderboard

Model	LOI	DEP	SLE	ENE	EAT	LSE	CON	MOV	SUI	Avg
DistilBERT	**.64**	.88	.67	.58	.60	.90	.50	.67	.81	.69
BERT-Base	.55	.88	.66	.55	.63	.90	.46	.66	.79	.68
RoBERTa-Base	.54	.88	.70	.57	.57	.90	**.51**	.69	**.85**	.69
RoBERTa-Large	.57	.86	**.75**	.63	**.65**	**.91**	.52	.71	**.85**	**.72**
DeBERTa-Base	.58	**.91**	.69	.52	.42	.90	.36	.61	.81	.64
DeBERTa-Large	.60	.90	.68	**.64**	.47	**.91**	.50	**.73**	.83	.70

Table 9: Symptom-wise F1-scores on the validation set.

suicidal thoughts (SUI). Additionally, the depressed mood (DEP) symptom showed slight improvement with DeBERTa models; however, decreased performance for eating disorder (EAT) symptom. RoBERTa models performed better for the sleeping disorder (SLE) and suicidal thoughts (SUI) prediction. Nevertheless, DistilBERT performed on par with larger models overall, setting a strong baseline. Finally, anhedonia (LOI) showed a decrease in performance for all the models compared to the DistilBERT.

5.2. Reannotation of PRIMATE

Weak performance across models of different sizes prompted us to put the annotations from the PRIMATE dataset under the magnifying glass. Specifically, we focused on the lack of interest (LOI) symptom. According to the DSM-5, anhedonia (LOI) is one of the core symptoms of depression. In addition, the results from Table 9 showed diminished and unstable performance for anhedonia (LOI). Furthermore, the cross-evaluation in Figure 8 revealed that if we used the predictions of the DistilBERT baseline for the lack of interest (LOI) symptom as the predictions for the depressed mood (DEP) and lack of self-esteem (LSE) symptoms, we would get the F1-scores of 0.68 and 0.66 correspondingly, which is higher than then F1-score of 0.64 for the lack of interest (LOI) symptom itself.

Reannotation. We investigated the diminished performance of the anhedonia (LOI) symptom by reannotating a subset of the validation set. A total of 170 texts from the validation set have been chosen for reannotation based on the predictions of the DistilBERT-based model; if at least one symptom was predicted incorrectly, the text was added to the reannotation subset.

The annotations were carried out based on the symptom description in the Montgomery-Åsberg Depression Rating Scale (MADRS) (Montgomery & Åsberg, 1979). MADRS is a ten-item clinician-rated questionnaire to assess the severity of the symptoms. The DSM-5 loss of interest (LOI) symptom is captured by one of the questions in MADRS, which is called “Inability to feel” and is described as “representing the subjective experience of reduced interest in the surroundings, or activities that normally give pleasure. The ability to react with adequate emotion to circumstances or people is reduced”.

A mental health professional (MHP) read all the posts in the subset and labeled

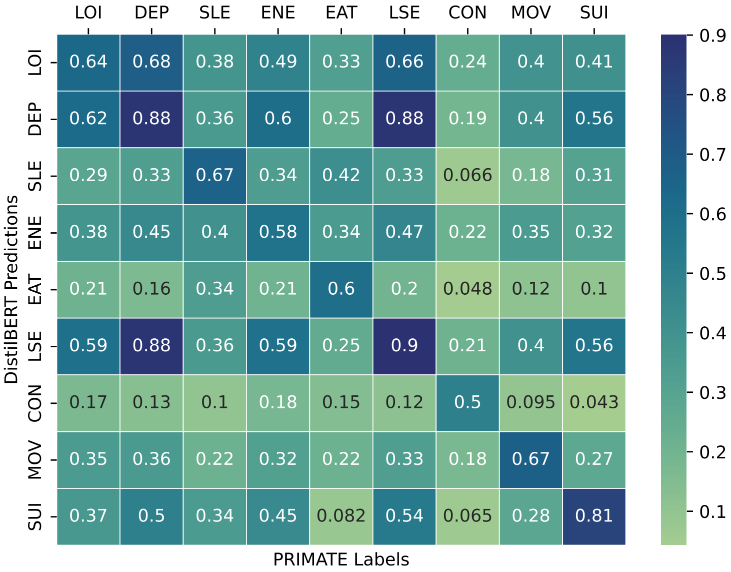


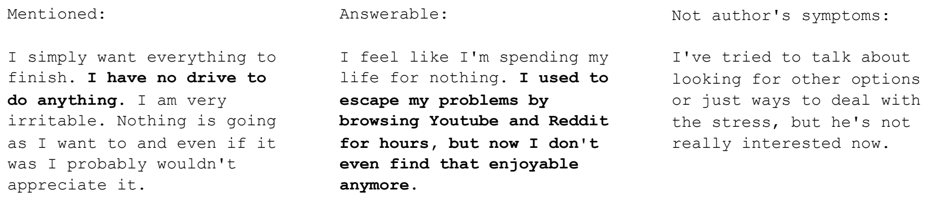
Figure 8: Cross-evaluation of DistilBERT predictions against PRIMATE labels on the validation set. The values inside of each cell represent F1-scores. Example of reading the graph: the value in the intersection of the first row and the second column represents the F1-score between the predictions of the DistilBERT baseline for the lack of interest (LOI) symptom and the PRIMATE labels for the depressed mood (DEP) symptom.

them for the presence of loss of interest or pleasure (anhedonia) following the MADRS symptom description. The MHP assigned four labels to each post: a) “mentioned” if the symptom is talked about in the text, but it is not possible to infer its duration or intensity; b) “answerable” if there is clear evidence of anhedonia; c) “writer’s symptoms” which shows whether the author of the post discusses themselves or a third person; d) “absence” if there is no mention of the symptom in the text. Additionally, the MHP selected the part of the text that supports the positive label.

Figure 9 shows examples for the reannotated posts.[3] Here, in the first example, it is not clear from the text whether the highlighted sentence is about lack of interest (LOI) or lack of energy (ENE). Hence, it is annotated as mentioned but is not answerable. The second example contains a clear indication that the person had the activities that they found enjoyable previously and not anymore, thus suggesting the loss of interest (LOI) in particular.

To compare the annotations on the reannotated subset, we measured DistilBERT against the “mentioned” and “answerable” labels from the new annotation and the original PRIMATE labels. As seen from Table 10, the model fine-tuned on the original labels performed considerably worse on our labels than against the

[3] All example posts are paraphrased for privacy.



Mentioned:

I simply want everything to finish. **I have no drive to do anything.** I am very irritable. Nothing is going as I want to and even if it was I probably wouldn't appreciate it.

Answerable:

I feel like I'm spending my life for nothing. **I used to escape my problems by browsing Youtube and Reddit for hours, but now I don't even find that enjoyable anymore.**

Not author's symptoms:

I've tried to talk about looking for other options or just ways to deal with the stress, but he's not really interested now.

Figure 9: Examples of reannotated posts. Evidences are highlighted in **bold**.

Predictions	Against PRIMATE				Against "mentioned"				Against "answerable"			
	A	P	R	F1	A	P	R	F1	A	P	R	F1
DistilBERT	.58	.56	.62	.58	.56	.30	.71	.42	.51	.10	.75	.18
PRIMATE Labels	-	-	-	-	.56	.27	.58	.37	.54	.09	.58	.15

Table 10: Results on the reannotated part of the validation set. Here, **A** stands for Accuracy, **P** for Precision, **R** for Recall, and **F1** for F1-score for the positive class.

original labels from PRIMATE. At the same time, when the original PRIMATE labels were used as predictions, they performed worse against our annotations than the predictions of the model fine-tuned on the original labels. This result was unsurprising given the extremely low agreement between these sets of labels with Cohen's kappa of 9% and 3%, respectively. Furthermore, the most common error type was a false positive, i.e., a symptom marked as present in PRIMATE when our MHP found no evidence of it in the text. This difference is also reflected in Table 11, where the number of positive labels is considerably smaller in our reannotated subset than in the original PRIMATE annotation.

Discussion. Our findings are consistent with the original results presented by Gupta et al. (2022). Similar to our experiment, they also trained a classifier based on the BERT-Base model and reported low prediction scores for the LOI symptom. We found that the size and underlying performance of the base model did not have an effect, and the best performance on this symptom was obtained by fine-tuning the smallest DistilBERT model. The subset of data reannotated by an MHP obtained very low agreement scores with the original annotations, showing that unreliable annotations can be the cause of poor prediction results. Additionally, we noticed that many posts that were mistakenly labeled with LOI are more closely related to the "inner tension" symptom from the MADRS.

While we agree that our reannotated test set is also susceptible to errors to some extent, we believe it serves as a more reliable benchmark for the anhedonia symptom. A more fine-grained labeling scheme reduces the risk of mislabelling and is more transparent for further verification. Finally, it lays the foundation for future collaboration to produce a higher-quality Reddit-based dataset for depression symptom estimation.

Labels	Positive	Negative
PRIMATE	81	89
Mentioned	38	132
Answerable	12	158

Table 11: Number of positive and negative labels for the lack of interest (LOI) for PRIMATE annotations and our “mentioned” and “answerable” annotations.

5.3. Conclusions and Future Work

In this chapter, we presented a detailed study of PRIMATE, one of the few publicly available social-media depression datasets with symptom-based annotations. First, we carried out a comparative study of different pre-trained language models by fine-tuning them on the PRIMATE dataset. This benchmarking showed that irrespective of the PLM chosen for fine-tuning, they failed to improve the results. This behavior brought us to reannotate the lack of interest (LOI) symptom with the help of a mental health professional. During the reannotation process, we found that the original PRIMATE annotations for the lack of interest (LOI) symptom are inconsistent with the symptom definition. As a result, we produced a new annotation for a subset of 170 texts from the PRIMATE dataset.

With this chapter, we advocate for a more rigorous and standardized approach to mental health dataset annotation, emphasizing the need for greater involvement of domain experts in the annotation process. We also show, on the example of the lack of interest (LOI) symptom, that a clear symptom definition is crucial to reliably annotate depression-related textual data (**RQ3**).

Furthermore, after the publication of this paper, we plan to continue to work on the annotations and increase the number of annotated posts. We released the annotations under free access (Dataset I); however, corresponding texts must be obtained from the authors of the original PRIMATE dataset. We plan to expand this topic and apply our experience to producing expert-annotated datasets in French and Estonian.

6. CONCLUSION

Major Depressive Disorder (MDD) is a prevalent psychiatric condition worldwide, significantly contributing to disability and increasing the risk of suicide. Recent studies have indicated a rise in depression levels in countries like France and Estonia and globally, particularly after the COVID-19 pandemic. Despite this, mental illnesses often face stigma, limiting access to psychiatric treatment and diagnosis. Early detection of depression is crucial for effective prevention and treatment, highlighting the need for automatic depression detection systems.

Automatic detection of depression from texts has long been a focus of NLP and linguistic research. Studies have demonstrated distinct linguistic patterns between depressed and non-depressed individuals. Methods have evolved from simple linguistic analysis to sophisticated machine and deep learning models applied to social media texts and clinical interview transcriptions. The common strategy of approaching automatic depression estimation from text as a binary classification task is widely used for depression assessment. Although it simplifies the diagnostic picture, it potentially overlooks critical symptomatic details. Furthermore, high-quality data for depression detection is scarce, with clinical datasets often restricted by regulations. Social media data, while abundant, typically lacks professional oversight in labeling, raising concerns about data validity and the need for expert involvement in the annotation process.

In Chapter 2 of this work, we aimed to connect the two worlds: NLP and clinical research. The study of recent related works showed a disconnection between the two domains. On one side, the NLP community treats depression as a binary problem. In addition, the collaboration between the NLP researchers and mental health professionals is often absent in the data annotation process. On the other side, mental health research advocates for a symptom-based approach to depression, i.e., treating depression not as a binary diagnosis but rather as a network of symptoms.

6.1. Main Conclusions

Symptom-based depression prediction. We began our research by exploring how predicting depression as a collection of symptoms compares to the binary classification approach. As described in Chapter 3, we developed a neural architecture that achieved state-of-the-art results in symptom-based depression estimation. This architecture also served as the foundation for the experiments conducted in Chapter 4. We found that the symptom-prediction model performed on par or better compared to binary classification or single regression depression severity models while simultaneously providing more descriptive and personalized symptom profiles (**RQ1**).

External knowledge integration. In Chapter 4, we continued our work on symptom-based depression prediction. First, we introduced incremental improvements to the neural architecture to better model text in dialog format. Second, we demonstrated that some pre-trained language models (PLM) can still gain advantages from existing lexical resources for symptom-based depression estimation. Specifically, we found that—for the DAIC-WOZ dataset—the selection of the base model is important; while MentalBERT benefited consistently from the included lexicon information, BERT did not (**RQ2**). As often happens in research, not all the results were conventionally positive. In particular, PRIMATE, the social-media-based dataset, demonstrated no improvement. In search of the reason behind this poor performance, we addressed the annotation quality of this dataset, prompting the research detailed in Chapter 5.

Annotation validity. In Chapter 5, we showed, on the example of the lack of interest or pleasure in doing things (anhedonia) symptom, the importance of a clear symptom definition to reliably annotate depression-related textual data (**RQ3**). As a result, we built a higher-quality social-media text dataset for anhedonia detection, which is one of the core symptoms of depression. We have made these annotations freely accessible as Dataset I.

6.2. Limitations and Ethical Considerations

This work also has several limitations. First, our work is limited to the DAIC-WOZ and PRIMATE datasets, one of the few datasets with symptom-based labels easily obtainable from their authors. However, DAIC-WOZ is relatively small to use for training powerful models, making results analysis challenging. The dataset also has a quite rigid structure, as all interview prompts are sampled from a closed set of prompts. Thus, we cannot assume the generalizability of the presented results to other datasets, limiting our model's applicability. By maintaining high standards of the code used in our experiments and making it publicly available, we hope that the research community will be able to replicate our experiments on different datasets.

The main motivation for predicting symptoms instead of binary diagnostic classes, total depression severity, or discrete severity class, as has been custom in previous works, is to align the computational task with the depression diagnosis definition defined in popular psychiatric nosologies such as DSM-5 or ICD-11

We also acknowledge the limitations of the re-annotated subset of the PRIMATE dataset presented in Chapter 5. First, the manually annotated explanations only show what information a clinician might find in the content of a Reddit post. This information does not necessarily assess the real mental state of the author of the post, which would require a true clinical setting. Furthermore, our re-annotation was carried out by only one mental health professional, which does not allow for calculating an inter-annotator agreement analysis. Finally, anhedonia, or lack of interest in doing things, is extremely challenging to conceptualize (Winer et al., 2019), and binary labels may not be the best choice when the difference between the presence and absence of the symptom is marginal.

We acknowledge the potential ethical aspects of the work that studies the methods to detect someone's mental health status unobtrusively. Here, we are using publicly available datasets collected for research purposes. Also, the lexicons we use are publicly available and have not been composed based on private confidential material. If such a system that could predict the presence of depression symptoms based on actual clinical interviews would be deployed in practice, it would require the informed consent of all participants involved as well as the understanding of the validity boundaries of such systems, meaning that the predictions of such systems cannot replace the assessment of trained clinicians, but rather assist them in their activities.

6.3. Future work

Thus far, we have researched and answered all the research questions of this thesis. Nevertheless, we can clearly see several paths to continue this research. First, with the rising popularity of Large Language Models (LLM), their application to depression estimation also gains traction in research (Y. Wang et al., 2024; Xu et al., 2024; K. Yang et al., 2023; K. Yang et al., 2024). Such properties as longer context length and the ability to generate explanations might seem advantageous for this domain. However, their bias and proneness to hallucinations have to be seriously taken into account (Heston, 2023). One rather obvious direction of applying LLMs to the depression estimation task is to estimate the depression symptoms intensity from text. Furthermore, the generative capabilities of the LLMs can be exploited to produce more data or to assist in data annotation (Pérez et al., 2023). It can also be leveraged to generate explanations, as it has been recently done for suicide risk estimation at the CLPsych 2024 shared task (Chim et al., 2024). Finally, rigorous evaluation of safety and potential ethical and health risks for using LLMs in clinical scenarios is highly important.

Second, other approaches to external knowledge introduction have yet to be explored. For example, external knowledge could be infused directly into the attention mechanism of the transformer model (Bai et al., 2022; Z. Li et al., 2021; S. Wang et al., 2022). Alternatively, the loss function can be tweaked during training such that it penalizes the model if its attention score on specific spans of text is low (Stacey et al., 2022). These methods could be adapted for depression symptom estimation and compared to the approach proposed in this thesis to further solidify the hypothesis that PLMs could still benefit from the domain-specific external knowledge for automatic depression symptom estimation.

Finally, cooperating with mental health professionals to produce high-quality and publicly available datasets is extremely important for the field. So far, we have annotated a small-scale dataset for one symptom. Undoubtedly, annotating more texts with other symptoms and collecting data for languages other than English is the direction to take. We plan to continue working with the A^{2}M^{2}P Hospital-University Federation to annotate more data in French. Additionally, annotating depression data in Estonian is planned to be carried out in collaboration with the University of Tartu.

BIBLIOGRAPHY

Agarwal, N., Dias, G., & Dollfus, S. (2024a). Analysing relevance of discourse structure for improved mental health estimation. In A. Yates, B. Desmet, E. Prud'hommeaux, A. Zirikly, S. Bedrick, S. MacAvaney, K. Bar, M. Ireland, & Y. Ophir (Eds.), *Proceedings of the 9th workshop on computational linguistics and clinical psychology (clpsych 2024)* (pp. 127–132). Association for Computational Linguistics. https://aclanthology.org/2024.clpsych-1.9

Agarwal, N., Dias, G., & Dollfus, S. (2024b). Multi-view graph-based interview representation to improve depression level estimation. *Brain Informatics*.

Agarwal, N., **Milintsevich, Kirill**, Metivier, L., Rotharmel, M., Dias, G., & Dollfus, S. (2024). Analyzing symptom-based depression level estimation through the prism of psychiatric expertise. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)* (pp. 974–983). ELRA; ICCL. https://aclanthology.org/2024.lrec-main.87

Al-Mosaiwi, M., & Johnstone, T. (2018). In an absolute state: elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, *6*(4), 529–542.

American Psychiatric Association. (2022). *Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR*.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bai, J., Wang, Y., Sun, H., Wu, R., Yang, T., Tang, P., Cao, D., Zhang1, M., Tong, Y., Yang, Y., Bai, J., Zhang, R., Sun, H., & Shen, W. (2022). Enhancing self-attention with knowledge-assisted attention maps. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 107–115). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.8

Bathina, K. C., Ten Thij, M., Lorenzo-Luaces, L., Rutter, L. A., & Bollen, J. (2021). Individuals with depression express more distorted thinking on social media. *Nature human behaviour*, *5*(4), 458–466.

Beck, A. T. (1979). *Cognitive therapy and the emotional disorders*. Penguin.

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. F. (1996). Comparison of beck depression inventories-ia and-ii in psychiatric outpatients. *Journal of personality assessment*, *67*(3), 588–597.

Beck, A. T., Steer, R. A., & Carbin, M. G. (1988). Psychometric properties of the beck depression inventory: twenty-five years of evaluation. *Clinical psychology review*, *8*(1), 77–100.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: the long-document transformer. *arXiv preprint arXiv:2004.05150*.

Belvederi Murri, M., Amore, M., Respino, M., & Alexopoulos, G. S. (2020). The symptom network structure of depressive symptoms in late-life: results from a european population study. *Molecular psychiatry*, *25*(7), 1447–1456.

Borchani, H., Varando, G., Bielza, C., & Larranaga, P. (2015). A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *5*(5), 216–233.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, 1–47.

Burdisso, S., Reyes-Ramírez, E., Villatoro-Tello, E., Sánchez-Vega, F., López-Monroy, P., & Motlicek, P. (2024). Daic-woz: on the validity of using the therapist's prompts in automatic depression detection from clinical interviews. *arXiv preprint arXiv:2404.14463*.

Burdisso, S., Villatoro-Tello, E., Madikeri, S., & Motlicek, P. (2023). Node-weighted Graph Convolutional Network for Depression Detection in Transcribed Clinical Interviews. *Proc. INTERSPEECH 2023*, 3617–3621. https://doi.org/10.21437/Interspeech.2023-1923

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, *11*, 2079–2107.

Chim, J., Tsakalidis, A., Gkoumas, D., Atzil-Slonim, D., Ophir, Y., Zirikly, A., Resnik, P., & Liakata, M. (2024). Overview of the CLPsych 2024 shared task: leveraging large language models to identify evidence of suicidality risk in online posts. In A. Yates, B. Desmet, E. Prud'hommeaux, A. Zirikly, S. Bedrick, S. MacAvaney, K. Bar, M. Ireland, & Y. Ophir (Eds.), *Proceedings of the 9th workshop on computational linguistics and clinical psychology (clpsych 2024)* (pp. 177–190). Association for Computational Linguistics. https://aclanthology.org/2024.clpsych-1.15

Chua, H., Caines, A., & Yannakoudakis, H. (2022). A unified framework for cross-domain and cross-task learning of mental health conditions. *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, 1–14.

Chung, C., & Pennebaker, J. (2011). The psychological functions of function words. In *Social communication* (pp. 343–359). Psychology Press.

Coppersmith, G., Dredze, M., & Harman, C. (2014a). Quantifying mental health signals in twitter. *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, 51–60.

Coppersmith, G., Dredze, M., & Harman, C. (2014b). Quantifying mental health signals in Twitter. In P. Resnik, R. Resnik, & M. Mitchell (Eds.), *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality* (pp. 51–60). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3207

Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). From ADHD to SAD: analyzing the language of mental health on Twitter through self-reported diagnoses. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 1–10. https://doi.org/10.3115/v1/W15-1201

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). CLPsych 2015 shared task: depression and PTSD on Twitter. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 31–39. https://doi.org/10.3115/v1/W15-1204

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Social media as a measurement tool of depression in populations. *Proceedings of the 5th annual ACM web science conference*, 47–56.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., et al. (2014). Simsensei kiosk: a virtual human interviewer for healthcare decision support. *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 1061–1068.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., & Smith, N. A. (2020). Fine-tuning pretrained language models: weight initializations, data orders, and early stopping. *CoRR*, *abs/2002.06305*.

Edinger, J. D., Means, M. K., Carney, C. E., & Krystal, A. D. (2008). Psychomotor performance defcits and their relation to prior nights' sleep among individuals with primary insomnia. *Sleep*, *31*(5), 599–607.

Ferrando, J., Gállego, G. I., & Costa-jussà, M. R. (2022). Measuring the mixing of contextual information in the transformer. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 8698–8714). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.595

Fried, E. I., Epskamp, S., Nesse, R. M., Tuerlinckx, F., & Borsboom, D. (2016). What are'good'depression symptoms? comparing the centrality of dsm and non-dsm symptoms of depression in a network analysis. *Journal of affective disorders*, *189*, 314–320.

Fried, E. I., & Nesse, R. M. (2015a). Depression is not a consistent syndrome: an investigation of unique symptom patterns in the star* d study. *Journal of affective disorders*, *172*, 96–102.

Fried, E. I., & Nesse, R. M. (2015b). Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC medicine*, *13*(1), 1–11.

Girshick, R. (2015). Fast R-CNN. *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Gkotsis, G., Oellrich, A., Hubbard, T., Dobson, R., Liakata, M., Velupillai, S., & Dutta, R. (2016). The language of mental health problems in social media. In K. Hollingshead & L. Ungar (Eds.), *Proceedings of the third workshop on computational linguistics and clinical psychology* (pp. 63–73). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-0307

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, S., & Morency, L.-P. (2014). The distress analysis interview corpus of human and computer interviews. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 3123–3128). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf

Gupta, S., Agarwal, A., Gaur, M., Roy, K., Narayanan, V., Kumaraguru, P., & Sheth, A. (2022). Learning to automate follow-up question generation using process knowledge for depression triage on Reddit posts. In A. Zirikly, D. Atzil-Slonim, M. Liakata, S. Bedrick, B. Desmet, M. Ireland, A. Lee, S. MacAvaney, M. Purver, R. Resnik, & A. Yates (Eds.), *Proceedings of the eighth workshop on computational linguistics and clinical psychology* (pp. 137–147). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.clpsych-1.12

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: adapt language models to domains and tasks. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8342–8360). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.740

Habermas, T., Ott, L.-M., Schubert, M., Schneider, B., & Pate, A. (2008). Stuck in the past: negative bias, explanatory style, temporal order, and evaluative perspectives in life narratives of clinically depressed individuals. *Depression and Anxiety*, *25*(11), E121–E132.

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

Harrigian, K., Aguirre, C., & Dredze, M. (2021). On the state of social media data for mental health research. In N. Goharian, P. Resnik, A. Yates, M. Ireland, K. Niederhoffer, & R. Resnik (Eds.), *Proceedings of the seventh workshop on computational linguistics and clinical psychology: improving access* (pp. 15–24). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.clpsych-1.2

He, P., Gao, J., & Chen, W. (2021). Debertav3: improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Heston, T. F. (2023). Safety of large language models in addressing depression. *Cureus*, *15*(12).

Hong, S., Cohn, A., & Hogg, D. C. (2021). Using graph representation learning with schema encoders to measure the severity of depressive symptoms. *International conference on learning representations*.

Hong, S., Cohn, A., & Hogg, D. C. (2022). Using graph representation learning with schema encoders to measure the severity of depressive symptoms. *International Conference on Learning Representations (ICLR)*.

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2022). MentalBERT: publicly available pretrained language models for mental healthcare. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 7184–7190). European Language Resources Association. https://aclanthology.org/2022.lrec-1.778

Kabir, M., Ahmed, T., Hasan, M. B., Laskar, M. T. R., Joarder, T. K., Mahmud, H., & Hasan, K. (2023). DEPTWEET: a typology for social media texts to detect depression severities. *Computers in Human Behavior*, *139*, 107503.

Kim, S. J., Kim, S., Jeon, S., Leary, E. B., Barwick, F., & Mignot, E. (2019). Factors associated with fatigue in patients with insomnia. *Journal of psychiatric research*, *117*, 24–30.

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: a new depression diagnostic and severity measure.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine*, *16*(9), 606–613.

Lau, C., Zhu, X., & Chan, W.-Y. (2023). Automatic depression severity assessment with deep learning using parameter-efficient tuning. *Frontiers in Psychiatry*, *14*, 1160291.

Léon, C., du Roscoät, E., & Beck, F. (2023). Prévalence des épisodes dépressifs en france chez les 18-85 ans: résultats du baromètre santé 2021. *Bull Épidemiol Hebd*, *2*, 28–40.

Li, C., Braud, C., & Amblard, M. (2022). Multi-task learning for depression detection in dialogs. In O. Lemon, D. Hakkani-Tur, J. J. Li, A. Ashrafzadeh, D. H. Garcia, M. Alikhani, D. Vandyke, & O. Dušek (Eds.), *Proceedings of the 23rd annual meeting of the special interest group on discourse and dialogue* (pp. 68–75). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sigdial-1.7

Li, Z., Zhou, Q., Li, C., Xu, K., & Cao, Y. (2021). Improving BERT with syntax-aware local attention. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 645–653.

Lin, L., Chen, X., Shen, Y., & Zhang, L. (2020). Towards automatic depression detection: a bilstm/1d cnn-based model. *Applied Sciences*, *10*(23), 8701.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: a robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Losada, D. E., & Crestani, F. (2016). A test collection for research on depression and language use. *International conference of the cross-language evaluation forum for European languages*, 28–39.

Losada, D. E., Crestani, F., & Parapar, J. (2017). ERISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations. *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*, 346–360.

Losada, D. E., Crestani, F., & Parapar, J. (2019). Overview of erisk 2019 early risk prediction on the internet. *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*, 340–357.

Losada, D. E., Crestani, F., & Parapar, J. (2020). Overview of eRisk 2020: early risk prediction on the internet. In A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, & N. Ferro (Eds.), *Experimental ir meets multilinguality, multimodality, and interaction* (pp. 272–287). Springer International Publishing.

Losada, D. E., & Gamallo, P. (2020). Evaluating and improving lexical resources for detecting signs of depression in text. *Language Resources and Evaluation*, *54*(1), 1–24.

Mallol-Ragolta, A., Zhao, Z., Stappen, L., Cummins, N., & Schuller, B. (2019). A hierarchical attention network-based approach for depression detection from transcribed clinical interviews. *Proc. Interspeech 2019*, 221–225. https://doi.org/10.21437/Interspeech.2019-2036

McCall, W. V., Blocker, J. N., D'Agostino Jr, R., Kimball, J., Boggs, N., Lasater, B., & Rosenquist, P. B. (2010). Insomnia severity is an indicator of suicidal ideation during a depression clinical trial. *Sleep medicine*, *11*(9), 822–827.

Mehl, M. R. (2004). *The sounds of social life: exploring students' daily social environments and natural conversations*. The University of Texas at Austin.

Milintsevich, Kirill & Agarwal, N. (2023). Calvados at MEDIQA-chat 2023: improving clinical note generation with multi-task instruction finetuning. In T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts, & A. Rumshisky (Eds.), *Proceedings of the 5th clinical natural language processing work-*

shop (pp. 529–535). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.clinicalnlp-1.56

Milintsevich, Kirill, Dias, G., & Sirts, K. (2024). Evaluating lexicon incorporation for depression symptom estimation. *Proceedings of the 6th Clinical Natural Language Processing Workshop*, 529–535.

Milintsevich, Kirill, Sirts, K., & Dias, G. (2023). Towards automatic text-based estimation of depression through symptom prediction. *Brain Informatics*, *10*(1), 1–14.

Milintsevich, Kirill, Sirts, K., & Dias, G. (2024). Your model is not predicting depression well and that is why: a case study of PRIMATE dataset. In A. Yates, B. Desmet, E. Prud'hommeaux, A. Zirikly, S. Bedrick, S. MacAvaney, K. Bar, M. Ireland, & Y. Ophir (Eds.), *Proceedings of the 9th workshop on computational linguistics and clinical psychology (CLPsych 2024)* (pp. 166–171). Association for Computational Linguistics. https://aclanthology.org/2024.clpsych-1.13

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, *29*(3), 436–465.

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

Naseem, U., Dunn, A. G., Kim, J., & Khushi, M. (2022). Early identification of depression severity levels on reddit using ordinal classification. *Proceedings of the ACM Web Conference 2022*, 2563–2572.

Naseem, U., Lee, B. C., Khushi, M., Kim, J., & Dunn, A. (2022). Benchmarking for public health surveillance tasks on social media with a domain-specific pretrained language model. In T. Shavrina, V. Mikhailov, V. Malykh, E. Artemova, O. Serikov, & V. Protasov (Eds.), *Proceedings of nlp power! the first workshop on efficient benchmarking in nlp* (pp. 22–31). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.nlppower-1.3

Neuman, Y., Cohen, Y., Assaf, D., & Kedma, G. (2012). Proactive screening for depression through metaphorical and automatic text analysis. *Artificial intelligence in medicine*, *56*(1), 19–25.

Nielsen, F. Å. (2011). A new ANEW: evaluation of a word list for sentiment analysis in microblogs, 93–98.

Niu, M., Chen, K., Chen, Q., & Yang, L. (2021). HCAG: a hierarchical context-aware graph attention model for depression detection. *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4235–4239.

Parapar, J., Martín-Rodilla, P., Losada, D. E., & Crestani, F. (2021). Overview of eRisk 2021: early risk prediction on the internet. In K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, & N. Ferro (Eds.), *Experimental ir meets multilingual-*

ity, multimodality, and interaction (pp. 324–344). Springer International Publishing.

Park, M., Cha, C., & Cha, M. (2012). Depressive moods of users portrayed in twitter. *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining, SIGKDD 2012*, 1–8.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: our words, our selves. *Annual review of psychology*, *54*(1), 547–577.

Pérez, A., Fernández-Pichel, M., Parapar, J., & Losada, D. E. (2023). DepreSym: a depression symptom annotated corpus and the role of LLMs as assessors of psychological markers. *arXiv preprint arXiv:2308.10758*.

Pigeon, W. R., Hegel, M., Unützer, J., Fan, M.-Y., Sateia, M. J., Lyness, J. M., Phillips, C., & Perlis, M. L. (2008). Is insomnia a perpetuating factor for late-life depression in the impact cohort? *Sleep*, *31*(4), 481–488.

Pirina, I., & Çöltekin, Ç. (2018). Identifying depression on Reddit: the effect of training data. In G. Gonzalez-Hernandez, D. Weissenbacher, A. Sarker, & M. Paul (Eds.), *Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task* (pp. 9–12). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5903

Pyszczynski, T., & Greenberg, J. (1987). Self-regulatory perseveration and the depressive self-focusing style: a self-awareness theory of reactive depression. *Psychological bulletin*, *102*(1), 122.

Qureshi, S. A., Dias, G., Hasanuzzaman, M., & Saha, S. (2020). Improving depression level estimation by concurrently learning emotion intensity. *IEEE Computational Intelligence Magazine*, *15*(3), 47–59.

Reed, G. M., Sharan, P., Rebello, T. J., Keeley, J. W., Elena Medina-Mora, M., Gureje, O., Luis Ayuso-Mateos, J., Kanba, S., Khoury, B., Kogan, C. S., et al. (2018). The ICD-11 developmental field study of reliability of diagnoses of high-burden mental disorders: results among adult patients in mental health settings of 13 countries. *World psychiatry*, *17*(2), 174–186.

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., & Kupfer, D. J. (2013). DSM-5 field trials in the United States and Canada, part II: test-retest reliability of selected categorical diagnoses. *American journal of psychiatry*, *170*(1), 59–70.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: sentence embeddings using siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. http://arxiv.org/abs/1908.10084

Riedel, B. W., & Lichstein, K. L. (2000). Insomnia and daytime functioning. *Sleep medicine reviews*, *4*(3), 277–298.

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., Schmitt, M., Alisamir, S., Amiriparian, S., Messner, E.-M., et al. (2019). AVEC

2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, 3–12.

Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, *18*(8), 1121–1133.

Safa, R., Bayat, P., & Moghtader, L. (2022). Automatic detection of depression symptoms in twitter using multimodal analysis. *The Journal of Supercomputing*, *78*(4), 4709–4744.

Sampath, K., & Durairaj, T. (2022). Data set creation and empirical analysis for detecting signs of depression from social media postings. *International Conference on Computational Intelligence in Data Science*, 136–151.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Shen, Y., Yang, H., & Lin, L. (2022). Automatic depression detection: an emotional audio-textual corpus and a gru/bilstm-based model. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6247–6251.

Spielberger, C. D., Gonzalez-Reigosa, F., Martinez-Urrutia, A., Natalicio, L. F., & Natalicio, D. S. (1971). The state-trait anxiety inventory. *Revista Interamericana de Psicologia/Interamerican journal of psychology*, *5*(3 & 4).

Stacey, J., Belinkov, Y., & Rei, M. (2022). Supervising model attention with human explanations for robust natural language inference. *Proceedings of the AAAI conference on artificial intelligence*, *36*(10), 11349–11357.

Syarif, I., Ningtias, N., & Badriyah, T. (2019). Study on mental disorder detection via social media mining. *2019 4th International conference on computing, communications and security (ICCCS)*, 1–6.

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in reddit social media forum. *Ieee Access*, *7*, 44883–44893.

Toto, E., Tlachac, M., & Rundensteiner, E. A. (2021). Audibert: a deep transfer learning multimodal classification framework for depression screening. *Proceedings of the 30th ACM international conference on information & knowledge management*, 4145–4154.

Trifu, R. N., Nemeş, B., Bodea-Haţegan, C., & Cozman, D. (2017). Linguistic indicators of language in major depressive disorder (MDD). an evidence based research. *Journal of Evidence-Based Psychotherapies*, *17*(1).

van Borkulo, C., Boschloo, L., Borsboom, D., Penninx, B. W., Waldorp, L. J., & Schoevers, R. A. (2015). Association of symptom network structure with the course of depression. *JAMA psychiatry*, *72*(12), 1219–1226.

van Rooijen, G., Isvoranu, A.-M., Meijer, C. J., van Borkulo, C. D., Ruhé, H. G., de Haan, L., et al. (2017). A symptom network structure of the psychosis spectrum. *Schizophrenia research*, *189*, 75–83.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Villatoro-Tello, E., Ramírez-de-la-Rosa, G., Gática-Pérez, D., Magimai.-Doss, M., & Jiménez-Salazar, H. (2021). Approximating the mental lexicon from clinical interviews as a support tool for depression detection. *Proceedings of the 2021 international conference on multimodal interaction*, 557–566.

Wang, S., Chen, Z., Ren, Z., Liang, H., Yan, Q., & Ren, P. (2022). Paying more attention to self-attention: improving pre-trained language models via attention guiding. *arXiv preprint arXiv:2204.02922*.

Wang, Y., Inkpen, D., & Kirinde Gamaarachchige, P. (2024). Explainable depression detection using large language models on social media data. In A. Yates, B. Desmet, E. Prud'hommeaux, A. Zirikly, S. Bedrick, S. MacAvaney, K. Bar, M. Ireland, & Y. Ophir (Eds.), *Proceedings of the 9th workshop on computational linguistics and clinical psychology (clpsych 2024)* (pp. 108–126). Association for Computational Linguistics. https://aclanthology.org/2024.clpsych-1.8

Wardle-Pinkston, S., Slavish, D. C., & Taylor, D. J. (2019). Insomnia and cognitive performance: a systematic review and meta-analysis. *Sleep medicine reviews*, *48*, 101205.

Watson, D., & Clark, L. A. (1994). The PANAS-X: manual for the positive and negative affect schedule-expanded form.

Williamson, J. R., Godoy, E., Cha, M., Schwarzentruber, A., Khorrami, P., Gwon, Y., Kung, H.-T., Dagli, C., & Quatieri, T. F. (2016). Detecting depression using vocal, facial and semantic communication cues. *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 11–18.

Winer, E. S., Jordan, D. G., & Collins, A. C. (2019). Conceptualizing anhedonias and implications for depression treatments. *Psychology Research and Behavior Management*, 325–335.

World Health Organization et al. (2017). *Depression and other common mental disorders: global health estimates* (tech. rep.). World Health Organization.

World Health Organization et al. (2022). World mental health report: transforming mental health for all.

Xezonaki, D., Paraskevopoulos, G., & Potamianos, A. (2020). Affective conditioning on hierarchical attention networks applied to depression detection from transcribed clinical interviews. *INTERSPEECH 2020*, 4556–4560.

Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A. K., & Wang, D. (2024). Mental-LLM: leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *8*(1), 1–32.

Yadav, S., Chauhan, J., Sain, J. P., Thirunarayan, K., Sheth, A., & Schumm, J. (2020). Identifying depressive symptoms from tweets: figurative language enabled multitask learning framework. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics* (pp. 696–709). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.61

Yang, K., Ji, S., Zhang, T., Xie, Q., Kuang, Z., & Ananiadou, S. (2023). Towards interpretable mental health analysis with large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 6056–6077). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.370

Yang, K., Zhang, T., & Ananiadou, S. (2022). A mental state knowledge–aware and contrastive network for early stress and depression detection on social media. *Information Processing & Management*, *59*(4), 102961.

Yang, K., Zhang, T., Kuang, Z., Xie, Q., Huang, J., & Ananiadou, S. (2024). MentaLLaMA: interpretable mental health analysis on social media with large language models. *Proceedings of the ACM on Web Conference 2024*, 4489–4500.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: human language technologies* (pp. 1480–1489). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1174

Yates, A., Cohan, A., & Goharian, N. (2017). Depression and self-harm risk assessment in online forums. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2968–2978). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1322

Yazdavar, A. H., Al-Olimat, H. S., Ebrahimi, M., Bajaj, G., Banerjee, T., Thirunarayan, K., Pathak, J., & Sheth, A. (2017). Semi-supervised approach to monitoring clinical depressive symptoms in social media. *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, 1191–1198.

Yazdavar, A. H., Mahdavinejad, M. S., Bajaj, G., Romine, W., Sheth, A., Monadjemi, A. H., Thirunarayan, K., Meddar, J. M., Myers, A., Pathak, J., et al. (2020). Multimodal mental health analysis in social media. *Plos one*, *15*(4), e0226248.

Zhang, Z., Chen, S., Wu, M., & Zhu, K. (2022). Symptom identification for interpretable detection of multiple mental disorders on social media. *Proceedings of the 2022 conference on empirical methods in natural language processing*, 9970–9985.

Zhou, W., & Chen, M. (2022). An improved baseline for sentence-level relation extraction. *2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*, 161–168. https://aclanthology.org/2022.aacl-short.21

Zirikly, A., & Dredze, M. (2022). Explaining models of mental health via clinically grounded auxiliary tasks. In A. Zirikly, D. Atzil-Slonim, M. Liakata, S. Bedrick, B. Desmet, M. Ireland, A. Lee, S. MacAvaney, M. Purver, R. Resnik, & A. Yates (Eds.), *Proceedings of the eighth workshop on computational linguistics and clinical psychology* (pp. 30–39). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.clpsych-1.3

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisors, Dr. Kairit Sirts and Dr. Gaël Dias, for their support and guidance during my PhD. Also, Dr. Sonia Dollfus for her invaluable insights into the world of clinical psychology. Along with Dr. Sonia Dollfus, I would like to thank Dr. Maud Rotharmel and Lucie Métivier for participating in a data annotation effort that eventually became a part of this thesis. Finally, I would like to express my sincerest gratitude to all my friends and close ones who supported me during these challenging times.

SISUKOKKUVÕTE

Depressioonitaseme hindamine tekstist: sümptomipõhine lähenemine, väliste infoallikate kasutamine, andmete valiidsus

Depressioon on üks levinumaid psüühikahäireid maailmas, põhjustades sageli töövõimetust ja suurendades enesetapu riski. Hiljutine COVID-19 pandeemia on depressioonimäärasid veelgi tõstnud nii Prantsusmaal, Eestis kui ka kogu maailmas. Samas takistavad vaimse tervise häiretega seotud stigma ja piiratud psühhiaatrilise ravi kättesaadavus paljudel inimestel õige diagnoosi ja ravi saamist.

Loomuliku keele töötluse valdkonna uurijad on juba pikka aega uurinud meetodeid automaatseks depressiooni tuvastamiseks tekstiandmetest. Varasemad lingvistilised uuringud on näidanud sõnavara kasutuse erinevusi depressioonis ja ilma depressioonita inimeste vahel. Masin- ja süvaõppe arengud on nüüdseks võimaldanud depressiooni tuvastamist nii sotsiaalmeedia tekstide kui ka kliiniliste intervjuude transkriptsioonide põhjal.

Enamik varasemaid uuringuid on aga käsitlenud depressiooni automaatset tuvastamist tekstist kui binaarse klassifitseerimise ülesannet. Siiski põhineb tõenäoliselt kõige laialdasemalt kasutatav depressiooni definitsioon Vaimsete Häirete Diagnostilise ja Statistilise Käsiraamatu (DSM-5) määratlusel. DSM-5 kohaselt defineeritakse depressioon spetsiifiliste sümptomite samaaegse esinemismustri alusel. Seega võivad sama diagnoosisildi taga peituda mitmesugused erinevad sümptomiprofiilid. Järelikult oleks sümptomipõhine lähenemine depressiooni automaatseks tuvastamiseks tekstist oluliselt informatiivsem ja läbipaistvam kui binaarne diagnostilise staatuse ennustamine.

Automaatsete meetodite arendamist depressiooni tuvastamiseks tekstist raskendab kvaliteetsete andmestike puudumine. Kliinilisi andmestikke, nagu näiteks patsiendi ja terapeudi vaheliste vestluste salvestused, kogutakse haiglates, kus kehtivad tavaliselt ranged konfidentsiaalsusnõuded, mis keelavad andmete jagamise. Üheks harvaks erandiks on DAIC-WOZ, mis on lõppkasutaja litsentsilepingu alusel avalikult kättesaadav dialoogipõhine intervjuude andmestik. Selles andmestikus täitis iga intervjueeritav enne vestlust PHQ-8 küsimustiku, mis hindab depressiooni raskusastet DSM-5 kriteeriumide põhjal sümptomite sageduse järgi. Seda andmestikku on kasutatud paljudes eelnevates uurimistöödes ning see on ka käesoleva väitekirja aluseks.

Teisalt on sotsiaalmeedia avalikult kättesaadavate andmestike nn "kullakaevandus". Mitmed uuringud on kasutanud automaatseks depressiooni tuvastamiseks andmeid, mis on kogutud sotsiaalmeediaplatvormidelt, nagu Reddit ja X (endine Twitter). Samas on suurem osa neist andmetest märgendatud kas automaatselt või siis tavakasutajate abiga, kellel on vähene või puuduv väljaõpe kliinilises psühholoogias või psühhiaatrias. Kahtlemata on vaimse tervise spetsialistide kaasamine märgendamisprotsessi keeruline. Siiski seab nende puudumine või vähene osalus selliste andmestike kehtivuse kahtluse alla.

Lisaks märgendatud tekstiandmetele võib automaatsel depressiooni tuvastamisel olla kasu erinevatest leksikonidest. Mitmed uuringud on näidanud erinevusi keelekasutuses depressioonis ja ilma depressioonita inimeste vahel. Need erinevused väljenduvad muu hulgas depressioonile kalduvate inimeste suuremas negatiivse varjundiga terminite, esimese isiku asesõnade ning emotsionaalsete sõnade kasutamises. Aja jooksul on loodud mitmeid leksikone, mis sisaldavad emotsioonidega seotud sõnu (*NRC EmoLex*), meelsusega seotud sõnu (*AFINN Sentiment Lexicon*) või depressioonispetsiifilist sõnavara (*Social-media Depression Detector*). Kuna leksikone on varemgi kasutatud depressiooni tuvastamiseks tekstist, võivad ka automaatse depressiooni tuvastamise mudelid leksikonidest sisalduvast infost potentsiaalselt kasu saada.

Selle doktoritöö peamine eesmärk oli arendada sümptomipõhiseid mudeleid depressiooni automaatseks hindamiseks tekstist ning uurida võimalusi leksikonides sisalduvad info integreerimiseks tehisnärvivõrkudesse. Töö eesmärk viis järgmiste uurimisküsimusteni: (**UK1**) Kuidas erineb depressiooni ennustamine sümptomite kogumina võrreldes depressiooni ennustamisega binaarse diagnoosina? (**UK2**) Kas väliste teadmiste kaasamine tänapäevastesse tehisnärvivõrkudesse parandab depressiooni automaatset hindamist? UK2 kallal töötades märkasime, et kasutatud sotsiaalmeedia andmestikul ei näidanud ühegi mudeli ennustused märkimisväärset paranemist, eriti anhedoonia sümptomi osas, mistõttu uurisime, kuivõrd selle andmestiku märgendid vastavad antud sümptomi kliinilisele definitsioonile (**UK3**).

Sümptomipõhine depressiooni ennustamine. Töös uuriti kõigepealt, kuidas erineb depressiooni ennustamine sümptomite kogumina binaarse klassifitseerimise lähenemisest. Arendati välja närvivõrgu arhitektuur, mis saavutas tipptasemel tulemused sümptomipõhises depressiooni hindamises. See arhitektuur oli aluseks ka teistele katsedele selles doktoritöös. Tulemused näitasid, et sümptomitel põhinev mudel ennustas depressiooni esinemist samal tasemel või paremini kui diagnostilist staatust ennustav binaarse klassifitseerimise mudel või depressiooni raskusastet ennustav regressioonimudel, lisaks väljastades samal ajal detailsemaid ja personaliseeritumaid sümptomiprofiile (**UK1**).

Väliste infoallikate kasutamine. Kuivõrd sümptomipõhine lähenemine õigustas ennast, jätkati tööd sümptomeid ennustavate mudelitega. Esiteks täiustati mudelite aluseks oleva närvivõrgu arhitektuuri, et paremini modelleerida dialoogiformaadis teksti. Teiseks näidati, et leksikonides sisalduva info lisamine sümptomipõhisele mudelile aitab mõnede baasmudelite puhul parandada sümptomite ennustamise täpsust. Tulemused näitasid, et eriti DAIC-WOZ andmestiku puhul on baasmudeli valik oluline; kui MentalBERTi puhul, mis on domeenispetsiifiline eeltreenitud keelemudel, ennustustulemused leksikonide info lisades paranesid, siis BERT, mis on üldkasutatav eeltreenitud keelemudel, leksikonide info lisamisest kasu ei saanud (**UK2**). Nagu sageli teadustöös juhtub, ei vii kõik katsetused oodatud tulemusteni. Sotsiaalmeediapõhise PRIMATE andmestiku puhul ei aidanud leksikonide info lisamine ennustustulemusi parandada kummagi katsetatud baasmudeli puhul. Selle negatiivse tulemuse põhjuste uurimisel keskenduti PRIMATE andmestiku märgendamise kvaliteedile.

Märgenduse valiidsus. Anhedoonia (huvipuudus või asjade tegemise naudingu kadumine) sümptomi näitel näitasime töös, et selle sümptomi märgendused ei vastanud PRIMATE andmestikus usaldusväärselt sümptomi kliinilisele kirjeldusele (**UK3**). Töös loodi sotsiaalmeedia tekstide andmestik anhedoonia tuvastamiseks, mis on üks depressiooni peamisi sümptomeid. Selle andmestiku märgendamine vastab rangemalt anhedoonia kliinilisele määratlusele. Need märgendused on tehtud vabalt kättesaadavaks ka teistele uurijatele.

Töö lõpus tõstatati ka mitmeid uurimissuundi tulevikuks. Suurenev huvi suurte generatiivsete keelemudelite vastu avab uusi võimalusi depressiooni hindamiseks, samas tuleb hoolikalt arvesse võtta nende mudelite kallutatust ja kalduvust hallutsineerida. Erinevate võimaluste uurimine väliste infoallikate integreerimiseks mudelitesse pakub samuti uusi suundi tuleviku teadusuuringuteks. Lisaks on vajalik täiendavate tekstide märgendamine erinevate sümptomitega ja andmete kogumine teistes keeltes kui inglise keel, et edendada valdkonna arengut.

RÉSUMÉ

Estimation du niveau de dépression à partir de données textuelles : approche basée sur les symptômes, utilisation de ressources externes, validité des jeux de données

Le trouble dépressif majeur (TDM) est l'un des troubles psychiatriques les plus répandus au monde, entraînant souvent une incapacité et un risque accru de suicide. La récente pandémie de COVID-19 a encore aggravé les taux de dépression dans des pays comme la France, l'Estonie et dans le monde entier. Cependant, la stigmatisation entourant les maladies mentales et la disponibilité limitée des traitements psychiatriques empêchent de nombreuses personnes de recevoir un diagnostic et des soins appropriés.

La communauté scientifique en traitement automatique du langage naturel (TALN) s'intéresse depuis longtemps à la détection automatique de la dépression à travers les textes. Les premières études linguistiques ont identifié des différences dans l'utilisation du vocabulaire entre les individus déprimés et non déprimés. Depuis, les avancées en apprentissage automatique et en apprentissage profond ont permis de détecter la dépression à partir des textes publiés sur les réseaux sociaux et des transcriptions d'entretiens cliniques.

Il est important de noter que la plupart des travaux précurseurs ont abordé la détection automatique de la dépression à partir de textes comme une tâche de classification binaire. Cependant, la définition du TDM la plus largement utilisée provient potentiellement de la version 5 du Manuel diagnostique et statistique des troubles mentaux (DSM-5). Selon le DSM-5, le diagnostic de la dépression est défini comme un schéma de cooccurrence de symptômes spécifiques. Ainsi, il existe de nombreux profils symptomatiques différents derrière une même étiquette diagnostique. Par conséquent, l'adoption d'une approche basée sur les symptômes pour la détection automatique de la dépression à partir des textes fournira plus d'informations et de transparence qu'une simple prédiction binaire du diagnostic.

Le manque de données de haute qualité est un autre défi pour l'estimation automatique de la dépression. Les jeux de données cliniques, tels que les enregistrements de conversations entre patients et thérapeutes, sont recueillis dans les hôpitaux qui sont généralement soumis à des réglementations strictes interdisant tout partage de données. L'une des rares exceptions est le DAIC-WOZ, un jeu de données d'entretiens basés sur des dialogues qui est disponible publiquement sous l'accord de licence utilisateur final. Dans ce jeu de données, avant la conversation, chaque interviewé a rempli le PHQ-8, un questionnaire qui mesure la gravité de la dépression en fonction de la fréquence des symptômes selon les critères du DSM-5. Ce jeu de données est donc devenu la base de nombreuses initiatives de recherche, dont cette thèse.

D'un autre côté, les réseaux sociaux sont une mine d'or de données accessibles au public. De nombreux travaux exploitent les données collectées sur des plate-

formes de réseaux sociaux comme Reddit et X (anciennement Twitter) pour la détection automatique de la dépression. Cependant, la plupart de ces données sont étiquetées soit automatiquement, soit avec l'aide d'annotateurs non spécialisés ayant peu ou pas de formation en psychologie clinique. Il est évident que l'implication des professionnels en santé mentale dans le processus d'annotation est difficile. Néanmoins, leur absence ou leur faible participation à ce processus remet en question la validité de ces données.

Un autre type de données qui peut être utilisé pour la détection automatique de la dépression à partir de textes est constitué de différents lexiques. Plusieurs études ont montré des différences dans l'usage de la langue entre les personnes déprimées et non déprimées. Ces différences se reflètent, entre autres, dans l'utilisation accrue de termes à connotation négative, de pronoms à la première personne ou de mots émotionnels par les personnes dépressives. Parallèlement, plusieurs lexiques codifiant les émotions (*NRC EmoLex*), les sentiments (*AFINN Sentiment Lexicon*) ou le vocabulaire spécifique à la dépression (*Social-media Depression Detector*) ont été créés au fil du temps. Étant donné que les lexiques seuls ont été utilisés précédemment pour détecter la dépression à partir des textes, les modèles de détection automatique de la dépression à partir de textes peuvent potentiellement bénéficier de ces ressources externes.

L'objectif principal de cette thèse est de développer des modèles basés sur les symptômes pour l'estimation automatique de la dépression à partir de textes et d'explorer des moyens d'intégrer les connaissances existantes du domaine dans les modèles neuronaux. Cet objectif a conduit aux questions de recherche suivantes : (**QdR1**) Comment la prédiction de la dépression en tant que collection de symptômes se compare-t-elle à la prédiction de la dépression en tant que diagnostic binaire ? (**QdR2**) L'inclusion de ressources externes dans les architectures neuronales de pointe améliore-t-elle l'estimation automatique de la dépression ? En travaillant sur QdR2, nous avons remarqué que le jeu de données des réseaux sociaux ne montrait aucune amélioration, en particulier pour le symptôme de manque d'intérêt. Ce constat nous a amenés à étudier si les annotations de cet ensemble de données correspondaient à la définition de ce symptôme (**QdR3**).

Prédiction de la dépression basée sur les symptômes. Nous avons commencé notre recherche en explorant comment la prédiction de la dépression en tant que collection de symptômes se compare à l'approche de classification binaire. Nous avons développé une architecture neuronale qui a obtenu des résultats de l'état de l'art dans l'estimation de la dépression basée sur les symptômes. Cette architecture a également servi de base à d'autres expériences dans cette thèse. Nous avons constaté que le modèle de prédiction des symptômes fonctionnait aussi bien voire mieux que les modèles de classification binaire ou de régression unique de la gravité de la dépression tout en fournissant simultanément des profils symptomatiques plus descriptifs et personnalisés (**QdR1**).

Intégration de ressources externes. Nous avons poursuivi notre travail sur la prédiction de la dépression basée sur les symptômes. Tout d'abord, nous avons introduit des améliorations progressives à l'architecture neuronale afin de mieux modéliser les textes sous forme de dialogue. Deuxièmement, nous avons démontré que certains modèles de langage pré-entraînés (PLM) peuvent encore tirer parti des ressources lexicales existantes pour l'estimation de la dépression basée sur les symptômes. En particulier, nous avons constaté que, pour le jeu de données DAIC-WOZ, le choix du modèle de base est important. MentalBERT, un PLM spécifique au domaine, a bénéficié de manière constante des informations du lexique inclus, alors que BERT, un PLM à domaine général, n'en a pas bénéficié (**QdR2**). Comme c'est souvent le cas dans la recherche, tous les résultats n'ont pas nécessairement été positifs. En particulier, PRIMATE, un jeu de données basé sur les réseaux sociaux, n'a montré aucune amélioration. En cherchant les raisons de cette mauvaise performance, nous avons examiné la qualité des annotations de ce jeu de données.

Validité des annotations. Sur l'exemple du symptôme de manque d'intérêt ou de plaisir à faire les choses (anhédonie), nous avons montré que les annotations des symptômes ne correspondaient pas de manière fiable à la description clinique du symptôme (**QdR3**). En conséquence, nous avons construit un jeu de données textuelles issu des réseaux sociaux pour la détection de l'anhédonie, qui est l'un des principaux symptômes de la dépression. L'annotation de ce jeu de données est plus rigoureusement conforme à la définition clinique de l'anhédonie. Nous avons rendu ces annotations librement accessibles sous le nom du Jeu de Données I.

Nous avons également proposé plusieurs pistes pour les travaux futurs. L'augmentation de la popularité des grands modèles de langage (LLM) offre de nouvelles possibilités pour l'estimation de la dépression, bien que leurs biais et leur tendance à l'hallucination nécessitent une attention particulière. L'exploration plus poussée de l'intégration des connaissances externes dans les modèles représente une autre direction pour la recherche future. De plus, l'annotation de plus de textes avec divers symptômes et la collecte de données dans d'autres langues que l'anglais sont nécessaires pour faire progresser le domaine.

PUBLICATIONS

I

K. Milintsevich, K. Sirts, G. Dias
Towards automatic text-based estimation of depression through symptom prediction
Brain Informatics, 10(1), pages 1–14

Brain Informatics

RESEARCH Open Access

Towards automatic text-based estimation of depression through symptom prediction

Kirill Milintsevich[1,2*], Kairit Sirts[1] and Gaël Dias[2]

Abstract

Major Depressive Disorder (MDD) is one of the most common and comorbid mental disorders that impacts a person's day-to-day activity. In addition, MDD affects one's linguistic footprint, which is reflected by subtle changes in speech production. This allows us to use natural language processing (NLP) techniques to build a neural classifier to detect depression from speech transcripts. Typically, current NLP systems discriminate only between the depressed and non-depressed states. This approach, however, disregards the complexity of the clinical picture of depression, as different people with MDD can suffer from different sets of depression symptoms. Therefore, predicting individual symptoms can provide more fine-grained information about a person's condition. In this work, we look at the depression classification problem through the prism of the symptom network analysis approach, which shifts attention from a categorical analysis of depression towards a personalized analysis of symptom profiles. For that purpose, we trained a multi-target hierarchical regression model to predict individual depression symptoms from patient–psychiatrist interview transcripts from the DAIC-WOZ corpus. Our model achieved results on par with state-of-the-art models on both binary diagnostic classification and depression severity prediction while at the same time providing a more fine-grained overview of individual symptoms for each person. The model achieved a mean absolute error (MAE) from 0.438 to 0.830 on eight depression symptoms and showed state-of-the-art results in binary depression estimation (73.9 macro-F1) and total depression score prediction (3.78 MAE). Moreover, the model produced a symptom correlation graph that is structurally identical to the real one. The proposed symptom-based approach provides more in-depth information about the depressive condition by focusing on the individual symptoms rather than a general binary diagnosis.

Keywords Computational methods for mental health, Automated depression estimation, Natural language processing, Symptom network analysis, Multi-target regression

1 Introduction

Major Depressive Disorder (MDD) is one of the most common mental disorders, with over 300 million people being affected by it [1]. Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [2] defines MDD by nine symptoms: (1) depressed mood; (2) markedly diminished interest or pleasure; (3) increase or decrease in either weight or appetite; (4) insomnia or hypersomnia; (5) psychomotor agitation or retardation; (6) fatigue or loss of energy; (7) feelings of worthlessness or inappropriate guilt; (8) diminished ability to think or concentrate; (9) recurrent thoughts of death or recurrent suicidal ideation. According to DSM-5, the diagnosis of MDD is warranted if the person has experienced at least 5 of those symptoms every day or almost every day for the last two weeks, and one of those symptoms must be either depressed mood (1) or the loss of interest (2). These diagnostic criteria indicate that behind the same diagnostic

*Correspondence:
Kirill Milintsevich
kirill.milintsevich@unicaen.fr
[1] Institute of Computer Science, University of Tartu, Tartu, Estonia
[2] Groupe de Recherche en Informatique, Image et Instrumentation (GREYC), National Graduate School of Engineering and Research Center (ENSICAEN), Université de Caen Normandie (UNICAEN), 14000 Caen, France

label, there can be many different symptom constellations or sub-types [3, 4].

1.1 Background

In recent years, considerable interest has emerged in using natural language processing (NLP) and artificial intelligence (AI) techniques for inferring the mental health status of a person unobtrusively based on their speech or writing (see for instance [5, 6] for reviews). A large majority of studies have focused on predicting depression [6], which is only to be expected considering its prevalence. However, most NLP and AI-based systems have treated the task as a discrete binary classification problem [7–9], predicting the presence or absence of the diagnosis, which does not appreciate the variability of the clinical phenomena of depression.

Although psychiatric diagnostic systems like DSM-5 still mostly operate with categorical diagnoses, there is a shift towards richer representations of psychiatric syndromes that can take into account the dimensional and heterogeneous nature of the clinical pictures of the same psychiatric diagnosis. One particular approach that is gaining attention concerns symptom network analysis (SNA) [10, 11]. According to the SNA, the symptoms of mental health disorders are not indicators of an underlying disease (an assumption of a traditional medical model), but it rather views the disorder itself as a causal system of interacting symptoms. The advantage of the SNA is that it also provides a natural way of analyzing and modeling the comorbidity between different disorders (see, for instance, [12] and [13] for examples), which is a norm rather than an exception for mental disorders. Depression, in particular, has been studied quite a lot from the perspective of SNA [14–16]. One way of depicting the SNA graphically is to use correlation graphs, such as the one shown in Figure 1. Although the symptom graph constructed based on correlations does not show the causal links between symptoms[1], it does show the strength of the co-occurrence relations between each pair of symptoms. The SNA view of the diagnosis prescribes a more thorough analysis of specific depression symptoms in clinical studies [17]. Thus, it seems only natural to extend the research based on NLP and AI to reflect these advances in psychiatry and start focusing on predicting the presence or degree of particular depression symptoms instead of the categorical diagnosis.

Developing predictive systems for mental health comes with the challenge of obtaining clinical data for training models. Getting patient speech or textual data is challenging due to ethical and legal reasons. Therefore, many studies have resorted to analyzing social media data [6] or other auxiliary data resources. In order to train predictive models, the clinical data needs to be supplied with diagnostic labels. One way of acquiring labels is asking people to fill in self-report questionnaires assessing the presence and/or severity of depression symptoms [18]. There are several questionnaires that assess the presence or absence of MDD based on depression symptom severity, such as the Beck Depression Inventory (BDI) [19], Hamilton Rating Scale for Depression (HRSD) [20], and Patient Health Questionnaire (PHQ) [21]; the last one shadowing the symptoms defined by DSM-5.

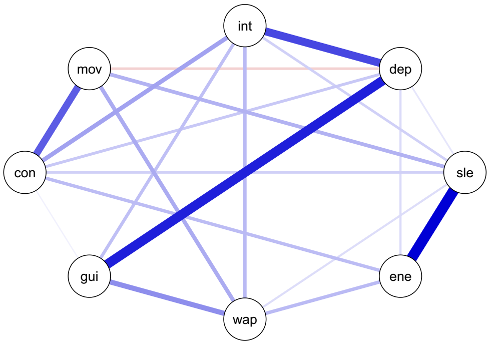


Fig. 1 Correlation graph of symptoms computed on the training set of the DAIC-WOZ data set. Thicker edges show a stronger correlation. Blue edges show a positive correlation, and red edges show a negative correlation. The nodes represent the following symptoms: *int*: markedly diminished interest or pleasure; *dep*: depressed mood; *sle*: insomnia or hypersomnia; *ene*: fatigue or loss of energy; *wap*: increase or decrease in either weight or appetite; *gui*: feelings of worthlessness or inappropriate guilt; *con*: diminished ability to think or concentrate; *mov*: psychomotor agitation or retardation

1.2 Problem

Previous approaches that have used the data with self-report questionnaire scores typically obtain the labels by first summing the scores of all the questions and then dichotomizing the sum at a predefined cutoff point, which results in a binary diagnostic status. This approach, however, has several problems. First of all, using the sum of scores of these questionnaires might not be a good basis for establishing the diagnostic status of a person [17], as identical labels can hide a set of very different symptom severity values. Second, the difference in depression level between two persons with the same label can end up being larger than the difference between two persons with differing labels. For instance, in the boundary cases, one person with the non-depressed label might have obtained a sum-score of only one point lower than another person who was labeled as depressed. At the

[1] This requires a longitudinal analysis over time.

same time, two people, both having the depression label, might have a very large score difference—one having a sum-score near the cutoff and the other one at the high end of the scale. These within- and between-group characteristics can make it hard for systems to learn true patterns about depression.

1.3 Methods

To create a model which is able to produce more fine-grained predictions we treat automatic depression prediction as a multi-target regression problem, predicting the severity score of each symptom from a common interview representation. We show that predicting each symptom individually not only gives more insight into a person's mental state but also allows to infer the binary, 5-class, and regression scores with gains in performance in most of the experimental configurations.

In this paper, we use DAIC-WOZ [22], a data set widely used for automatic depression prediction. It consists of interviews between a person and a human-controlled virtual assistant, Ellie. Each interview has facial features from the video, audio recording, and text transcription. Each interview is also accompanied by the answers to the PHQ-8 screening questionnaire—an eight-symptom version of the PHQ, which does not include the suicidality/self-harm question from the depression diagnostic criteria. The data set is relatively small, featuring only less than 200 interviews. However, it is closer to the domain of clinical interviews than the social media data often used for developing predictive systems for mental health. Even though the DAIC-WOZ data set provides severity scores for each individual question, previous works using this data for developing automated systems have predicted either a binary label, i.e., depressed or non-depressed, [7–9, 23], or a regression score based on the total sum of individual PHQ-8 question scores [23–27]. Few other studies have discretized the range of PHQ-8 scores into five categories and have thus predicted a label within a set of five possible classes, i.e., no symptoms, mild, moderate, moderately severe, severe depression [25, 28].

1.4 Contributions

Our goal in this study is twofold. First, we want to highlight the importance of the advances in the clinical field when developing NLP and AI-based mental health prediction models. In particular, we want to emphasize the turning away from the medical latent disease model with its categorical diagnostic predictions and more toward dimensional and symptom-level analyses. Second, we aim to demonstrate that by adopting the symptom-level prediction, the models do not lose accuracy also on the categorical diagnosis level and can add a more fine-grained representation of the clinical picture for each person, thus better capturing the heterogeneity of the clinical phenomena.

2 Related work

Most studies on MDD that make use of NLP and AI methods over clinical data have been developed over the DAIC-WOZ [22] data set, although some marginal works have been carried out on the General Psychotherapy Corpus (GPC) from Alexander Street Press [8, 29]. In particular, the GPC comprises a large collection of transcripts of patient–provider conversations, but as it is not easily available[2], most researchers have been focusing on the DAIC-WOZ for reproducibility purposes. DAIC-WOZ is a multimodal data set containing interviews accompanied with facial features from the videos, audio recordings, and text transcriptions. Therefore, various previous works have tackled the multimodal aspect of this data set.

In our work, we only make use of the textual transcriptions; thus, we limit our review to those works that have also focused on the textual modality of this data set. One line of work has concentrated on exploring various neural network architectures to best model the interviews, including hierarchical attention-based networks [7] and deep neural graph structures [27]. Other studies have experimented with multi-task modeling, aiming to improve the performance by simultaneously predicting both binary diagnostic and the overall depression severity regression scores [24]. Finally, some studies have explored the utility of enriching the models with additional, in particular affective, information from external sources. In this regard, Xezonaki et al. [8] experimented with explicitly modeling the affective features of words extracted from various affective lexicons. Qureshi et al. [25] employed an additional emotion data set and experimented with a multi-task classification model to concurrently predict both the depression severity level of the DAIC-WOZ data and the emotional intensity of the emotion data set.

All these previous studies concerning predicting depression based on clinical data of patient–therapist interviews have developed categorical models to predict the binary, multi-class, or continuous diagnostic status. The only previous work we are aware of that has used the DAIC-WOZ data set for symptom prediction is by Delahunty et al. [30]. However, as their focus was on modeling the comorbidity between depression and anxiety, they only predicted the two main depression symptoms

[2] Our contacts with Alexander Street Press were unfruitful to get the GPC corpus.

(lowered mood and loss of interest) instead of the full symptom profile. Next, we will review some studies based on social media data that have adopted symptom prediction either instead of or for aiding the diagnostic classification.

Studies based on social media data have used Twitter [31, 32], Reddit [33] or other depression-related internet forums [34–36] as their data source. Even though these works collect their data from public sources, the data sets themselves are not publicly available. Some authors [31–33], however, stated that their data sets can be accessed by other researchers who agree to follow the ethical guidelines put forward by the corresponding authors. A challenge with working with social media data is obtaining the labels necessary for training classification models. One option is to manually label the symptoms in the data. This approach was adopted by Yadav et al. [31], who annotated the symptoms in tweets using a mental health lexicon constructed by mental health professionals. The main focus of this work was to use an auxiliary classification task to detect figurative speech that might be used to express symptoms and can be hard to detect via lexicon lookup. Yao et al. [34] analyzed a Chinese depression forum for depression symptom prediction. Their work aimed to develop a comprehensive annotation scheme for a list of symptoms that goes beyond the diagnostic symptoms of DSM-5. Davcheva et al. [36] developed a symptom-based classification system using internet forum data. The data were manually annotated with the symptom lexicon constructed based on DSM-5 symptom descriptions and topic modeling. The overall goal of the model was to provide a categorical diagnosis based on the predicted symptoms. Several diagnoses were addressed in this work, also targeting schizophrenia and attention deficit hyperactivity disorder in addition to depression.

An alternative to manual labeling is to use lexicons or rules to automatically extract the symptom mentions. This approach was adopted by Karmen et al. [35], who used lexicons to detect the mention of symptoms in the posts of an internet forum. The goal of their work was to simulate assessing the depression severity score with a self-report assessment measure by aggregating the symptom scores with the frequency of symptom mentions. Similarly, Yazdavar et al. [32] used a lexicon-based approach on tweets to compile user-specific depression lexicons and adopted a semi-supervised topic modeling approach to model the symptom progression over time. Recently, Nguyen et al. [33] adopted Reddit data to train models to predict depression diagnosis grounded in PHQ-9 symptoms. In their work, the symptoms were automatically annotated using manually constructed symptom patterns. The symptom mentions found that this way thus serves as weak labels that were used to constrain the model to predict the binary diagnosis.

Table 1 Number of interviews for each depressive symptom severity category in the DAIC-WOZ data set, distributed by train, validation and test sets

Depression severity	Data split		
	Train	Validation	Test
No symptoms [0..4]	47	17	22
Mild [5..9]	29	6	11
Non-depressed Total	76	23	33
Moderate [10..14]	20	5	5
Moderately severe [15..19]	7	6	7
Severe [20..24]	4	1	2
Depressed Total	31	12	14
Total	107	35	47

3 Method

While previous works that tackle patient–therapist interviews have been developing automated systems that either predict a categorical label or a regression score, the SNA approach aims at scoring each symptom individually. As a consequence, shifting to the paradigm of multi-target regression architectures is necessary. In this section, we overview the DAIC-WOZ data set and present the experimented learning architectures.

3.1 Data

The DAIC-WOZ data set contains 189 clinical interviews in a dialog format. Each interview has two actors: the virtual assistant Ellie and a participant. The utterances of Ellie come from a predefined set of prompts, although the exact subset of prompts and their ordering can vary for each interview. The data set is distributed in pre-determined splits, such that 107 interviews are used for training, 35 for validation, and 47 for testing (see Table 1). Each interview in the data set is accompanied with a PHQ-8 assessment, which consists of eight questions inquiring about diagnostic depression symptoms. Each question is scored from 0 to 3, and the total PHQ score, which is the sum of the scores of all eight questions, ranges from 0 to 24. According to the standard cutoff score of 10, the interviews can be divided into diagnostic classes, where the subjects whose PHQ-8 total score is less than 10 are considered non-depressed, and those whose score is at least 10 are categorized as depressed. Based on the total score, the interviews can be further divided into five classes according to the depressive symptom severity [21]. From the overall layout of the DAIC-WOZ data set shown in Table 1, it is evident that

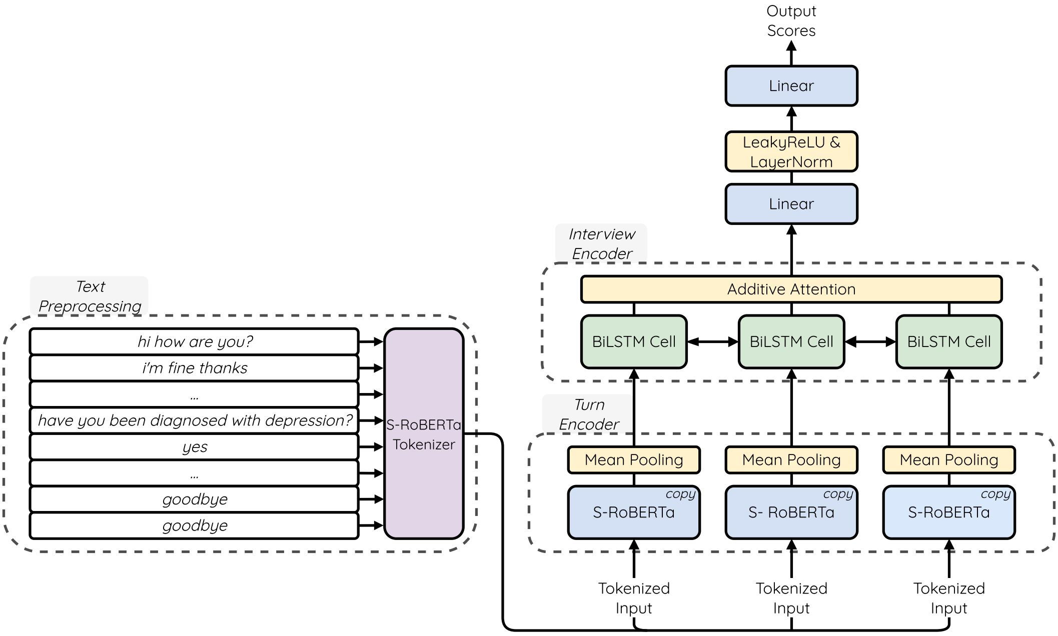


Fig. 2 Overview of the model. On the turn-level, the same instance of S-RoBERTa is used to encode each turn. Mean Pooling is the operation that averages all the token representations output by S-RoBERTa

the classes are imbalanced, and the imbalance is even stronger in the high PHQ score range.

3.2 Model

To efficiently encode the interviews, we employed a hierarchical architecture [37]. Since we aim at predicting scores for individual symptoms, we adopted a prediction head that produces eight regression outputs, effectively making it a multi-target regression model.

The model has two encoders: $\mathbf{Enc}^{\mathbf{turn}}$ and $\mathbf{Enc}^{\mathbf{int}}$. Figure 2 shows an overview of the model. First, the dialog turn encoder $\mathbf{Enc}^{\mathbf{turn}}$ encodes each interview $D = \{t_1, \ldots, t_{n-1}, t_n\}$, where $t_i = \{w^i_1, \ldots, w^i_{m-1}, w^i_m\}$ is a dialog turn and w^i_j is a jth token in turn t_i, on the word level, producing an embedding $\boldsymbol{h}^{\mathbf{turn}}_i$ for each turn (1). Then, the dialog turn embeddings are processed on a higher level of the hierarchy with the interview-level encoder $\mathbf{Enc}^{\mathbf{int}}$ to produce the interview representation $\boldsymbol{h}^{\mathbf{int}}$ (2). Finally, the interview embedding is passed to a feed-forward network that maps the interview representation to a label vector $\hat{\boldsymbol{l}} = [\hat{l}_1, \hat{l}_2, \ldots, \hat{l}_7, \hat{l}_8]$ (3, 4, 5), where each predicted label $\hat{l}_k \in [0, 3]$ represents a symptom score for a corresponding question in PHQ-8. The feed-forward classifier consists of two linear layers (W_1, W_2) with biases $(\boldsymbol{b}_1, \boldsymbol{b}_2)$, with a LeakyReLU activation function and a LayerNorm layer [38] in-between.

$$\boldsymbol{h}^{\mathbf{turn}}_i = \mathbf{Enc}^{\mathbf{turn}}(t_i) \text{ for } i = 1, \ldots, |D| \tag{1}$$

$$\boldsymbol{h}^{\mathbf{int}} = \mathbf{Enc}^{\mathbf{int}}(\{\boldsymbol{h}^{\text{turn}}_1, \ldots, \boldsymbol{h}^{\text{turn}}_{|D|}\}) \tag{2}$$

$$\boldsymbol{z}' = \text{LeakyReLU}(\boldsymbol{h}^{\text{int}} W_1^T + \boldsymbol{b}_1) \tag{3}$$

$$\boldsymbol{z} = \text{LayerNorm}(\boldsymbol{z}') \tag{4}$$

$$\hat{\boldsymbol{l}} = \boldsymbol{z} W_2^T + \boldsymbol{b}_2 \tag{5}$$

The word-level turn encoder $\mathbf{Enc}^{\mathbf{turn}}$ uses a distilled RoBERTa-based model from the SentenceTransformers (S-RoBERTa)[3]. SentenceTransformers is a collection of pre-trained Transformer-based language models that have been tuned to produce better sentence embeddings [39]. RoBERTa is a Transformer-based language model which has been pre-trained on a large collection

[3] https://huggingface.co/sentence-transformers/all-distilroberta-v1.

of common-domain corpora for the masked language modeling (MLM) task [40]. During MLM pre-training, some of the input tokens are masked, and the model's objective is to predict the token that has been masked [41]. Finally, in SentenceTransformers, the model is further fine-tuned on the sentence similarity task, where the sentence embedding is produced by averaging all its respective token embeddings [39]. Furthermore, the S-RoBERTa model used in our experiments has been distilled. Knowledge distillation is a process of training a smaller student model which learns to copy the larger pre-trained teacher model [42]. Distilled models keep most of the capabilities of their full-sized counterparts while being almost twice as small and fast. Decreasing the computational complexity of our model is crucial due to the fact that all turns of the interviews have to be processed in parallel, i.e., several copies of $\mathbf{Enc}^{\mathbf{turn}}$ are created, and their respective computational graphs stored during training. The turn-level interview encoder $\mathbf{Enc}^{\mathbf{int}}$ deploys a single layer BiLSTM with a hidden dimension of 300 and an additive attention layer on top of it.

As a training objective for the symptom prediction task, the Smooth L_1 loss [43] was used, which is defined as in (6) for multi-target regression:

$$\text{Smooth}_{L_1}(\hat{\boldsymbol{l}}, \boldsymbol{l}) = \frac{1}{K}\sum_{k=1}^{K}\text{Smooth}_{L_1}(\hat{l}_k, l_k) \tag{6}$$

where $\hat{l}_k$ and l_k are the predicted and true scores for the kth symptom respectively, $K = 8$ is the number of symptoms, and with

$$\text{Smooth}_{L_1}(\hat{l}_k, l_k) = \begin{cases} 0.5(\hat{l}_k - l_k)^2, & \text{if}|\hat{l}_k - l_k| < 1 \\ |\hat{l}_k - l_k| - 0.5, & \text{otherwise} \end{cases} \tag{7}$$

Since distinct random seeds can lead to substantially different results [44], each model was trained five times using different random seeds, and the average of the five runs is reported. Each model was trained for 200 epochs using AdamW optimizer with the learning rate of $3e^{-5}$ and a linear warm-up scheduler. A model checkpoint was saved after each epoch, and the checkpoint with the highest micro-averaged F1-score on the development set was chosen as the final model.

3.3 Baseline models

To provide some validity to the symptom prediction approach, we compare the results of our model to three baseline tasks adopted in previous works: 1) binary diagnostic classification, where a patient is said to be depressed if their PHQ-8 score is at least 10, and non-depressed otherwise, 2) multi-class classification into five classes with differing severity as depicted in Table 1, i.e., no symptoms, mild, moderate, moderately severe and severe depression, and 3) depression severity prediction modeled as PHQ-8 total score regression ranging from 0 to 24.

The outputs of our multi-target regression model predicting symptom scores can be recast to a suitable format for these three tasks. For the depression severity prediction task, the symptom scores are summed up to give the estimate of the final PHQ-8 value. For the binary and multi-class classification tasks, the summed total score can be converted either into a binary label at a cutoff of 10 for the binary diagnostic classification or converted into five classes for the multi-class classification, such that [0..5) stands for no symptoms, [5..10) mild, [10..15) moderate, [15..20) moderately severe and [20..24] severe depression estimate.

For comparison, we train three baseline models that predict the three tasks directly, i.e., the model predicts one of two classes for the binary diagnostic prediction, one class out of five for the multi-class severity prediction, and a continuous score for the total depression severity regression. All baseline models use the same hierarchical architecture shown in Fig. 2; only the output layer of the feed-forward classifier network is different. Whereas the output layer for the symptom prediction model has multiple regression heads, the depression severity prediction model has a single regression head, and the models for the binary and the multi-class classifiers have a classification head that predicts one of the two or five classes, respectively.

3.4 Evaluation

For evaluating the regression tasks (symptom scores regression and PHQ-8 total score regression), we use the mean absolute error (MAE) as defined in equation (8), where y_i is the correct PHQ-8 score, and $\hat{y}_i$ is the predicted PHQ-8 value, which in case of the symptom prediction model is obtained by summing up all the predicted symptom scores. N is the number of interviews in the evaluation set.

$$\text{MAE} = \frac{\sum_{i=1}^{N}|\hat{y}_i - y_i|}{N} \tag{8}$$

In order to better take into account the imbalance in scores and especially the scarcity of interviews with higher PHQ-8 total score values, we also use a macro-averaged version of the MAE (*ma*MAE), where the MAE is first computed separately for each class/score range, and then the resulting MAE-s are averaged. The computation is defined in Eq. (9), where C is the set of classes, MAE^c denotes the MAE for the class c.

$$maMAE = \frac{\sum_{c \in C} \mathrm{MAE}^c}{|C|} \tag{9}$$

For evaluating the classification tasks, we use the micro-averaged F1-score (miF_1) and the macro-averaged F1-score (maF_1) defined in Eqs. (10) and (11) respectively. For computing the precision and recall for the miF_1, the true positive, false positive, and false negative counts are accumulated over all classes. For maF_1, the class-specific F1-score F_1^c is first computed for each class c separately from the class-specific precision and recall measures, and then the F1-scores for all classes are averaged.

$$miF_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{10}$$

$$maF_1 = \frac{\sum_{c \in C} F_1^c}{|C|} \tag{11}$$

4 Results

In this section, we present the results of the multi-target architecture compared to baselines for the binary, multi-class, and regression tasks. We then show the performance of our method for each symptom individually and illustrate the symptom-based decisions for the binary and multi-class cases with radar plots. Finally, we present the results of the symptom network analysis based on non-dynamic data.

4.1 Comparison to baselines

The top section of Table 2 shows the comparison of our Symptom Prediction model to the three baselines outlined in “3.3” section—the Binary Diagnostic model, the 5-class Severity prediction model, and the PHQ-8 Severity prediction model. Overall, the Symptom Prediction model performed better or in the same range compared to the baseline models in all evaluation tasks. In particular, the Symptom Prediction model performed considerably better than other models when evaluated on the Binary Diagnosis and the PHQ-8 Score Severity evaluation tasks. On the 5-Class Severity evaluation task, the 5-Class Severity classification model that was explicitly trained to predict these five severity classes performed better on the micro-F1 evaluation score, while on the macro-F1 evaluation score, which weighs all classes equally, both models performed similarly. We also noticed that the PHQ-8 Score Severity model, which was trained to predict the total PHQ-8 score, performed considerably worse than other models on both classification tasks.

The bottom part of Table 2 shows the results of the previous works on DAIC-WOZ data for comparison. All these works have used only text modality as input, as is also the case in our work. Overall, the Symptom Prediction model shows results that are in a similar range compared to previously published results. The only notable exception is the 5-Class Severity Evaluation task, where Qureshi et al. [25] obtained considerably higher results.

Table 3 shows the results on the development set that was used for selecting the final model. Slight overfitting on the development set can be observed for the Binary Diagnosis model. The standard deviations of the reported scores for the development set were higher than for the test set. Finally, the Symptom-based Diagnosis model was more robust than the rest of the baseline models.

4.2 Symptom prediction analysis

Next, we will look at the performance of the Symptom Prediction model for each symptom separately. The performance of each symptom was evaluated with both MAE and micro- and macro-averaged F1-scores. To compute the F1-scores, the predicted symptom scores were converted into binary labels with a cutoff of 1.5 points, such that scores lower than 1.5 were considered as symptom absent (negative class), and the scores starting from 1.5 were considered as symptom present (positive class).

While MAE is generally an effective and easily interpretable metric for evaluating regression tasks, it can give artificially low error scores when the data set is imbalanced, and the model tends to predict scores close to the mean value. Relative root mean square error (RRMSE) [45] can give a better view of the performance in those cases, as it penalizes more the model that tends to predict scores close to the mean value of the training set. RRMSE is defined in Eq. (12)

$$RRMSE = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}} \tag{12}$$

where $\bar{y}$ is the mean score of the training set, $\hat{y}_i$ is the model’s prediction, and y_i is the correct score. RRMSE is a normalized measure where desirable values lie in the range of $[0 \ldots 1)$. RRMSE value 1 means that the evaluated model is equivalent to a naive model that always predicts the mean score of the training set, and the RRMSE value greater than 1 shows that the evaluated model is even worse than predicting the average score.

Table 4 shows the symptom prediction performances. First of all, one can observe that the scores for the two main depression symptoms—depressed mood and lack of interest—are among the most accurately predicted ones across all evaluation measures; this indicates that those symptoms can be inferred sufficiently well from

Table 2 Experimental results on the test set of the DAIC-WOZ data set

Model	**Binary Diagnosis Eval**		**PHQ-8 Score Severity Eval**		**5-Class Severity Eval**	
	***miF*$_1$ ± σ**	***maF*$_1$ ± σ**	**MAE ± σ**	***ma*MAE ± σ**	***miF*1-5c ±σ**	***maF*1-5c ±σ**
Binary Diagnosis	0.719 ± 0.016	0.701 ± 0.010	–	–	–	–
5-Class Severity	0.711 ± 0.026	0.683 ± 0.024	–	–	**0.468** ± 0.023	**0.270** ± 0.025
PHQ-8 Score Severity	0.681 ± 0.019	0.584 ± 0.024	5.03 ± 0.09	5.69 ± 0.12	0.289 ± 0.029	0.135 ± 0.014
Symptom Prediction	**0.766** ± 0.023	**0.739** ± 0.025	**3.78** ± 0.13	**4.19** ± 0.13	0.426 ± 0.014	**0.270** ± 0.019
HCAN [7]	–	0.630	–	–	–	–
HAN+L [8]	–	0.700	–	–	–	–
ASP MT. DLC+DLR+EIR [25]	–	–	3.69	–	0.600	–
HCAG-T [23]	–	0.770‡	3.73‡	–	–	–
SGNN [27]	–	–	3.76	–	–	–

Top Section: results of our model and the baselines. All models were run five times with different seed values, and the average values with standard deviation are presented; *miF*1-5c (resp. *maF*1-5c) stands for the 5-class micro-averaged F1-score (resp. macro-averaged F1-score). Bottom Section: previously published results on the same DAIC-WOZ test set using only text modality; all results are given for the best model and not based on the average performance of several runs.

Bold values indicates the best results for each model

‡ indicates that the results are given for the validation set only

Table 3 Experimental results on the development set of the DAIC-WOZ data set

Model	**Binary Diagnosis Eval**		**PHQ-8 Score Severity Eval**		**5-Class Severity Eval**	
	***miF*$_1$±σ**	***maF*$_1$±σ**	**MAE±σ**	***ma*MAE±σ**	***miF*1-5c ±σ**	***maF*1-5c ±σ**
Binary Diagnosis	**0.806** ± 0.031	**0.798** ± 0.031	-	-	-	-
5-Class Diagnosis	0.739 ± 0.049	0.713 ± 0.058	-	-	**0.503** ± 0.049	0.237 ± 0.017
PHQ-8 Score Diagnosis	0.600 ± 0.030	0.507 ± 0.026	5.51 ± 0.06	6.01 ± 0.08	0.255 ± 0.024	0.159 ± 0.018
Symptom-based Diagnosis	0.752 ± 0.035	0.719 ± 0.047	**3.61** ± 0.12	**4.11** ± 0.18	0.442 ± 0.106	**0.286** ± 0.063

All models were run five times with different seed values, and the average values with standard deviation are presented; *miF*1-5c (resp. *maF*1-5c) stands for the 5-class micro-averaged F1-score (resp. macro-averaged F1-score). Bold values indicates the best results for each model

the interview texts. Similarly, symptoms related to sleep and feeling of being a failure show good performance relative to the other symptoms according to all measures. According to MAE and miF_1, the most accurately predicted symptom is movement related, but this is misleading. In our sample, the moving symptom has a relatively low score for most participants, biasing the model towards always predicting low scores. Indeed, the RRMSE score reveals that most of the predictions were close to the average value for this symptom in the data set. Furthermore, a high miF_1 score and a low maF_1 score show that the model mostly predicts scores in a very similar range—in our case, it is the symptom score in the lower end of values that will be binarized into the negative, i.e., symptom absent, class.

Figure 3 shows a graphical view of the symptom predictions against the ground truth symptom scores averaged for the five-class depression severity scale (the main view) and non-depressed and depressed participants (bottom-right corner). The overall shape of the predictions generally follows the one of the ground truth scores for all groups. However, the model tends to predict scores closer to moderate ranges, thus overestimating the scores for non-depressed participants and underestimating for moderately and severely depressed participants.

4.3 Symptom network analysis

The symptom scores for all the participants in the test set can also be represented as a correlation graph, a representation that is in line with the SNA approach. In our case, we can test whether the graph with predicted values is structurally equivalent to the graph with the real scores. We followed the method by van Borkulo et al. [14]. We used a permutation-based hypothesis test where network structures are estimated with sparse, L_1 regularized partial correlations. The test is implemented in the NCT[4] package for R.

Two hypotheses were tested: about the invariant network structure, and the invariant global strength [46]. For the invariant network structure, the null hypothesis is that given the connection strength matrices A_1 and A_2 for graphs G_1 and G_2, all edge weights in A_1 are identical

[4] https://cran.r-project.org/web/packages/NetworkComparisonTest/.

Table 4 Test scores for each symptom

Symptom	MAE $\pm\sigma$	RRMSE $\pm\sigma$	*miF*1 $\pm\sigma$	*maF*1 $\pm\sigma$
No interest	0.529 ± 0.047	0.877 ± 0.067	0.800 ± 0.024	0.669 ± 0.043
Depressed mood	0.550 ± 0.027	0.733 ± 0.022	0.821 ± 0.019	0.729 ± 0.024
Insomnia or hypersomnia	0.753 ± 0.073	0.805 ± 0.060	0.774 ± 0.055	0.757 ± 0.047
Feeling tired	0.638 ± 0.031	0.816 ± 0.030	0.745 ± 0.030	0.709 ± 0.035
Eating too little or too much	0.811 ± 0.049	0.972 ± 0.064	0.762 ± 0.035	0.685 ± 0.026
Feeling of being a failure	0.620 ± 0.018	0.796 ± 0.012	0.817 ± 0.024	0.779 ± 0.021
Problems with concentrating	0.830 ± 0.040	0.878 ± 0.012	0.681 ± 0.034	0.557 ± 0.029
Moving too slowly or too fast	0.438 ± 0.022	0.976 ± 0.035	0.936 ± 0.000	0.484 ± 0.000

All models were run five times with different seed values, and the average values with standard deviation are presented. For computing the F1 scores, the predicted scores were binarized, such that the scores < 1.5 were treated as negative class instances, and the scores ≥ 1.5 were treated as positive class instances

to those in A_2. The test statistic M is the largest difference between all connection strengths. For invariant global strength, the null hypothesis states that the overall connectivity is the same across the two graphs. The test statistic is the distance S that is defined as:

$$S(G_1, G_2) = |\sum_{i,j \in V} |A_{1ij}| - \sum_{i,j \in V} |A_{2ij}|| \tag{13}$$

where V is the set of nodes in networks G_1 and G_2. On the test set, the invariant network structure test results were in $M = 0.3648$ and p value $= 0.75$, and the invariant global strength test in $S = 0.0307$ and p value $= 0.96$. Thus, we accept the null hypotheses of both tests and conclude that the symptom networks with predicted and real symptom scores are, indeed, structurally equivalent.

5 Discussion

In this work, we address the automatic prediction of depression based on text transcripts. Instead of predicting the binary diagnostic label, as has been common in previous works, we propose to predict the fine-grained profile of symptoms that underlie the diagnosis of depression. According to our knowledge, such symptom-based approach has not been attempted before on the DAIC-WOZ data set, which has been used in many previous studies to develop clinical prediction models for depression. The predicted PHQ-8 symptom scores can be easily represented in various ways: as a total sum score representing the overall depression severity, and as both binary diagnostic and multi-class severity categories, thus also allowing for comparison with other systems. The experimental results showed that the symptom prediction approach is relatively robust and is on par with the previously published systems while at the same time giving a fine-grained overview of the person's symptoms that the previous automatic diagnostic classification systems lack.

The models were able to predict some symptoms better than the others. In particular, Table 4 shows that such symptoms as lack of interest, depressed mood, feeling of being a failure, and feeling tired are among the most accurately predicted symptoms. This reflects the nature of the DAIC-WOZ data since these topics are discussed the most during each interview. Some of the symptoms may be addressed directly, e.g., by asking if the person was diagnosed with depression or PTSD in the past. The other symptoms are given attention as well, even though they are less direct, e.g., assessing the feeling of being a failure by asking what the interviewee's friends and family think about them. The sleep-related symptom is also predicted relatively accurately; there are indeed questions about the person's sleep problems, but they are not present in every interview. Finally, the symptoms related to eating, problems with concentration, and slowed down or overly agitated movement are not detected accurately by the model. Interestingly, the results show a RRMSE score close to 1 for these symptoms, which can indicate that there is little textual evidence of these symptoms in the data and thus, the model just learns an average score for these symptoms across the training data set.

The radar plots on Fig. 3 showed that the model's predictions are close to the real ones for people with the depressive symptoms in the mild, and moderate severity range. However, the model tends to overevaluate the cases in the absent severity range and underevaluate the cases in the moderately severe and severe range. The underevaluation in the high range can be explained by the lack of data in this region: only seven interviews are available for training for the moderately severe subclass and four for severe one and even less for testing, with seven and two interviews, respectively. Additionally, Fig. 3 shows that the moving-related symptom consistently receives low scores across the whole depression severity spectrum. This is also reflected in the interviews;

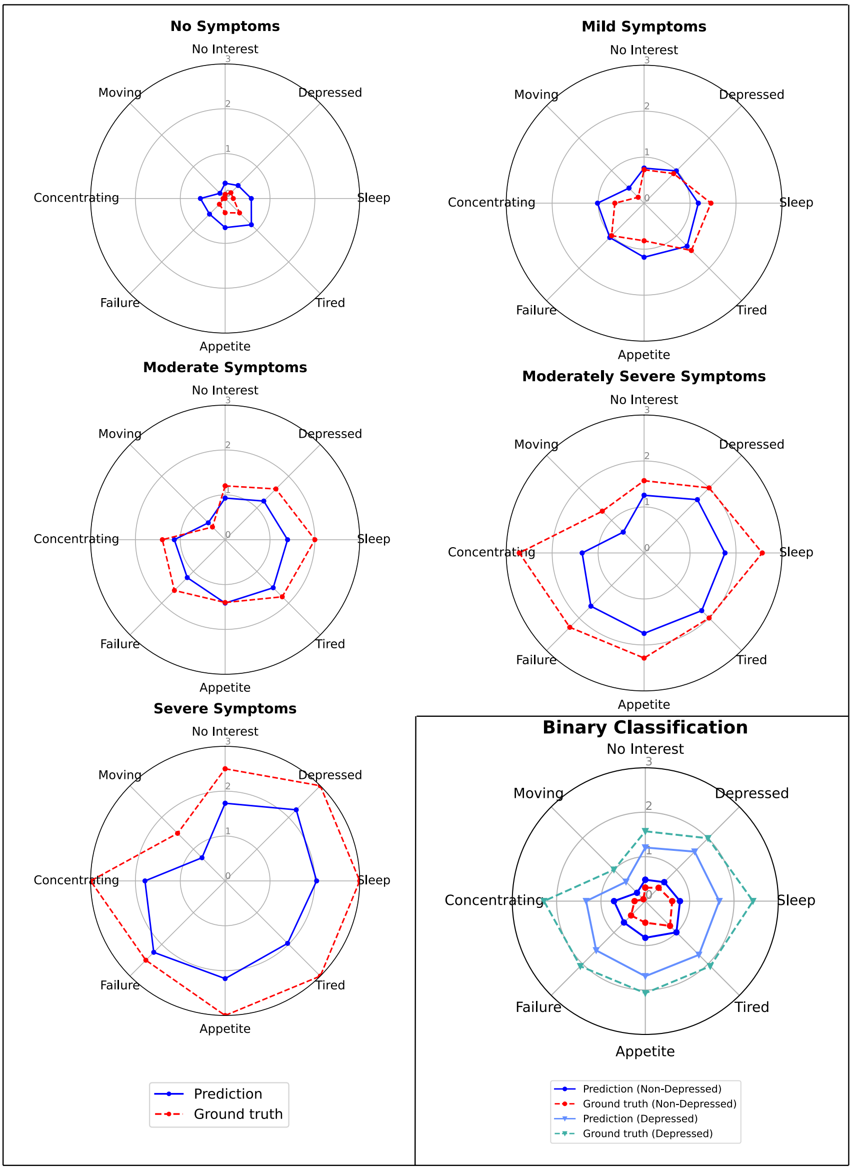


Fig. 3 Averaged predictions and ground truth symptom scores for the test set for five fine-grained classes and binary classification. Predictions are averaged across five models trained with the same parameters and different seed values

the moving-related symptoms are almost never verbally discussed, which can explain the high RRMSE score. We believe that the indicators of this symptom are mostly non-verbal; thus, a multi-modal setting that includes visual input might improve the results.

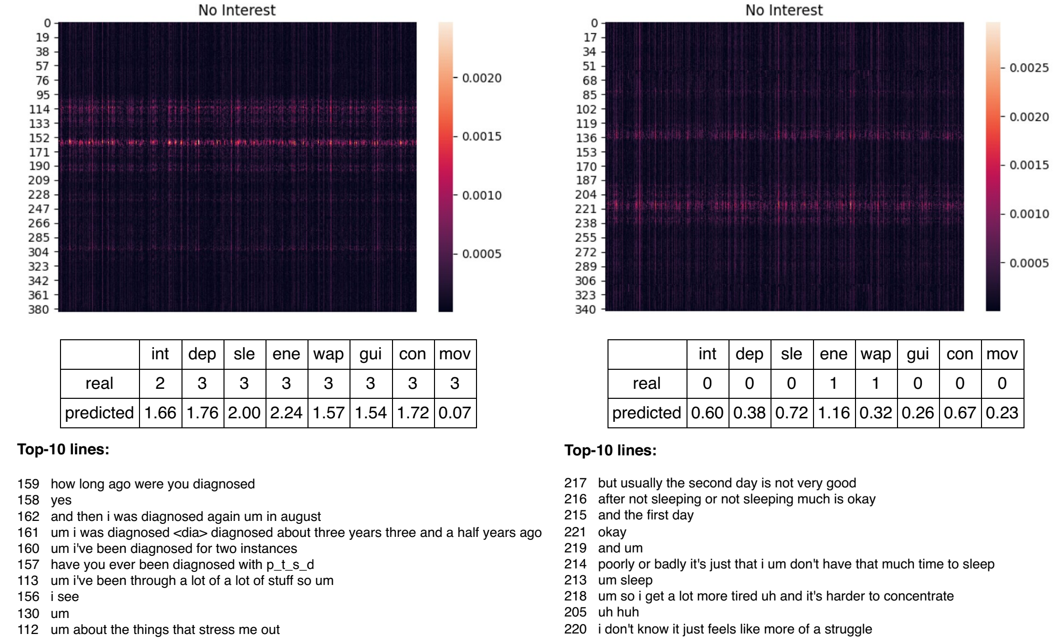
Interpreting the model predictions may help to understand the data itself better. To gain an understanding of the model's decision-making, we extracted the saliency maps that track the prediction of each symptom back to the inputs, which in our case, are the dialogue turns. Saliency mapping is a gradient-based method that computes the importance of an input feature, i.e., an interview turn, based on the first-order derivative with respect to that feature [47]. Although saliency maps can be noisy [48], they can still provide useful information. They are also relatively easy to extract from most neural architectures. We extracted saliency maps for each symptom prediction and observed that the areas corresponding to the high importance are almost identical for each symptom and point to the area close to the middle of the interview. Indeed, each interview is structured in a way that the interviewer asks general non-depression-related questions in the beginning in order to establish trust with the interviewee. Similarly, at the end of the interview, the interviewer moves away from depression-related topics to wind the interviewee down.

Figure 4 shows an enhanced view of the gradients tracked back from the same symptom (lack of interest) to the input features for two different persons in the test set. The lines to which the highest absolute gradient value was attributed are "diagnostic"-related in the case of the person with a high PHQ-8 score indicating severe depressive symptoms (left in Fig. 4); for the non-depressed person (in the right), the model attributed high importance to the sleep-related utterances. After having studied the feature attributions across the whole test set, we observe that the model assigns importance to the symptom-related turns of the interview most of the time.

Every interview also includes the question "Have you been diagnosed with depression?". Thus, it is plausible that the model can extract information relevant to predictions only from the answer to this question, thus using it as a shortcut. Although inspecting the saliency scores showed that the turn involving this question was not among the most important ones for most of the interviews, we investigated more thoroughly whether this question strongly correlates with the model's predictions. First, we classified the answers to this question into three categories: "yes", "no", and "other." "Yes" and "no" categories were assigned to the answers that can be clearly interpreted as positive or negative. If a participant tried to avoid the question or started to give extra information about their condition, the answer was classified as "other". Fisher's exact test at the p value < 0.05 was used to decide whether the depressed and non-depressed participant groups were different in their "yes" and "no" answers to this question. Similar analyses were conducted for every symptom with the groups formed by the symptom scores. Based on these analyses, we can conclude that the answers to the question "Have you been diagnosed with depression?" differ significantly between the groups formed based on different symptom scores, and thus, the model is suspect in utilizing these differences when making predictions. To estimate how dependent the model is on these answers, we replaced all the "yes" answers with a random answer variation from the "no" answer set and vice versa. Additionally, we replaced each "other" answer with another random answer from the "other" answer set as well. The same model was run on this perturbed test set, showing no drop in the miF_1 score (− 0.00%) and an insignificant minor drop in the maF_1 score (− 0.52%). Similar pattern was observed for MAE (+ 0.06) and *ma*MAE (+ 0.11). Thus, we can conclude that the model did not use this question with its explicit answers as a shortcut for making complex predictions.

This work also has several limitations. First, our work is limited to the DAIC-WOZ data set, which is, to our knowledge, the only high-quality data set that is easily obtainable from its authors. This data set is, however, quite small which might lead the models to overfit; nonetheless, the comparison of the development and test set results showed that the symptom-based model is fairly robust to overfitting. The data set also has a quite rigid structure, as all interview prompts are sampled from a closed set of prompts. Thus, we cannot assume the generalizability of the presented results to other data sets, which limits the applicability of our model. Furthermore, the transcribed interviews are long and require using a hierarchical architecture as one way to encode them. This entails a lot of computational power for training such a model due to its high computational complexity, thus limiting us in the choice of pre-trained contextual embeddings that are a foundation of most of the NLP neural architectures.

The main motivation for predicting symptoms instead of binary diagnostic classes, total depression severity or discrete severity class as has been custom in previous works, is the understanding of the need to keep up with advances in psychiatry, which moves towards more dimensional and descriptive diagnostic profiles. In our work, we also followed the general ideas of symptom network analysis and conducted analyses on the graphs of predicted and real symptom relations. However, because our data is cross-sectional, we are constrained to correlational analyses, whereas the real aim and strength of symptom network analysis rely on following the causal relations between symptoms, e.g., which symptoms cause which other symptoms. However, modeling these complex causal relationships with predictive models would require longitudinal data with several points of measurement in time.



	int	dep	sle	ene	wap	gui	con	mov
real	2	3	3	3	3	3	3	3
predicted	1.66	1.76	2.00	2.24	1.57	1.54	1.72	0.07

Top-10 lines:

159 how long ago were you diagnosed
158 yes
162 and then i was diagnosed again um in august
161 um i was diagnosed <dia> diagnosed about three years three and a half years ago
160 um i've been diagnosed for two instances
157 have you ever been diagnosed with p_t_s_d
113 um i've been through a lot of a lot of stuff so um
156 i see
130 um
112 um about the things that stress me out

	int	dep	sle	ene	wap	gui	con	mov
real	0	0	0	1	1	0	0	0
predicted	0.60	0.38	0.72	1.16	0.32	0.26	0.67	0.23

Top-10 lines:

217 but usually the second day is not very good
216 after not sleeping or not sleeping much is okay
215 and the first day
221 okay
219 and um
214 poorly or badly it's just that i um don't have that much time to sleep
213 um sleep
218 um so i get a lot more tired uh and it's harder to concentrate
205 uh huh
220 i don't know it just feels like more of a struggle

Fig. 4 Saliency maps showing which parts of the interview are used for symptom predictions

6 Conclusions

The main contribution of this work is highlighting the importance of keeping up with the advances in psychiatry and clinical psychology in the computational modeling and automatic prediction domain by moving away from predicting static diagnostic categories that contain limited information towards more descriptive and personalized symptom profiles. Towards this goal, we trained a multi-target hierarchical regression model to predict the severity scores of individual depression symptoms from patient–psychiatrist interview transcripts from the DAIC-WOZ corpus. The model achieved a mean absolute error (MAE) from 0.438 to 0.830 on eight depression symptoms and showed state-of-the-art results in binary depression estimation (0.739 maF_1) and total depression score prediction (3.78 MAE). Moreover, our model produced a symptom correlation graph that is structurally identical to the real one based on the static data. The applicability of the presented model is limited because it was trained and evaluated on the relatively small DAIC-WOZ data set. Despite this limitation, we believe that the proposed symptom-based approach should be developed further as it provides more in-depth information about the depressive condition than a general binary diagnosis. Moreover, it aligns with the symptom network analysis which is a recently proposed diagnostic approach in psychiatry.

Abbreviations

BERT	Bidirectional encoder representations from transformers
DAIC-WOZ	Distress analysis interview corpus wizard-of-oz
DSM-5	Diagnostic and statistical manual of mental disorders
GNN	Graph neural network
GPC	General psychotherapy corpus
HCAN	Hierarchical contextual attention network
HRSD	Hamilton rating scale for depression
MAE	Mean absolute error
MDD	Major depressive disorder
MLM	Masked language modeling
NCT	Network comparison test
NLP	Natural language processing
PHQ	Patient health questionnaire
RRMSE	Relative root mean square error
SNA	Symptom network analysis

Acknowledgements
Not applicable.

Author contributions

KM contributed in designing the study, did all the implementations and experiments, and contributed in writing and approving the final manuscript. Both KS and GD supervised the study, contributed in designing the study. All authors read and approved the final manuscript.

Funding
This research was supported by the Estonian Research Council Grant PSG721.

Availability of data and materials
The DAIC-WOZ data set that supports the findings of this study is available from The University of Southern California Institute for Creative Technologies (https://dcapswoz.ict.usc.edu/) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data set is, however, available from the University of Southern California Institute for Creative Technologies upon reasonable request [22]. The code of the model is available in the author's repository: https://git.unicaen.fr/kirill.milintsevich/hierarchical-depression-symptom-classifier.

Declaration

Ethics approval and consent to participate
Not applicable.

Consent for publication
All authors have seen and approved the current version of the paper.

Competing interests
The authors declare that they have no competing interests.

Received: 23 September 2022 Accepted: 18 January 2023
Published online: 13 February 2023

References

1. WHO (2017) Depression and other common mental disorders: global health estimates. Technical report, World Health Organization
2. American Psychiatric Association (2013) Diagnostic and Statistical Manual of Mental Disorders, 5th edn. American Psychiatric Pub, Arlington
3. Rodgers S, Holtforth MG, Müller M, Hengartner MP, Rössler W, Ajdacic-Gross V (2014) Symptom-based subtypes of depression and their psychosocial correlates: a person-centered approach focusing on the influence of sex. Journal of Affective Disorders 156:92–103
4. Ten Have M, Lamers F, Wardenaar K, Beekman A, de Jonge P, van Dorsselaer S, Tuithof M, Kleinjan M, de Graaf R (2016) The identification of symptom-based subtypes of depression: A nationally representative cohort study. J Affect Disord 190:395–406
5. Calvo RA, Milne DN, Hussain MS, Christensen H (2017) Natural language processing in mental health applications using non-clinical texts. Natural Language Engineering 23(5):649–685
6. Chancellor S, De Choudhury M (2020) Methods in predictive techniques for mental health status on social media: a critical review. NPJ Digital Med 3(1):1–11
7. Mallol-Ragolta A, Zhao Z, Stappen L, Cummins N, Schuller B (2019) A hierarchical attention network-based approach for depression detection from transcribed clinical interviews
8. Xezonaki D, Paraskevopoulos G, Potamianos A, Narayanan S (2020) Affective conditioning on hierarchical attention networks applied to depression detection from transcribed clinical interviews. In: Interspeech (INTERSPEECH), pp. 4556–4560
9. Dai Z, Zhou H, Ba Q, Zhou Y, Wang L, Li G (2021) Improving depression prediction using a novel feature selection algorithm coupled with context-aware analysis. J Affect Disord 295:1040–1048
10. Borsboom D, Cramer AO (2013) Network analysis: an integrative approach to the structure of psychopathology. Ann Rev Clin Psychol 9:91–121
11. McNally RJ (2016) Can network analysis transform psychopathology? Behav Res Ther 86:95–104
12. Kaiser T, Herzog P, Voderholzer U, Brakemeier E-L (2021) Unraveling the comorbidity of depression and anxiety in a large inpatient sample: Network analysis to examine bridge symptoms. Depression and anxiety 38(3):307–317
13. Price M, Legrand AC, Brier ZM, Hébert-Dufresne L (2019) The symptoms at the center: examining the comorbidity of posttraumatic stress disorder, generalized anxiety disorder, and depression with network analysis. J Psychiatr Res 109:52–58
14. van Borkulo C, Boschloo L, Borsboom D, Penninx BW, Waldorp LJ, Schoevers RA (2015) Association of symptom network structure with the course of depression. JAMA Psychiatry 72(12):1219–1226
15. Fried EI, Epskamp S, Nesse RM, Tuerlinckx F, Borsboom D (2016) What are'good'depression symptoms? comparing the centrality of dsm and non-dsm symptoms of depression in a network analysis. J Affect Disord 189:314–320
16. Park S-C, Kim D (2020) The centrality of depression and anxiety symptoms in major depressive disorder determined using a network analysis. J Affect Disord 271:19–26
17. Fried EI, Nesse RM (2015) Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. BMC Med 13(1):72. https://doi.org/10.1186/s12916-015-0325-4
18. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC (2017) Detecting depression and mental illness on social media: an integrative review. Curr Opin Behav Sci 18:43–49
19. Beck AT, Steer RA, Carbin MG (1988) Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. Clinical Psychology Review 8(1):77–100. https://doi.org/10.1016/0272-7358(88)90050-5
20. Hamilton M (1986) The hamilton rating scale for depression. In: Assessment of Depression, pp. 143–152. Springer, Berlin
21. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH (2009) The PHQ-8 as a measure of current depression in the general population. Journal of Affective Disorders 114(1–3):163–173. https://doi.org/10.1016/j.jad.2008.06.026. Accessed 2021-06-03
22. Gratch J, Artstein R, Lucas G, Stratou G, Scherer S, Nazarian A, Wood R, Boberg J, DeVault D, Marsella S, et al. (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), pp. 3123–3128
23. Niu M, Chen K, Chen Q, Yang L (2021) Hcag: A hierarchical context-aware graph attention model for depression detection. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4235–4239
24. Qureshi SA, Saha S, Hasanuzzaman M, Dias G (2019) Multitask representation learning for multimodal estimation of depression level. IEEE Intelli Systems 34(5):45–52. https://doi.org/10.1109/MIS.2019.2925204
25. Qureshi SA, Dias G, Hasanuzzaman M, Saha S (2020) Improving depression level estimation by concurrently learning emotion intensity. IEEE Comput Intell Mag 15(3):47–59
26. Qureshi SA, Dias G, Saha S, Hasanuzzaman M (2021) Gender-aware estimation of depression severity level in a multimodal setting. In: International Joint Conference on Neural Networks (IJCNN), pp. 1–8
27. Hong S, Cohn A, Hogg DC (2022) Using graph representation learning with schema encoders to measure the severity of depressive symptoms. In: International Conference on Learning Representations (ICLR)
28. Mao K, Zhang W, Wang DB, Li A, Jiao R, Zhu Y, Wu B, Zheng T, Qian L, Lyu W, Ye M, Chen J (2022) Prediction of depression severity based on the prosodic and semantic features with bidirectional lstm and time distributed cnn. IEEE Transactions on Affective Computing. https://doi.org/10.1109/TAFFC.2022.3154332
29. Gaut G, Steyvers M, Imel ZE, Atkins DC, Smyth P (2015) Content coding of psychotherapy transcripts using labeled topic models. IEEE journal of biomedical and health informatics 21(2):476–487
30. Delahunty F, Johansson R, Arcan M (2019) Passive diagnosis incorporating the phq-4 for depression and anxiety. In: Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task, pp. 40–46
31. Yadav S, Chauhan J, Sain JP, Thirunarayan K, Sheth A, Schumm J (2020) Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 696–709
32. Yazdavar AH, Al-Olimat HS, Ebrahimi M, Bajaj G, Banerjee T, Thirunarayan K, Pathak J, Sheth A (2017) Semi-supervised approach to monitoring clinical depressive symptoms in social media. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pp. 1191–1198

33. Nguyen T, Yates A, Zirikly A, Desmet B, Cohan A (2022) Improving the generalizability of depression detection by leveraging clinical questionnaires. arXiv preprint arXiv:2204.10432
34. Yao X, Yu G, Tang J, Zhang J (2021) Extracting depressive symptoms and their associations from an online depression community. Computers in human behavior 120:106734
35. Karmen C, Hsiung RC, Wetter T (2015) Screening internet forum participants for depression symptoms by assembling and enhancing multiple nlp methods. Computer Methods Programs Biomed 120(1):27–36
36. Davcheva E (2019) Classifying mental health conditions via symptom identification: A novel deep learning approach. In: International Conference of Information Systems, 2019
37. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical Attention Networks for Document Classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489. Association for Computational Linguistics, San Diego, California . https://doi.org/10.18653/v1/N16-1174. http://aclweb.org/anthology/N16-1174 Accessed 2021-06-03
38. Ba JL, Kiros JR, Hinton GE (2016) Layer normalization. arXiv preprint arXiv:1607.06450
39. Reimers N, 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. http://arxiv.org/abs/1908.10084
40. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692
41. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
42. Sanh V, Debut L, Chaumond J, Wolf T (2019) Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108
43. Girshick R (2015) Fast R-CNN. arXiv:1504.08083
44. Dodge J, Ilharco G, Schwartz R, Farhadi A, Hajishirzi H, Smith NA (2020) Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. CoRR **abs/2002.06305**
45. Borchani H, Varando G, Bielza C, Larranaga P (2015) A survey on multi-output regression. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 5(5):216–233
46. Van Borkulo CD, van Bork R, Boschloo L, Kossakowski JJ, Tio P, Schoevers RA, Borsboom D, Waldorp LJ (2022) Comparing network structures on three aspects: A permutation test. Psychological methods
47. Li J, Chen X, Hovy E, Jurafsky D (2015) Visualizing and understanding neural models in nlp. arXiv preprint arXiv:1506.01066
48. Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: International Conference on Machine Learning, pp. 3319–3328. PMLR

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

II

K. Milintsevich, G. Dias, K. Sirts
Evaluating Lexicon Incorporation for Depression Symptom Estimation
Proceedings of the 6th Clinical Natural Language Processing Workshop (Clinical NLP 2024), pages 529–535

Evaluating Lexicon Incorporation for Depression Symptom Estimation

Kirill Milintsevich[1,2] and **Gaël Dias**[1] and **Kairit Sirts**[2]
[1]Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, France
[2]Institute of Computer Science, University of Tartu, Estonia
{first_name}.{last_name}@{unicaen.fr[1]|ut.ee[2]}

Abstract

This paper explores the impact of incorporating sentiment, emotion, and domain-specific lexicons into a transformer-based model for depression symptom estimation. Lexicon information is added by marking the words in the input transcripts of patient-therapist conversations as well as in social media posts. Overall results show that the introduction of external knowledge within pre-trained language models can be beneficial for prediction performance, while different lexicons show distinct behaviours depending on the targeted task. Additionally, new state-of-the-art results are obtained for the estimation of depression level over patient-therapist interviews.

Illustration of the lexicon-based input marking
a) i'm pretty much good because see by me being a bus operator you run into circumstances and situations you gotta remain calm and still remain professional at the same time
b) i'm @ pretty @ much @ good @ because see by me being a bus operator you run into circumstances and situations you gotta remain @ calm @ and still remain professional at the same time
c) i'm @ pretty @ much @ good @ because see by me being a bus operator you run into circumstances and situations you gotta remain @ calm @ and still remain @ professional @ at the same @ time @

Table 1: Example of input marking. Text a) is the original text without markings, b) and c) show text with terms from AFINN and NRC lexicons.

1 Introduction

Considerable interest has emerged in using natural language processing to unobtrusively infer one's mental health condition (Chancellor and De Choudhury, 2020). A majority of studies have focused on predicting major depressive disorder (MDD) either as a symptom-based estimation (Yadav et al., 2020; Milintsevich et al., 2023) or a binary classification problem (Burdisso et al., 2023; Xezonaki et al., 2020). Both clinically motivated research initiatives and social media studies have emerged. In the latter case, Twitter (Zhang et al., 2023a), Reddit (Gupta et al., 2022) and depression-related forums (Yao et al., 2021) have fostered attention. In the former case, recorded patient-therapist conversations are transcribed and associated with self-assessment depression questionnaires, such as PHQ-8 (Kroenke et al., 2009) or BDI (Beck et al., 1988).

The DAIC-WOZ dataset (Gratch et al., 2014) has mostly been studied within the context of clinical research. Different works have been proposed to automatically infer depression level on this dataset: multi-modal (Qureshi et al., 2019; Wei et al., 2022) and text-based architectures (Li et al., 2023; Agarwal et al., 2022). The PRIMATE dataset (Gupta et al., 2022) has also received recent attention within the context of early symptom prediction on social media posts. The most comprehensive work on this dataset is proposed by Zhang et al. (2023a), which defines a context- and PHQ-aware transformer-based architecture.

People with MDD have shown increased use of negative emotional words and decreased use of positive emotional words (Rude et al., 2004; Savekar et al., 2023). In this line, Xezonaki et al. (2020) and Qureshi et al. (2020) used feature-level and task fusion of emotion and sentiment knowledge and showed improved performance for depression estimation. However, these works, along with other studies on social media mental health data (Zhang et al., 2023b), have used pre-transformer era neural architectures. Recent state-of-the-art approaches that rely on transformer-based pre-trained language models (PLMs) have not explored external knowledge fusion (Milintsevich et al., 2023).

In this paper, we investigate whether pre-trained language models could benefit from

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 322–328
June 21, 2024 ©2024 Association for Computational Linguistics

Lexicon	PHQ-8	Train	Dev	Test
AFINN	≥ 10	8.4	7.6	8.0
	< 10	8.2	7.6	7.9
NRC	≥ 10	7.6	†6.8	†7.1
	< 10	7.7	†7.6	†7.6
SDD	≥ 10	†0.6	0.4	0.5
	< 10	†0.4	0.3	0.4

Table 2: Proportion of marked words for each lexicon over the DAIC-WOZ. Reported values are in percentage. † shows if the difference between the depressed and non-depressed populations is statistically significant.

the introduction of emotional, sentimental, and domain-specific external knowledge from the lexicons: AFINN (Nielsen, 2011), NRC (Mohammad and Turney, 2013) and SDD (Yazdavar et al., 2017). Introducing this external knowledge into a transformer-based model is feature-level and is achieved by modifying the input with specific markers that highlight spans of text, as shown in Table 1, inspired by the works of Wang et al. (2021) and Zhou and Chen (2022). This approach does not require any modification to the model's architecture, such as changing attention mechanism (Li et al., 2021; Wang et al., 2022) or adding new layers (Bai et al., 2022); it also keeps the model's vocabulary unchanged unlike Zhong and Chen (2021).

Results on the DAIC-WOZ dataset show that the performance of transformer-based models is impacted by the added lexicon information (especially sentiment), and new state-of-the-art values can be obtained from the combination of the three lexicons. However, such results are less expressive for the PRIMATE dataset, with slight improvements induced by the introduction of external information. Overall, the improvement in predicting particular symptoms evidences that lexicon information can be helpful, provided that its content closely corresponds to the targeted task.

2 Methodology

Data. In this work, we use two depression datasets: DAIC-WOZ (Gratch et al., 2014) and PRIMATE (Gupta et al., 2022). The DAIC-WOZ dataset contains 189 clinical interviews in a dialogue format. Each interview has two actors: a human-controlled virtual therapist and a participant. The dataset is distributed in pre-determined splits, such that 107 interviews are used for training, 35 for validation, and 47 for testing. Each interview in the dataset is accompanied with a PHQ-8 assessment, which consists of eight questions inquiring about symptoms. Each question is scored from 0 to 3 on a Likert scale, and the total PHQ score ranging from 0 to 24 is the sum of the eight symptom scores. According to a standard cutoff score of 10, the interviews can be divided into diagnostic classes, where subjects with PHQ-8 total score < 10 are considered non-depressed, and those with score ≥ 10 are categorized as depressed. The eight listed symptoms are: `LOI` (lack of interest), `DEP` (feeling down), `SLE` (sleeping disorder), `ENE` (lack of energy), `EAT` (eating disorder), `LSE` (low self-esteem), `CON` (concentration problem), `MOV` (hyper/lower activity).

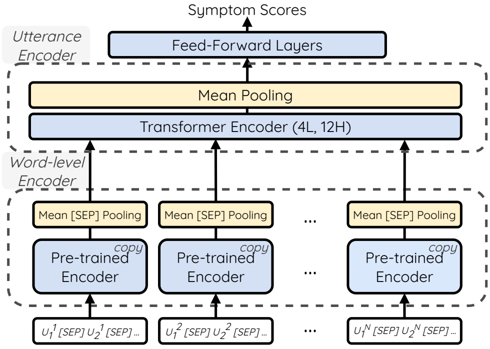


Figure 1: Overview of the model architecture. U_i^N stands for i-th utterance of N-th input. *Symptom Scores* are $||L||$ real numbers, where $||L||$ is the number of symptoms to predict.

The PRIMATE dataset is based on Reddit posts from depression-related communities, or subreddits, in which people describe their health conditions. A total of 2003 posts were manually annotated with binary labels for each individual symptom from the PHQ-9 (Kroenke et al., 2001), each label signifying whether the corresponding symptom is discussed in the post or not. PHQ-9 has the same first eight symptoms as PHQ-8 and one additional `SUI` (suicidal thoughts). The data was labeled by five crowd workers and verified by a mental health professional. The dataset is not pre-split into the train, validation, and test sets, so we randomly take 1601, 201, and 201 posts for each split accordingly.

Model architecture. To encode the interview transcripts, we adopt the hierarchical model from (Milintsevich et al., 2023). In their model, the interview is first split utterance-by-utterance, with each utterance processed by a word-level encoder.

Model	LOI	DEP	SLE	ENE	EAT	LSE	CON	MOV	PHQ-8
BERT	$0.56_{\pm.05}$	$\mathbf{0.63_{\pm.02}}$	$0.77_{\pm.05}$	$0.87_{\pm.04}$	$\mathbf{0.81_{\pm.03}}$	$0.78_{\pm.06}$	$0.74_{\pm.01}$	$0.34_{\pm.01}$	$4.38_{\pm.21}$
+SDD	$0.70_{\pm.02}$	$0.88_{\pm.05}$	$0.94_{\pm.05}$	$0.94_{\pm.04}$	$1.00_{\pm.07}$	$0.97_{\pm.04}$	$0.87_{\pm.02}$	$0.34_{\pm.00}$	$5.60_{\pm.18}$
+AFINN	$\mathbf{0.50_{\pm.03}}$	$0.70_{\pm.03}$	$0.79_{\pm.03}$	$0.81_{\pm.04}$	$0.85_{\pm.03}$	$0.72_{\pm.02}$	$0.77_{\pm.02}$	$0.34_{\pm.00}$	$4.56_{\pm.22}$
+NRC	$\mathbf{0.50_{\pm.03}}$	$0.66_{\pm.05}$	$\mathbf{0.73_{\pm.05}}$	$0.77_{\pm.03}$	$\mathbf{0.81_{\pm.05}}$	$0.71_{\pm.07}$	$\mathbf{0.73_{\pm.05}}$	$0.34_{\pm.00}$	$\mathbf{4.31_{\pm.18}}$
+ALL	$\mathbf{0.50_{\pm.04}}$	$0.69_{\pm.03}$	$0.81_{\pm.12}$	$\mathbf{0.74_{\pm.06}}$	$\mathbf{0.81_{\pm.07}}$	$\mathbf{0.69_{\pm.05}}$	$0.74_{\pm.03}$	$0.34_{\pm.00}$	$4.56_{\pm.42}$
MEBERT	$0.59_{\pm.02}$	$0.64_{\pm.06}$	$0.91_{\pm.05}$	$0.92_{\pm.04}$	$0.89_{\pm.04}$	$0.71_{\pm.02}$	$0.71_{\pm.04}$	$0.35_{\pm.01}$	$4.71_{\pm.23}$
+SDD	$0.69_{\pm.07}$	$0.72_{\pm.08}$	$0.89_{\pm.07}$	$0.92_{\pm.02}$	$0.93_{\pm.07}$	$0.85_{\pm.07}$	$0.78_{\pm.06}$	$0.34_{\pm.00}$	$5.07_{\pm.38}$
+AFINN	$0.48_{\pm.04}$	$0.62_{\pm.02}$	$0.71_{\pm.05}$	$0.78_{\pm.04}$	$0.79_{\pm.03}$	$0.70_{\pm.03}$	$0.74_{\pm.03}$	$0.34_{\pm.00}$	$4.27_{\pm.22}$
+NRC	$0.60_{\pm.05}$	$0.68_{\pm.03}$	$0.71_{\pm.05}$	$0.78_{\pm.04}$	$0.80_{\pm.08}$	$0.74_{\pm.02}$	$0.71_{\pm.05}$	$0.34_{\pm.00}$	$4.35_{\pm.26}$
+ALL	$\mathbf{0.44_{\pm.06}}$	$\mathbf{0.55_{\pm.04}}$	$\mathbf{0.63_{\pm.06}}$	$\mathbf{0.72_{\pm.07}}$	$\mathbf{0.69_{\pm.03}}$	$\mathbf{0.67_{\pm.04}}$	$\mathbf{0.67_{\pm.03}}$	$0.34_{\pm.00}$	$\mathbf{3.59_{\pm.31}}$
SOTA	$0.53_{\pm.05}$	$\mathbf{0.55_{\pm.03}}$	$0.75_{\pm.07}$	$\mathbf{0.64_{\pm.03}}$	$0.81_{\pm.05}$	$\mathbf{0.62_{\pm.02}}$	$0.83_{\pm.04}$	$0.44_{\pm.02}$	$3.78_{\pm.13}$

Table 3: Results for the DAIC-WOZ test set. The mean MAE and standard deviation are reported for five runs. The best MAE for each symptom is **in bold**. SOTA means current state-of-the-art results in the literature (Milintsevich et al., 2023).

All utterance representations are then concatenated into one sequence, later processed by an utterance-level encoder. In the end, the classification head produces a real number in the range from 0 to 3 for each symptom. Several changes are made to the original architecture to gain training efficiency. First, the BiLSTM utterance-level encoder is replaced with a randomly initialized 4-layer 12-head transformer encoder. Second, we change the way the input data is represented. In the original model, each utterance of the interview is encoded separately by a word-level encoder. This is far from optimal since most of the utterances are short (<10 tokens), thus, a lot of computation is wasted on padding tokens. Instead, the utterances are concatenated into one input text separated by the [SEP] special token. This way, the number of passes through the encoder is reduced from the number of utterances K to $\bar{K}$, defined as in Equation 1, where $|U_i|$ is the number of tokens in an utterance and m is the maximum input length of the word-level encoder.

$$\bar{K} = \left\lceil \frac{\sum(|U_i| + 1)}{m} \right\rceil \tag{1}$$

In practice, it reduces the number of word-level encoder passes by $\sim$40 times for each input. After, we perform the *Mean* [SEP] *pooling* on the tokens representing each utterance to get the final utterance representation. The overview of the model architecture is presented in Figure 1.

Lexicons. To incorporate the external knowledge into the model, we use three lexicons: AFINN (Nielsen, 2011), NRC (Mohammad and Turney, 2013), and SDD (Yazdavar et al., 2017). AFINN is a sentiment lexicon that includes a list of 2,477 terms manually rated for the sentiment valence with a value between -5 (negative) and $+5$ (positive). Nielsen (2011) used Twitter postings together with different word lists as a source for the lexicon. NRC is a word-emotion association lexicon that is a list of 14,182 words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). Mohammad and Turney (2013) compiled terms from Macquarie Thesaurus (Bernard, 1986), WordNet Affect Lexicon (Strapparava and Valitutti, 2004), and General Inquirer (Stone et al., 1966) and labeled them with the help of crowd-sourced workers. SDD is a part of the Social-media Depression Detector and is a lexicon of more than 1,620 depression-related words and phrases created in collaboration with a psychologist clinician.

Input marking. In particular, we employ the technique proposed by Zhou and Chen (Zhou and Chen, 2022) to identify and annotate the lexicon words in the input text. It involves marking a lexicon word using the "@" token on either side (see Table 1 for examples). We chose the "@" token for marking since it is not present in the data but included in the model's vocabulary. This way, the pretrained model's architecture remains unchanged[1]. The proportion of marked words within the DAIC-WOZ is illustrated in Table 2, where the statistical test is Student's t-test with p-value < 0.05.

[1] Typed marking strategies that include emotion and sentiment values have also been tested and provided no additional insights compared to the simple input marking.

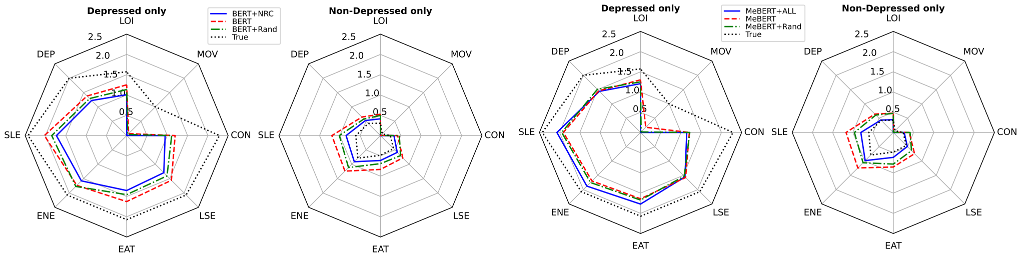


Figure 2: Average predicted values for depressed and non-depressed patients of the DAIC-WOZ test set.

Model	LOI	DEP	SLE	ENE	EAT	LSE	CON	MOV	SUI
BERT	$\mathbf{0.59_{\pm.03}}$	$\mathbf{0.65_{\pm.03}}$	$0.81_{\pm.01}$	$0.62_{\pm.02}$	$0.75_{\pm.06}$	$0.60_{\pm.02}$	$\mathbf{0.65_{\pm.01}}$	$0.81_{\pm.01}$	$0.82_{\pm.01}$
+SDD	$0.58_{\pm.03}$	$0.62_{\pm.02}$	$0.81_{\pm.01}$	$\mathbf{0.64_{\pm.03}}$	$0.74_{\pm.03}$	$\mathbf{0.63_{\pm.03}}$	$0.63_{\pm.03}$	$\mathbf{0.82_{\pm.02}}$	$0.82_{\pm.01}$
+AFINN	$0.57_{\pm.03}$	$0.60_{\pm.03}$	$0.80_{\pm.02}$	$0.62_{\pm.02}$	$0.76_{\pm.02}$	$0.59_{\pm.03}$	$0.64_{\pm.01}$	$0.81_{\pm.02}$	$\mathbf{0.83_{\pm.01}}$
+NRC	$0.55_{\pm.04}$	$0.62_{\pm.04}$	$\mathbf{0.82_{\pm.01}}$	$0.60_{\pm.02}$	$0.79_{\pm.04}$	$0.59_{\pm.03}$	$0.61_{\pm.04}$	$0.80_{\pm.01}$	$0.82_{\pm.02}$
+ALL	$0.56_{\pm.05}$	$0.63_{\pm.02}$	$0.79_{\pm.02}$	$0.61_{\pm.02}$	$\mathbf{0.80_{\pm.02}}$	$0.58_{\pm.03}$	$0.61_{\pm.01}$	$\mathbf{0.82_{\pm.01}}$	$0.82_{\pm.02}$
MEBERT	$\mathbf{0.58_{\pm.03}}$	$0.58_{\pm.02}$	$0.82_{\pm.02}$	$0.62_{\pm.01}$	$0.78_{\pm.03}$	$0.60_{\pm.04}$	$0.62_{\pm.03}$	$\mathbf{0.82_{\pm.01}}$	$0.84_{\pm.01}$
+SDD	$0.53_{\pm.04}$	$\mathbf{0.60_{\pm.02}}$	$\mathbf{0.83_{\pm.01}}$	$0.62_{\pm.02}$	$0.79_{\pm.01}$	$0.60_{\pm.02}$	$0.61_{\pm.03}$	$0.81_{\pm.02}$	$\mathbf{0.86_{\pm.01}}$
+AFINN	$0.57_{\pm.03}$	$0.55_{\pm.04}$	$\mathbf{0.83_{\pm.01}}$	$0.62_{\pm.02}$	$0.79_{\pm.01}$	$\mathbf{0.63_{\pm.02}}$	$0.58_{\pm.02}$	$0.81_{\pm.02}$	$0.85_{\pm.02}$
+NRC	$0.57_{\pm.03}$	$0.58_{\pm.03}$	$0.82_{\pm.02}$	$\mathbf{0.63_{\pm.03}}$	$0.79_{\pm.02}$	$\mathbf{0.63_{\pm.01}}$	$0.61_{\pm.03}$	$0.80_{\pm.02}$	$0.85_{\pm.01}$
+ALL	$0.56_{\pm.03}$	$0.59_{\pm.04}$	$0.80_{\pm.02}$	$0.62_{\pm.02}$	$\mathbf{0.80_{\pm.02}}$	$0.61_{\pm.01}$	$\mathbf{0.63_{\pm.02}}$	$\mathbf{0.82_{\pm.02}}$	$0.84_{\pm.01}$

Table 4: Results for the PRIMATE test set. The mean macro-F1 score is reported for five runs. The best macro-F1 for each symptom is **in bold**. As standard splits are not provided, we cannot present SOTA results. As standard splits are not provided, we cannot present SOTA results.

Experimental setup. We used two pre-trained models in the word-level encoder of our architecture: BERT-Base model (Devlin et al., 2018) and MentalBERT (Ji et al., 2022). We refer to them as **BERT** and **MeBERT** further on. Both models share the same architecture; however, BERT was pre-trained on general domain data, while MeBERT used mental health-related data, mostly based on Reddit. Each model is finetuned with the same hyperparameters (mostly following Mosbach et al., 2020) and different input markings. For example, the BERT+SDD model uses BERT as a pre-trained model and SDD lexicon for input marking. +ALL models use a union of all three lexicons. All models are trained with a mini-batch size of 16, PyTorch realization of AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of $2 \cdot 10^{-5}$ and linear scheduler with a warm-up ratio of 0.1. For the word-level PLMs, only their attention layers are finetuned. The utterance-level encoder is randomly initialized based on the transformer encoder architecture with the following hyperparameters: 4 layers, 12 attention heads, hidden dimensions of encoder and pooler layers of 768, intermediate hidden dimension of 1536. The rest of the hyperparameters follow the default `BertConfig` from the HuggingFace Transformers library (Wolf et al., 2020). For the DAIC-WOZ dataset, results are evaluated with micro-averaged mean absolute error (MAE). Symptom-based errors are calculated for each symptom individually. PHQ-8 score is obtained by summing the eight symptom scores, and MAE for PHQ-8 is calculated on this summation. We evaluate results on the PRIMATE dataset with a macro-averaged F1 score.

3 Results and Discussion

Table 3 shows the results for the DAIC-WOZ test set. For the BERT model, the lexicon-based input marking brings slight overall improvement when AFINN or NRC lexicons are introduced. Most notably, the NRC input marking shows improved or equal MAE for all symptom scores except `DEP`. The combination of all lexicons is marginally beneficial overall, and results have deteriorated with the exclusive introduction of the SDD lexicon. On the other hand, for the MeBERT model, the combination of all the lexicons produces the best results overall, both symptom-wise and for the global PHQ-8 score. Furthermore, both AFINN and NRC lexicons improve the prediction for the MeBERT model, similar to the BERT model. Also, when only the SDD lexicon is used for input marking, the model shows worse performance than the baseline setting.

Figure 2 depicts a more detailed overview of the best-performing models: BERT+NRC and MeBERT+ALL. Additionally, we finetune the +Rand version of both BERT and MeBERT to verify if the improvement comes only from the input marking by randomly marking 8% of the words in each interview. From the results, the improvement for the BERT+NRC model comes from the non-depressed population. MeBERT+All model, however, improves for both depressed and non-depressed populations and is less sensitive to the marking bias. Interestingly, +Rand models show some improvement for the non-depressed population, suggesting that input markings alone act as a regularizer.

Table 4 shows the results for the PRIMATE test set. Contrary to the results from Table 3, introducing external knowledge does not clearly improve performances. The models that use the lexicon input marking show signs of improvement for some symptoms, but it is largely inconsistent. Unlike for the DAIC-WOZ, the SDD-based input marking provides the best F1 score for three symptoms, both for BERT and MentalBERT models, while the benefits of AFINN and NRC are limited or absent and spread over symptoms.

The results from the DAIC-WOZ show that PLMs can indeed benefit from the introduction of external knowledge about the sentiment and emotional value of the words. Surprisingly, the introduction of the depression-specific lexicon had the opposite effect. We hypothesize that two reasons could cause it. First, as seen in Table 2, SDD covers less than 0.5% of words in the interview, almost 15 times less than AFINN and NRC. Thus, the introduced signal might be too weak for the model to learn. Second, the SDD lexicon was based on Twitter data, while DAIC-WOZ contains transcripts of real conversations. From our observations, the people describe their problems more explicitly in their social media posts. At the same time, DAIC-WOZ conversations are more generally themed, and the PHQ-8 scores are based on the person's self-assessment test rather than the conversations themselves. This brings us back to the conceptual difference between the DAIC-WOZ and PRIMATE datasets. While the first one aims at establishing the link between the underlying person's mental condition and their speech, the latter one sets a goal of detecting whether a particular symptom is mentioned in the text. In addition, the PRIMATE dataset is annotated by layman crowd workers, and the labels are not consistent and contain inevitable mistakes (Milintsevich et al., 2024). This might explain the reason behind the greater impact of the AFINN and NRC lexicons for modeling the DAIC-WOZ dataset.

4 Conclusion

This paper targets lexicon incorporation in transformer-based models for symptom-based depression estimation. The external information is supplied through a marking strategy, which avoids any modification to the model's architecture. The set of endeavoured experiments shows that introducing sentimental, emotional and/or domain-specific lexicons can correlate with overall performance improvement if adapted to the targeted task[2].

Limitations

The main limitation in automated clinical mental health assessment with natural language processing is the difficulty of acquiring and accessing large quantities of data. DAIC-WOZ and PRIMATE are rare exceptions as it is publicly available and clinically verified. However, DAIC-WOZ, in particular, suffers from a small number of data points that makes it hard to train and validate hypotheses, as both validation and test sets are particularly small. As a consequence, this piece of research requires further validation on a larger body of clinical data.

Ethical Considerations

We acknowledge the potential ethical aspects of the work that studies the methods to unobtrusively detect someone's mental health status. Here, we are using publicly available datasets collected for research purposes. Also, the lexicons we use are publicly available and have not been composed based on private confidential material. If such a system that could predict the presence of depression symptoms based on actual clinical interviews would be deployed in practice, it would require the informed consent of all participants involved

[2] Source code is available here: https://github.com/501Good/dialogue-classifier.

as well as the understanding of the validity boundaries of such systems, meaning that the predictions of such systems cannot replace the assessment of trained clinicians, but rather assist them in their activities.

Acknowledgements

This research was supported by the Estonian Research Council Grant PSG721 and the FHU A^2M^2P project funded by the G4 University Hospitals of Amiens, Caen, Lille and Rouen (France). The calculations for model's training and inference were carried out in the High Performance Computing Center of the University of Tartu (University of Tartu, 2018).

References

Navneet Agarwal, Gaël Dias, and Sonia Dollfus. 2022. Agent-based splitting of patient-therapist interviews for depression estimation. In *PAI4MH @ 36th Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, USA.

Jiangang Bai, Yujing Wang, Hong Sun, Ruonan Wu, Tianmeng Yang, Pengfei Tang, Defu Cao, Mingliang Zhang1, Yunhai Tong, Yaming Yang, Jing Bai, Ruofei Zhang, Hao Sun, and Wei Shen. 2022. Enhancing self-attention with knowledge-assisted attention maps. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 107–115, Seattle, United States. Association for Computational Linguistics.

Aaron T Beck, Robert A Steer, and Margery G Carbin. 1988. Psychometric properties of the beck depression inventory: Twenty-five years of evaluation. *Clinical psychology review*, 8(1):77–100.

J.R.L. Bernard. 1986. *The MacQuarrie Thesaurus: The Book of Words*. Macquarie Library.

Sergio Burdisso, Esaú Villatoro-Tello, Srikanth Madikeri, and Petr Motlicek. 2023. Node-weighted graph convolutional network for depression detection in transcribed clinical interviews. In *INTERSPEECH*.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Shrey Gupta, Anmol Agarwal, Manas Gaur, Kaushik Roy, Vignesh Narayanan, Ponnurangam Kumaraguru, and Amit Sheth. 2022. Learning to automate follow-up question generation using process knowledge for depression triage on Reddit posts. In *Eighth Workshop on Computational Linguistics and Clinical Psychology (CLPSY)*, pages 137–147, Seattle, USA. Association for Computational Linguistics.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. Mental-BERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of LREC*.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B.W. Williams, Joyce T. Berry, and Ali H. Mokdad. 2009. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1-3):163–173.

Mingzheng Li, Xiao Sun, and Meng Wang. 2023. Detecting depression with heterogeneous graph neural network in clinical interview transcript. *IEEE Transactions on Computational Social Systems*.

Zhongli Li, Qingyu Zhou, Chao Li, Ke Xu, and Yunbo Cao. 2021. Improving BERT with syntax-aware local attention. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP)*, pages 645–653. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Kirill Milintsevich, Kairit Sirts, and Gaël Dias. 2023. Towards automatic text-based estimation of depression through symptom prediction. *Brain Informatics*, 10(1):1–14.

Kirill Milintsevich, Kairit Sirts, and Gaël Dias. 2024. Your model is not predicting depression well and that is why: A case study of PRIMATE dataset. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 166–171, St. Julians, Malta. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Syed Arbaaz Qureshi, Gael Dias, Mohammed Hasanuzzaman, and Sriparna Saha. 2020. Improving depression level estimation by concurrently learning emotion intensity. *IEEE Computational Intelligence Magazine*, 15(3):47–59.

Syed Arbaaz Qureshi, Sriparna Saha, Mohammed Hasanuzzaman, and Gaël Dias. 2019. Multitask representation learning for multimodal estimation of depression level. *IEEE Intelligent Systems*, 34(5):45–52.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

Anbu Savekar, Shashikanta Tarai, and Moksha Singh. 2023. Structural and functional markers of language signify the symptomatic effect of depression: A systematic literature review. *European Journal of Applied Linguistics*, 11(1):190–224.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

University of Tartu. 2018. UT rocket.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP)*, pages 1405–1418. Association for Computational Linguistics.

Shanshan Wang, Zhumin Chen, Zhaochun Ren, Huasheng Liang, Qiang Yan, and Pengjie Ren. 2022. Paying more attention to self-attention: Improving pre-trained language models via attention guiding. *arXiv preprint arXiv:2204.02922*.

Ping-Cheng Wei, Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. 2022. Multi-modal depression estimation based on sub-attentional fusion. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 623–639.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 38–45, Online. Association for Computational Linguistics.

Danai Xezonaki, Georgios Paraskevopoulos, Alexandros Potamianos, and Shrikanth Narayanan. 2020. Affective Conditioning on Hierarchical Attention Networks Applied to Depression Detection from Transcribed Clinical Interviews. In *INTERSPEECH*, pages 4556–4560.

Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. 2020. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. In *28th International Conference on Computational Linguistics (COLING)*, pages 696–709, Barcelona, Spain.

Xiaoxu Yao, Guang Yu, Jingyun Tang, and Jialing Zhang. 2021. Extracting depressive symptoms and their associations from an online depression community. *Computers in human behavior*, 120:106734.

Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 1191–1198.

Tianlin Zhang, Kailai Yang, Hassan Alhuzali, Boyang Liu, and Sophia Ananiadou. 2023a. Phq-aware depressive symptoms identification with similarity contrastive learning on social media. *Information Processing & Management*, 60(5):103417.

Tianlin Zhang, Kailai Yang, Shaoxiong Ji, and Sophia Ananiadou. 2023b. Emotion fusion for mental illness detection from social media: A survey. *Information Fusion*, 92:231–246.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 50–61. Association for Computational Linguistics.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In *2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*, pages 161–168. Association for Computational Linguistics.

III

N. Agarwal, K. Milintsevich, L. Métivier, M. Rothärmel, G. Dias, S. Dollfus
Analyzing Symptom-based Depression Level Estimation through the Prism of Psychiatric Expertise
Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 974–983

Analyzing Symptom-based Depression Level Estimation through the Prism of Psychiatric Expertise

Navneet Agarwal*[1], Kirill Milintsevich*[1,2], Lucie Metivier[3], Maud Rotharmel[4], Gaël Dias[1], Sonia Dollfus[5]
[1]Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France.
[2]Institute of Computer Science, University of Tartu, Tartu, Estonia.
[3]Normandie Univ, UNICAEN, PhIND, UMR-S 1237, GIP CYCERON, 14000 Caen, France
[4]Service hospitalo-universitaire de psychiatrie de l'adulte, Centre Thérapeutique d'Excellence,
Centre Hospitalier du Rouvray, Sotteville-les-Rouen, France.
[5]CHU de Caen, Service de Psychiatrie; Normandie Univ, UNICAEN, ISTS, GIP Cyceron;
Normandie Univ, UNICAEN, UFR de Médecine, 14000 Caen, France.
{navneet.agarwal,kirill.milintsevich}@unicaen.fr

Abstract

The ever-growing number of people suffering from mental distress has motivated significant research initiatives towards automated depression estimation. Despite the multidisciplinary nature of the task, very few of these approaches include medical professionals in their research process, thus ignoring a vital source of domain knowledge. In this paper, we propose to bring the domain experts back into the loop and incorporate their knowledge within the gold-standard DAIC-WOZ dataset. In particular, we define a novel transformer-based architecture and analyze its performance in light of our expert annotations. Overall findings demonstrate a strong correlation between the psychological tendencies of medical professionals and the behavior of the proposed model, which additionally provides new state-of-the-art results.

Keywords: Depression estimation, psychiatrist annotations, external knowledge introduction.

1. Introduction

Mental illness is a serious issue with high social and economic costs, yet a significant number of mental illness cases go undetected. Up to half of the patients with psychiatric disorders are not diagnosed as having mental illness by their primary care physicians (Higgins, 1994), a situation made worse due to a shortage of medical professionals (Butryn et al., 2017). As a consequence, artificial intelligence in psychiatry has been emerging as a general term that implies the use of computerized techniques and algorithms for the diagnosis, prevention, and treatment of mental illnesses (Fakhoury, 2019). Within clinical settings, semi-structured interviews are the common practice for evaluating a person's mental health. These interviews usually act as inputs for training automated models with self-assessment scores being used as the final ground truth (e.g. Patient Health Questionnaire PHQ-8 for depression estimation). Throughout the literature, different strategies have been proposed for the automated estimation of depression. Multimodal models combine inputs from different modalities (Ray et al., 2019; Qureshi et al., 2019). Multitask architectures simultaneously learn related tasks (Qureshi et al., 2019, 2020). Gender-aware models explore the impact of gender on depression estimation (Bailey and Plumbley, 2021; Qureshi et al., 2021). Hierarchical models process transcripts at different granularity levels (Mallol-Ragolta et al., 2019; Xezonaki et al., 2020). Attention models integrate external knowledge from lexicons (Xezonaki et al., 2020). Feature-based strategies compute multimodal characteristics (Dai et al., 2021). Graph-based systems aim to study complex structures within interview transcripts (Hong et al., 2022; Niu et al., 2021). Multiview architectures treat the input transcripts as a combination of different text views (Agarwal et al., 2022). Symptom-based models treat depression estimation as an extension of the symptom prediction problem (Milintsevich et al., 2023). Domain-specific language models are built (Ji et al., 2022) and large language models are prefix-tuned to automate depression level estimation (Lau et al., 2023).

Despite the multidisciplinary nature of the problem, most previous research initiatives have failed to include medical professionals in the learning process, except Yadav et al. (2020), who asked a psychiatrist to label tweets in terms of PHQ-9 symptoms. In this paper, we propose to follow this line of research by providing a clinically annotated version of the gold-standard DAIC-WOZ dataset[1] (Gratch et al., 2014) to allow the integration of domain expertise in artificial models. We also define a novel transformer-based model and examine ways to utilize psychiatric annotations within its learning

[*] These authors contributed equally to this work

[1] The Distress Analysis Interview Corpus (DAIC) is the only publicly available resource for interview-based distress analysis.

process. Finally, we analogize the psychological tendencies of medical professionals against the proposed model in an attempt to validate its reliability as a predictive model in clinical settings. Overall results show that our model successfully aligns with medical experts thus being a trustful source of predictions for clinicians in psychiatry. Additionally, the proposed model provides new state-of-the-art results over the DAIC-WOZ test set.

2. Related Work

Different architectures and strategies have been used throughout the literature to build models capable of estimating patients' depression level based on patient-therapist interviews. One promising research area is to leverage inputs from different modalities into one learning modal. Qureshi et al. (2019) explore the possibility of combining audio, visual, and textual input features into a single architecture using attention fusion networks. They further show that training the model for regression and classification simultaneously on the same dataset provides improvements in results. Ray et al. (2019) present a similar framework that invokes attention mechanisms at several layers to identify and extract important features from different modalities. The network uses several low-level and mid-level features from audio, visual and textual modalities of the participants' inputs. Another interesting approach aims at combining different tasks that share some common traits thus following the multi-task paradigm. Qureshi et al. (2020) propose to simultaneously learn both depression level estimation and emotion recognition on the basis that depression is a disorder of impaired emotion regulation. They show that this combination provides improvements in performance for the multiclass problem as well as the regression of the PHQ-8 score. Building on the success of hierarchical models for document classification, different studies (Mallol-Ragolta et al., 2019; Xezonaki et al., 2020) propose to encode patient-therapist interviews with hierarchical structures, showing boosts in performance. Xezonaki et al. (2020) further extend their proposal and integrate affective information (emotion, sentiment, valence, and psycho-linguistic annotations) from existing lexicons in the form of specific embeddings. Exploring a different research direction, Qureshi et al. (2021) study the impact of gender on depression level estimation and build four different gender-aware models that show steady improvements over gender-agnostic models. In particular, an adversarial multi-task architecture provides the best results overall. Along the same line, Bailey and Plumbley (2021) study gender bias from audio features as compared to (Qureshi et al., 2021), who target textual information. They find that deep learning models based on raw audio are more robust to gender bias than ones based on other common hand-crafted features, such as mel-spectrogram. Although most strategies rely on deep learning architectures, a different research direction is proposed by Dai et al. (2021), who build a topic-wise feature vector based on a context-aware analysis over different modalities (audio, video, and text). Niu et al. (2021) use graph structures within their architecture to grasp relational contextual information from audio and text modality. They propose a hierarchical context-aware model to capture and integrate contextual information among relational interview questions at word and question-answer pair levels. Milintsevich et al. (2023) treat binary classification as a symptom profile prediction problem and train a multi-target hierarchical regression model to predict individual depression symptoms from patient-therapist interview transcripts. Agarwal et al. (2022) highlight the importance of retaining discourse structure and define multi-view architectures that divide the input transcript into views based on sentence identities. The two views are processed both independently and co-dependently in order to account for intra-view and inter-view interactions. Building upon the success of language models in understanding textual data, Ji et al. (2022) fine-tune different BERT-based models on mental health data and provide a pre-trained masked language model for generating domain-specific text representations. Lau et al. (2023) further account for the lack of large-scale high-quality datasets in the mental health domain and propose the use of prefix-tuning as a parameter-efficient way of fine-tuning language models for mental health.

The gathering and assimilation of external knowledge into neural networks have garnered substantial attention in research endeavors in the domain of mental health. For the former case, Arseniev-Koehler et al. (2018) asked crowd workers to read excerpts of de-identified interview data from the DAIC-WOZ and rate how likely they thought a speaker had depression based on the transcribed utterances. Similarly, Yadav et al. (2020) work with Twitter data and employ four native English speakers from multiple disciplines to independently annotate tweets into the 9 categories of PHQ-9. For the latter case, various strategies have been proposed for the integration of external knowledge into neural network training. Outside the mental health domain, Soares et al. (2019) and Boualili et al. (2020) use special tokens to highlight information directly within the input text and rely on fine-tuning pre-trained language models to understand the importance of marked text. Deshpande and Narasimhan (2020), (Stacey et al., 2022) and Wang et al. (2022) introduce additional loss terms during training as a means to guide the attention mechanism within the

Depression severity	Data split		
	Train	Val.	Test
No symptoms $[0..4]$	47	17	22
Mild $[5..9]$	29	6	11
Non-depressed Total	76	23	33
Moderate $[10..14]$	20	5	5
Moderately severe $[15..19]$	7	6	7
Severe $[20..24]$	4	1	2
Depressed Total	31	12	14
Total	107	35	47

Table 1: Number of interviews for each depressive class severity in the DAIC-WOZ dataset, distributed by train, validation and test sets.

neural networks towards the desired distributions. Within the mental health domain, only Xezonaki et al. (2020) generate custom context vectors using information from different lexicons, which are concatenated to word level representations.

3. Dataset and Psychiatric Annotations

3.1. Dataset

The Distress Analysis Interview Corpus (DAIC) is a multimodal corpus of semi-structured clinical interviews designed to simulate standard protocols for identifying people at risk of depression. Within our research, we focus on the textual input from the publicly available Wizard-of-Oz part of the corpus (DAIC-WOZ), which contains 189 interviews, where patients interact with an animated virtual agent controlled by a human therapist from a different room. Each session ranges from 7 to 33 minutes with an average time of 16 minutes. The dataset contains valuations for eight specific symptoms that are part of the PHQ-8 questionnaire: loss of interest, feeling of depression, sleeping habits, tiredness, loss of appetite, feeling of failure, lack of concentration and lack of movement. Table 1 shows the data splits between train, development and test sets, along with the class imbalance within the DAIC-WOZ dataset.

3.2. Psychiatrist Annotations

In our attempt to reintroduce domain expertise into the learning process, we carried out the clinical annotation of the DAIC-WOZ dataset[2]. In contrast to previous works that use crowd workers (Arseniev-Koehler et al., 2018) or native English speakers (Yadav et al., 2020) as annotators, we select mental health professionals for the annotation process. In particular, three psychiatrists from public hospitals were employed to undertake two major tasks: (1) span-based annotation of the transcripts and (2) PHQ-8 scoring based on interview transcripts.

Span-based annotation: This task consists of highlighting information within transcripts that influences a psychiatrist's decision during an interview. Since it is a subjective task that lacks a definitive right or wrong answer, a common consensus on the importance of various utterances within the transcripts does not exist. Even within the field of medicine, professionals do not universally agree on the significance of various pieces of information, and subtle differences in opinion exist between psychiatrists based on their individual knowledge and experience. As such, after various meetings and discussions with the psychiatrists, it was agreed that the medical annotators should have complete freedom to annotate the transcripts without any constraints in order to capture their true judgment. As a consequence, we forgo defining detailed annotation protocols and rely on the annotator's judgment as experts in the field for the reliability of their annotations. However, they were encouraged not only to identify information that suggests the presence of depression, but also to pinpoint clues that indicate its absence. Furthermore, the inherent lack of consensus within the task eliminates the need for inter-annotator agreements. In case multiple annotators are assigned per transcript, a simple union of annotated spans would be used to capture knowledge from all assigned annotators. Unfortunately, at this stage of our research, only one annotator per transcript could be assigned due to the workload experienced by the annotators, particularly due to the radical increase of mental care demand after the covid pandemic coupled with the shortage of mental health professionals. The current annotation process lasted nearly 5 months and we anticipate this time frame to scale linearly with the increase in the number of annotators per transcript.

For the annotation purpose, we designed an online tool based on the doccano[3] project which was hosted on servers from the herokou platform[4] enabling the entire annotation process to take place remotely for the convenience of the psychiatrists. The tool was designed to allow the psychiatrists to annotate any span of text (word, phrase, sentence, text) within the transcript and assign a label of importance to each span: highly important, important (default) or minimally important. Upon analysis, it was found that these labels did not provide any information since more than 99% of the spans were marked with the default label (important), and were therefore not used in any further analysis. The annotation process gave rise to an

[2] The annotations can be accessed at https://github.com/navneet-agarwal/DAIC-WOZ-Annotations

[3] https://github.com/doccano/doccano

[4] https://www.heroku.com/

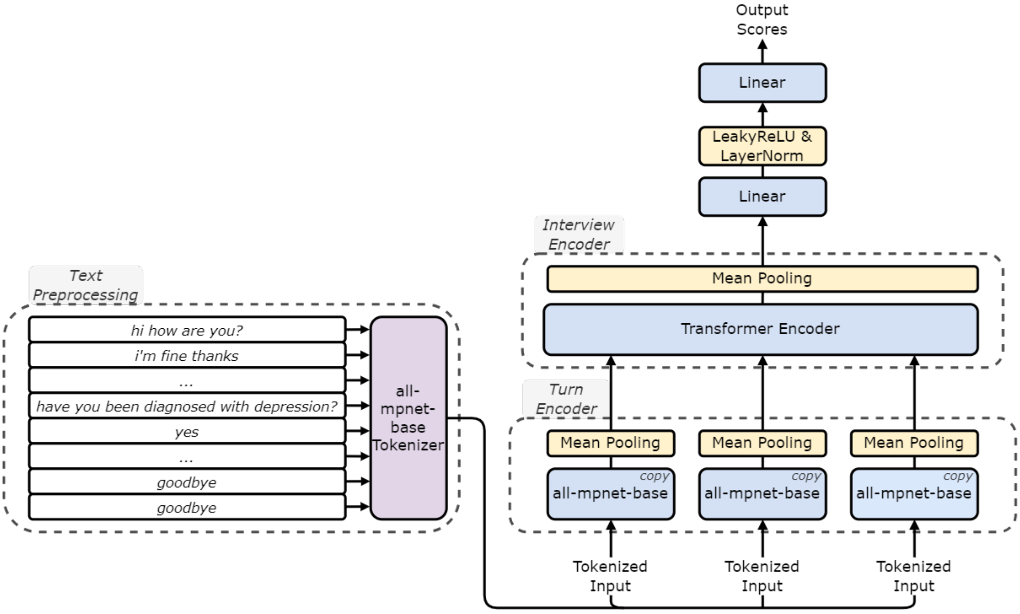


Figure 1: Hierarchical neural architecture for symptom-based prediction.

Span Level	Non-Depressed	Depressed
Word	467 (3.53)	227 (3.98)
Phrase	4101 (31.06)	1913 (33.56)
Sentence	0	0
Multi-sentences	77 (0.58)	42 (0.73)
Total	4645 (35.18)	2182 (38.28)

Table 2: Number of annotations for different levels of annotation spans. Figures in bracket indicate the average number of annotations per transcript.

average of 36.12 annotations per transcript (35.18 for the non-depressed class and 38.28 for the depressed class) with a mean length of 7.45 words (7.74 for the non-depressed class and 7.17 for the depressed class). The distribution of the annotations by patient class and span level is given in Table 2. Interestingly, complete sentences were not annotated by any of the psychiatrists, who mostly followed a ngram-based strategy, with a small number of annotations focusing on multiple sentences. Furthermore, none of the psychiatrists highlighted questions within the dataset with all the annotations contained within patient responses.

PHQ-8 scoring: This task involves completing the self-assessment PHQ-8 questionnaire on behalf of each patient only based on their interview transcripts. Although the PHQ-8 screening tool is widely used as a measure of depression and has been found to be precise (Shin et al., 2019), it relies on the subjective assessment of the patient about his/her condition outside the context of the interview. As such, an interview transcript might not contain enough information to accurately express the intensity of individual symptoms. Furthermore, since the interviews are conducted with the aim of depression estimation and not specifically for fulfilling the PHQ-8 questionnaire, information on some symptoms might be missing altogether within individual transcripts depending on the questions asked during the interview. In order to verify these propositions, we asked the clinicians to fulfill the PHQ-8 questionnaires on behalf of each patient based on their understanding of the given transcripts. This task consists of evaluating each of the 8 symptoms within the PHQ-8 questionnaire on a Likert scale ranging from 0 to 3. The statistics about this task, illustrated in Table 3, show that 5 out of 8 symptoms (i.e. loss of interest, feeling of depression, sleeping habits, feeling of tiredness, and feeling of failure) are steadily mentioned in most transcripts, while 3 of them (i.e. loss of appetite, lack of concentration and lack of movement) could not be measured reliably by the psychiatrists. This confirms our claims regarding the lack of symptom-level information within individual interviews. This annotation task also acts as a human expert performance baseline, that defines an achievable learning goal for correctly inferring PHQ-8 scores for each symptom based on information present within the transcripts.

Symptoms	No interest	Depressed	Sleep	Tired	Appetite	Failure	Concentration	Movement
# annotations	178	188	179	160	47	176	48	10

Table 3: Nb. of psychiatrist scorings for each PHQ-8 symptom over the 189 interviews of the DAIC-WOZ.

ELLIE: *how close are you to your family*
PARTICIPANT: *@@ very close @@ even though i don't live with them @@ i try to see them as much as possible @@*
ELLIE: *mhm*
ELLIE: *how do you like your living situation*
PARTICIPANT: *uh it's ok*

Figure 2: Example of annotation marking.

4. Model and Mark-up Strategy

4.1. Neural Network Architecture

To learn the 8 symptom values of the PHQ-8, we design the transformer-based hierarchical model illustrated in Figure 1. The architecture is based on the model defined by Milintsevich et al. (2023), which has been updated to have access to sentence-level attention and take advantage of recent sentence representation models. In particular, the architecture has undergone two significant alterations compared to the definition in §3.2 of (Milintsevich et al., 2023): (1) the BiLSTM cells are replaced by a transformer-based encoder at the interview level (interview encoder), and (2) the pre-trained turn encoder is based on the all-mpnet-base model[5] in place of *S-RoBERTa*[6], both using a contrastive learning objective (Reimers and Gurevych, 2019). In particular, the model consists of two encoders: the turn encoder that encodes each sentence and the interview encoder that encodes sentence level representations into an interview level embedding. The interview level embedding is then passed through a feed-forward network that maps it to a prediction vector $m = [m_1, m_2, ..., m_8]$, where each predicted label $m_k \in [0, 3]$ represents a symptom score for the corresponding question in the PHQ-8 questionnaire. The interview encoder contains 4 layers containing 12 attention heads each with an intermediate size of 1536 and an hidden size of 768. This model acts as the base architecture for the different experiments and model configurations explored within our research and is referred to as the **Baseline model**.

Model	MAE Dev.	MAE Test
SOTA		
ASP MT. DLC+DLR+EIR (Qureshi et al., 2020)		3.69
HCAG-T (Niu et al., 2021)	3.73	-
SGNN (Hong et al., 2022)	3.76	-
Symptom prediction (Milintsevich et al., 2023)	3.61	3.78
Dual encoder (warm start) (Lau et al., 2023)	2.76	3.80
Our Configurations		
Baseline model	4.08	**3.52**
Marked-up model	3.49	3.60

Table 4: Comparison of overall model performance against current state-of-the-art results. The results are averaged over 5 random initializations.

4.2. External Knowledge Integration

In our effort to reintroduce domain expertise into depression estimation tasks, we incorporate psychiatrist annotations into the learning process of our neural network model. We align our work with the research approach taken by Soares et al. (2019) and Boualili et al. (2020), and introduce special markers into the input text to directly highlight clinical annotations within the transcripts. The underlying idea is that explicitly marking spans in the input text may allow the model to carefully identify the annotations and make a more informed prediction. Consequently, all annotations provided by the psychiatrists are encompassed in between the @@ markers within the transcripts, giving rise to a marked-up corpus (example in figure 2). We use the Baseline architecture defined earlier and fine-tune it using the marked-up corpus. Specifically, the pre-trained *all-mpnet-base* model is fine-tuned by unfreezing only the final layer. The resulting model is referred to as the **Marked-up model**.

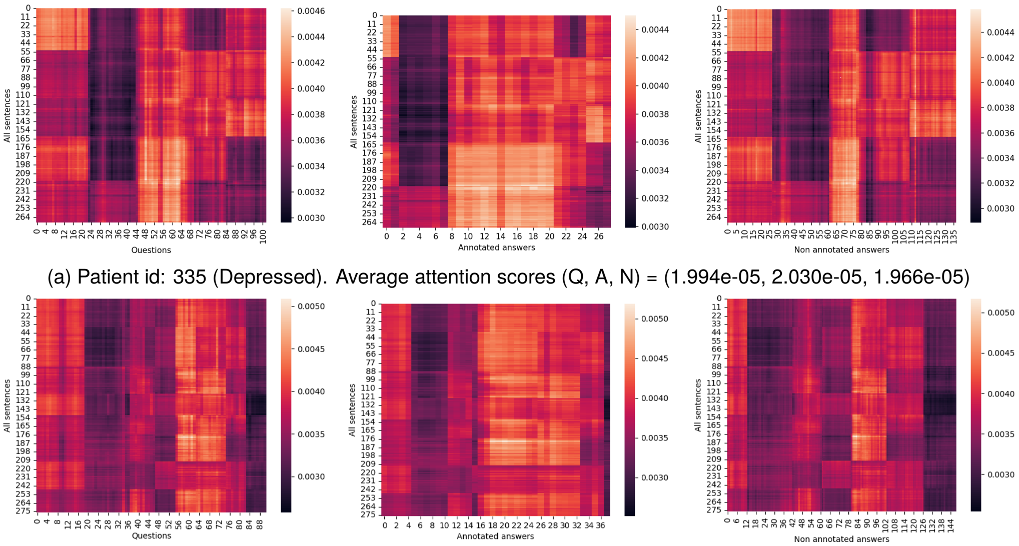
5. Overall Results

Table 4 provides overall results for the various model configurations considered in our experiments and puts them into perspective by comparison against current state-of-the-art results. Our baseline model provides new state-of-the-art performance for the Mean Absolute Error (MAE) metric on the test set of the DAIC-WOZ on an average over 5 runs. It is interesting to notice that the marked-up model does not improve over the baseline model despite containing extra information, although it does outperform all previous research initiatives. This issue is further discussed in detail in §7.

Ablation study: We conduct an ablation study to analyze the amount of information contained within

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[6] https://huggingface.co/sentence-transformers/all-distilroberta-v1



(a) Patient id: 335 (Depressed). Average attention scores (Q, A, N) = (1.994e-05, 2.030e-05, 1.966e-05)

(b) Patient id: 307 (Non-Depressed). Average attention scores (Q, A, N) = (1.955e-05, 2.044e-05, 1.905e-05)

Figure 3: Heat maps of sentence level attention scores from the Baseline model for two different patients.

Ablation	MAE on Test set
Baseline model	**3.52**
Baseline$_{\text{ann.}}$ inference	4.02
Baseline$_{\text{non-ann.}}$ inference	3.84

Table 5: Ablation study with baseline model for exclusively non-annotated and annotated sentences.

the clinical annotations. Given the complete set of information required for estimating depression, we seek to understand the role played by our clinical annotations within this set. For that purpose, we define two new input configurations and use them with the trained baseline model at the inference stage to generate new predictions over the modified inputs. The two versions in this input ablation study are defined as follows:

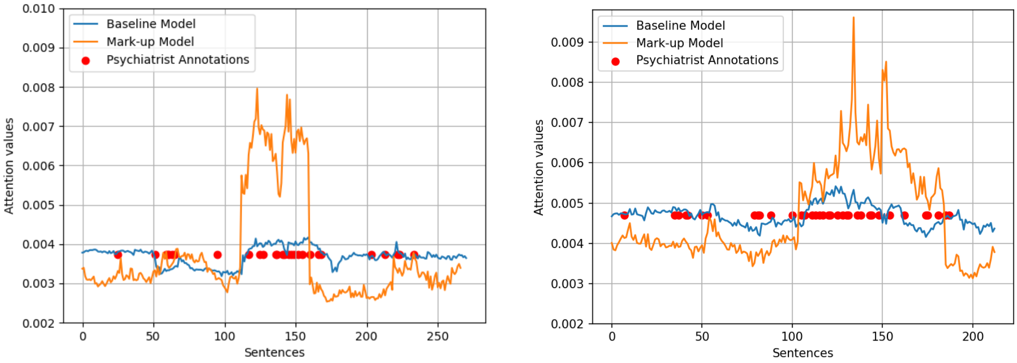
Baseline$_{ann}$ inference: only question-answer pairs with at least one annotation are kept within the input transcripts.

Baseline$_{non-ann}$ inference: only question-answer pairs without any annotation are retained within the input transcripts.

Results of the ablation study are shown in table 5. We see a significant drop in performance on removing annotated question-answer pairs from the input transcripts, highlighting the validity of the psychiatrists' annotations. Surprisingly, we also see a drop in performance when only annotated question-answer pairs are used as inputs. This behavior can be attributed to the fact that in this case the number of sentences within the interviews is severely reduced and as such the coherence of the discourse is undermined, affecting the performance of the automated models.

6. Attention and Annotated Spans

Psychiatrist annotations highlight text spans that hold relevance for depression estimation as per clinicians' knowledge and medical guidelines. Given their importance from the medical point of view, we propose to verify whether automated models attend to the same annotated text spans or look for information that complements clinical knowledge. Psychiatrist annotations are analyzed against sentence-level attention scores from the model, the sentence being the atomic textual element for this analysis. In particular, we focus on 3 different sentence types: questions (Q), non-annotated turns (N) that contain answers without any annotations, and clinically-annotated turns (A) that contain patient responses with at least one annotation. Thus, each attention head $H^{s \times s}$ of the interview encoder is converted into three attention sub-matrices $H^{s \times q}$, $H^{s \times n}$ and $H^{s \times a}$, where s is the number of sentences in a given transcript, q the number of questions, a the number of annotated turns and n the number of non-annotated turns, such that $s = q + n + a$. For each interview, we average the sentence-level attention scores for Q, N and A sentence types for all attention heads contained in the interview encoder as defined in equation 1, where h and l stand for the number of



(a) Patient id 335 (Depressed)

(b) Patient id 307 (Non-Depressed)

Figure 4: Attention scores for the baseline and marked-up models plotted against clinical annotations.

Class	Metric	Q	N	A
Non-depressed	min.	12.84	12.93	13.60
	max.	137.50	136.76	135.35
	med.	42.03	42.10	**42.25**
	avg.	30.85	31.01	**31.25**
Depressed	min.	15.29	15.02	15.37
	max.	103.88	102.83	110.89
	med.	37.96	38.50	**38.82**
	avg.	12.18	12.18	**12.29**

Table 6: Sentence-level attention scores calculated over the DAIC-WOZ dataset for **Q**uestions, **N**on-annotated and **A**nnotated turns. Values are with a precision of 10^{-4}. Med. and avg. stand for median and arithmetic mean.

heads and layers respectively.

$$\overline{X} = \frac{1}{l.h}\sum_{l,h}\frac{1}{i.j}\sum_{i,j} H_{i,j}^{s\times x}, \forall x \in \{q,n,a\} \qquad (1)$$

Finally, we average these values over the 189 interviews of the DAIC-WOZ to get the overall picture. Results with the baseline model are given in Table 6 and show that the transformer-based model focuses more on clinically annotated spans compared to other parts of the transcripts, independently of the patient class. This provides the first evidence that the baseline model targets clinically motivated spans for its decision process without the introduction of any external knowledge or use of specific architectures tuned towards guiding the attention values.

To complement this analysis, figure 3 plots three attention heatmaps $\overline{Q}$, $\overline{A}$ and $\overline{N}$ with brighter regions representing higher attention scores. Plots are provided for a depressed patient as well as a non-depressed patient. This illustration exemplifies overall results and shows that although model attention is distributed over all three categories, clinically-annotated turns receive higher average attention as compared to non-annotated turns and questions. Finally, figure 4 illustrates the attention scores in perspective of the psychiatrists' annotations for the same patients. Following the blue line corresponding to the baseline model, we observe an increase in attention scores in the vicinity of psychiatrist annotations, while the opposite is true in the absence of annotations. These plots represent a general trend observed throughout the dataset with some exceptions.

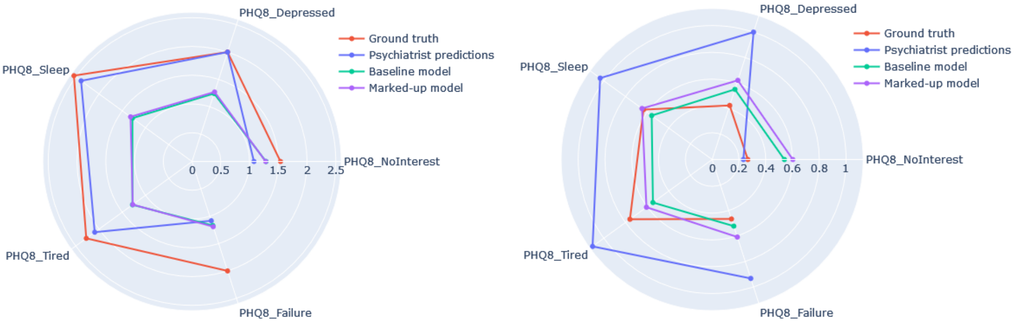
7. Performance Analysis against Knowledge Introduction

Although the baseline model attends to parts of the interviews that psychiatrists find relevant, we explore the impact of the introduction of clinician expertise directly in the learning process and analyze the performance of the marked-up model. Overall results are illustrated in Table 7 and do not evidence gains in performance resulting from the knowledge added by the psychiatrist annotations. Indeed, the baseline model outperforms the marked-up model 5 times out of 8 for both the depressed and non-depressed classes. This confirms our previous findings from section §6, showing that the baseline architecture already attends to clinically annotated sentences, thus reducing the impact of the marked-up strategy. Figure 4 compares both baseline and marked-up models, with plots showing similar behaviors of attending to the annotated sentences although with different amplitude. In particular, the marked-up model tends to pay high attention to the middle of the transcripts thus failing to highlight important information from other regions. This is not the case for the baseline model, which has more evenly distributed attention values, while still being consistent with psychiatrist annotations.

In order to put prediction results into perspective, we calculate the Mean Absolute Error (MAE) between the psychiatrists' PHQ-8 scores and pa-

Symptoms	**Psychiatrist Pred.**		**Baseline model**		**Marked-up model**	
	Depr.	Non-Depr.	Depr.	Non-Depr.	Depr.	Non-Depr.
Loss of interest	0.615	0.366	**0.611**	**0.431**	0.699	0.485
Feeling of depression	0.571	0.696	**0.884**	**0.443**	0.939	0.465
Sleeping habits	0.615	0.533	0.761	**0.691**	**0.651**	0.808
Tiredness	0.727	0.689	**0.797**	0.711	0.812	**0.666**
Feeling of failure	1.083	0.800	0.820	**0.543**	**0.786**	0.573
Lack of concentration	-	-	**1.332**	0.521	1.361	**0.475**
Loss of appetite	-	-	**0.932**	0.745	1.037	**0.628**
Lack of movement	-	-	1.008	**0.105**	**0.964**	0.125

Table 7: MAE calculated against patients' self-assessments scores by symptoms over the DAIC-WOZ test set. Results are averaged over 5 runs for the automated models. Psychiatrist prediction evidences the difference between the patients' assessments and the psychiatrists' ones.



(a) Depressed class

(b) Non-Depressed class

Figure 5: Radar plots showing symptom-wise average scores for the different automated models, the patient self-assessments and the psychiatrists' ratings over the test set of the DAIC-WOZ. Note that only 5 symptoms are illustrated, which refer to the ones that psychiatrists could reliably annotate.

tients' self-assessments. Results in Table 7 show that psychiatrist predictions outperform automated models in most cases, albeit by a small margin for most of the symptoms (feeling of failure being an exception where the baseline model performs better). Further analysis of psychiatrist scoring confirms findings from the medical domain (Domken et al., 1994), showing that clinicians tend to underevaluate the PHQ-8 scores for the depressed class while over-evaluating those for the non-depressed class. Intriguingly, we observe the same behavior for the automated models as illustrated in Table 8. The figures show that both the baseline model and the marked-up model exhibit the same behavior as psychiatrists, which further strengthens our claim of shared psychological tendencies between our proposed model and psychiatrists. As expected, the number of transcripts misdiagnosed by the automated models far exceeds those misdiagnosed by psychiatrists. This is due to the fact that models generate floating point predictions whereas psychiatrists' predictions are based on a Likert scale ranging from 0 to 3.

Symptoms	Depr.		Non-Depr.	
	Over	Under	Over	Under
Psychiatrist Prediction				
Loss of Interest	1	5	3	6
Feeling of depression	3	3	16	2
Sleeping habits	3	3	10	2
Tiredness	2	3	12	5
Feeling of failure	1	8	13	5
Baseline Model				
Loss of Interest	4	9	24	5
Feeling of depression	2	12	24	9
Sleeping habits	1	12	19	10
Tiredness	1	10	14	14
Feeling of failure	1	11	20	9
Marked-up model				
Loss of Interest	4	9	27	3
Feeling of depression	3	11	26	7
Sleeping habits	1	12	19	11
Tiredness	1	10	15	14
Feeling of failure	2	10	23	7

Table 8: Number of over- and under-evaluated transcripts in the test set for the baseline model, the marked-up model and the psychiatrists' scorings.

In order to further analyze the behavior of over and under-evaluation, we plot the symptom-wise average scores for the different automated models, the patient self-assessments and the psychiatrists' ratings in figure 5. The illustrations show a high correlation between the results from the two automated models. Both baseline and marked-up models generate the same average scores for the depressed class, while for the non-depressed class the values are very close. This confirms that the introduction of annotations into the learning process through the markup strategy does not provide significant performance gain. These plots also support the claims of over and under-evaluation of PHQ-8 scores, and showcase a similar pattern as seen in table 8.

8. Conclusion

In this paper, we examine automated depression estimation through the prism of psychiatric expertise and compare the behavior of automated models against clinical annotators. The analysis of sentence-level attention scores shows that the baseline model learns to analyze the transcripts in ways similar to trained psychiatrists despite the lack of medical knowledge in the training process. Our analysis further establishes a strong correlation between the psychological tendencies of our automated model and medical professionals, thus validating its role as a credible source of predictions for clinicians in psychiatry. Additionally, the proposed architecture provides new state-of-the-art results over the DAIC-WOZ test set. The source code and the clinically annotated DAIC-WOZ dataset will be publicly released upon acceptance.

9. Acknowledgements

This research is supported by the FHU A^{2}M^{2}P project funded by the G4 University Hospitals of Amiens, Caen, Lille and Rouen (France). The calculations for model's training and inference were carried out in the High Performance Computing Center of the University of Tartu (University of Tartu, 2018).

10. Bibliographical References

Navneet Agarwal, Gaël Dias, and Sonia Dollfus. 2022. Agent-based splitting of patient-therapist interviews for depression estimation. In *Workshop on Participatory Approach to AI for Mental Health (PAI4MH) associated to 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Alina Arseniev-Koehler, Sharon Mozgai, and Stefan Scherer. 2018. What type of happiness are you looking for? - a closer look at detecting mental health from language. In *5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (CLPSYCH) associated to 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Andrew Bailey and Mark D. Plumbley. 2021. Gender bias in depression detection using audio features. In *29th European Signal Processing Conference (EUSIPCO)*, pages 596–600.

Lila Boualili, José G. Moreno, and Mohand Boughanem. 2020. Markedbert: Integrating traditional IR cues in pre-trained language models for passage retrieval. In *43rd International ACM SIGIR conference on research and development in Information Retrieval (SIGIR)*, pages 1977–1980.

Tracy Butryn, Leah Bryant, Christine Marchionni, and Farhad Sholevar. 2017. The shortage of psychiatrists and other mental health providers: causes, current state, and potential solutions. *International Journal of Academic Medicine*, 3(1):5–9.

Zhijun Dai, Heng Zhou, Qingfang Ba, Yang Zhou, Lifeng Wang, and Guochen Li. 2021. Improving depression prediction using a novel feature selection algorithm coupled with context-aware analysis. *Journal of Affective Disorders*, 295:1040–1048.

Ameet Deshpande and Karthik Narasimhan. 2020. Guiding attention for self-supervised learning with transformers. In *Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4676–4686. Association for Computational Linguistics.

Marc Domken, Jan Scott, and Peter Kelly. 1994. What factors predict discrepancies between self and observer ratings of depression? *Journal of Affective Disorders*.

Marc Fakhoury. 2019. *Artificial Intelligence in Psychiatry*, pages 119–125. Springer Singapore, Singapore.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David Devault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *9th International Conference on Language Resources and Evaluation (LREC)*.

Edmund S Higgins. 1994. A review of unrecognized mental illness in primary care: prevalence, natural history, and efforts to change the course. *Archives of family medicine*, 3(10):908.

Simin Hong, Anthony Cohn, and David Crossland Hogg. 2022. Using graph representation learning with schema encoders to measure the severity of depressive symptoms. In *International Conference on Learning Representations (ICLR)*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *13th Language Resources and Evaluation Conference (LREC)*, pages 7184–7190.

Clinton Lau, Xiaodan Zhu, and Wai-Yip Chan. 2023. Automatic depression severity assessment with deep learning using parameter-efficient tuning. *Frontiers in Psychiatry*, 14:1160291.

Adria Mallol-Ragolta, Ziping Zhao, Lukas Stappen, Nicholas Cummins, and Björn W. Schuller. 2019. A hierarchical attention network-based approach for depression detection from transcribed clinical interviews. In *Interspeech (INTERSPEECH)*, pages 221–225.

Kirill Milintsevich, Kairit Sirts, and Gaël Dias. 2023. Towards automatic text-based estimation of depression through symptom prediction. *Brain Informatics*, 10(1):1–14.

Meng Niu, Kai Chen, Qingcai Chen, and Lufeng Yang. 2021. Hcag: A hierarchical context-aware graph attention model for depression detection. In *IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4235–4239.

Syed Arbaaz Qureshi, Gaël Dias, Mohammed Hasanuzzaman, and Sriparna Saha. 2020. Improving depression level estimation by concurrently learning emotion intensity. *IEEE Computational Intelligence Magazine*, 15(3):47–59.

Syed Arbaaz Qureshi, Gaël Dias, Sriparna Saha, and Mohammed Hasanuzzaman. 2021. Gender-aware estimation of depression severity level in a multimodal setting. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Syed Arbaaz Qureshi, Sriparna Saha, Mohammed Hasanuzzaman, and Gaël Dias. 2019. Multitask representation learning for multimodal estimation of depression level. *IEEE Intelligent Systems*, 34(5):45–52.

Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg. 2019. Multi-level attention network using text, audio and video for depression prediction. In *9th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, page 81–88.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3980–3990.

Cheolmin Shin, Seung-Hoon Lee, Kyu-Man Han, Ho-Kyoung Yoon, and Changsu Han. 2019. Comparison of the usefulness of the phq-8 and phq-9 for screening for major depressive disorder: Analysis of psychiatric outpatient data. *Psychiatry Investigation*, 16(4):300–305.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *57th Conference of the Association for Computational Linguistics (ACL)*, pages 2895–2905.

Joe Stacey, Yonatan Belinkov, and Marek Rei. 2022. Supervising model attention with human explanations for robust natural language inference. In *AAAI conference on artificial intelligence (AAAI)*, volume 36, pages 11349–11357.

University of Tartu. 2018. UT rocket.

Shanshan Wang, Zhumin Chen, Zhaochun Ren, Huasheng Liang, Qiang Yan, and Pengjie Ren. 2022. Paying more attention to self-attention: Improving pre-trained language models via attention guiding. *arXiv preprint arXiv:2204.02922*.

Danai Xezonaki, Georgios Paraskevopoulos, Alexandros Potamianos, and Shrikanth Narayanan. 2020. Affective conditioning on hierarchical attention networks applied to depression detection from transcribed clinical interviews. In *Interspeech (INTERSPEECH)*, pages 4556–4560.

Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. 2020. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. In *28th International Conference on Computational Linguistics (COLING)*, pages 696–709.

K. Milintsevich, K. Sirts, G. Dias
Your Model Is Not Predicting Depression Well And That Is Why: A Case Study of PRIMATE Dataset
Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 166–171

Your Model Is Not Predicting Depression Well And That Is Why: A Case Study of PRIMATE Dataset

Kirill Milintsevich[1,2] and **Kairit Sirts**[2] and **Gaël Dias**[1]
[1]Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, France
[2]Institute of Computer Science, University of Tartu, Estonia
{first_name}.{last_name}@{unicaen.fr[1]|ut.ee[2]}

Abstract

This paper addresses the quality of annotations in mental health datasets used for NLP-based depression level estimation from social media texts. While previous research relies on social media-based datasets annotated with binary categories, i.e. depressed or non-depressed, recent datasets such as D2S and PRIMATE aim for nuanced annotations using PHQ-9 symptoms. However, most of these datasets rely on crowd workers without the domain knowledge for annotation. Focusing on the PRIMATE dataset, our study reveals concerns regarding annotation validity, particularly for the lack of interest or pleasure symptom. Through reannotation by a mental health professional, we introduce finer labels and textual spans as evidence, identifying a notable number of false positives. Our refined annotations, to be released under a Data Use Agreement, offer a higher-quality test set for anhedonia detection. This study underscores the necessity of addressing annotation quality issues in mental health datasets, advocating for improved methodologies to enhance NLP model reliability in mental health assessments.

1 Introduction

Applying various NLP techniques to automatically estimate the depression level from social media texts has been a widely researched topic in the field of NLP applied for mental health. Most of these datasets consist of online posts gathered from popular social media platforms, such as Twitter or Reddit. These posts are usually annotated by crowd workers who had only a brief training with a mental health professional (MHP) or sometimes only had access to the annotation instructions.

While there exist multiple depression-related datasets based on social media texts, most of them only present binary annotation, i.e. whether the user is depressed or not. The most common sources of data are Reddit (Losada and Crestani, 2016; Yates et al., 2017; Pirina and Çöltekin, 2018) and X (former Twitter) (Coppersmith et al., 2014; Syarif et al., 2019). Most of the studies use automatic methods of annotations, such as regular expression matching of self-reported terms, like "I have been diagnosed with depression". Some of them perform manual verification and annotation either via layman crowd workers (Yates et al., 2017) or by the authors themselves (Coppersmith et al., 2014; Losada and Crestani, 2016).

Recently, the interest in more fine-grained depression annotation has emerged. In particular, the two recent datasets D2S (Yadav et al., 2020) and PRIMATE (Gupta et al., 2022), identify depressed social media posts from X and Reddit, respectively and annotate them with PHQ-9 symptoms (Kroenke and Spitzer, 2002). Both datasets have been annotated with the help of crowd workers and later verified by MHPs. However, the verification process was different. For D2S, conflicting annotations were resolved with the majority voting, and the psychiatrist resolved the ties. After that, 100 random samples were selected for quality control and verified by a psychiatrist. Additionally, Zirikly and Dredze (2022) annotated a random sample of D2S with the explanations for each symptom with the help of two MHPs[1], increasing the validity of the data. In the case of PRIMATE, no information is given on the quality control procedure. This raises concerns about the validity of the annotations; thus, we selected PRIMATE for our case study.

In this study, on the example of the PRIMATE dataset, we show that the validity of the annotations for the mental health data is a concern when performed by layman crowd workers. Our MHP reannotated 170 posts from the PRIMATE dataset for the lack of interest or pleasure (anhedonia) symp-

[1] Zirikly and Dredze (2022) did not report any conflicts between their annotation and the labels provided with D2S.

Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 166–171
March 21, 2024 ©2024 Association for Computational Linguistics

tom. The MHP is the second author of the paper, who is also a practising clinical psychology intern. Our annotations include more fine-grained labels (“mentioned” vs “answerable”, as well as an additional “writer’s symptom” label) as well as spans of texts that serve as evidence of the labels. We observe a high number of false positives in the PRIMATE labels, which can be related to the high difficulty of conceptualizing anhedonia (Rizvi et al., 2016). The annotations are to be released under a Data Use Agreement (DUA), and we believe that it can serve as a higher-quality test set for anhedonia detection.

2 Dataset

PRIMATE (Gupta et al., 2022) is a dataset based on the Reddit posts from the r/depression_help subreddit. Each post is annotated with binary labels for each PHQ-9 question, where “yes” means that a post contains the answer to a PHQ-9 question and “no” otherwise. The nine symptoms are shortly described as follows: lack of interest or pleasure in doing things (LOI), feeling down or depressed (DEP), sleeping disorder (SLE), lack of energy (ENE), eating disorder (EAT), low self-esteem (LSE), problems with concentrating (CON), hyper or lower activity (MOV), suicidal thoughts (SUI).

The annotation was performed by five crowd workers with additional quality control by an MHP. The information about the annotation procedure or crowd worker training, as well as how exactly the MHPs were involved in the quality control, are not provided in the paper. The only metric on the annotation process is an annotator agreement using Fleiss’ kappa, which is reported to be 67% for initial annotation and 85% after involvement of the MHPs.

In total, the dataset consists of 2003 posts. Table 1 shows the distribution of the labels[2]. Note that the exact numbers of labels are slightly different from the ones presented by Gupta et al. (2022). The dataset is not pre-split into train, validation and test sets; thus, we randomly sample 200 posts for validation and another 200 posts for testing.

PHQ-9 Symptom	Number of Posts	
	Present	Absent
LOI	949	1054
DEP	1664	339
SLE	374	1629
ENE	688	1315
EAT	194	1809
LSE	1680	323
CON	195	1808
MOV	527	1476
SUI	743	1260

Table 1: Label distribution in PRIMATE.

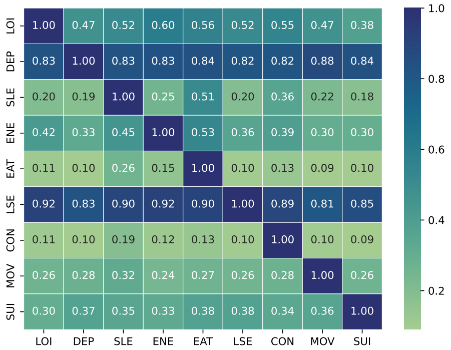


Figure 1: Symptom label co-occurrence matrix of the PRIMATE training set. Each value is normalized column-wise by dividing it by the highest value in the column.

Figure 1 shows the label co-occurrence matrix of the training set. Two symptoms, DEP and LSE, co-occur the most with all the other symptoms, which can be explained by their general prevalence in the dataset. The connection between the lack of interest or pleasure (LOI) and lack of energy (ENE) is also seen in the dataset, which reflects high comorbidity of these symptoms (van Borkulo et al., 2015; Park and Kim, 2020).

3 Experimental Setup

In our experiments, we aimed to test how well current pre-trained language models can model the depression symptom detection problem using the PRIMATE dataset. We first chose DistilBERT (Sanh et al., 2019) as a baseline and BERT-Base (Devlin et al., 2018), RoBERTa-Base, RoBERTa-Large (Liu et al., 2019), DeBERTa-Base, and DeBERTa-Large (He et al., 2020) as higher-performing models. In particular, DeBERTa has shown constant improvements in various NLP tasks and replaced BERT and RoBERTa as the state-of-the-art model for many of them[3].

For fine-tuning, we used the implementation from Transformers library (Wolf et al., 2020). Each

[2] The order of the symptoms in the original work by Gupta et al. (2022) is different from the one of PHQ-9. In our work, we reordered the symptoms to match PHQ-9.

[3] https://gluebenchmark.com/leaderboard

Model	LOI	DEP	SLE	ENE	EAT	LSE	CON	MOV	SUI
DistilBERT	.64	.88	.67	.58	.60	.90	.50	.67	.81
BERT-Base	.55	.88	.66	.55	.63	.90	.46	.66	.79
RoBERTa-Base	.54	.88	.70	.57	.57	.90	.51	.69	.85
RoBERTa-Large	.57	.86	.75	.63	.65	.91	.52	.71	.85
DeBERTa-Base	.58	.91	.69	.52	.42	.90	.36	.61	.81
DeBERTa-Large	.60	.90	.68	.64	.47	.91	.50	.73	.83

Table 2: Symptom-wise F1-scores on the validation set.

`Mentioned:`

`I simply want everything to finish.` **`I have no drive to do anything.`** `I am very irritable. Nothing is going as I want to and even if it was I probably wouldn't appreciate it.`

`Answerable:`

`I feel like I'm spending my life for nothing.` **`I used to escape my problems by browsing Youtube and Reddit for hours, but now I don't even find that enjoyable anymore.`**

`Not author's symptoms:`

`I've tried to talk about looking for other options or just ways to deal with the stress, but he's not really interested now.`

Figure 2: Examples of reannotated posts. Evidences are highlighted in **bold**.

Predictions	Against PRIMATE				Against "mentioned"				Against "answerable"			
	A	P	R	F1	A	P	R	F1	A	P	R	F1
DistilBERT	.58	.56	.62	.58	.56	.30	.71	.42	.51	.10	.75	.18
PRIMATE Labels	-	-	-	-	.56	.27	.58	.37	.54	.09	.58	.15

Table 3: Results on the reannotated part of the validation set. Here, **A** stands for Accuracy, **P** for Precision, **R** for Recall, and **F1** for F1-score for the positive class.

model consists of a pre-trained encoder with a classification head on the top of the `[CLS]` token. The classification head is represented by a linear layer; in the case of DeBERTa, another linear layer followed by GELU (Hendrycks and Gimpel, 2016) is added before the classification head. We trained each model for 20 epochs using AdamW optimizer with the learning rate of $2e^{-5}$, ϵ of $1e^{-6}$, β_1, β_2 of $(0.9, 0.999)$, and weight decay λ of 0.01. Additionally, a linear learning rate scheduler is applied with a warmup ratio of 0.1. Finally, the training batch size was set to 16.

4 Results and Discussion

Table 2 shows that larger models, such as RoBERTa-Large and DeBERTa-Large, perform better for `ENE`, `LSE`, `MOV`, and `SUI`. Additionally, `DEP` shows slight improvement with DeBERTa models, however, decreased performance for `EAT`. RoBERTa models perform better for `SLE` and `SUI` prediction. Nevertheless, DistilBERT sets a strong baseline and performs on par with larger models overall. Finally, `LOI` shows a decrease in performance for all the models compared to the DistilBERT.

We investigate the diminished performance of the `LOI` symptom since it is a core symptom of a major depressive disorder (Association, 2013) and shows unstable results for our models. Furthermore, `LOI` is one of the symptoms of schizophrenia (Association, 2013) and is associated with both anxiety and depression (Winer et al., 2017). Thus, we selected a subset of 170 posts from the validation set based on the DistilBERT predictions: if at least one symptom was predicted incorrectly, the post was selected. Next, an MHP read all the posts in the subset and labelled them for the presence of loss of interest or pleasure (`LOI`). The MHP assigned three labels to each post: a) “mentioned” if the symptom is talked about in the text, but it is not possible to infer its duration or intensity; b) “answerable” if there is clear evidence of anhedonia; c) “writer’s symptoms” which shows whether the author of the post discusses themselves or a third person. Additionally, the MHP selected the part of the text that supports the positive label.

Figure 2 shows examples for the reannotated posts[4]. The first example is labelled as “mentioned” since it contains evidence of a symptom but does not contain information about the *loss* of interest. The second example is labelled as “answerable” because it is possible to infer that the person used to have interest in what they were doing before but lost it at some point in time. Finally, the last example shows the post without signs of LOI that describes the condition of another person.

Table 3 shows accuracy, precision, recall and F1-score for positive class against different sets of labels on our manually reannotated subset. DistilBERT, when measured against “mentioned” and “answerable” labels, performs considerably worse than against original labels from PRIMATE. It is unsurprising given the extremely low agreement between these sets of labels with Cohen’s kappa of 9% and 3%, respectively. Furthermore, the most common error type is a false positive, i.e., a symptom marked as present in PRIMATE when our MHP found no evidence of it in the text. Additionally, using PRIMATE labels as predictions and comparing their performance against our labels shows lower performance than the DistilBERT model.

Considering the “writer’s symptom” label, in 18 out of 170 selected posts, the author describes a symptom of another person rather than themselves. This raises the question of how these posts should be annotated and whether they should be included in the dataset at all. We suspect that the language of describing one’s condition or feelings in the first person is different from the third person. We leave this question for future debate and assign “mentioned” and “answerable” labels to the posts describing a third person in the same manner as to the personal posts.

Our findings are consistent with the original results presented by Gupta et al. (2022). Similar to our experiment, they also trained a classifier based on the BERT-Base model and reported low MCC for LOI. However, we provided the evidence that this might be caused by annotation errors. Additionally, we noticed that many posts that were mistakenly labelled with LOI are more closely related to the “inner tension” symptom from the Montgomery-Åsberg Depression Rating Scale (MADRS) (Montgomery and Åsberg, 1979).

While we agree that our reannotated test set is also, to some extent, susceptible to errors, we believe that it serves as a more reliable benchmark for the anhedonia symptom. A more fine-grained, evidence-based labelling scheme reduces the risk of mislabelling and is more transparent for further verification. Finally, it lays the foundation for future collaboration to produce a higher-quality Reddit-based dataset for depression symptom estimation.

5 Conclusion

In conclusion, this study highlights the importance of evaluating and enhancing the quality of annotations in mental health datasets, particularly within the context of automated depression level estimation from social media texts. While recent datasets such as PRIMATE introduce commendable efforts toward nuanced annotations using PHQ-9 symptoms, our examination of the PRIMATE dataset reveals concerns about annotation validity, specifically regarding the lack of interest or pleasure symptom. Through careful reannotation by a mental health professional, we discerned a considerable number of false positives among the original labels indicative of challenges in conceptualizing anhedonia.

The findings presented here advocate for a more rigorous and standardized approach to mental health dataset annotation, emphasizing the need for greater involvement of domain experts in the annotation process. The release of our refined annotations under a Data Use Agreement (DUA) contributes a valuable resource for future research, offering a higher quality test set for anhedonia detection. Moving forward, a concerted effort toward refining annotation methodologies and promoting collaboration between domain experts and NLP practitioners is imperative to foster advancements in this crucial intersection of technology and mental health research.

6 Availability of Data

The instructions for accessing the annotations presented in this paper can be found here: https://github.com/501Good/primate-anhedonia.

[4] All example posts are paraphrased for privacy.

7 Ethical Considerations

According to Benton et al. (2017), studies involving user-generated content are exempt from Institutional Review Board (IRB) requirements if the data source is public and user identities are not identifiable. We access and use the data according to the Data Use Agreement provided with the PRIMATE dataset. Finally, we are going to release our annotations under another Data Use Agreement and separate them from the original PRIMATE data. We also acknowledge that no automatic system can replace a real mental health professional and cannot be used as a sole instrument of diagnostics.

8 Limitations

We acknowledge the limitations inherent in our work and findings. First, the manually annotated explanations serve as a proxy for what clinicians might find informative in assessing Reddit posts flagged as depressive. While evaluating the informativeness of explanations in a true clinical setting would provide more insight, it falls beyond the scope of this paper. Furthermore, our reannotation was carried out by only one mental health professional, which does not allow for performing an inter-annotator agreement analysis. However, we believe that our evidence-based labelling scheme partially mitigates this problem. Finally, anhedonia is extremely challenging to conceptualize and binary labels may not be the best choice in situations when the difference between the presence or absence of the symptom is marginal. In this case, labels based on the Likert scale, as in PHQ-9, would be more appropriate and allow us to capture the intensity of the symptom more accurately. Furthermore, different demographics, for example, adolescents and adults, express signs of anhedonia differently (Watson et al., 2020).

Acknowledgements

This research was supported by the Estonian Research Council Grant PSG721 and the FHU A^2M^2P project funded by the G4 University Hospitals of Amiens, Caen, Lille and Rouen (France). The calculations for model's training and inference were carried out in the High Performance Computing Center of the University of Tartu (University of Tartu, 2018).

References

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5™ (5th ed.)*. American Psychiatric Publishing, Inc.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shrey Gupta, Anmol Agarwal, Manas Gaur, Kaushik Roy, Vignesh Narayanan, Ponnurangam Kumaraguru, and Amit Sheth. 2022. Learning to automate follow-up question generation using process knowledge for depression triage on Reddit posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 137–147, Seattle, USA. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Kurt Kroenke and Robert L Spitzer. 2002. The PHQ-9: a new depression diagnostic and severity measure.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

David E Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *International conference of the cross-language evaluation forum for European languages*, pages 28–39. Springer.

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

Seon-Cheol Park and Daeho Kim. 2020. The centrality of depression and anxiety symptoms in major depressive disorder determined using a network analysis. *Journal of affective disorders*, 271:19–26.

Inna Pirina and Çağrı Çöltekin. 2018. Identifying depression on Reddit: The effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 9–12, Brussels, Belgium. Association for Computational Linguistics.

Sakina J Rizvi, Diego A Pizzagalli, Beth A Sproule, and Sidney H Kennedy. 2016. Assessing anhedonia in depression: Potentials and pitfalls. *Neuroscience & Biobehavioral Reviews*, 65:21–35.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Iwan Syarif, Nadia Ningtias, and Tessy Badriyah. 2019. Study on mental disorder detection via social media mining. In *2019 4th International conference on computing, communications and security (ICCCS)*, pages 1–6. IEEE.

University of Tartu. 2018. UT rocket.

Claudia van Borkulo, Lynn Boschloo, Denny Borsboom, Brenda WJH Penninx, Lourens J Waldorp, and Robert A Schoevers. 2015. Association of symptom network structure with the course of depression. *JAMA psychiatry*, 72(12):1219–1226.

Rebecca Watson, Kate Harvey, Ciara McCabe, and Shirley Reynolds. 2020. Understanding anhedonia: A qualitative study exploring loss of interest and pleasure in adolescent depression. *European Child & Adolescent Psychiatry*, 29:489–499.

E Samuel Winer, Jessica Bryant, Gregory Bartoszek, Enrique Rojas, Michael R Nadorff, and Jenna Kilgore. 2017. Mapping the relationship between anxiety, anhedonia, and depression. *Journal of affective disorders*, 221:289–296.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. 2020. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 696–709, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Ayah Zirikly and Mark Dredze. 2022. Explaining models of mental health via clinically grounded auxiliary tasks. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 30–39, Seattle, USA. Association for Computational Linguistics.

CURRICULUM VITAE

Personal data

Name:	Kirill Milintsevich
Date of birth:	March 7, 1995
Citizenship:	Russian Federation
Languages:	Russian, English, French
Contact:	milintsevich@gmail.com

Education

2020–2024	University of Tartu & University of Caen Normandy, PhD in Computer Science
2018–2020	University of Tartu, MSc in Computer Science
2016–2018	Higher School of Economics, MA in Computational Linguistics
2012–2016	Far Eastern Federal University, BA in Applied Linguistics

Employment

2023–2024	University of Caen Normandy, Teaching and Research Assistant (ATER)
2020–2023	University of Caen Normandy, Doctoral researcher
2017–2020	Medialogia, NLP Engineer

Scientific work

Main fields of interest:

- natural language processing
- language generation
- mental health

Publications

1. **Milintsevich, K.**, Sirts, K., & Dias, G. (2024). Evaluating Lexicon Incorporation for Depression Symptom Estimation. In *Proceedings of the 6th Clinical Natural Language Processing Workshop (**Clinical NLP**)* at *NAACL 2024*.
   </> Code: https://github.com/501Good/dialogue-classifier
2. Agarwal, N.*, **Milintsevich, K.***, Métivier, L., Rothärmel, M., Dias, G., & Dollfus, S. (2024). Analyzing Symptom-based Depression Level Estimation

through the Prism of Psychiatric Expertise. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (**LREC-COLING 2024**).*
**equal contributions*

3. **Milintsevich, K.**, Sirts, K., & Dias, G. (2024). Your Model Is Not Predicting Depression Well And That Is Why: A Case Study of PRIMATE Dataset. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (**CLPsych**)* at *EACL 2024.*
4. **Milintsevich, K.** & Agarwal, N. (2023). Calvados at MEDIQA-Chat 2023: Improving Clinical Note Generation with Multi-Task Instruction Fine-tuning. In *Proceedings of the 5th Clinical Natural Language Processing Workshop (**Clinical NLP**)* at *ACL 2023.*
   </> Code: https://github.com/501Good/MEDIQA-Chat-2023-Calvados
5. **Milintsevich, K.**, Sirts, K., & Dias, G. (2023). Towards automatic text-based estimation of depression through symptom prediction. ***Brain Informatics**, 10(1), 1-14.*
   </> Code: https://tinyurl.com/3ssw4rcf
6. **Milintsevich, K.** and Sirts K. (2021). Enhancing Sequence-to-Sequence Neural Lemmatization with External Resources. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* at ***EACL 2021**.* Association for Computational Linguistics.
   </> Code: https://github.com/501Good/lexicon-enhanced-lemmatization
7. **Milintsevich, K.** and Sirts K. (2020). Lexicon-Enhanced Neural Lemmatization for Estonian. In *Proceedings of the Ninth International Conference of **Baltic HLT 2020**.* Frontiers in Artificial Intelligence and Applications.
8. Kittask, C., **Milintsevich, K.** & Sirts K. (2020). Evaluating Multilingual BERT for Estonian. In *Proceedings of the Ninth International Conference of **Baltic HLT 2020**.* Frontiers in Artificial Intelligence and Applications.
9. Chernyak, E., Ponomareva, M., & **Milintsevich, K.** (2019). Char-RNN for Word Stress Detection in East Slavic Languages. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects (**VarDial**)* at *NAACL-HLT 2019.* Association for Computational Linguistics.
10. Ponomareva, M., **Milintsevich, K.**, Chernyak, E., & Starostin, A. (2017). Automated word stress detection in Russian. *In Proceedings of the First Workshop on Subword and Character Level Models in NLP (**SCLeM**)* at *EMNLP 2017.* Association for Computational Linguistics.

ELULOOKIRJELDUS

Isikuandmed

Nimi:	Kirill Milintsevich
Sünniaeg:	07.03.1995
Kodakondsus:	Venemaa Föderatsioon
Keelteoksus:	vene, inglise, prantsuse
Kontaktandmed:	milintsevich@gmail.com

Haridus

2020–2024	Tartu Ülikool & Caen Normandia Ülikool, informaatika doktorant
2018–2020	Tartu Ülikool, informaatika magister
2016–2018	Rahvuslik Uurimisülikool “Kõrgem Majanduskool”, arvutilingvistika magister
2012–2016	Kaug-Ida Föderaalse Ülikool, rakenduslingvistika bakalaureus

Teenistuskäik

2023–2024	Caen Normandia Ülikool, õppe- ja teadusassistent (ATER)
2020–2023	Caen Normandia Ülikool, doktorantuuriteadur
2017–2020	Medialogia, NLP insener

Teadustegevus

Peamised uurimisvaldkonnad:

- loomuliku keele töötlemine
- keele genereerimine
- vaimne tervis

Publikatsioonid

1. **Milintsevich, K.**, Sirts, K., & Dias, G. (2024). Evaluating Lexicon Incorporation for Depression Symptom Estimation. In *Proceedings of the 6th Clinical Natural Language Processing Workshop (**Clinical NLP**)* at *NAACL 2024*.
   </> Kood: https://github.com/501Good/dialogue-classifier
2. Agarwal, N.*, **Milintsevich, K.***, Métivier, L., Rothärmel, M., Dias, G., & Dollfus, S. (2024). Analyzing Symptom-based Depression Level Estimation

through the Prism of Psychiatric Expertise. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (**LREC-COLING 2024**).*
*võrdne panus

3. **Milintsevich, K.**, Sirts, K., & Dias, G. (2024). Your Model Is Not Predicting Depression Well And That Is Why: A Case Study of PRIMATE Dataset. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (**CLPsych**)* at *EACL 2024.*
4. **Milintsevich, K.** & Agarwal, N. (2023). Calvados at MEDIQA-Chat 2023: Improving Clinical Note Generation with Multi-Task Instruction Fine-tuning. In *Proceedings of the 5th Clinical Natural Language Processing Workshop (**Clinical NLP**)* at *ACL 2023.*
   </> Kood: https://github.com/501Good/MEDIQA-Chat-2023-Calvados
5. **Milintsevich, K.**, Sirts, K., & Dias, G. (2023). Towards automatic text-based estimation of depression through symptom prediction. ***Brain Informatics,*** *10(1), 1-14.*
   </> Kood: https://tinyurl.com/3ssw4rcf
6. **Milintsevich, K.** and Sirts K. (2021). Enhancing Sequence-to-Sequence Neural Lemmatization with External Resources. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* at ***EACL 2021***. Association for Computational Linguistics.
   </> Kood: https://github.com/501Good/lexicon-enhanced-lemmatization
7. **Milintsevich, K.** and Sirts K. (2020). Lexicon-Enhanced Neural Lemmatization for Estonian. In *Proceedings of the Ninth International Conference of **Baltic HLT 2020***. Frontiers in Artificial Intelligence and Applications.
8. Kittask, C., **Milintsevich, K.** & Sirts K. (2020). Evaluating Multilingual BERT for Estonian. In *Proceedings of the Ninth International Conference of **Baltic HLT 2020***. Frontiers in Artificial Intelligence and Applications.
9. Chernyak, E., Ponomareva, M., & **Milintsevich, K.** (2019). Char-RNN for Word Stress Detection in East Slavic Languages. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects (**VarDial**)* at *NAACL-HLT 2019.* Association for Computational Linguistics.
10. Ponomareva, M., **Milintsevich, K.**, Chernyak, E., & Starostin, A. (2017). Automated word stress detection in Russian. *In Proceedings of the First Workshop on Subword and Character Level Models in NLP (**SCLeM**)* at *EMNLP 2017.* Association for Computational Linguistics.

Estimation du niveau de dépression à partir de données textuelles : approche basée sur les symptômes, utilisation de ressources externes, validité des jeux de données

Mots-clés : traitement automatique des langues, intelligence artificielle, santé mentale

Le trouble dépressif majeur (TDM) est l'un des troubles mentaux les plus répandus au monde, entraînant souvent une incapacité et un risque accru de suicide. La récente pandémie de coronavirus (COVID-19) a fait grimper le taux de dépression dans le monde entier. De plus, la stigmatisation et l'accès limité aux traitements entravent le diagnostic et les soins appropriés pour de nombreuses personnes.

Des études préliminaires ont montré que les personnes déprimées et non déprimées utilisent un vocabulaire différent. Par exemple, les personnes déprimées ont tendance à utiliser davantage de mots négatifs ou émotionnels. Plus récemment, des modèles d'apprentissage profond ont été développés pour détecter la dépression à partir de textes. Cependant, la plupart des chercheurs ont traité la détection de la dépression comme une tâche de classification simple avec seulement deux étiquettes possibles : « déprimé » et « non déprimé ». Lorsqu'on considère deux personnes atteintes de dépression, il est important de noter qu'elles peuvent présenter des symptômes sous-jacents différents. Une personne peut souffrir d'insomnie et de difficultés de concentration, tandis qu'une autre peut présenter des changements d'appétit et une faible estime de soi. Ces personnes nécessitent des traitements différents, donc disposer d'informations sur les symptômes est essentiel.

Dans cette thèse, nous avons développé une architecture de réseau neuronal qui prédit les symptômes de la dépression à partir de textes. Nous avons constaté que la prédiction des symptômes, plutôt qu'un simple diagnostic, était plus précise, tout en nous fournissant plus de détails. Nous avons encore amélioré le réseau de neurones en y introduisant des connaissances externes provenant de lexiques de sentiments et d'émotions. Nous avons utilisé une approche simple mais efficace qui consiste à marquer directement les mots des lexiques dans le texte. Enfin, en travaillant sur un jeu de données provenant des réseaux sociaux, nous avons constaté que le processus d'annotation était erroné. En conséquence, nous avons réannoté une partie de ce jeu de données avec l'aide d'un professionnel en santé mentale, démontrant ainsi l'importance de suivre les définitions médicales des symptômes et d'établir des directives claires pour l'annotation.

Estimation of Depression Level from Text: Symptom-Based Approach, External Knowledge, Dataset Validity

Keywords: natural language processing, artificial intelligence, mental health

Major Depressive Disorder (MDD) is one of the most prevalent mental disorders globally, often resulting in disability and an increased risk of suicide. The recent COVID-19 pandemic has made depression rates go up around the world. Moreover, stigma and limited treatment access hinder proper diagnosis and care for many.

Early studies have found that depressed and non-depressed people use different vocabulary. For example, depressed people tend to use more negative or emotional words. More recently, deep learning models have been developed to detect depression from text. However, most researchers have treated depression detection as a simple classification task with only two possible labels: depressed and non-depressed. When considering two individuals with depression, it is important to note that they may exhibit different underlying symptoms. One person may experience insomnia and difficulty concentrating, while another may struggle with changes in appetite and low self-esteem. These people would require different treatments, so having information about the symptoms is essential.

In this work, we developed a neural network that predicts depression symptoms from text. We found that predicting symptoms instead of a simple diagnosis was more accurate while giving us more details at the same time. We further improved the neural network by introducing external knowledge from existing sentiment and emotion lexicons. We used a simplistic yet effective approach of directly marking the words from the lexicons in the text. Finally, while working with a social media dataset, we discovered it was poorly annotated. As a result, we reannotated a part of this dataset with the help of a mental health professional, showing the importance of following medical symptom definitions and establishing clear annotation guidelines.