DISSERTATIONES
INFORMATICAE
UNIVERSITATIS
TARTUENSIS
58

KIRILL MILINTSEVICH

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

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS
58

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS
58

KIRILL MILINTSEVICH

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

UNIVERSITY OF TARTU
Press
1632

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.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in Computer Science on October 1, 2024 by the Council of the Institute of Computer Science, University of Tartu and on August 27, 2024 by the Doctoral School MIIS, University of Caen Normandy.

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.

ISSN 2613-5906 (print) ISSN 2806-2345 (pdf)
ISBN 978-9916-27-752-2 (print) ISBN 978-9916-27-753-9 (pdf)

University of Tartu Press
`http://www.tyk.ee/`

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 **10**

List of original publications **12**

List of original resources **13**

1. Introduction **14**

2. Background **17**
2.1. Language of Depression 18
2.2. Language Resources 19
2.2.1. Lexicons 19
2.2.2. Depression Datasets 20
2.3. Automatic Depression Estimation from Text 24
2.3.1. Approaches for Automatic Depression Estimation 24
2.3.2. Evaluation Metrics 27
2.3.3. Published Results 29

3. Symptom-Based Automatic Depression Estimation (Publication I) **31**
3.1. Methodology 31
3.2. Data and Experimental Setup 33
3.3. Results and Discussion 33
3.4. Conclusions and Future Work 36

4. External Knowledge Incorporation for Depression Symptom Estimation (Publications II and III) **37**
4.1. External Knowledge Incorporation via Input Marking 37
4.2. Model Modifications 38
4.3. Results and Discussion 39
4.4. Exploring Model's Attention 42
4.5. Conclusions and Future Work 45

5. Social-Media-Based Depression Datasets Validity (Publication IV and Dataset I) **46**
5.1. Benchmarking Pre-Trained Models on PRIMATE 46
5.2. Reannotation of PRIMATE 47
5.3. Conclusions and Future Work 50

6. Conclusion **51**
6.1. Main Conclusions 52
6.2. Limitations and Ethical Considerations 52
6.3. Future work 53

Bibliography **55**

Acknowledgements **67**

Sisukokkuvõte (Summary in Estonian) **68**

Résumé (Summary in French) **71**

Publications **75**

Towards Automatic Text-Based Estimation of Depression through Symptom Prediction 77

Evaluating Lexicon Incorporation for Depression Symptom Estimation . 93

Analyzing Symptom-based Depression Level Estimation through the Prism of Psychiatric Expertise 103

Your Model Is Not Predicting Depression Well And That Is Why: A Case Study of PRIMATE Dataset 115

Curriculum Vitae **123**

Elulookirjeldus (Curriculum Vitae in Estonian) **124**

LIST OF ABBREVIATIONS

Acronyms

BDI Beck Depression Inventory. 23, 24

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

DAIC-WOZ Distress Analysis Interview Corpus Wizard-of-Oz. 14–16, 19–22, 24–26, 29–31, 33, 35, 39–44, 46, 52

DSM-5 Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. 14, 17, 47, 53

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. 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, 25, 31, 46, 51

PHQ-8 Patient Health Questionnaire. 14, 29–31, 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, 54

PTSD Post-Traumatic Stress Disorder. 21

RMSE Root Mean Square Error. 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

[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.

us to pursue the case study regarding the validity of the annotations in 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. In this contribution, we adapted a hierarchical neural classifier to build a multi-target regression model to predict each depression symptom based on the DAIC-WOZ dataset. We showed that by predicting each depression symptom individually, the model outperformed a binary classification approach and, at the same time, provided a much more informative symptomatic overview. 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**). To incorporate external resources in the form of lexicons and psychiatrists' annotations, we applied a simplistic input marking approach. It allowed us to improve the performance of the model from Chapter 4 on the DAIC-WOZ dataset. However, on the PRIMATE dataset, which is based on social-media texts, conflicting results led us to carry out a case study described in the next chapter of this thesis. Chapter 5 describes a case study of PRIMATE, 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 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; M. 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.

PRIMATE dataset was collected from the `r/depression_help` subreddit. Initially, Gupta et al. (2022) scraped a set of approximately 21,000 posts that were marked with "advice", "help" or "support" flair tags. Then, a subset of 2,003 posts that had at least one interrogative comment was created out of the initial set. Finally, each of 2,003 posts was annotated with nine labels, each signifying whether a corresponding question form the PHQ-9 could be answered based on the content

of the post. The dataset is distributed without any predefined train/validation/test splits.

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] expanding the dataset.

Finally, several initiatives were carried out to collect social-media-based depression datasets in languages other than English. For example, in Spanish (Mármol Romero et al., 2024), Portuguese (dos Santos et al., 2023), and Chinese (Cai et al., 2023), to name a few. However, since this work focuses on English-based data, providing a detailed overview of datasets in other languages is out of its scope.

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-

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

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 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:

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

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

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

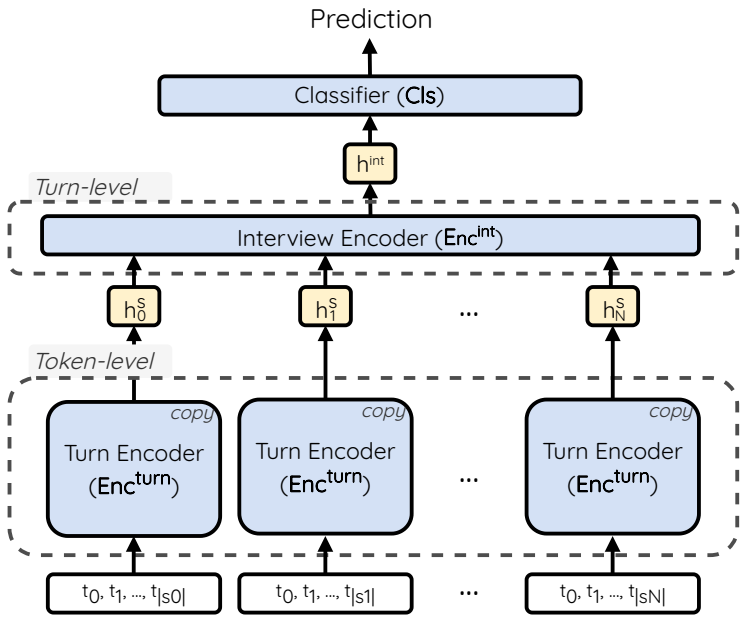


Figure 1: General architecture of a hierarchical classifier model.

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

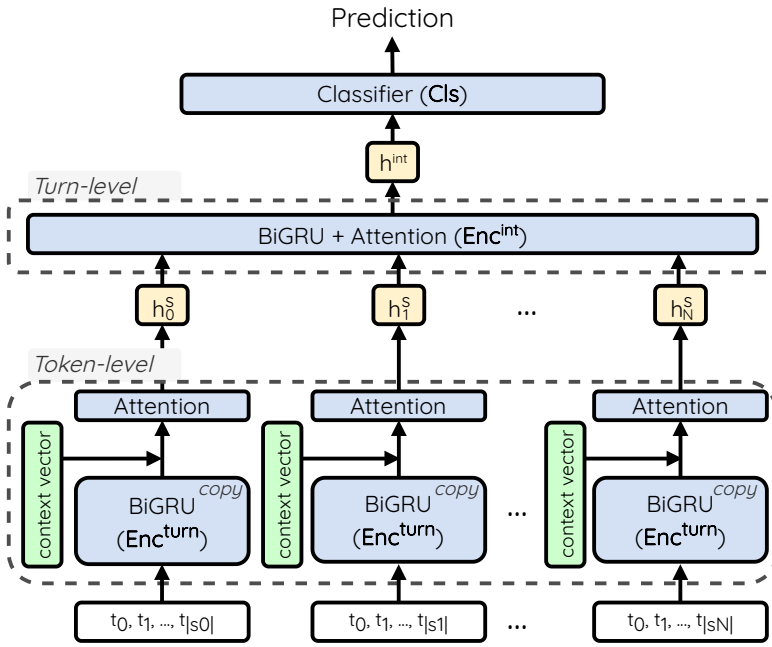


Figure 2: Hierarchical model with attention conditioning proposed by Xezonaki et al. (2020).

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:

[10] A thematic community on Reddit.

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

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 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.

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

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

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

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.

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

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.

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.

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

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 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 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.

[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.

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 $\mathbf{Enc}^{\mathbf{turn}}$ 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 $\mathbf{Enc}^{\mathbf{turn}}$ are created, and their respective computational graphs are stored during training. The turn-level interview encoder $\mathbf{Enc}^{\mathbf{int}}$ deploys a single

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

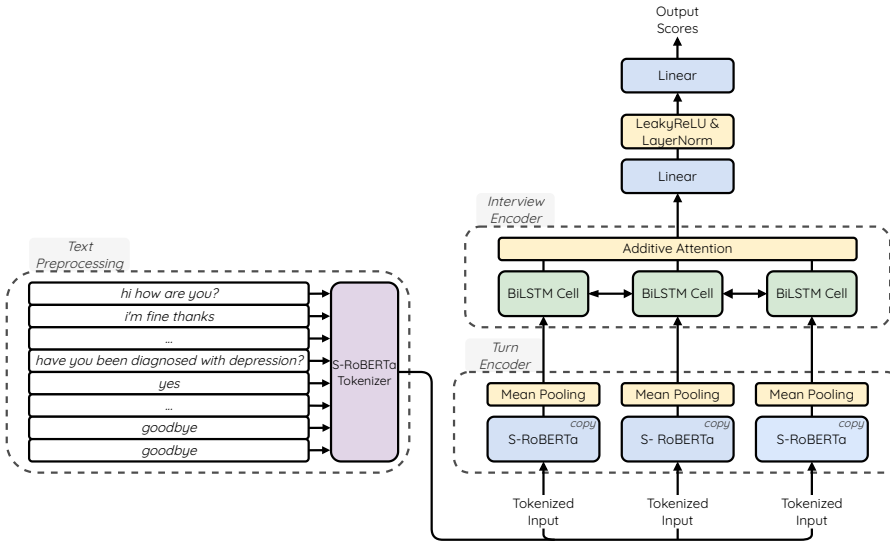


Figure 3: Overview of the model. On the turn level, the same instance of S-RoBERTa is used to encode each turn. Mean Pooling is the operation that averages all the token representations output by S-RoBERTa.

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	
	$miF_{1\ \pm\sigma}$	$maF_{1\ \pm\sigma}$	$mi\text{MAE}_{\pm\sigma}$	$ma\text{MAE}_{\pm\sigma}$
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	$\mathbf{0.766}_{\pm 0.023}$	$\mathbf{0.739}_{\pm 0.025}$	$\mathbf{3.78}_{\pm 0.13}$	$\mathbf{4.19}_{\pm 0.13}$

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 $\pm\sigma$	$miF1\ \pm\sigma$	$maF1\ \pm\sigma$
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,

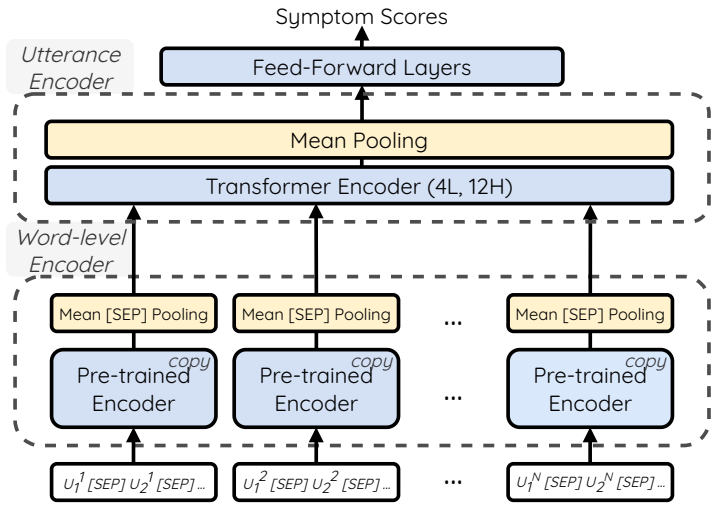


Figure 4: Overview of the model architecture. U_i^N stands for i-th utterance of N-th input. *Symptom Scores* are $|L|$ real numbers, where $|L|$ is the number of symptoms to predict.

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

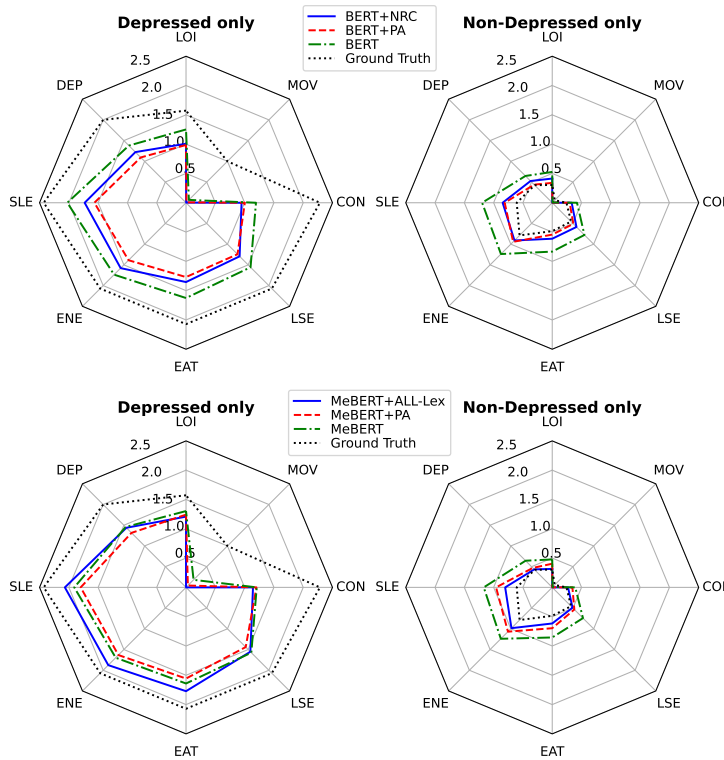


Figure 5: Average predicted values for depressed and non-depressed patients of the DAIC-WOZ test set.

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.

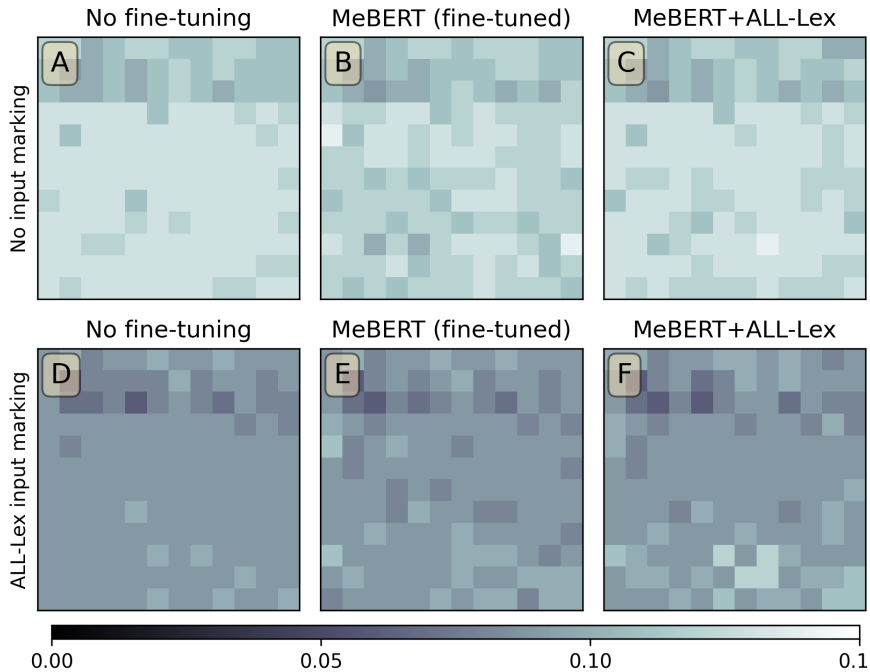
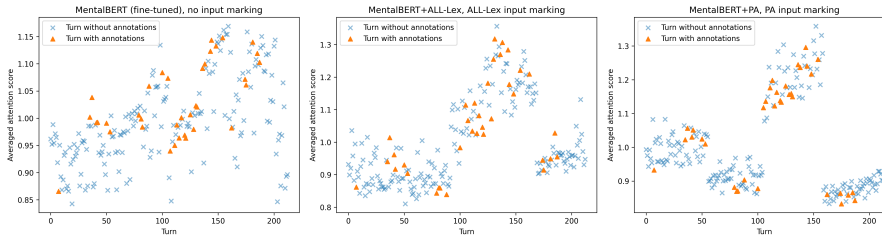


Figure 6: Relative lexicon attention scores. For each heatmap, the rows and columns correspond to layers and attention heads, respectively. The top row `[A-C]` shows the relative attention scores for the models tested on the inputs without any markings, and the bottom row `[D-F]` shows the scores tested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

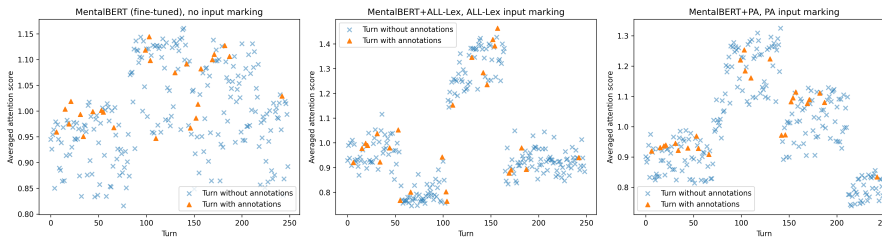
more comfortable, followed by more depression-targeted questions, and finishes with a cool-down phase to make the patient feel at ease again (Gratch et al., 2014). While both MEBERT fined-tuned without input markings, MEBERT+PA and MEBERT+ALL-LEX generally assign higher attention scores to the middle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

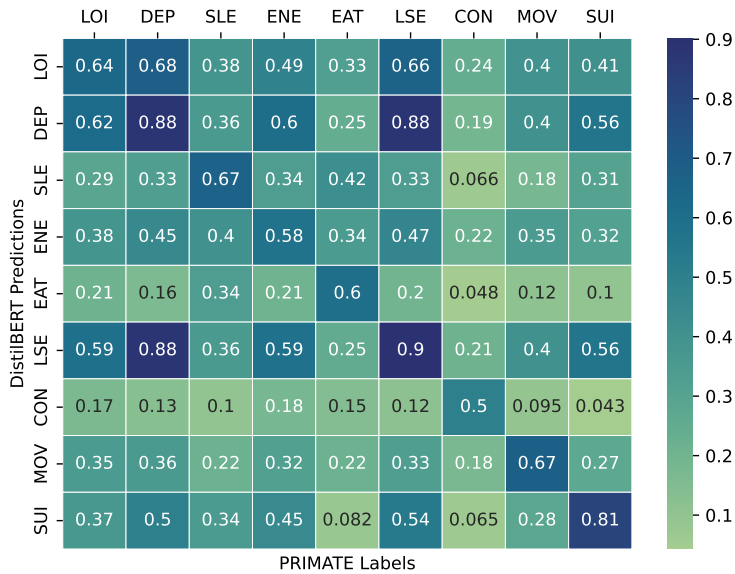


Figure 8: Cross-evaluation of DistilBERT predictions against PRIMATE labels on the validation set. The values inside of each cell represent F1-scores. Example of reading the graph: the value in the intersection of the first row and the second column represents the F1-score between the predictions of 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 text 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.

The outcomes of this work could be applied to other interdisciplinary NLP research. The inclusion of domain experts into the process of developing NLP models is crucial not only for mental health domain, but also in others like, for example, finance or legal texts (S. Park et al., 2021). Developing domain-specific NLP systems without involvement of domain experts might lead to a misalignment in the task definition and as a result to a poor applicability of such systems.

Finally, collaboration with domain experts in data annotation is important for producing high-quality data. Arguably, data is the core of most deep learning NLP models. As shown in this work, the use of layperson annotators for the labeling of depression symptoms produced annotations that were not consistent with the clinical definition of the symptom. We believe that other interdisciplinary research might also benefit from higher-quality data labeling produced by domain experts.

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^2M^2P 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

Cai, Y., Wang, H., Ye, H., Jin, Y., & Gao, W. (2023). Depression detection on online social network with multivariate time series feature of user depressive symptoms. *Expert Systems with Applications*, *217*, 119538. https://doi.org/10.1016/j.eswa.2023.119538

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*.

dos Santos, W. R., de Oliveira, R. L., & Paraboni, I. (2023). SetembroBR: a social media corpus for depression and anxiety disorder prediction. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-022-09633-0

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

Mármol Romero, A. M., Moreno Muñoz, A., Plaza-del-Arco, F. M., Molina González, M. D., Martín Valdivia, M. T., Ureña-López, L. A., & Montejo Ráez, A. (2024). MentalRiskES: a new corpus for early detection of mental disorders in Spanish. *LREC-COLING 2024*, 11204–11214.

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 workshop* (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 multilinguality, 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.

Park, S., Wang, A. Y., Kawas, B., Liao, Q. V., Piorkowski, D., & Danilevsky, M. (2021). Facilitating knowledge sharing from domain experts to data scientists for building nlp models. *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 585–596.

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

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

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

DISSERTATIONES INFORMATICAE PREVIOUSLY PUBLISHED IN DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 lk.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.

74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.

113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

1. **Abdullah Makkeh**. Applications of Optimization in Some Complex Systems. Tartu 2018, 179 p.
2. **Riivo Kikas**. Analysis of Issue and Dependency Management in Open-Source Software Projects. Tartu 2018, 115 p.
3. **Ehsan Ebrahimi**. Post-Quantum Security in the Presence of Superposition Queries. Tartu 2018, 200 p.
4. **Ilya Verenich**. Explainable Predictive Monitoring of Temporal Measures of Business Processes. Tartu 2019, 151 p.
5. **Yauhen Yakimenka**. Failure Structures of Message-Passing Algorithms in Erasure Decoding and Compressed Sensing. Tartu 2019, 134 p.
6. **Irene Teinemaa**. Predictive and Prescriptive Monitoring of Business Process Outcomes. Tartu 2019, 196 p.
7. **Mohan Liyanage.** A Framework for Mobile Web of Things. Tartu 2019, 131 p.
8. **Toomas Krips.** Improving performance of secure real-number operations. Tartu 2019, 146 p.
9. **Vijayachitra Modhukur.** Profiling of DNA methylation patterns as biomarkers of human disease. Tartu 2019, 134 p.
10. **Elena Sügis.** Integration Methods for Heterogeneous Biological Data. Tartu 2019, 250 p.
11. **Tõnis Tasa.** Bioinformatics Approaches in Personalised Pharmacotherapy. Tartu 2019, 150 p.
12. **Sulev Reisberg.** Developing Computational Solutions for Personalized Medicine. Tartu 2019, 126 p.
13. **Huishi Yin.** Using a Kano-like Model to Facilitate Open Innovation in Requirements Engineering. Tartu 2019, 129 p.
14. **Faiz Ali Shah.** Extracting Information from App Reviews to Facilitate Software Development Activities. Tartu 2020, 149 p.
15. **Adriano Augusto**. Accurate and Efficient Discovery of Process Models from Event Logs. Tartu 2020, 194 p.
16. **Karim Baghery.** Reducing Trust and Improving Security in zk-SNARKs and Commitments. Tartu 2020, 245 p.
17. **Behzad Abdolmaleki.** On Succinct Non-Interactive Zero-Knowledge Protocols Under Weaker Trust Assumptions. Tartu 2020, 209 p.
18. **Janno Siim.** Non-Interactive Shuffle Arguments. Tartu 2020, 154 p.
19. **Ilya Kuzovkin.** Understanding Information Processing in Human Brain by Interpreting Machine Learning Models. Tartu 2020, 149 p.
20. **Orlenys López Pintado.** Collaborative Business Process Execution on the Blockchain: The Caterpillar System. Tartu 2020, 170 p.
21. **Ardi Tampuu.** Neural Networks for Analyzing Biological Data. Tartu 2020, 152 p.

22. **Madis Vasser.** Testing a Computational Theory of Brain Functioning with Virtual Reality. Tartu 2020, 106 p.
23. **Ljubov Jaanuska.** Haar Wavelet Method for Vibration Analysis of Beams and Parameter Quantification. Tartu 2021, 192 p.
24. **Arnis Parsovs.** Estonian Electronic Identity Card and its Security Challenges. Tartu 2021, 214 p.
25. **Kaido Lepik.** Inferring causality between transcriptome and complex traits. Tartu 2021, 224 p.
26. **Tauno Palts.** A Model for Assessing Computational Thinking Skills. Tartu 2021, 134 p.
27. **Liis Kolberg.** Developing and applying bioinformatics tools for gene expression data interpretation. Tartu 2021, 195 p.
28. **Dmytro Fishman.** Developing a data analysis pipeline for automated protein profiling in immunology. Tartu 2021, 155 p.
29. **Ivo Kubjas.** Algebraic Approaches to Problems Arising in Decentralized Systems. Tartu 2021, 120 p.
30. **Hina Anwar.** Towards Greener Software Engineering Using Software Analytics. Tartu 2021, 186 p.
31. **Veronika Plotnikova.** FIN-DM: A Data Mining Process for the Financial Services. Tartu 2021, 197 p.
32. **Manuel Camargo.** Automated Discovery of Business Process Simulation Models From Event Logs: A Hybrid Process Mining and Deep Learning Approach. Tartu 2021, 130 p.
33. **Volodymyr Leno.** Robotic Process Mining: Accelerating the Adoption of Robotic Process Automation. Tartu 2021, 119 p.
34. **Kristjan Krips.** Privacy and Coercion-Resistance in Voting. Tartu 2022, 173 p.
35. **Elizaveta Yankovskaya.** Quality Estimation through Attention. Tartu 2022, 115 p.
36. **Mubashar Iqbal.** Reference Framework for Managing Security Risks Using Blockchain. Tartu 2022, 203 p.
37. **Jakob Mass.** Process Management for Internet of Mobile Things. Tartu 2022, 151 p.
38. **Gamal Elkoumy.** Privacy-Enhancing Technologies for Business Process Mining. Tartu 2022, 135 p.
39. **Lidia Feklistova.** Learners of an Introductory Programming MOOC: Background Variables, Engagement Patterns and Performance. Tartu 2022, 151 p.
40. **Mohamed Ragab.** Bench-Ranking: A Prescriptive Analysis Approach for Large Knowledge Graphs Query Workloads. Tartu 2022, 158 p.
41. **Mohammad Anagreh.** Privacy-Preserving Parallel Computations for Graph Problems. Tartu 2023, 181 p.
42. **Rahul Goel.** Mining Social Well-being Using Mobile Data. Tartu 2023, 104 p.

43. **Anti Ingel.** Algorithms using information theory: classification in brain-computer interfaces and characterising reinforcement-learning agents. Tartu 2023, 142 p.
44. **Shakshi Sharma.** Fighting Misinformation in the Digital Age: A Comprehensive Strategy for Characterizing, Identifying, and Mitigating Misinformation on Online Social Media Platforms. Tartu 2023, 158 p.
45. **Kristiina Rahkema.** Quality Analysis of iOS Applications with Focus on Maintainability and Security Aspects. Tartu 2023, 182 p.
46. **Ivan Slobozhan.** Studying Online Social Media Engagement in CIS Countries during Protests, Mass Demonstrations and War. Tartu 2023, 81 p.
47. **Nurlan Kerimov.** Building a catalogue of molecular quantitative trait loci to interpret complex trait associations. Tartu 2023, 248 p.
48. **Pavlo Tertychnyi.** Machine Learning Methods for Anti-Money Laundering Monitoring. Tartu 2023, 117 p.
49. **Abasi-amefon Obot Affia.** A Framework and Teaching Approach for IoT Security Risk Management. Tartu 2023, 180 p.
50. **Raimond-Hendrik Tunnel.** Video Game Design and Development Bachelor's Curriculum for Estonia. Tartu 2024, 137 p.
51. **Ahto Salumets.** Bioinformatics analysis of various aspects in immunology. Tartu 2024, 198 p.
52. **Mohammed Abdulhameed Shaif Ali.** Deep Learning Methods for Cell Microscopy Image Analysis. Tartu 2024, 143 p.
53. **Pille Pullonen-Raudvere**. Foundations of Efficient and Secure Algorithm Development for Secure Multiparty Computation. Tartu 2024, 265 p.
54. **Marili Rõõm.** Multiple approaches to learners' success and factors affecting it in computer programming MOOCs. Tartu 2024, 170 p.
55. **Shivananda Rangappa Poojara.** Design and Orchestration of Scalable, Event-Driven Serverless Data Pipelines for Internet of Things (IoT) Applications. Tartu 2024, 172 p.
56. **Hassan Abdulgaleel Hassan Salim Eldeeb.** Empowering Machine Learning Pipelines with Automated Feature Engineering. Tartu 2024, 121 p.
57. **Muhammad Uzair.** Soft decision making for agri-food 4.0. Tartu 2024, 158 p.