UNIVERSIDADE DA CORUÑA

Depression Severity Estimation on the Internet: New Models and Resources

Doctoral Thesis

Anxo Pérez Vila

2023

Anxo Pérez: *Depression Severity Estimation on the Internet: New Models and Resources* , Doctoral Thesis, Universidade da Coruña, 2023.

Depression Severity Estimation on the Internet: New Models and Resources

Anxo Pérez

Doctoral Thesis / 2023

Advisors:

Álvaro Barreiro

Javier Parapar

PhD in Computer Science

UNIVERSIDADE DA CORUÑA

Álvaro Barreiro García, Professor at the Department of Computer Science of Universidade da Coruña,

and

Javier Parapar López, Associate Professor at the Department of Computer Science of Universidade da Coruña,

HEREBY CERTIFY

that the present Doctoral Thesis entitled ***Depression Severity Estimation on the Internet: New Models and Resources*** , submitted to the Universidade da Coruña by Anxo Pérez Vila, has been carried out under our supervision and fulfills all the requirements for the award of the degree of *PhD in Computer Science with International Mention*.

Álvaro Barreiro García
Advisor

Javier Parapar López
Advisor

A Mamá y Mariña,
sin las cuales esta tesis nunca habría visto la luz (Ni yo tampoco).

A Papá, Javi y David,
muchas veces me he perdido. Por suerte, siempre tengo quien me recuerde
el camino de vuelta.

Como no estás experimentado en las cosas del mundo, todas las cosas que tienen dificultad te parecen imposibles. Confía en el tiempo, que suele dar dulces salidas a muchas amargas dificultades.

— Miguel de Cervantes

Since you are not experienced in the world's ways, all challenging things seem impossible. Trust in time, which often provides sweet endings to many bitter challenges.

— Miguel de Cervantes

ACKNOWLEDGMENTS

The journey that has accompanied this thesis has been a long one. Like every important decision in life, it is not made without a level of uncertainty that, at times, can confuse you. Through the ups and downs, I can affirm that every step was worth it. After all this time, one thing is clear: it has all been possible thanks to the people I have around. I want to start by thanking my advisors, Álvaro and Javi. My father inculcated me from a young age the importance of having good mentors in one's life. Thanks to you, I have truly understood the meaning of that advice. Words cannot express all that you have given me.

To Mom, Dad, and Mariña. Mom and Dad, the path was dark, but you guided me through. Mariña, you have been my best example of how to face your fears and becoming the best doctor in the world. To my grandmother, for filling me with energy during each conversation. To all the members of the IRLab, for forming the best possible group anyone could dream of working with. Alfonso, Eliseo, Manu, Juan, Jorge, Brais, among others, you have made this stage of my life one of the best. Today, I can say that I have formed friendships with incredible people, and I am very grateful for all the hours we have spent and will spend together. A special mention to David, with whom I have shared this journey from day one. Your honesty has always allowed me to see my mistakes, even when my ego closed my eyes. To my friends, for letting me feel their warmth every single day. Being there during the tough times is not easy, but being present during the good times can be just as challenging. Thank you for being there in both scenarios. For supporting me during the lows, and for sharing your joy during the highs. Desi, Andre, Abraham, once again, and it will never be enough.

I wish to express my gratitude to the members of UKPLab in Germany for the great opportunity they offered me during my research stay. Equally, I am deeply thankful to my USC colleagues, David and Marcos, for their consistent support during all our collaborations. I would also like to thank the thesis defence committee, the external reviewers, and all the researchers who have supported me during this stage. Lastly, I also acknowledge the financial support offered by Xunta de Galicia through grant ED481 A 2021/034.

I have lost my way countless times. Luckily, whenever I forgot who I was, I always had someone to remind me.

ABSTRACT

On the one hand, there is extensive evidence from medicine and psycholinguistics fields of changes in language usage from people suffering from mental health problems. On the other hand, social media platforms provide a vast repository of written language. There is a recent trend in computational linguistics where researchers aim to exploit social posts to detect individuals at risk. In this thesis, we follow that line in the field of depression detection. A shortcoming in actual research efforts is the need for more interpretability of the models' decisions. To mitigate that problem, we investigate the development of models based on validated clinical symptoms to identify depressive signs.

The contributions of this thesis are three-fold: (i) new models for depression severity estimation based on symptom markers, (ii) the creation of new datasets for helping the development of new symptom-based approaches, and (iii) the exploration of recent massive large language models for helping with the scaling up of the datasets construction. As a final step, we incorporate the above contributions into a demonstrative platform to be used by health professionals. This thesis contributes to advancing the understanding and detection of depression through symptom markers, and lays the foundation for future research in this critical area of depression detection on social media.

RESUMEN

Existe evidencia proveniente de los campos de la medicina y psicolingüística sobre cambios en el uso del lenguaje de personas que sufren problemas de salud mental. Por otro lado, las redes sociales proporcionan un amplio repositorio de lenguaje escrito. Hay una tendencia reciente en lingüística computacional donde los investigadores buscan explotar publicaciones en las redes para detectar usuarios en riesgo. En esta tesis, seguimos esa línea en el campo de detección de la depresión. Sin embargo, un defecto de estas investigaciones es la necesidad de una mayor interpretabilidad de las decisiones de los modelos. Para mitigar ese problema, investigamos el desarrollo de modelos basados en síntomas validados clínicamente.

Las contribuciones de esta tesis tienen tres enfoques: *i*): nuevos modelos para la estimación de la gravedad basados en marcadores de síntomas, *ii*) creación de colecciones para ayudar al desarrollo de métodos basados en síntomas, y *iii*) la exploración de los recientes modelos masivos de lenguaje para ayudar en la creación de colecciones. Buscando una integración práctica de los modelos de detección de la depresión, incorporamos nuestras aportaciones a una plataforma demostrativa para su uso por parte de clínicos. Esta tesis contribuye a avanzar en la comprensión y detección de la depresión a través de sus síntomas, y sienta bases para futuras investigaciones en el área de la detección de la depresión en las redes sociales.

RESUMO

Hai unha evidencia extensa que provén dos campos da medicina e a psicolingüística sobre os cambios no uso do lenguaxe das persoas que sofren problemas de saúde mental. Por outro lado, as redes sociais proporcionan un amplo repositorio de linguaxe escrito. Existe unha tendencia recente na lingüística computacional onde os investigadores buscan explotar publicacións nas redes para detectar usuarios en risco. Nesta tese, seguimos esa liña no campo da detección da depresión. Porén, un defecto das investigacións previas é a necesidade dunha maior interpretabilidade das decisións dos modelos. Para mitigar ese problema, investigamos o desenvolvemento de modelos baseados en síntomas validados clínicamente para identificar sinais de depresión.

As contribucións desta tese teñen tres enfoques diferentes: i) novos modelos para a estimación da gravidade baseados en marcadores de síntomas, ii) a creación de coleccións para axudar ao desenvolvemento de métodos baseados en síntomas, e iii) a exploración dos recentes modelos masivos de linguaxe para axudar a escalar a creación destes datasets. Como último paso, e na nosa procura dunha integración práctica dos modelos de detección da depresión, incorporamos as nosas aportacións anteriores a unha plataforma demostrativa para o seu uso por parte de clínicos. Esta tesis contribúe a avanzar na comprensión e detección da depresión a través de marcadores de síntomas, e asenta as bases para futuras investigacións nesta área crítica da detección da depresión nas redes sociais.

CONTENTS

i PRELIMINARIES

ii DEPRESSION SEVERITY ESTIMATION BASED ON SYMPTOMS INFORMATION

III RESOURCES AND APPLICATIONS FOR DEPRESSION DETECTION BASED ON SYMPTOM MARKERS

IV CLOSING

APPENDICES

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

BDI-II	Beck Depression Inventory-II
kNN	k-Nearest-Neighbour
IR	Information Retrieval
NLP	Natural Language Processing
LM	Language Model
LLM	Large Language Model
KLD	Kullback-Leibler Divergence
MAP	Mean Average Precision
TF	Term Frequency
IDF	Inverse Document Frequency
SVM	Support Vector Machine
LR	Logistic Regression
BERT	Bidirectional Encoder Representations from Transformers
T5	Text-To-Text Transfer Transformer
LIWC	Linguistic Inquiry and Word Count
GPT	Generative Pre-trained Transformer
eRisk	Early Risk Prediction On The Internet
CLPsych	Computational Linguistic and Clinical Psychology
CLEF	Conference and Labs of the Evaluation Forum
QA	Question Answering
DCHR	Depression Category Hit Rate
DODL	Difference Between Overall Depression Levels
ADODL	Average Difference Between Overall Depression Levels
HR	Hit Rate
AHR	Average Hit Rate
ACR	Average Closeness Rate
CR	Closeness Rate

RMSE	Root-Mean-Square Error
SBERT	Sentence Transformers
BM25	Okapi Best Matching 25
LLM	Large Language Model
PHQ-9	9-Question Patient Health Questionnaire
DSM-V	Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
MDD	Major Depressive Disorder
CSV	Comma-Separated Values
JSON	JavaScript Object Anotation
PoS	Part-of-Speech
LDA	Latent Dirichlet Allocation
API	Application Programming Interface
SQuAD	Stanford Question Answering Dataset
WH	Writing History
KDE	Kernel Density Estimation
AUC	Area Under the Curve
WHO	World Health Organization
ML	Machine Learning
NB	Naive Bayes
RF	Random Forest
BoW	Bag-of-Words
RNN	Recurrent Neural Network
PTSD	Post-traumatic Stress Disorder
CNN	Convolutional Neural Network

Part I

PRELIMINARIES

1

INTRODUCTION

Mental disorders, including depression, are among the most prevalent public health issues. According to the World Health Organization (WHO), approximately 332 million individuals globally are affected by major depressive disorder [1]. Mental health plays a pivotal role in fostering happiness, promoting social interaction, and contributing to individual and population health. Good mental health is a fundamental prerequisite for success in all aspects of life. Moreover, it significantly impacts national output and labour force productivity [2]. Early intervention in depressive disorders is critical in mitigating their impact and consequences (Picardi et al. 2016). However, due to the stigma surrounding mental disorders, more than 60% of individuals affected do not seek professional support (Gulliver et al. 2010), which is particularly concerning considering the increasing number of cases among young people (Thapar et al. 2022). To help with this problem, governments and agencies have launched programs to raise awareness of the importance of mental health in their citizens. Nevertheless, the limited resources of public health systems severely constrain their capacity for case detection and diagnosis (Arango et al. 2018).

As an alternative to public health systems, social platforms are a promising channel to assess risks in an unobtrusive manner (Coppersmith et al. 2015). The proliferation of social media constitutes a valuable resource for detecting early signs of depression. Individuals experiencing depression often find comfort in expressing their thoughts and emotions on these platforms, motivated by factors such as privacy and anonimity (Callahan and Inckle 2012; Kauer et al. 2014). Consequently, social media provides a unique opportunity to access valuable information about individuals' health risks that would otherwise remain impossible to obtain. Researchers in Information Retrieval (IR), Natural Language Processing (NLP) and Machine Learning (ML) have leveraged the vast resources

[1] https://apps.who.int/iris/handle/10665/254610

[2] https://www.who.int/europe/health-topics/mental-health

of social networking to obtain considerable advancements in detecting signs of depression (Ríssola et al. 2021). However, one shortcoming in current research efforts is the need for more interpretability of the models' decisions (Walsh et al. 2020a). In the domain of mental health detection, where reliable interpretation of outcomes is crucial for clinicians, it becomes essential for models to produce trustworthy and interpretable classification outcomes (Ernala et al. 2019).

In line with this objective, this doctoral thesis focuses on developing models based on validated clinical symptoms to identify depressive signs on social media. By incorporating clinical markers into the models' decisions, we aim to improve the interpretability of their outputs by health professionals. For this reason, we adhere to established clinical protocols, considering the 21 symptoms covered in the Beck Depression Inventory (BDI-II), a questionnaire widely used to measure depression. The BDI-II encompasses a range of depressive symptoms such as irritability, pessimism or sleep problems. The BDI-II not only serves as a tool to detect depression but also as a grading tool for severity estimation. In adhering to clinical protocols, we aspire to build predictive models that not only detect but also provide detailed severity estimations, thereby equipping health professionals with robust tools for a complete diagnosis.

1.1 MOTIVATION

Depressive disorders have numerous harmful effects. Nevertheless, there are validated and effective treatments available, which can be boosted with therapies and intervention programs (Duarte et al. 2009). An early and accurate detection significantly reduces the negative impact of the disorder (Halfin 2007; Picardi et al. 2016). In clinical practice, the diagnosis and severity of depression relies on validated psychometric tests. These questionnaires have a satisfactory performance in diagnosing individuals (Smarr and Keefer 2011). Relevant examples are the Patient Health Questionnaire 9 (PHQ-9) (Kroenke et al. 2001), the Center for Epidemiologic Studies Depression Scale (Eaton et al. 2004) or the Hamilton Rating Scale for Depression (Hamilton 1980). Among these, the BDI-II is one of the most widely recognized and reliable instruments, with existing extensive empirical evidence supporting its efficacy (Dozois et al. 1998).

However, self-reporting and family reporting often serve as the primary methods for detecting cases of depressive illnesses (Sanchez-Villegas et al. 2008). Population-level analysis via traditional methods often requires substantial resources. For instance, phone surveys are a common approach that can lead to significant delays in obtaining practical results [3]. For this reason, both public and private health organizations have made these questionnaires available to users for self-completion. In certain instances, online tests based on these questionnaires even provide recommendations for individuals to seek professional medical help according to their scores. Per contra, when aiming for a precise diagnosis, conventional procedures have certain limitations. Beyond the societal stigma associated with mental health problems, which can influence individuals' willingness to provide accurate questionnaire responses, studies have examined how these responses can drastically fluctuate based on variable factors (Cameron et al. 2011). The final scores can be easily manipulated, as they can be minimized or exaggerated. Bowling (2005) studied the quality results' variations based on the administration of these tests. Social expectations, such as doing a test in front of a doctor, would change the results drastically compared to doing it in a friendly environment like your room.

The development of these instruments stems from extensive prior work into understanding the underlying causes of depression. Studies on topics related to depressive conditions have been performed within the medicine and psycholinguistics fields (Campbell and Pennebaker 2003; Rude et al. 2004). All of them have tried to identify the presence of symptoms, causes and how to perform a precise diagnosis. A large body of this research has delved into understanding the connection between language and mental health. Such works underscore the impact that words can have on our emotional state and cognitive states. Pennebaker's pioneering work explored the subtle nuances of language use in daily life, demonstrating that certain patterns of language, such as the frequent use of first-person pronouns, can serve as indicators of an individual's mental well-being (Pennebaker et al. 2003).

Consequently, social media offers a complementary opportunity to obtain valuable information about individuals' mental states, supplementing traditional professional therapy. The combination of computational linguistics with the extensive data derived from social networks has produced significant progress in detecting depression indicators (Garg 2023; Ríssola et al. 2021). Recognizing the great importance of this domain, substantial efforts have been dedicated to creating curated experimental benchmarks (Parapar et al. 2023; Zirikly et al. 2022). These resources have facilitated the development and evaluation of numerous new predictive models, which will be elaborated upon in the related work section.

[3] https://www.cdc.gov/brfss/about/index.htm

While researchers in this field do not aim to replace mental health professionals, they seek to support their work. Clinicians play an indispensable role in validating the predictions made by computational models and taking appropriate action with individuals when necessary. Careful application of computational models can amplify the reach and efficacy of these professionals and facilitate their workflows. Nevertheless, most current models exhibit several limitations in achieving this objective (Walsh et al. 2020a). A significant barrier is their limited capacity to explain their predictions, often resulting in scepticism among professionals. For practitioners to trust and incorporate these models into their daily work, they require reliable interpretations of the models' decisions (Hauser et al. 2022). One approach to address this involves designing new models incorporating trustworthy and reliable explanations (Ernala et al. 2019). Following that path, emerging research has explored using symptoms obtained from validated clinical questionnaires. Most of these proposals, particularly in the field of depression, leverage symptom markers from the BDI-II (Beck et al. 1996b) or the PHQ-9 (Kroenke et al. 2001) inventories, which cover a variety of depressive symptoms such as irritability, pessimism, and sleep disturbances. The application of such symptom markers has been shown to enhance the explainability, generalization, and overall performance of depression detection models (Nguyen et al. 2022a; Zhang et al. 2022a,b).

1.2 AIM AND SCOPE

Our principal aim is to exploit the language used in social media to construct computational models for detecting and estimating the severity of depression. One pre-requisite of the models presented here is that they follow clinical schemas to provide interpretable and practical outputs. For this reason, our solutions leverage user-generated content to develop models that can effectively predict the evidence of depressive symptoms. To do so, we focus on the applicability of techniques within the fields of IR, NLP and ML.

Using an established tool for diagnosing depression, such as the BDI-II, is vital to ensure that our models offer trustworthy diagnostic support. Recognizing the limited resources available for identifying depressive symptoms, our second goal focuses on constructing datasets centred on symptom markers. To achieve this, we leverage the descriptions of the BDI-II to employ various text mining techniques to filter candidate linguistic expressions associated with depressive symptoms. Moreover, we include symptom-by-symptom analyses of our resources, and perform experiments to validate its practical utility in different classification tasks. Parallel to our resource-creation efforts, we also analyze the importance of a robust annotation methodology for constructing resources in this complex domain. Following this idea, we explore the capabilities of recent conversational Large Language Models (LLMs) in dataset creation and augmentation. Bringing all these efforts together, the final part of the thesis introduces these contributions in the form of a demonstrative platform designed for healthcare professionals.

Evaluation plays a crucial role in experimental sciences such as IR and NLP. A reliable evaluation is especially critical in the mental health detection domain since model outcomes can directly influence clinical decisions and overall mental health assessments. In this thesis, our approaches and resources are building upon the foundation laid by a well-known experimental benchmark. Specifically, the Early Risk Prediction on the Internet (eRisk [4]) (Losada et al. 2017, 2018, 2019, 2020; Parapar et al. 2021b, 2022). By evaluating our models on the eRisk collections, we not only ensure a consistent basis of comparison with other leading researchers but also align our work with an established framework known for its reliability and clinical relevance.

1.3 STRUCTURE AND CONTRIBUTIONS OF THE THESIS

This doctoral thesis is divided into four parts with nine chapters. Contribution chapters are meant to be as self-contained as possible. Below, we present the organization of this dissertation in more detail:

PART I The initial part contains three foundational chapters: Chapter 1, which is the introduction to this thesis, and Chapter 2, which presents the related work on the main themes covered in our study. The introduction is structured in three parts: *i*) the context and motivation, *ii*) the aim and scope and *iii*) the structure and main contributions of our work. In the related work, we review the most relevant advancements in the field of depression detection on the Internet and their evolution to the present day. The last chapter (Chapter 3) presents the research methods and experimental guidelines followed in our work. Initially, we contextualize the eRisk experimental benchmark and how it aligns with our proposals, since all our contributions are related to the

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

eRisk framework. Following this, we provide an overview of the tasks and collections that guide our approaches, explaining how our research aligns with them. Concluding this part, we present the main metrics adopted in our contributions, elaborating on how we evaluate the efficacy of our models.

PART II We present here two different classification frameworks that automatically estimate the 21 symptoms of the BDI-II questionnaire. On the one hand, Chapter 4 uses word embeddings to explore the presence of these symptoms depending on their sensitivity. We refer to symptom sensitivity in terms of users' inclination to openly discuss them (i.e., there are more intimate symptoms, and users avoid explicitly talking about them). For this reason, we analyzed the sensitivity of each symptom and devised two different methods to capture the different characteristics that each symptom may have. On the other hand, Chapter 5 uses sentence transformers to create a classification pipeline that estimates depression severity through semantic similarities. In this work, we focus on selecting users' posts related to depressive symptoms by exploring different data selection strategies. Once we selected the most risky posts from the test user, we produce a semantic ranking that gives us training labelled sentences. These training sentences are previous labelled and we know they are associated with depressive symptoms. Subsequently, we utilize the sentences derived from these rankings as evidence for predicting symptom severities.

PART III First, this part contains the work related to the construction of symptom-based resources for depressive disorder. We introduce two pivotal resources: *BDI-Sen*, discussed in Chapter 6, and *DepreSym*, explained in Chapter 7. Both resources comprise symptom-annotated sentence datasets for depression, offering manual annotations related to the 21 symptoms covered in the BDI-II. In Chapter 6 (BDI-Sen), we begin by describing the retrieval strategy we employed to obtain candidate sentences for annotation. Subsequently, three assessors decided on the actual relevance of the candidates. Progressing further into this part, we provide an in-depth analyses of this resource, studying the linguistic style, emotional attributes and other psycholinguistic markers of the sentences. Additionally, we conducted a series of experiments investigating the utility of BDI-Sen for various tasks, including the detection and severity classification of symptoms. Finally, we also examine their generalization when considering symptoms from other mental diseases.

In Chapter 7 (DepreSym), we study alternative forms of symptom labelling. For this purpose, we leverage the eRisk 2023 ranking tasks, which is centred on developing ranking methods to find sentences associated with depressive symptoms. The construction of DepreSym is based on the ranking methods of the task participants. In this case, the labelled sentences come from a pool of diverse ranking methods, and the final candidate sentences were obtained using top-k pooling from them. Due to the complex nature of the relevance annotation, we designed a robust assessment methodology carried out by three expert assessors. To validate the effectiveness of this methodology, we calculate the inter-rater agreement and conduct further analysis of the resulting set of judgements. Additionally, we also explore the feasibility of employing recent conversational LLMs (ChatGPT and GPT-4) to assist in this complex task. We undertake a comprehensive examination of their performance, determine their main limitations and analyze their role as complement or replacement for human annotators.

Finally, in Chapter 8, we introduce PsyProf, a demonstrative platform designed for the task of assessing depression severity. The platform is intended to be used by health professionals to demonstrate effective depression screening capabilities. We integrate in this platform the previous models that estimate the presence of BDI-II symptoms. Moreover, we have complemented our tool with user profiling methods to bring wider context when measuring at-risk users. Finally, we also included the functionality for collecting the data from social media users, which can help to create symptom-based datasets with the inspection coming from health professionals.

PART IV Concluding this dissertation, we present the primary conclusions of our research and discuss the potential direction for future work. Additionally, we look into the ethical considerations and challenges surrounding the detection of mental health indicators on social media.

2

RELATED WORK

In this chapter, we will review the most relevant work and advancements in the field of depression detection on the Internet. We begin by describing the initial efforts of this research domain, with a strong focus on detecting patterns in language use. Subsequently, we will discuss the main methods and techniques employed for mental health assessment and their evolution to the present day.

2.1 THE DIGITAL FOOTPRINT OF MENTAL HEALTH

Language has long been investigated as a mirror reflecting our mental and emotional well-being. Before we started sharing our lives on social media, experts in psychology and medicine explored that the words we choose can offer indicators about our mental well-being. People who are depressed, for instance, might talk about the world in a particular way, choosing certain words or phrases more often than others.

Historically, expert clinicians analyzed these subtle changes in speech or writing during therapy sessions or clinical interviews. Pioneering work by Pennebaker et al. (2003) revealed that words people use can indeed provide clues about their emotional state. By analyzing personal diaries, he found patterns in language that connected with mental health problems in people's lives. Following this idea, clinical practitioners relied on patients' words as essential tools for diagnosis and therapy. However, the proliferation of social media platforms in the last years has amplified this connection between language and mental health. The vast digital footprints left behind by individuals on these platforms provide an unprecedented corpus of linguistic data.

People started expressing themselves more openly and frequently on social platforms. Traditional research about social media and mental health activity has thoroughly addressed how the users' self-disclosure of their illnesses and symptoms affects their mental health. Many studies

Self-disclosure refers to any communication about oneself communicated by an individual to others.

have analyzed the benefit of self-disclosure on social media users (Hayes et al. 2016; Luo and Hancock 2020) . Mental health-related self-disclosure on social media may help users perceive higher levels of social support and reduce psychological distress (Seo et al. 2016). In addition, positive disclosure tends to produce more positive feedback for the community, improving the connectedness feelings (Metzler and Scheithauer 2017). Moreover, people being authentic in the disclosure shows improvements in self-esteem (Yang et al. 2017). These and other reasons (like anonymity (Andalibi et al. 2017) or reduced perception of vulnerability (Lin et al. 2021) explain why many social media users are prone to talk about their inner selves in those public forums.

Computational researchers saw an opportunity in this vast amount of information available. A growing body of work has revealed that there are indeed specific linguistic markers and patterns of online behaviour that are indicative of mental health issues. These patterns, often unnoticed in daily conversations, become more evident when analyzed with computational tools over large datasets. From the frequency and timing of posts to the choice of words and sentiments expressed, computational studies have explored numerous nuances in the online behavior of individuals. Next, we are going to delve into some of these pivotal studies.

De Choudhury et al. (2013a,b, 2016) have conducted extensive studies on how social media data can be leveraged to understand and potentially predict mental health conditions, including depression and suicidal ideation. Their pioneering work introduced linguistic and social network analysis into traditional machine learning classifiers, thereby developing systems capable of predicting online risks. In the first study on depression and social media (De Choudhury et al. 2013a), the authors used Twitter data to explore features such as text content, posting frequency, and user interactions. To obtain the data, they used crowdsourcing on Twitter users who reported having been diagnosed with depression, with 476 users. They observed a higher frequency of words related to emotional states commonly associated with depression, such as sadness and loneliness. Additionally, the language patterns suggested fewer social interactions and greater use of past-tense verbs, possibly indicating regrets or reflections about past events.

In a subsequent study (De Choudhury et al. 2016), authors used Reddit data to identify shifts in mental states that could lead to suicidal ideation. This work exemplifies how computational techniques can be applied to gain deeper insights into mental health risks. The study observed notable changes in linguistic style indicative of suicidal thoughts, including increased use of first-person pronouns and a more negative emotional tone. Specifically, they found a rise in terms related to emotional pain, self-harm, or existential dread as individuals moved closer to expressing suicidal ideation. Both studies concluded that predicting mental risks through social media is feasible and aligned well with existing psychological literature. Moreover, they demonstrated the potential for early diagnosis and intervention via non-intrusive means, emphasizing the potential of this line of research.

Coppersmith's research has also influenced machine learning and natural language processing applications in mental health studies (Coppersmith et al. 2015; Loveys et al. 2018b). In one of their first works (Coppersmith et al. 2015), they analysed Twitter data to identify linguistic markers related to four distinct mental health conditions: post-traumatic stress disorder (PTSD), depression, bipolar disorder, and seasonal affective disorder (SAD). To identify users for each condition, the researchers relied on online self-disclosures, such as tweets stating, *"I was diagnosed with depression."* Meanwhile, the control group was randomly selected. Three primary methods were employed in the study to quantify linguistic differences: 1) Linguistic Inquiry and Word Count (LIWC) categories, 2) activity patterns on Twitter (e.g., number of mentions, tweets per day, and interactions), and 3) language modelling. The LIWC (Pennebaker et al. 2001) is a text analysis software categorising words into psychologically meaningful groups. It can measure the frequency of words that fall under specific categories such as emotional tone, cognitive processes or social references (e.g., friends, family). The study found that language models outperformed the other methods, suggesting that the complexity of language holds more nuanced signals than what is captured by LIWC or activity patterns alone. Combining the two types of features, language models and LIWC, further improved the classification results. On the analytical front, the study identified linguistic terms associated with emotional states like anger, swearing, and anxiety more frequently in tweets from individuals with depression. Interestingly, Twitter activity patterns were less effective than the other methods in distinguishing users with mental health conditions from the control group.

An important yet often overlooked aspect in the study of linguistic markers for mental health conditions is the role of demographic variables, particularly culture and gender. These factors can profoundly influence how mental illnesses are perceived and manifested. Language, as a vehicle of culture and identity, can reveal these nuances in attitudes towards mental health and even the symptoms themselves. De Choudhury and Coppermisth have contributed to recognizing these differences by examining their data across different cultures and genders (De Choudhury et al. 2017; Loveys et al. 2018b).

In De Choudhury et al. (2017), the authors analyzed tweets from users who had self-reported some form of mental illness. They divided the users into two groups: 'Western' (comprising users from the United States and the United Kingdom) and 'Non-Western' (users from South Africa and India). For linguistic analysis, they employed both LIWC and topic modelling methods. The study found that 'Non-Western' users tended to inhibit their depression, using more positive and fewer negative terms compared to their 'Western' counterparts. Furthermore, 'Western' users were more expressive about their emotional processes and social experiences in their language. When considering gender, the study revealed that women exhibited greater sadness and anxiety in their language and showed a higher concern for social and family issues.

Loveys et al. (2018b) conducted an exploratory analysis to investigate linguistic differences in the expression of depression among various cultural groups within the United States. The study specifically compared the language used by White, Asian, Pacific Islander, Black or African American, and Hispanic or Latino individuals discussing their mental health on an online support forum. Consistent with Choudhury's findings, Coppermisth also observed cross-cultural variations in the online language used by individuals with depression. Asian and Pacific Islander users tended to inhibit the expression of negative emotions, while White and Black or African American users displayed more negativity in their language. Hispanic and Latino users exhibited a wide range of positive and negative emotions compared to other groups. This made the effects of depression less evident among Asian and Pacific Islanders. With regard to topic modelling, the study did not find many differences across cultures. For instance, the 'friendship' topic was prevalent across all cultural groups, suggesting that loneliness or the 'need for a friend' is a common concern for individuals with depression.

2.2 KEY FEATURES

In this section, we will comment the pivotal extracted features commonly employed for building mental health detection systems. Most traditional methods in the field employed linguistic features, as the ones commented previously. Specifically, these features often serve as input for standard machine learning classifiers such as Support Vector Machine (SVM) (Jamil et al. 2017; Ortega-Mendoza et al. 2018), Naive Bayes (NB) (Sadeque et al.

2017; Villegas et al. 2017), Random Forest (RF) techniques (Almeida et al. 2017; Cacheda et al. 2018) and Logistic Regression (LR) (Preoţiuc-Pietro et al. 2015; Ramiandrisoa et al. 2018).

Lexicon-based approaches have always been applied due to their straightforward yet effective methodology. These approaches rely on counting the frequency of words that belong to predefined, manually-created categories, linking them to psychological variables. Several lexicons and dictionaries have been used to facilitate this connection. The LIWC are the most commonly used categories for this purpose, both in English (Guntuku et al. 2017; Trotzek et al. 2018) and also in other languages such as Spanish (Ramírez-Cifuentes et al. 2020). Other works also complemented LIWC with emotion (Uban and Rosso 2020b) or depression-specific dictionaires (Al-Mosaiwi and Johnstone 2018). For instance, Nguyen et al. (2014) employed the ANEW lexicon (Bradley and Lang 1999) to extract sentiments conveyed in the publications. It consists of 1034 words rated in terms of valence and arousal, suitable for quantitative estimation. The categories provided by these lexicons have also been explored for building topic modelling techniques to detect depression risks (Resnik et al. 2013; Schwartz et al. 2014).

While lexicon-based approaches focus on the psychological categorization of individual words, other research employed more text-agnostic models like Bag-of-Words (BoW) and n-grams to capture the nuances of language usage. The Bag-of-Words model represents text data by counting the frequency of each word in a document, disregarding the order. This provides a high-dimensional vector, sparse representation that is particularly effective for tasks that capture the overall thematic content.

On the other hand, n-grams extend this concept by considering sequences of 'n' consecutive words or characters. For example, bi-grams (two-word sequences) and tri-grams (three-word sequences) are often used to capture common phrases that may offer deeper insights into the individual's psychological state. All of their publications are aggregated into a single document to create these high-dimensional vectors that represent a user's language. Given the substantial size of the vocabulary, these methods often utilize word filtering techniques such as eliminating stopwords or restricting the vocabulary to words from specific lexicons to form the vectors that represent the user (Almeida et al. 2017; Cacheda et al. 2018; Coppersmith et al. 2014a; Oliveira 2020; Ortega-Mendoza et al. 2018; Trotzek et al. 2018).

Metadata associated with user profiles and their online activity can also be instrumental in identifying mental health risks. Researchers have begun to explore a variety of metadata attributes, such as posting frequency, time of posts, and user interaction metrics, to supplement language-based analyses. Posting activity, for instance, could reveal patterns indicative of insomnia or increased anxiety (Almeida et al. 2017; Cacheda et al. 2018; Tsugawa et al. 2015). The frequency and nature of interactions with other users can also offer clues into social withdrawal or emotional states (Coppersmith et al. 2014a; De Choudhury et al. 2013b; Park et al. 2015; Preoţiuc-Pietro et al. 2015). The length of posts, often measured in terms of characters or words, can be another insightful feature. Long, detailed posts may indicate a higher level of engagement or emotional investment, while shorter posts might suggest a more reserved or cautious approach to sharing personal information (Cacheda et al. 2019; De Choudhury et al. 2013a; Ramiandrisoa et al. 2018; Trotzek et al. 2018). By incorporating metadata, researchers can more accurately model the complexities mental health disorders, often improving the predictive performance of their machine learning algorithms (Coppersmith et al. 2014b).

As machine learning techniques continue to evolve, the field has witnessed a shift towards more advanced deep learning methods that capture the language in a different manner. Unlike traditional features that treat each word independently, deep learning approaches often encode posts as sequences to capture the semantic and contextual relationships between words. One of the most prevalent techniques is word embeddings, where each word or phrase from the vocabulary is mapped to vectors of real numbers, encapsulating much richer semantic information (Orabi et al. 2018).

Orabi et al. (2018) pre-trained optimised word embedding models in a depression collection from CLPsych to learn better features representation of health-specific tasks. For this optimisation, their Word2Vec model not only predicts words but also the possible disease related to that word (i.e., depressed, Post-traumatic Stress Disorder (PTSD) or neither). In particular, they explored convolutional and recurrent neural networks (CNNs and RNNs) based on different word embedding models and compared their performance. The experiments showed that the best deep learning models were the RNN-based ones. Moreover, optimising the word embedding models showed good generalisation abilities on a different dataset.

In Trotzek et al. (2018), the authors addressed the task of early detection of depression using various word embedding models and compared their performance against linguistic metadata features. In this study, the researchers utilized data from the eRisk 2017 edition, specifically from the depression task, for training and evaluation. For linguistic features, they employed LIWC categories along with an additional set of features, such as text length, title length, and terms related to depression, anxiety, therapy, or diagnosis. In terms of deep learning techniques, they evaluated a Convolutional Neural Network (CNN) that utilized different word embeddings, including pre-trained Word2Vec (Mikolov et al. 2013) and GloVe (Pennington et al. 2014) embeddings. They also trained a custom Word2Vec model using Reddit data from 2007 to 2015. For the classification task, CNNs were deployed with word embeddings serving as the input. Additionally, metadata features were used in conjunction with a logistic regression classifier. Interestingly, the metadata features outperformed the word embeddings. An ensemble approach that combined both methods yielded the best results, suggesting that using ensembles comprising more than two models with different objectives holds promise.

Following the use of word embeddings, the advent of transformer architectures has marked another milestone, offering even more sophisticated ways to encode the syntactic and semantic structures inherent in natural language. These architectures have made a significant impact in the domain of mental health research, providing nuanced understanding and improved classification performance. Initially introduced by Vaswani et al. (2017) in the paper 'Attention is All you Need', the transformer architecture has revolutionized the field of NLP. Transformers utilize self-attention mechanisms to process each token in relation to all other tokens in the input sequence simultaneously. This enables the model to capture long-range dependencies and relationships between words or sub-words. One of the most significant advancements brought about by transformers is the notion of 'contextualized embeddings.' Traditional word embeddings like the ones we commented earlier (Word2Vec (Mikolov et al. 2013) and GloVe (Pennington et al. 2014)) represent each word with a static vector, meaning the same word has the exact representation regardless of its context. In contrast, contextualized embeddings provide a dynamic representation of each word based on its surrounding context. This allows for a more nuanced understanding of word meanings, as the same word can have different representations based on how it is used, making it highly beneficial for tasks that require understanding the complexity of human language, including mental health analysis.

In Jiang et al. (2020), the authors studied eight mental health conditions, gathering datasets through self-identification methods on social media platforms. This research is one of the first works that leveraged contextual representations for mental health detection, comparing its efficacy with a baseline logistic regression model trained on LIWC features. The study introduced an attention-based model utilizing BERT representations for input, and a REALM-like model (Guu et al. 2020), influenced by the recent progress in open-domain question-answering. All models were designed for performing user-level classification tasks, determining whether an individual has a specific diagnosis by evaluating an aggregated representation of their posts. The classification models based on transformers demonstrated that they outperformed LIWC by a large margin. Moreover, the results showed that linguistic traits for mental health detection are more easily recognized at the user-level and thus, aggregating post-level signals can be helpful for a more accurate diagnosis.

In the last years, the majority of research in mental health detection has shifted towards leveraging transformer models, underscoring their adaptability and efficacy for the mental health domain (Martínez-Castaño et al. 2020; Pérez et al. 2023b; Uban and Rosso 2020b). Even Large Language Models (LLMs) were specifically pre-trained in diverse collections related to mental health forums to help the research community (Ji et al. 2022). This evolution in model design is evident in works proposed to the leading experimental benchmarks in this domain, which will be elaborated upon in the following Subsection.

2.3 EXPERIMENTAL BENCHMARKS IN MENTAL HEALTH DETECTION ON SOCIAL MEDIA

As the field of mental health detection via social media has grown in popularity, the research community has produced experimental benchmarks. These initiatives serve as collaborative platforms, enabling researchers to innovate and develop effective methods for identifying and assessing online risks. By offering task definitions, test collections, and standardized evaluation methodologies, these benchmarks facilitate the comparison of computational approaches on a global scale.

In this context, eRisk (Losada et al. 2017, 2018, 2019, 2020; Parapar et al. 2021b, 2022, 2023) and the Computational Linguistics and Clinical Psychology (CLPsych [1]) (Goharian et al. 2021; Loveys et al. 2018a; Niederhoffer et al. 2019; Zirikly et al. 2022) are the two most popular experimental frameworks in the field. Both initiatives have operated as distinct workshops within broader conferences. Both specialized benchmarks are dedicated to fostering interdisciplinary collaboration between computational linguistics, information retrieval and clinical psychology researchers. Through the years, they have presented a diverse range of methodologies and innovative solutions, encouraging the incorporation

[1] https://clpsych.org

of features that we discussed before: linguistic features, social network analysis, and other computational strategies into studying and treating mental health issues. The outcomes of the research presented at these workshops offer novel insights into mental health diagnostics and set the benchmark for future work in this rapidly evolving interdisciplinary field.

The CLPsych and the eRisk initiatives both serve as important platforms for research on mental health detection and risk assessment. However, they differ in a few aspects. CLPsych addresses the intersection of clinical psychology and computational linguistics in a wide array of topics, not just risk assessment. On the other hand, eRisk focuses on early risk prediction and is more centred on computational methods for detecting specific types of mental health disorders such as depression, pathological gambling or anorexia. While CLPsych may involve multiple data sources, including clinical records and social media posts, eRisk primarily targets user-generated online data (all the tasks proposed use Reddit as data source). In summary, while both initiatives have made significant contributions to mental health research, they differ in focus, methodologies, and community engagement, complementing the other in enriching the landscape of computational approaches to mental health.

Next, we will delve into the major tasks and contributions presented in the context of the eRisk. This focus aligns closely with our own research objectives, as the models and resources we introduce in this dissertation are grounded within the eRisk framework.

2.4 ERISK INITIATIVE

The eRisk initiative is part of the Conference and Labs of the Evaluation Forum (CLEF) evaluation campaign, which contributes to the systematic evaluation of information access systems, primarily through experimentation on shared tasks. eRisk explores the evaluation methodology, effectiveness metrics and practical applications related to health and safety on the Internet.

eRisk organizers propose different tasks every edition, each one focusing on a specific aspect of social media risk assessment. These tasks are designed to explore different dimensions of risk, providing a comprehensive evaluation framework for researchers and practitioners in the field. Starting with a pilot task in the 2017 edition (Losada et al. 2017), which focused on giving early alerts of depression tendencies, eRisk has been continued with tasks covering a wide range of mental illnesses. We can find challenges covering diseases such as eating disorders, pathological

Table 2.1: Depressive levels related to the BDI-II score.

Depression level	BDI-II Score
Minimal depression	(0-9)
Mild depression	(10-18)
Moderate depression	(19-29)
Severe depression	(30-63)

gambling or self-harm. One particular value of eRisk tasks is that they introduce additional challenges beyond binary classification scenarios, as organizers recognize the complexities associated with mental health-related content in social media. For instance, many eRisk tasks involve intensity or severity estimation, where participants are required to quantify the severity of mental health concerns. By incorporating this scenario, researchers are encouraged to develop models that capture more nuanced mental health risks.

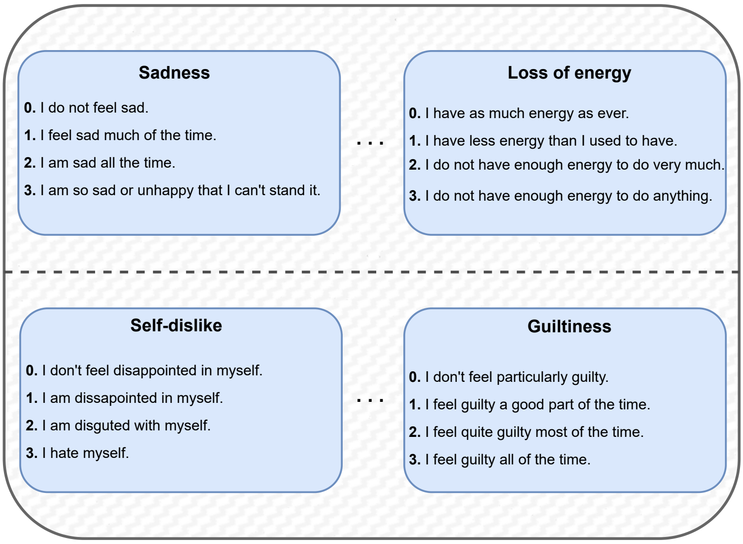
2.5 ERISK TASK: MEASURING THE SEVERITY OF THE SIGNS OF DEPRESSION

In the recent three editions of eRisk (2019, 2020 and 2021), a new task came up aimed at estimating the level of depression, called *Measuring the Severity of the Signs of Depression*. The Chapters 4 and 5 are directly related with this task, as we present a variety of methods under this experimental benchmark. The remaining chapters are also strongly related, describing methods and resources focused on identifying BDI-II symptoms.

The task aims to estimate the level of depression based on users' publications from social media. The levels correspond to the BDI-II (Beck et al. 1996a) questionnaire scores. The participants were given their history of posting for all the users, and they had to predict the BDI-II responses for each user. The history of each user covered a total of 2 years of publications. The predicted answers are based on the evidence found in the history of publications. Therefore, for each user, the collection provides two main elements: (1) their real responses to the symptoms of the BDI-II (ground truth) and (2) their entire Writing History (WH).

BDI-II [2] is a self-report instrument designed to assess the severity of depressive symptoms in adolescents and adults. It consists of 21 symptoms

[2] The full BDI-II can be consulted at `https://erisk.irlab.org/2019/` (Task 3)



Sadness

0. I do not feel sad.
1. I feel sad much of the time.
2. I am sad all the time.
3. I am so sad or unhappy that I can't stand it.

. . .

Loss of energy

0. I have as much energy as ever.
1. I have less energy than I used to have.
2. I do not have enough energy to do very much.
3. I do not have enough energy to do anything.

Self-dislike

0. I don't feel disappointed in myself.
1. I am dissapointed in myself.
2. I am disguted with myself.
3. I hate myself.

. . .

Guiltiness

0. I don't feel particularly guilty.
1. I feel guilty a good part of the time.
2. I feel quite guilty most of the time.
3. I feel guilty all of the time.

Figure 2.1: Four BDI-II symptoms and their corresponding options.

that measure attitudes and symptoms of clinical depression (Beck et al. 1996b; Steer et al. 1986). Each symptom contains four answer options accompanied by a sentence explaining its meaning. The options are rated from 0 to 3 according to the Likert scale (Joshi et al. 2015). Options scale in terms of severity, from the total absence of the symptom to a total identification. It takes approximately 10 to 15 minutes to complete it. BDI-II has shown high internal consistency, and its validity has been established through extensive research.

Figure 2.1 shows four BDI-II symptoms and their possible options. The symptoms are: *Sadness, Loss of energy, Self-dislike and Guiltiness*. The accumulating result of all the 21 symptoms is associated with a scale of depression manifestation. Table 2.1 shows these levels, where the maximum score is 63 (i.e., all the answers from the questionnaire are replied with a 3).

In what follows, we highlight the primary contributions that were submitted to this shared task. The first edition in 2019 did not provide participants with training data. Consequently, a majority of participant teams leaned towards rule-based solutions. Several participants, as noted by Abed-Esfahani et al. (2019) and Trifan and Oliveira (2019), utilized hand-crafted features based on LIWC and depression-specific lexicons combined with logistic regression classifiers. The classification of risks associated with BDI-II symptoms was based on the presence and intensity of these identified features.

For example, Trifan and Oliveira (2019) proposed a rule-based approach, comprising the 21 symptoms into six broader categories. For each category, they generated estimated responses by combining various features, including the prevalence of specific lexical categories in posts and the use of personal pronouns, among others. In contrast, Abed-Esfahani et al. (2019) obtained external training data and designed a supervised machine learning methodology. Their feature extraction relied on embeddings generated by the GPT model. Subsequent vectorial representations of users integrated with both lexical and semantic features, facilitating a comparative analysis with BDI-II responses. Pioneering the adoption of word embeddings in this task was (Rijen et al. 2019), who employed GloVe word embeddings to compute the semantic similarity between symptom options and user publications.

In the next edition, 2020, the participating teams were provided with the data from 2019 for training their models. As a result, we could see more works related to designing supervised learning methods. Although some work continued to address the problem as a rule-based approach, most models elaborated on supervised learning methods as training data became available. Trifan et al. (2020) obtained the best results at the symptom-level using a rule-based approach. The authors captured different psycholinguistic patterns and behavioural features to model each rule. More specifically, they represent the user's writings history as a vector of several depression-specific features. Some examples are guilt emotions, sleep or irritation. For each category, they calculated a user score using the frequency of the categories for that user concerning the total number of occurrences over the training dataset.

These scores were then normalized to the interval [0-3] of the BDI-II. In particular, one common approach was to design a multi-classification task (Martínez-Castaño et al. 2020; Uban and Rosso 2020a), where they trained one individual classifier for each BDI-II symptom, with one class per option available. Martínez-Castaño et al. (2020) were the first in using a BERT-based classifier trained explicitly for the task. The authors fine-tuned the base language model with a head for multi-classification for every symptom. They also balanced the classes' weights due to the training data's sparsity. In inference, their system predicts the answer for a given user, obtaining the softmax prediction for every publication. The class with the highest accumulated value is the estimated answer by the system. They experimented with different LLMs, and RoBERTa achieved the greatest performance.

In the 2021 edition, participants had access to the training data from the two prior years. This year, there was a predominant shift towards fine-tuned LLMs in submitted works. Notably, the top three articles deployed BERT-based classifiers (Basile et al. 2021; Shih-Hung and Qiu 2021; Spartalis et al. 2021). The work by Basile et al. (2021) had the best results. They trained multiple neural models supplemented with data annotated by psychologists, adhering to the DSM-V schema. Their BERT-based approach adopted a two-stage training methodology. Firstly, they curated posts from subreddits addressing mental health issues, including depression, self-harm, and anxiety, and then trained a classifier to distinguish these topics. In the subsequent phase, the authors extracted the [CLS] embedding for each post and computed the probability associated with the depression category. While the former served as the primary representation for classification tasks, the latter provided an insight into post relevance. Leveraging this representation, they trained a classifier for each of the 21 symptoms. Spartalis et al. (2021) used semantic features with Sentence Transformers (SBERT) to extract one dense representation per training user. These user vectors are then fed to standard ML classifiers, such as Linear SVM (Mammone et al. 2009) and Naive Bayes (Berrar 2018).

2.6 THE INTEGRATION OF CLINICAL SYMPTOMS IN NLP FOR MENTAL HEALTH

As the eRisk task centred on predicting depression via BDI-II symptoms gained traction, the use of symptoms for developing depression detection models gained more and more attention. Instead of solely relying on vast amounts of unstructured data, researchers started to see the merit in structured clinical questionnaires and symptom lists. This combination of traditional clinical wisdom with computational techniques promised greater accuracy and a richer context in understanding mental health through the lens of NLP. For this reason, apart from the eRisk proposals, a recent line of work focused on developing models that integrate depressive symptoms as reliable clinical markers. Recent studies have explored the creation of symptom-based prediction models for signs of depression. These models showed the importance of presenting reliable depression markers to aid health professionals in their diagnosis (Coppersmith et al. 2018).

Many of these studies employed Large Language Models (LLMs) as the base for their classifiers (Nguyen et al. 2022a; Zhang et al. 2022a,b). For example, Zhang et al. (2022a) introduced a psychiatric scale-guided method to screen risky posts related to the dimensions defined in clinical depression questionnaires. By using depression templates, which come from BDI-II, they obtained direct expressions of depressive symptoms. They optimized their approach by filtering out irrelevant posts, concentrating on those that align with the depression templates. A hierarchical network incorporating BERT further aggregates the selected posts of a user, and assigns higher weights to important contents related to depressive signs. To measure the similarity between posts and depression templates, authors used pretrained sentence transformers to get the sentence representations and calculate the cosine similarity.

A Hierarchical Attentional Network equipped with BERT (HAN-BERT) is proposed to advance in explainable predictions. In this work, the authors showed in different qualitative examples how leveraging the attention weights can serve as explanations accompanying the model predictions. Moreover, they established a methodology for early risk prediction, where they proposed an online algorithm based on an evolving queue of risky posts that can significantly reduce the number of model inferences to boost efficiency. In an early risk detection scenario, we need to incrementally make predictions each time a user posts instead of processing the whole posting history once. If the models' predicted probability exceeds a predefined threshold, it will report an early alert of depression and stop further calculations.

Their Hierarchical Attentional Network with BERT (HAN-BERT) was designed to provide enhanced, explainable predictions. The attention weights in their models were leveraged as valuable insights into the reasoning behind the predictions. Additionally, they designed a methodology for early risk detection, devising an online algorithm premised on a queue of evolving risky posts. This setup reduced the computational overhead drastically, especially in scenarios where predictions are sought each time a new post emerges. Their system sends an early depression alert if the predicted probability surpasses a predefined threshold.

In terms of evaluation, the eRisk2017 dataset served as their experimental benchmark. Zhang et al. (2022a) compared their proposal against various baselines, from traditional models like LR with TF-IDF and feature-rich models leveraging LDA topic distributions to neural baselines such as HAN-GRU. Their research showcased that screening risky posts and leveraging LLMs formed a potent combination for accurate predictions. Moreover, their framework's compatibility with early detection scenarios while maintaining high efficiency underscored the potency of their evolving queue algorithm in minimizing model inferences.

In a recent contribution, Nguyen et al. (2022a) also explored BERT-based methods using symptom classifiers and compared them against a standard depression classifier. They employed the nine symptoms from the PHQ-9 (Cameron et al. 2011) to design their symptom classifiers. Given the lack of datasets related to depressive symptoms at sentence level, they collected regular expression patterns and heuristics to construct weakly-supervised training data based on the Reddit platform.

Their approach consisted of two simple yet effective models: a questionnaire model trained to detect PHQ-9 symptoms and a broader depression detection model. The former relied solely on manually crafted patterns, while the depression model makes classification decisions by counting how often these patterns appear in a user's post. They compared this approach with a BERT classifier that served as an unconstrained depression detection model. Furthermore, they refined the questionnaire model, leveraging BERT-based symptom classifiers that utilized the manual patterns by considering symptom representations rather than counts.

They found that their constrained models perform competitively compared to a standard unconstrained BERT classifier when trained and evaluated on the same dataset. Moreover, these models can be more easily understood in terms of the presence of depressive symptoms. Their performance was also extended to a *dataset-transfer* evaluation, covering three different datasets (RSDD (MacAvaney et al. 2018), eRisk2018 (Losada et al. 2018) and the TRT corpus (Wolohan et al. 2018). These classifiers performed well compared to the standard depression classifier while generalizing better to other datasets.

Similarly to Zhang et al. (2022a), the authors found that when leveraging the weights from the attention architecture, these symptom classifiers provide a model that can highlight specific posts based on relevant symptoms, improving their interpretability. In a separate experiment focused on evaluating the performance of the symptom classifiers, they found that even though they were trained on weak labels, the symptom classifiers showed a good performance. However, performance variations were observed based on the specific symptom under evaluation.

A predominant limitation in the existing literature on symptom detection is the lack of quality training data. The absence of large-scale annotated corpus made the research effort to rely on unsupervised or weakly supervised methods, depending on pattern matching. As we commented, these methods have demonstrated potential, but they inevitably fall short of the precision and reliability that a well-annotated large corpus can provide.

In another recent contribution to this new trend of symptom-based models for mental health detection, Zhang et al. (2022b) introduced the PsySym dataset. This dataset is the first annotated symptom sentence dataset that covers multiple mental disorders, including depression. It includes manual annotations of 38 symptoms from 7 mental disorders. The authors established the symptom classes according to the DSM-V, with symptom descriptions on clinical questionnaires as supplementary.

First, the authors searched for candidate posts to annotate the symptoms, where they used the Reddit platform as data source. They only selected candidates from mental health-related subreddits, where most posts are likely to be relevant. Moreover, they leveraged embedding-based retrieval methods (Reimers and Gurevych 2019) instead of keyword matching to get the candidate sentences for annotation, aiming for a concise yet diverse selection of posts conducive to efficient annotation.

Once they obtained their final candidate publications, the annotation process was executed via crowdsourcing, which included professional psychiatrists. This collaborative effort included training sessions on annotation protocols and demonstrations of exemplary posts through virtual meetings. Advancing in their research, the authors also proposed a methodology for mental health detection using models trained on PsySym. This involved training symptom classifiers, with their predicted probabilities serving as the base for the symptom features, thereby constructing a 38-dimensional symptom feature vector (one per symptom they cover in the dataset).

To compare their proposed methods, they used two types of baselines: textual features with TF-IDF along with a logistic regressor classifier, and a pre-trained BERT for classification. Their proposed methods that considered symptom features outperform all pure-text methods, including the solid BERT model, suggesting the usefulness of symptom features for mental disease detection on social media. Moreover, their research also highlighted the potential of symptom-based interpretations to help with diagnostic precision and classification.

3

EVALUATION

This chapter presents and contextualizes the experimental guidelines followed in this doctoral thesis to evaluate our models. Through this thesis, we present novel methods and resources that contribute to social media risk assessment, building upon the foundation laid by eRisk initiatives. Our research aims to address the challenges and limitations faced by current approaches by proposing innovative algorithms and different resources. The decision to conduct our research using eRisk collections stems from a combination of factors. First, these collections provide a curated, validated and diverse set of data that captures the challenges inherent in mitigating risks associated with social media content. Second, by participating in this collaborative environment, we obtain a fair comparison, enhancing the rigor and validity of our research findings.

Next, we provide a detailed overview of the eRisk datasets that guided the approaches presented in this thesis, which were based on the eRisk task of depression severity estimation presented in the previous chapter. Additionally, we provide descriptions into the evaluation metrics employed in different chapters of the thesis.

3.1 DATASETS

For experimental purposes, this thesis uses the datasets provided by eRisk 2019, 2020 and 2021 editions (Losada et al. 2019, 2020; Parapar et al. 2021b). eRisk organizers contacted with users from Reddit. Reddit is an open-source platform where members can submit content such as links, text or images. With the users' agreement, they obtained their real responses to the BDI-II and extracted their complete WH from the platform. Both collections contain posts from English-speaking users. For each user, the dataset provides its real responses to the questionnaire and its complete postings history. The collections can be obtained on request from the eRisk organizers. Table 3.1 describes the main statistics of these datasets.

Table 3.1: Statistics of the eRisk collections.

eRisk Dataset	Level	2019	2020	2021
Users	*Minimal*	4	10	6
	Mild	4	23	13
	Moderate	4	18	27
	Severe	8	19	34
Total Users		20	70	80
Avg Posts/User		519	480	404
Avg Sentences/User		1688	1339	1123

In 2019, there were a total of 20 users. In 2020 and 2021, the number of users increased to 70 and 80 users, respectively. We can also observe the number of users per severity level.

The collections provide an XML file for each subject, which contains all user posts ordered by chronological order. Each post consists of the following elements: *id*, a unique identifier for each Reddit user; and *writing*, which represents a publication made on the platform. Simultaneously, within each writing, there are the next fields: *title, date, info and text*. The field title represents the Reddit thread title; date indicates the exact time of the publication, info designates the platform used (in our case, just Reddit), and text represents the user's publication text.

3.2 METRICS

Next, we present the metrics we use in this thesis to measure the effectiveness of our models. First, we consider the metrics used for the task of *Measuring the Severity of the Signs of Depression*. The metrics correspond to the official evaluation benchmarks for the task (Losada et al. 2019), and assess the quality of the BDI-II questionnaire estimated by a system compared to the real one reported by the user. By doing so, we can obtain a fair comparison among all the previous works that presented approaches to this task. We can divide them into two main evaluation metrics:

3.2.1 *Questionnaire Level*

There are two metrics that evaluate the total BDI-II score. These metrics do not take into account the individual responses to the symptom level. Instead, they evaluate the actual BDI-II score reported by the user with the estimated score by the system. In both cases, the higher values reported by the metric, the better performance.

3.2.1.1 *Depression Category Hit Rate (DCHR)*

As we commented in Table 2.1, the BDI-II score is associated with four depression levels: *Minimal, mild, moderate and severe*. The Depression Category Hit Rate (DCHR) computes the percentage of users where a system has estimated users' real category or level based on the BDI-II score.

3.2.1.2 *Average Difference Between Overall Depression Levels (ADODL)*

Difference Between Overall Depression Levels DODL calculates the absolute difference ($ad\_overall$) between the users' actual BDI-II score and the system estimated BDI-II score. For example, if the actual BDI-II score is 50, and the estimated one is 45, this difference is 5. Thus, it penalizes more the systems that deviate more from the actual BDI-II score. BDI-II score integers are between 0 and 63, and, thus, DODL is normalized to $[0, 1]$, as follows:

$$DODL = (63 - ad\_overall)/63 \tag{3.1}$$

The ADODL is simply the DODL averaged over all test users.

3.2.2 *Symptom Level*

eRisk official metrics also consider the evaluation at symptom level. Again, the higher values, the better.

3.2.2.1 *Average Hit Rate (AHR)*

Hit rate (HR) is a measure that computes the ratio of symptoms the system has estimated the same answer option as the user. If the HR for a user is 10/21, then for 10 of the 21 symptoms, the system estimated for the test user the same option as the actual user response. The AHR is the hit rate averaged for all the users.

3.2.2.2 *Average Closeness Rate (ACR)*

Closeness Rate CR calculates the absolute difference between the actual answer and the estimated answer, and subsequently, an effectiveness score is applied as follows:

$$CR = (mad - ad) \tag{3.2}$$

where *mad* represents the maximum absolute difference, and *ad* the actual difference. If a real user has answered '0' for a symptom, this metric will penalize a system that has estimated '3' more than one that has estimated '1', as the latter is closer to getting it right.

3.3 ERROR METRICS

Apart from the eRisk evaluation, we include one additional error metric for evaluating our systems.

3.3.1 *Root-Mean-Square Error (RMSE)*

We use the RMSE (Chai and Draxler 2014) to calculate the error of the model estimations of the BDI-II score. Thus, the lower the value reported by RMSE, the lower the differences between predictions and actual scores. RMSE is computed as follows:

$$RMSE = \sqrt{\Sigma_{i=1}^{N} \left(\frac{Predicted_i - Actual_i}{N} \right)^2} \tag{3.3}$$

3.4 STANDARD CLASSIFICATION METRICS

In Chapter 7, we introduce datasets focused on sentences related to depressive symptoms. We showcased their utility through a series of experiments pertaining to the classification and severity assessment of these symptoms. To evaluate the effectiveness of our methods, we employed standard classification metrics, including Precision (**P**), Recall (**R**), **F1 score**, and Area Under the Curve (AUC). These widely recognized metrics provided valuable insights into the performance and accuracy of our approaches, enabling a comprehensive evaluation of the models used for depressive symptom identification.

3.5 INTER-RATER AGREEMENT AND RANKING CORRELATION METRICS

Chapter 7 also introduces the DepreSym dataset, which contains sentences that have been labeled by experts in terms of their relevance to the BDI-II symptoms. The candidate sentences in the dataset were obtained through top-k pooling from relevance rankings created by participants in the task, with a total of 37 different ranking methods presented. To analyze the effectiveness of our methodology and annotation process, we present two types of metrics: (1) metrics related to the inter-annotator agreement among the assessors and (2) correlation metrics of the rankings generated by participating systems. In the following sections, we will provide a detailed description of the metrics used for each aspect:

3.5.1 *Inter-rater Agreement*

The **Cohen's Kappa Score** (κ) is a widely used metric for measuring the inter-annotator agreement. It assesses the level of agreement between two or more annotators, taking into account the agreement that could be expected by chance. The formula to calculate κ is as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3.4}$$

where p_o represents the observed proportion of agreement between the annotators. It is calculated as the number of agreements divided by the total of number of items being assessed. p_e denoted the expected proportion of agreement between the annotators by chance. It is calculated based on the distribution of the categories and the individual probabilies of each annotator labeling a specific category. The κ score ranges from -1 (no agreement) to 1 (perfect agreement). A higher Cohen's Kappa Score indicates a greater level of agreement among annotators, enhancing the confidence in the quality of the annotations and the overall reliability of the results.

In addition, we also computed the **Mean Krippendorff's Alpha** for our experiments (Krippendorff 2018) (α). Similarly, Krippendorff's is used to assess the reliability of agreement among multiple raters. It takes into account the agreement between each pair of raters, and then calculates the average agreement across all possible pairs. As a result, it provided a single value that represents the overall agreement among the raters, taking into consideration all the possible combination of annotators. The formulate to calculate it is as follows:

$$Krippendorf\alpha = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\alpha_{ij}}{\binom{N}{2}} \tag{3.5}$$

where N is the total number of annotators, and α_{ij} is the Krippendorff's Alpha between the $i-th$ and $j-th$ annotators.

3.5.2 *Ranking Correlation Metrics*

We compared the ranking of the participant systems with the official assessments against a hypothetical ranking based on assessments from each single annotator we considered. To that end, we ranked the systems by decreasing Mean Average Precision (MAP) and compared the rankings with Kendall's τ and AP Correlation (τ_{ap}) (Yilmaz et al. 2008).

Kendall's τ is a measure of correlation used to assess the association between two sets of rankings, commonly employed when dealing with ordinal data (only the position is important, not the exact score). Given two lists of length N, let C be the total number of concordant pairs (pairs that are in the same order in both ranking lists) and D the total number of discordant pairs (pairs that are ranked in opposite order). Then, the Kendall's τ is calculated as:

$$\tau = \frac{\text{C} - \text{D}}{\text{N(N-1)/2}} \tag{3.6}$$

Note that given a list with N elements, there are $\binom{N}{2} = N(N-1)/2$ pair of items. Similar to previous metrics, Kendall's τ ranges from -1 to 1, being 1 a perfect agreement between the rankings.

AP Correlation (τ_{ap}) (Yilmaz et al. 2008) is similar to Kendall's, but assigns greater weight to errors made to the systems positioned higher in the ranking. The motivation of this metric comes from the IR domain, since the documents retrieved towards the top of the list are assumed more important than others.

In this case, let *rank1* a rank of items of length N, and *rank2* the actual ranking for that same number of items. If we want to compute their correlation, consider the following random experiment:

- Pick any item from *rank1*, other than the top ordered item, at random.

- Pick another item from this list that is ranked above the current item, at random.

- Return 1 if this pair of documents are in the same relative order as in $rank2$; otherwise, return 0.

In mathematical terms, the expected outcome of this random experiment can be written as:

$$p' = \frac{1}{N-1} \cdot \sum_{i=2}^{N} \frac{C(i)}{(i-1)} \tag{3.7}$$

where $C(i)$ is the number of items above rank i and correctly ranked with respect to the item at rank i in $rank1$. The difference with Kendall's τ is that instead of comparing with any random item, it is only compared with random items that are above. Following this idea, Yilmaz et al. (2008) defined the AP correlation as a function of the expected outcome of the above random experiment (equation 3.7), so that its value will fall between -1 and $+1$. With this range of values commonly used by correlation metrics, the AP correlation (τ_{ap}) is finally defined as:

$$\tau_{ap} = p' - (1 - p') = 2p' - 1 = \frac{2}{N-1} \cdot \sum_{i=2}^{N} \left(\frac{C(i)}{i-1} \right) - 1 \tag{3.8}$$

Part II

DEPRESSION SEVERITY ESTIMATION MODELS BASED ON SYMPTOMS INFORMATION

4

DEPRESSION ESTIMATION DEPENDING ON SYMPTOMS SENSITIVITY

As previously analyzed in this thesis, current diagnostic methods for depression detection rely on self-report questionnaires that analyze different depressive symptoms. However, medical experts highlighted certain factors that severely limit the success of traditional methods (Barney et al. 2006; Lienemann et al. 2013). Social stigma or sensitivity is a main problem that impedes their effectiveness. In this chapter, we introduce an innovative method that considers the sensitivity associated with depressive symptoms to assess the level of depression in social media users.

In our scenario, we refer to symptom sensitivity in terms of users' inclination to openly discuss about them.

To explore this problem, we leveraged the semantic information of neural language models (word embeddings, in this case) to extract valuable insights from users' writings, focusing on specific symptoms related to depression. In line with this objective, we devised two distinct methods based on the sensitivity of symptoms. The first method analyzes users' general language patterns across their social media publications, while the second method identifies explicit mentions of symptom concerns in their publications. Both approaches provide automated estimations of the BDI-II score (Beck et al. 1996a).

To validate our approach, we conducted an evaluation of our methods using the eRisk 2020 task of *Measuring the Severity of the Signs of Depression*. We give a detailed description of this task in Section 3.1. Our study obtained competitive results, demonstrating the potential of neural language models in accurately estimating depression rating scales. In addition, we performed a detailed symptom-by-symptom analysis, exploring the variations in sensitivity among different symptoms based on the performance of our methods. The contributions presented in this chapter have been previously published (Pérez et al. 2022).

4.1 INTRODUCTION AND MOTIVATION

Social media platforms offer an excellent source of data that can provide valuable insights into emotions and experiences associated with mental health conditions, such as clinical depression (Ríssola et al. 2021). Drawing upon this notion, researchers from the fields of IR and NLP have analyzed user-generated content on platforms such as Reddit (Aragón et al. 2019), Twitter (Chen et al. 2018), and Facebook (Ophir et al. 2020) to develop predictive solutions for the early detection of mental illnesses.

In the domain of depression detection, De Choudhury et al. (De Choudhury et al. 2013b,c) made pioneer contributions to the field. Their studies extracted relevant features in the language of depressive individuals. Some examples are higher self-attentional focus and more negative emotions. Following this idea, several studies started to investigate a wide range of different features: linguistic, emotional expression, semantic, lexicon-based or social network properties and metadata (Ríssola et al. 2021). Additionally, related works focused on contact's networks structure (Cacheda et al. 2019) and the relevance of personal statements (i.e., information present in phrases with singular first pronoun) (Ortega-Mendoza et al. 2022).

In recent years, the investigation of semantic features has predominantly relied on the utilization of neural language models. In this chapter, we examine conventional word embedding architectures, such as Word2Vec (Mikolov et al. 2013) and GloVe (Pennington et al. 2014), to capture the semantic nuances associated with depressive symptoms. Our experiments entails an exploration of how users express themselves, focusing on individuals with varying degrees of symptom severity.

Binary classification considers only distinguishing between depressed and control users.

Traditionally, the majority of previous research has focused on addressing depression detection as a binary classification problem. However, there has been limited effort in conducting a more detailed analysis that distinguishes between the clinical symptoms that characterize depression. Our work represents a step forward in this direction. We aim to automatically estimate all 21 symptoms of the BDI-II questionnaire. The primary research objective of this chapter is to explore the capability of word embeddings to capture signals associated with individual symptoms. To achieve this, our proposed solutions involve the development of symptom-classifiers that leverage word embeddings as input features.

While working with the symptoms collected in the BDI-II, we observed a variation in their sensitivity levels. The more sensitive a symptom is, it influences the willingness of users to comment on it publicly. For instance, symptoms like fatigue and changes in appetite may be less intimate, leading users to discuss them more openly. In such cases, it becomes easier to search for direct mentions of these concerns. On the contrary, others are more sensitive (e.g., loss of interest in sex, crying), and users may avoid explicitly talking about them. In this case, we would need to explore alternative markers or signals within their language. This observation aligns with our intuition that users are more or less prone to talk about specific aspects of their lives explicitly. Our hypothesis suggests that the signals related to depressive symptoms differ depending on their sensitivity. Exploiting and analyzing this observation constitutes the second objective of this chapter.

Based on the previous considerations, we devised two methods. The first method involves capturing users' general language usage at sentence-level, searching for communication patterns. In contrast, the second method specifically targets direct mentions of symptom-related concerns, focusing on extracting information from specific responses rather than the rest of their publications. To validate this idea, we conducted a symptom by symptom analysis, which led us to develop a third hybrid solution. This hybrid approach dynamically selects either the first or second method depending on the specific symptom being predicted.

The remainder of this chapter is organized as follows: Section 4.2 presents the framework proposed for estimating the presence of BDI-II symptoms. The specific variants of this framework are described in Section 4.3. The results obtained from our methods are discussed in Section 4.4, with detailed explanations of the experimental settings provided in the same section. Section 4.5 provides a comparative symptom-by-symptom analysis. Our conclusions and suggestions for future research are outlined in Section 4.6.

4.2 THE SYMPTOM-CLASSIFIERS FRAMEWORK

Since we are introducing a new classification framework, instead of going straight to the specific methods, we have opted to present the general framework first and then, the three different methods that came up based on it. Our proposed approach is intended to be used as a versatile framework for depression estimation, considering it is (1) model-agnostic and (2) general to be easily adapted to different symptoms and questionnaires. Consequently, the following subsections shows our framework's main aspects along with intuitive examples highlighting the process.

We addressed our proposal as a classification task. Instead of relying on a unique classifier, we build 21 symptom-classifiers, each one corresponding to a different symptom present in the BDI-II. The symptom-classifiers are designed as a four-class problem, and each class is associated with one of the possible answer options (ranging from 0 to 3). With the use of these classifiers, we can infer the answered option for each one of BDI-II symptoms. Finally, to predict the user's total BDI-II score, we simply aggregate the decisions of the 21 classifiers. Keeping this in mind, our proposal relies on three critical components: data filtering (§ 4.2.1), users and BDI-II options feature extraction (§4.2.2) and the use of symptom-classifiers (§4.2.3).

4.2.1 *Data Selection Strategy*

The creation of reliable datasets with sufficient cases poses a significant challenge in research of depression from social media. Most existing collections contain labels at the user-level. This is also true for the collection used in this chapter, the eRisk *depression severity estimation* dataset, where the labels correspond to users' responses to the BDI-II symptoms. Thus, there is no explicit association between the symptoms and specific textual extracts from the users' publications. Manually annotating the symptoms at sentence-level would be a high-cost process. Therefore, in our framework, we extend the user-level labels ($0 - 3$) assigned to users to all their publications. This step introduces noise into many samples, as users often write many publications that are out of context or not topically related to any symptom.

In our selection strategy, we address the variability in content length of Reddit publications by first segmenting the publications of users into individual sentence-level units. To filter only relevant sentences from the extensive set of publications, we employ a measure of textual similarity, BM25. During the training phase, we employ the eRisk 2019 collection to gather all the sentences that answered each option. Using BM25, we can retrieve sentences that better characterize the option. For that, we use a query consisting of the statement that describes the respective option (we refer the reader to the Chapter 2 for information about the BDI-II statements). For instance, if we want to extract features for option 0 in symptom 14 *Loss of energy*, the query would be: *'I have as much energy as ever'*. Subsequently, we select the top k sentences from users who answered that option. This approach enables us to discard irrelevant publications that would not contribute for feature extraction.

In the case of a test user, we apply the same process for selecting its relevant sentences. However, during the inference phase, we do not have access to the test users' responses. Therefore, in this case, we use the queries to obtain relevant sentences of all the four options. This selection strategy uses four queries (from the four option of the BDI-II) to obtain the candidate sentences.

4.2.2 *Extraction of Users and Options Features*

Our main intention is to investigate the capture of the semantics of the BDI-II symptom options, $o \in \{0, 1, 2, 3\}$. This resulted in a total of 84 option vectors (21 symptoms with four options each). In all our experiments, we extracted the features using Word2Vec models. This process is carried out in the training phase, where we retrieve all the filtered sentences from the users that replied each option.

To obtain the features, we assume $e(s)$ as a function that takes a sentence s and maps it into its vector representation. Following this approach, the computation of the final features vector of a symptom i and an option j, defined by $\vec{o^i_j}$, follows the equation:

$$\vec{o^i_j} = \frac{1}{|S_j|} \sum_{s_j \in S_j} e(s) \tag{4.1}$$

where $|S_j|$ denotes the total number of filtered sentences for the option. $\vec{o^i_j}$ is then calculated by averaging the sentences embeddings of the users that answered the specific option j. In inference, we also summarized the semantic of a test user into a single vector embedding. For that, we average all the features from its filtered sentences.

4.2.3 *Use of Symptom-Classifiers*

(iii) Every symptom-classifier has two phases: training and inference. The training process consists in generating a feature vector representing each possible option. As a result, we train 21 symptom-classifiers by calculating the option vectors, where $C_n = \{\vec{o^n_0}, \vec{o^n_1}, \vec{o^n_2}, \vec{o^n_3}\}$, with $1 \le n \le 21$. We illustrate the options vector generations process in the Figure 4.1. It is exemplified for one of the symptom, 14 : (*Loss of energy*). The process for the rest of symptoms and their options follows the same procedure.

First, we retrieve all the training users that answered each option: (0. *I have much energy as ever,* 1. *I feel more discouraged about my future than I used to,* 2. *I do not expect things to work out for me,* 3. *I feel the future is hopeless and will only get worse*). From their WH, we apply the data selection strategy to filter the relevant sentences most related to

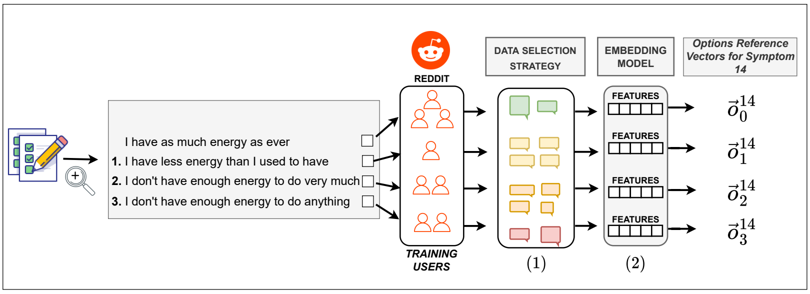

REDDIT
DATA SELECTION STRATEGY
EMBEDDING MODEL
Options Reference Vectors for Symptom 14
I have as much energy as ever □
1. I have less energy than I used to have □
2. I don't have enough energy to do very much □
3. I don't have enough energy to do anything □
TRAINING USERS
FEATURES
FEATURES
FEATURES
FEATURES
(1)
(2)
$\vec{o}_0^{14}$
$\vec{o}_1^{14}$
$\vec{o}_2^{14}$
$\vec{o}_3^{14}$

Figure 4.1: Training overview showing the extraction of option feature vectors of symptom 14: *Loss of energy*.

the symptom. At this moment, we already have k candidate sentences $\{s_{j1}, s_{j2}, ..., s_{jk}\}$ for each option o_j. Finally, we extract the features from these filtered sentences to compute the vectors o_j^{14} for the symptom 14.

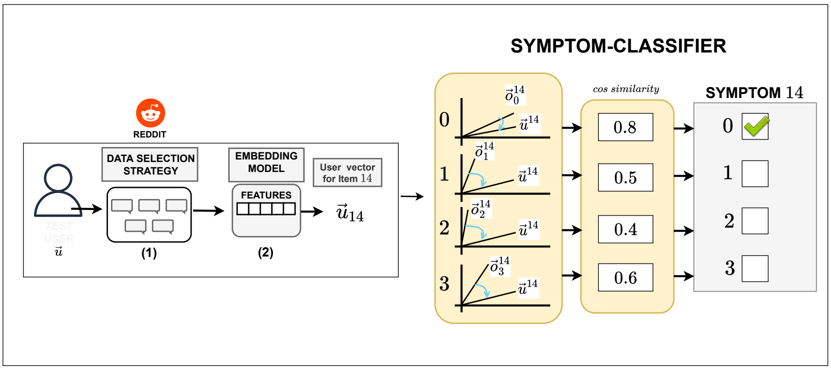


Figure 4.2: User feature vector extraction u_{14} for an symptom 14 and the posterior classification decision.

In inference, we classify the BDI-II answers of a test user after measuring the similarity with the previously calculated option vectors. We illustrate the classification process in Figure 4.2. It is exemplified for only one test user and the symptom 14. First, we filter its corresponding WH following the data selection strategy described in Subsection 4.2.1. Then, we apply the $e(s)$ function to extract features at sentence-level, and we compute the user vector $\vec{u^{14}}$ for the symptom 14 by averaging all the sentences features. We produce a different user vector for each symptom, containing only the sentences more related to it. Finally, to obtain the predictions, we simply compare the similarity of $\vec{u^{14}}$ with the option vectors: $\{\vec{o}_0^{14}, \vec{o}_1^{14}, \vec{o}_2^{14}, \vec{o}_3^{14}\}$. We use cosine similarity to obtain a result for

If no posts are found after filtering, we assume the answer estimated is 0 as there are no information traces from the user.

each option. In the last step, we classify the given test user representation u with the option that has the highest cosine similarity with the test user.

4.3 SYMPTOM-CLASSIFIERS VARIANTS

This section of the chapter describes the three different variants used for constructing the symptom-classifiers. All of these variants adhere to the classification framework discussed above, enabling the classifiers to produce decisions for each symptom.

4.3.1 *General symptom-classifiers*

We call the first approach *General symptoms-classifiers*, which objective is to capture the general language of the users. In this method, we look for communication patterns that may indicate the presence of the symptoms. The procedure for obtaining the representations of options and users is analogous to the pipeline as described in Section 4.2.3. During training, we obtain the top k sentences that satisfy the filtering process and extract all its features in training. In the inference phase, the symptom-classifiers take as input the vector of the test user. To compute the final BDI-II score of the test user, we aggregate the decisions made by all the symptom-classifiers.

4.3.2 *Direct answer symptom-classifiers*

We refer to the second approach as the *Direct answer symptoms-classifiers*. Contrarily to the general one, this variant focuses solely on direct mentions or expressions related to the symptoms. As a result, the extraction process specifically extracts the relevant portions of sentences that contain explicit answers regarding to the users feelings about a particular symptom. To achieve this, we employ a Question Answering (QA) model.

QA is one of the NLP tasks that has significantly disrupted. QA systems are based on triplets (P, Q, A), which can generate an answer A from a passage P and a question Q. In our case, we leverage the QA framework to extract potential mentions and answers for each symptom directly from the users' writings. Our model uses the users' publications as the passage P and the corresponding BDI-II item as the question Q to obtain the potential answer A. Behind the premise that particular items are more likely to be commented on in a more direct way than others, we can experiment if capturing only direct answers to symptoms can improve prediction performance.

At this point, it is necessary to mention that the BDI-II does not provide a question per symptom. Therefore, we had to manually construct the questions for each item as questions to our QA model. For this, we formulated a simple question containing keywords related to the symptom. Table 4.1 shows some of the questions we constructed for the items in the left column, as well as some extracted answers from those questions in the training collection.

The answers are paraphrased in accordance to the eRisk license.

Table 4.1: Example of four questions that we used for BDI-II symptoms along with an extracted answer from these same questions.

BDI-II Symptom	Model Question	Extracted Answer
Crying	Do you usually cry?	*"I've grown used to crying over anything that occurs"*
Social withdrawal	Have you lost interest in people or social life?	*"I've been increasingly estranged from most of my peers"*
Worthlessness	Do you feel worthless or insignificant?	*"Always feeling guilty and unworthy dude"*
Tiredness	Are you usually tired or fatigued?	*"It's difficult to be creative or even get out of bed these days"*

This method follows the same pipeline as the general symptom classifiers for feature extraction. However, there is a key difference in the sentence selection process. In this approach, we utilize the QA model to search for explicit answer related to the symptom. Consequently, the embedding model will only extract features from short and direct answers. This drastically reduces the total number of sentences to extract the features. The remaining steps of the classification process, both in training and inference, remain analogous to the general framework.

If the QA model does not output any answer for test sentence, we discard it.

4.3.3 *Mixed symptom-classifiers*

The final method, called Mixed symptoms-classifier, is a hybrid solution leveraging the two previous methods. Rather than introducing a new method, the mixed classifier dynamically selects between the general and direct symptom-classifiers based on the specific symptom being analyzed. Since the general and direct classifiers aim to capture different aspects of language, we conducted a symptom-by-symptom analysis to determine which model performed better for each symptom. To achieve this, we performed leave one-out cross-validation on the eRisk 2019 training set of 20 users. For this purpose, we used the AHR as the metric to maximize.

Next section describes a comprehensive understanding of the symptom-by-symptom analysis

Based on these results, in the test set, the mixed classifier decides to use the best of the two previous classifiers that best adjusts to each symptom. In the case of the symptoms which both classifiers had the same AHR result, we considered to use the general classifier as as it showed more consistency in the rest of the metrics. In Section 4.5, we describe which symptoms obtained better performance in our analysis.

4.4 EXPERIMENTS AND RESULTS

This section covers the experimental analysis of our symptom-classifiers on the depression estimation task. The evaluation was performed on the eRisk task of *Measuring the Severity of the Signs of Depression*. Specifically, we used the eRisk 2019 collection as training data and the eRisk 2020 as test data.

4.4.1 *Experimental Configuration*

We experimented with different embedding models to extract the features from the sentences based on word2vec. The first one is FastText, an incremental word2vec technique that also encodes the morphology of words (Mikolov et al. 2013). The second is sense2vec, a word2vec variant that uses supervised disambiguation to generate unique embeddings for each word sense (Trask et al. 2015). Furthermore, sense2vec was trained on Reddit comments from 2015, making it even more suitable for our collections.

Different settings to develop the classification framework were also investigated. Table 4.2 shows the hyperparameters that produced better results. We applied leave-one-out cross-validation using the training set. The following is a summary of them: (1) stopwords, we considered removing stopwords when gathering the set of posts. (2) Apply the data selection strategy or instead consider all the publications. In the training process, we found that not using any filter improves the direct classifier. In contrast, the filtering strategy improves the performance of the general classifier. (3) Apply the selection strategy in the user representations. In this case, we are processing a lower amount of sentences (only one user). Thus, not filtering the sentences of the test user obtained better results.

Furthermore, (4) show the number of the best top k sentences filtered with BM25. We tuned the k value from 100 to all the possible sentences matched in increments of 100. The training user with more k sentences is 2608. Hence, using 2608 corresponds to retrieving all the sentences with BM25 (BM25_ALL), which was the best value obtained in our experiments. Finally, (5) we tuned the text units considered from the set of publications selected. In the case of the general classifier, using sentences for embedding yielded better results. For this reason, we believe that the QA model works better with the whole publication as it has more context available to look for direct answers.

Table 4.2: Tuned hyperparameters experimented in the training process of our classifiers.

Hyperparameter	General Classifier	Direct Classifier
1) Stopwords	Remove	Remove
2) Data selection on options representations	BM25	Not filtered
3) Data selection on users representations	Not filtered	Not filtered
4) Best k on options representations	BM25_ALL (2608)	-
5) Text unit	Sentence level	Publication level

For the QA model, we trained a retrospective reader designed by Zhang et al. (Zhang et al. 2021). The collection used to train this model was the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al. 2018). SQuAD is a reading comprehension dataset consisting of crowdsourced question/answer pairs on a set of Wikipedia articles.

4.4.2 *Compared Methods*

Table 4.3 presents the performance of the selected baselines, the state-of-the-art approaches for each metric, and the methods we presented. The first three rows (upper block) include baselines proposed by the organizers in order to gain some perspective (Losada et al. 2020). The first and second rows are all 0s and all 1s, consisting of filling in the same option (0 or 1) for all the items. These methods represent good baselines for the metrics that consider the closeness of the predictions (ACR and ADODL). Visualizing their results, we can see how the options are distributed between 0's and

1's. As we included the participants' approaches with better results for each metric in 2020, we will summarily explain their approaches:

- The **BioInfo@UAVR** (Trifan et al. 2020) method used an external dataset to train a rule-based approach. The authors captured different psycholinguistic patterns and behavioral features to model each rule.

- **ILab** (Martínez-Castaño et al. 2020) obtained the best results in ACR. Their method used BERT-based classifiers trained explicitly for the task.

- **PRHLT-UPV** (Uban and Rosso 2020a) obtained the best results in DCHR. Sabina et al. experimented with different linguistic and emotional features with a SVM classifier.

- The **Relai** lab (Maupomé et al. 2020) addressed the problem as authorship attribution, which relies on decision models to predict the probability of a documents written by a reference user with depression tendencies.

4.4.3 *Results*

Our solutions include the general, direct-answers and mixed classifiers with the sense2vec and FastText embedding model variants. Next, we will briefly explain and compare the results of our methods with the baselines considered.

Looking at Table 4.3, it is possible to observe competitive results with respect to the top participant methods. The sense2vec variant was always slightly superior than the FastText. In the vast majority of metrics, both our general and mixed classifiers have improved all systems. Furthermore, while most participants struggled to perform well on all the four metrics, our methods showed a high level of consistency in performance.

In addition, the general classifiers also obtained better results than any participant overall. Using the sense2vec model, *General_S2V*, we are close to rank first in AHR and ACR. In DCHR, we outperform the best participant by a wide margin (nearly 30% improvement to the best solution). With the FastText model, *General_FT*, we are first in AHR and DCHR and third in ACR. However, this classifier has been found to perform worse in ADODL.

Finally, our direct-answers classifiers are the ones that performed more modest. Using the sense2vec model, we are in the top 5 of three metrics

Table 4.3: Results of our classifiers along with the baselines and the best runs of eRisk 2020. S2V and FT stand for our two embedding models, sense2vec and FastText, respectively. The numbers in parenthesis after the score corresponds to the position it would have obtained if our methods had participated in the task. Bold values highlight the best value obtained in the metric.

Run	AHR (%)	ACR (%)	ADODL (%)	DCHR (%)
all 0s	36.26	64.22	64.22	14.29
all 1s	29.18	73.38	81.95	25.71
random	23.94	58.44	75.22	26.53
BioInfo@UAVR	38.30	69.21	76.01	30.00
ILab run2	37.07	69.41	81.70	27.14
Relai	36.39	68.32	83.15	34.29
Prhlt-Upv	34.56	67.44	80.63	35.71
General_S2V	38.23$_{(2)}$	69.23$_{(2)}$	81.56$_{(5)}$	**44.29**$_{(1)}$
General_FT	38.57$_{(1)}$	69.16$_{(3)}$	80.54$_{(7)}$	38.57$_{(1)}$
Direct_S2V	36.94$_{(4)}$	69.39$_{(2)}$	81.41$_{(5)}$	28.57$_{(9)}$
Direct_FT	35.64$_{(9)}$	67.89$_{(9)}$	80.91$_{(6)}$	28.57$_{(9)}$
Mixed_S2V	**38.97**$_{(1)}$	70.10$_{(1)}$	82.61$_{(3)}$	37.14$_{(1)}$
Mixed_FT	38.51$_{(1)}$	70.00$_{(1)}$	81.80$_{(3)}$	30.00$_{(7)}$

(AHR, ACR, and ADODL). In DCHR is where we obtained the worst position. This drop in the performance may illustrate that, for most symptoms, capturing general language use rather than searching for direct answers to the item of the questionnaire is more appropriate. From the results, we can conclude that the proposed classification framework performs considerably well in the depression estimation scenario. Despite the simplicity of our approaches, they still show better performance than presented baselines. Moreover, we have to stress that, in contrast to the participant systems, our framework does not (i) use external datasets, (ii) apply an elaborated set of textual and hand-crafted features, or (iii) rely on complex decision models based on ensembles of different machine learning classifiers.

4.5 SYMPTOM-BY-SYMPTOM ANALYSIS

Our main hypothesis is that the manifestation of depressive symptoms in social media may vary depending on the sensitivity of the symptoms.

In this subsection, we conducted symptom-by-symptom comparative between the general and direct classifiers. We performed leave-one-out cross-validation on the training data to determine which classifier performed better for each symptom, with the objective of maximizing the AHR metric. Results of this analysis are illustrated in Figure 4.3. The (x, y) points represent the symptoms. The x-axis is the AHR value obtained for the symptom when considering the best classifier. The y-axis for that symptom is the AHR difference between the direct and general classifier, and there we can see the differences in the performance of the two symptom-classifiers. The positive y-axis indicates that the direct classifier outperformed the general one in terms of AHR (higher y-value indicates better performance). On the negative y-axis, we observe the symptoms that were better captured by the general classifier. The circle size on the graph represents the percentage of improvement of the best classifier over the lowest performing one, reflecting the variability of results across symptoms. Larger circles indicate a greater improvement in performance.

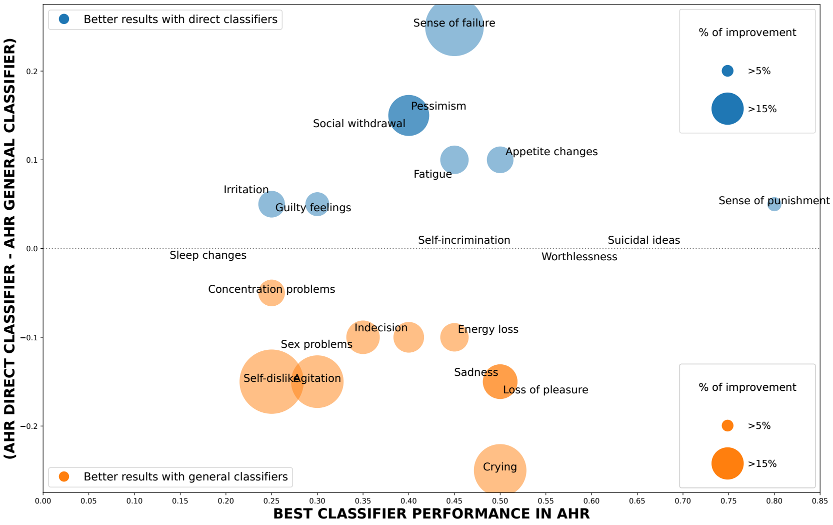


Figure 4.3: Results of the comparative study of the symptoms considering the general and direct classifiers.

Analyzing Figure 4.3, we can draw relevant conclusions: (i) The results indicate a great variability in performance depending on the symptom. For example, while we achieved approximately 80% accuracy for the *Sense of Punishment* symptom with the best classifier, the accuracy drops to less than 20% for *Sleep Changes*. This variability demonstrated that certain symptoms are particularly challenging to capture. In most cases, the best

AHR is below 50%. (*ii*) The two classifiers exhibit notable differences in their results, suggesting that there also variability in the performance depending on the method used. Only four symptoms showed no difference in AHR between the two classifiers (*Sleep changes, Self-incrimination, Worthlessness,* and *Suicidal ideas*). These findings align with the results obtained in the test collection, where the mixed classifier demonstrated improvement. This evidence validates our initial hypothesis regarding the diverse nature of depressive symptoms and exemplifies the substantial variation in performance depending on symptom characteristics.

To further analyze the results analysis for the symptoms, we examined the potential impact of the distribution of training and test users. Figure 4.4 presents the balance of answers (ranging from 0 to 3) for all symptoms. The first two rows correspond to the distribution of the options answered by the users in the training and test sets, respectively. The last rows illustrates the distribution of the general symptom-classifier estimations.

We display the results from the general classifier as it exhibited the most consistent results across various metrics.

Visualizing Figure 4.4, several observations can be made. First, we can see that for certain symptoms in the training set, there is an absence of users who selected all the possible options. For example, no training users provided responses for option 3 in the symptom of *Sadness* or *Suicidal Ideas*. Consequently, our approach was unable to generate representation for option 3. That is limitation of our approach that relies on the existence of training data for all options. Secondly, in terms of our results, the distributions of answers from our general classifier indicates a tendency to underestimate the severity of symptoms. The median of our decisions normally falls within options 0 and 1. However, there is a higher prevalence of more severe options in training and test sets. Lastly, no evidence suggests that our trained models may be over-fitting towards the majority class in the training data. This behavior is positive considering the limited amount of training data available.

4.6 CONCLUSIONS

In this chapter, we explored the potential of using neural language models, specifically word embeddings, to serve as a tool to estimate depressive states. For that, we designed a classification framework to estimate the severity level of symptoms associated with depression. We used the 21 symptoms covered in the BDI-II questionnaire as the base of our research. Our research was motivated by the recognition that these symptoms exhibit variations in sensitivity and openness for discussion. To assess

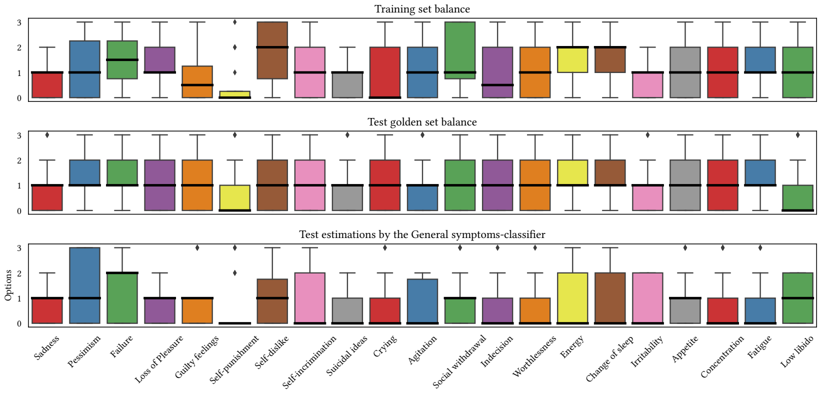


Figure 4.4: Distribution of the options in the training (first row), test set (second row) and general symptom-classifier decisions (third row) for all the BDI-II symptoms.

this idea, we proposed and evaluated two methods (1) General symptom-classifiers, which captures individuals' general language user and (2) Direct symptom-classifiers, which only captures direct answers related to the symptoms.

Our study encompassed a comprehensive symptom-by-symptom analysis of these methods, aiming to identify the superior performer for each symptom. The analysis revealed two key findings: (1) substantial performance variability exists between the two methods, and (2) certain symptoms are much more complex to capture than others. Building upon our comparative study, we introduced a mixed classifier that leverages the most suitable approach for each symptom. This hybrid classifier achieved state-of-the-art results, surpassing the performance of all previous methods. Our findings suggest that there may be an relevant connection between the sensitivity of the symptoms and the performance of predictive approaches.

5

SEMANTIC SIMILARITY MODELS FOR DEPRESSION SEVERITY ESTIMATION

In the previous chapter, we explored the potential of using word embeddings to estimate depressive states. Through the design of a classification framework, we estimated the severity level of the BDI-II symptoms. Building upon this research, the current chapter further advances this research line by introducing an efficient semantic pipeline for studying depression severity in individuals based on their social media writings. The experimental settings adopted in this chapter mirror those presented in the previous one.

Within this chapter, our focus consists in the selection of test user sentences that yield semantic rankings over an index of representative training sentences associated with depressive symptoms. Subsequently, we utilize the sentences derived from these rankings as evidence for predicting symptoms severity. To accomplish this, we explore different aggregation methods to answer one of the four BDI-II options per symptom. To evaluate the effectiveness of our methods, we utilize two benchmark datasets sourced from Reddit: the eRisk 2020 and 2021 collections. Diverging from the approaches described in the previous chapter, our work leverages pre-trained models with Sentence Transformers (SBERT), which are state-of-the-art models for semantic similarity tasks. By incorporating semantic rankings, our pipeline provides a more insightful interpretation of our models' decisions, enabling us to identify user posts potentially associated with depressive symptoms.

Under this new method, we achieved 30% improvement over state of the art in terms of measuring depression level. This includes comparing the methods presented in the previous chapter, where we improve the performance of our best prior method (Mixed_Sense2Vec), which we refer to in this chapter as *Sense2Vec*. The contributions presented here is currently available on arxiv (Pérez et al. 2023b) and has been accepted to the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2023.

5.1 INTRODUCTION AND MOTIVATION

As discussed in previous chapters, depression identification from social media posts faces challenges considering their integration into clinical settings (Walsh et al. 2020b). A notable example is the oversimplification of this challenge to a binary classification problem (Ríssola et al. 2021). Despite achieving remarkable results under these types of settings, ignoring different levels of depression limits the capacity to prioritize users with higher risks (Naseem et al. 2022). Moreover, most existing approaches have heavily relied on the use of engineered features, which may be more difficult to interpret than other clinical markers . Similarly, the black-box nature of deep learning models restricts the understanding of their decision-making processes, particularly for domain experts such as clinicians.

An observable sign indicative of a depressive tendency.

A recent line of work focused on developing models that integrate depressive symptoms as reliable clinical markers. In this context, the aforementioned eRisk depression severity task (Losada et al. 2019, 2020; Parapar et al. 2021b) made pioneer contributions to promote the integration of symptoms detection, as they were the first to release a dataset containing user-produced labels at symptom level. These new types of datasets allowed the leverage of depression markers from standard questionnaires (e.g., the BDI-II) to construct detection models. Consequently, the approaches presented for this task focused their efforts on designing models that predict the BDI-II symptom responses (Basile et al. 2021; Spartalis et al. 2021; Uban and Rosso 2020b).

Besides the works presented to eRisk, two recent studies explored the use of depressive symptoms to screen social media posts. Zhang et al. (2022a) aggregated symptoms from different questionnaires into a BERT-base model to calculate symptom risks at post level. Nguyen et al. (2022b) experimented with various methods using symptom markers to detect depression, demonstrating their potential to improve the generalization and interpretability of their approaches. In this case, authors considered the symptoms from the PHQ-9 questionnaire (Kroenke et al. 2001) to define manual pattern-based strategies and train symptom-classifiers at post level. Both approaches formulated their methods with a binary classification setting, while our approach considers different severity levels.

In this chapter, we perform a fine-grained analysis of depression severity using semantic features to detect the presence of symptom markers. Using a sentence-based pipeline, we build 21 different symptom-classifiers for estimating the user responses to the symptoms. For this purpose, we employ eRisk collections related to depression levels. In our pipeline, we explore selection algorithms to filter relevant sentences from training users to each BDI-II symptom. Once filtered, we index these training sentences with the user responses as labels ($0 - 3$) as examples of how people with different severity speak about the symptom. Then, to predict test users' responses, we select their relevant sentences, which serve as queries to produce a semantic ranking over the indexed training sentences. Finally, we construct two aggregation methods based on the ranking results to estimate the symptoms severity.

We differ from previous works in that we pre-compute dense representations of training posts, rather than relying on pre-trained language models, which may be slow for many practical cases (Reimers and Gurevych 2019). By doing so, our approach significantly enhances the efficiency of our solutions, as it only requires a few post encodings and cosine similarity calculations. The remainder of this chapter is organized as follows: Section 5.2 presents the semantic retrieval pipeline proposed to perform fine-grained classification of the severity of depressive symptoms. We detail the main components of this pipeline. In Section 5.3, we describe the experimental settings applied to our pipeline and present the corresponding results. Section 5.4 describes the construction of a manual and little dataset used in our methods. Furthermore, in Section 5.5, we provide a case study illustrating how our approach offers interpretability in its decision-making process. Finally, we summarize our findings and propose avenues for future research in Section 5.6.

5.2 METHOD

Our approach relies on two critical components: 1) a semantic retrieval pipeline (§5.2.1) and 2) silver sentences selection (§5.2.2). The symptom-classifiers follow a semantic retrieval pipeline to predict every symptom decision. This pipeline searches for semantic similarities over an index of silver sentences for a specific symptom s, denoted as Ag^s. These silver sentences are considered relevant to the symptom s, and each one has a label corresponding to the symptom options, o. Formally, Ag^s is the set containing the pairs of the silver sentences ag_i and their corresponding label o_i for the symptom s, where $Ag^s = \{(ag_i, o_i)\}$, and $o_i \in \{0, 1, 2, 3\}$.

Throughout the rest of the chapter, we will refer to these severity options as the labels of the symptoms.

To the best of our knowledge, there are no datasets in the literature where sentences are relevant to the symptom and labeled by their severity. For this reason, we propose a selection process to create the silver sentences, Ag^s, where we use as training data the eRisk collections (§5.2.2). In our experiments, we explore the performance of the semantic pipeline with our generated silver sentences. However, we could apply this pipeline to any similar datasets. In the following subsections, we explain both components in detail.

5.2.1 *Semantic Retrieval Pipeline*

Using the writing history from a test user as input, our semantic retrieval pipeline classifies the label severity for a specific symptom s. From the publications of the test user, we first select only the sentences that are relevant to s, which will serve as queries. We denote these relevant sentences as the symptom test queries Q^s, where $Q^s = \{q_1^s, ..., q_k^s\}$, since we select a top k of them. In the next subsection, we explain our sentence selection algorithms (§5.2.2). The top k queries are the input to our semantic pipeline. Figure 5.1 illustrates this process, exemplified for one test query, q_1^{energy}, corresponding to the symptom *Loss of energy*.

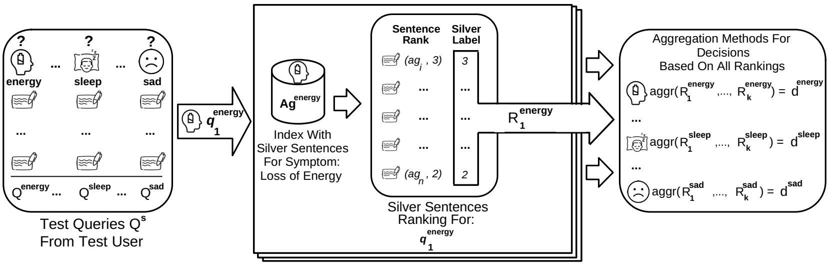


Figure 5.1: Retrieval pipeline to predict symptom options for a test user. R_1^{energy} is the list with the top ranked silver sentences for the query q_1^{energy}. Each silver sentence from the rank has a silver label associated (0-3). Finally, d^{energy} represents the option decision for that symptom based on the ranking retrieved for all the test queries, Q^{energy}.

1) The first step consists in calculating a semantic ranking for each test query for the symptom s, defined as q_i^s. To calculate that ranking, we encode q_i^s and all the silver sentences covered in the index Ag_s for the symptom as embeddings. Then, we use k-Nearest Neighbours (kNN) to compute the semantic similarity of each silver sentence w.r.t the test query q_i^s. The semantic similarity sm for a silver sentence ag_j, belonging to Ag^s,

and a test query q_i^s, is the cosine similarity between their embeddings (ϕ):

$$sm(ag_j, q_i^s) = cos\big(\phi(ag_j), \phi(q_i^s)\big) \tag{5.1}$$

Computing sm, we produce a ranking of silver sentences, R_i^s, for each test query q_i^s (q_1^{energy} in Figure 5.1). The silver sentences in the ranking have an associated silver label. For example, the position j of the ranking contains the pair: $R_i^s[j] = \{(ag_j, o_j)\}$, with $o_j \in \{0, 1, 2, 3\}$. To select the cut-off of the rankings R_i^s, we experimented with a varying number of similarity thresholds. To calculate the embeddings, we use a pre-trained model based on RoBERTa[1] using Sentence Transformers (SBERT).

2) In the second step, we apply aggregation methods to accumulate the score of the labels based on the ranking results. After processing all the test queries, the final decision predicted for the symptom s, d^s, is the label with the highest accumulated score. Specifically, we explore two aggregation methods:

Accumulative Voting: For each ranking R_i^s, we count the option labels from the n pairs that are in the rank: $\{(ag_j, o_j)\}$. The label of each silver sentence, o_j, represents a vote for that option. Then, return the sum of all the votes over the rankings. The final decision for the symptom s is the label with most votes, $d^s = \arg\max_o f_{av}(o)$, where:

$$f_{av}(o) = \sum_{i \in R_i^s} \sum_{j=1}^{n} \begin{cases} 1 & \{(ag_j, o_j) | o_j = o\} \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

Accumulative Recall: For each ranking R_i^s, compute the recall for each option label o. That is, the fraction of silver sentences in the ranking out of all the available silver sentences from that label, denoted as Ag_o^s, where $Ag_o^s = \{(ag_i, o_i) | o_i = o\}$. Then, we accumulate the recall over the rankings R_i^s. The rationale behind accumulating recall is to address the imbalance between sentences belonging to different options. This aggregation method helps prevent disadvantaging severity options with fewer relevant sentences by compensating for the disparity. The final decision is $d^s = \arg\max_o f_{ar}(o)$ with:

$$f_{ar}(o) = \sum_{i \in R_i^s} \frac{\sum_{j=1}^{n} \begin{cases} 1 & \{(ag_j, o_j) | o_j = o\} \\ 0 & \text{otherwise} \end{cases}}{|Ag_o^s|} \tag{5.3}$$

[1] huggingface.co/sentence-transformers/all-roberta-large-v1

5.2.2 *Silver Sentences Selection*

We design a process to select relevant sentences for each symptom s, and the severity labels o (previously denoted as Ag_o^s), defined as silver sentences. For this purpose, we use the eRisk 2019 collections as the source training data [2]. Therefore, the training labels are initially available at user level.

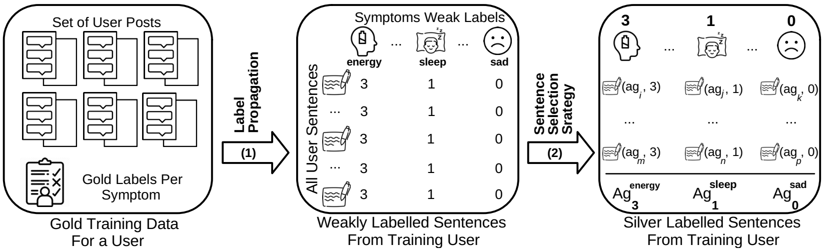


Figure 5.2: From the responses of the eRisk training users to the symptom options (0-3), the silver selection process creates one different set of silver sentences relevant to each symptom s and option o, denoted as Ag_o^s.

Figure 5.2 illustrates the sentence selection process for one training user and three symptoms. 1) In the first step, we propagate the user responses as labels for all the sentences from its writing history, resulting in weakly labelled sentences. For example, in the second component of Figure 5.2, the user replied with the option 3 for the symptom *Loss of energy* (first column). Thus, all the sentences from the user have that weak label assigned. However, since users tend to talk about different topics, most of their sentences are not relevant to any symptom. For this reason, the weak labels contain many false positives that introduce noise. 2) To reduce this noise, we propose two distant supervision strategies for sentence selection. These strategies aim to filter out the training sentences that may be non-informative w.r.t the assigned weak label. We implement two different strategies:

Option descriptions as queries: This strategy works in an unsupervised manner, since we consider the option descriptions from the BDI-II symptoms as queries to select the silver sentences. In Chapter 2, we show examples of the descriptions for the BDI-II symptoms. Similar to the strategy designed in PART II, we use each option description as one different query. Based on the sentences retrieved from these queries, we select a top of sentences from the eRisk training users who answered

[2] We refer to the reader to the Subsection 3.1 to obtain more information about this dataset.

the same option used as the query. Following this approach, we perform lexical and semantic retrieval variants. For lexical search, we use BM25 to retrieve relevant sentences for each training user. In the semantic variant, we calculate the similarity based on a semantic threshold, as described in the semantic ranking (§ 5.2.1), using the same RoBERTa model for the semantic search.

Few manually labelled sentences as queries: A drawback in using the option descriptions of the BDI-II symptoms as queries is that they only have subtle differences among one another. Consequently, previous queries struggle to capture their actual distinctions. To alleviate this problem, we hypothesize that using actual sentences from eRisk training users who answered each option may be better to differentiate between such options. In this second strategy, a small set of manually labelled sentences, referred to as *golden sentences*, serve as queries to generate an augmented silver set. The use of a larger, higher-quality set of queries allows us to cover more diverse expressions of symptom signals.

As we commented, we used the eRisk2019 training users to obtain the golden sentences. Following the approach by Karisani and Agichtein (2018a), three experts in the field conducted the annotation process. The number of golden sentences was low, averaging 35 per symptom. The data augmentation process consisted of, for every golden sentence, obtain silver sentences that correspond to the same option. For this purpose, we use the golden sentence as query, and we compute the semantic ranking over the rest of weakly-labelled sentences from that same option (§ 5.2.1). The final set of relevant sentences combines the golden and the silver sentences that surpass the similarity threshold. Table 5.1 shows an example of a golden sentence along with the top 3 augmented silver sentences. The golden sentence corresponds to the option 3 for the symptom *Pessimism in the future*, and the augmented silver sentences correspond to other training users who reported the same option [3].

5.3 EXPERIMENTS AND RESULTS

We evaluate the performance of our methods in the eRisk2020 and 2021 collections. In eRisk2020, we use 2019 as training data. In eRisk2021, we use the 2019 and 2020 collections as training. The competing methods used the same collection splits, while some of them also considered external datasets. In our experiments, we study the two components of our

[3] Information about the dataset construction and the annotation process can be found in the Section 5.4.

Table 5.1: Examples of augmented silver sentences with highest semantic similarity to the golden sentence.

Golden Sentence	**Silver Sentences Augmented**
(Option 3) *"I'm a stupid student with no intelligence/future."*	*"I know I'll never be like that; I'll be a stupid failure my entire life."* *"Used to be a stellar student, but I'm scared of opinions now that I received a C in a class."* *"It's actually starting to irritate me, and I'm starting to feel stupid."*

approach: i) the performance of the semantic retrieval pipeline (§ 5.2.1) and ii) the effectiveness of the sentence selection strategies (§ 5.2.2). For this reason, our methods consist of combinations of these components. We consider three hyperparameters: 1) The value k of the number of test queries, $Q^s = \{q^s_1, ..., q^s_k\}$. 2) The semantic threshold to select the cut-off of the rankings, R^s_i. 3) The number of silver sentences to generate the silver dataset, Ag^s. Next, we will comment the competing methods. Moreover, we will briefly describe the specific hyperparameters and tuning process.

5.3.1 *Experimental Settings*

Competing Methods. We consider the best prior works for each metric for the eRisk2020/2021 collections. We refer to the reader to the corresponding shared task surveys for a detailed analysis (Losada et al. 2020; Parapar et al. 2021b). In eRisk2020, BioInfo (Trifan et al. 2020) and Relai (Maupomé et al. 2020) methods obtained their own datasets to perform standard ML classifiers using engineered features as linguistic markers. Other deep learning approaches, such as ILab (Martínez-Castaño et al. 2020) and UPV (Uban and Rosso 2020b), focused their efforts on the use of Large Language Models (LLMs) explicitly trained for depression severity estimation. Finally, a recent work by Pérez et al. (2022) (Sense2vec) designed different word embedding models for each of the symptoms and achieved state-of-the-art results in this dataset. In eRisk2021, Symanto (Basile et al. 2021) team trained a neural model with additional data annotated by psychologists and combined it with a set of engineered features, whereas (Shih-Hung and Qiu 2021) (CYUT) experimented with different RoBERTa classifiers. Similar to our work, Spartalis et al. (2021) (DUTH) used semantic features with sentence transformers to extract one dense representation per user, which is then fed as input, experimenting with various classifiers. Although insightful, eRisk approaches cannot evidence the sentences that lead to symptom decisions.

Experimental Settings. We experimented with different hyperparameters to validate the results from our two main components: the semantic retrieval pipeline (§ 5.2.1) and the sentence selection process (§ 5.2.2). As we do not have a validation set, we performed leave-one-out cross-validation using the training set available to calculate the optimal values of all hyperparameters. The metric maximized was DCHR. When evaluating our methods in eRisk2020, the training set was the eRisk2019 dataset. When using as test collection the eRisk2021 dataset, the training set was the eRisk2019 and 2020 collections. Table 5.2 presents the hyperparameters and the optimal values for each method used in our experiments [4].

Semantic retrieval pipeline (§ 5.2.1). In the semantic pipeline, we experimented with two hyperparameters:

1) **The value k of the number of test queries.** We explored with selecting a different number of top k values of the user test queries, $Q^s = \{q_1^s, \ldots, q_k^s\}$. To select these test queries, we used the data selection strategies of using the option descriptions as queries (§ 5.2.2). Using BM25, the k values explored were: $[5, 10, 15, 20, 25, 30, 40]$. We also experimented with the same k values using the semantic variant. However, we did not include those results in the paper as they could not improve the use of BM25.

2) **The semantic threshold to select the cut-off of the rankings, R_i^s.** We experimented with different semantic thresholds to select the cut-off of the ranking of silver sentences, R_i^s. This semantic threshold, calculated as the cosine similarity, was explored with the next values: $[0.45, 0.50, 0.55, 0.60, 0.65]$. The higher the cosine similarity, the lower the number of silver sentences retrieved by the semantic ranking obtained by the test queries.

Silver sentences selection (§ 5.2.2). Additionally, we also experimented with a filtering hyperparameter for creating more or less restrictive filters when generating the silver dataset, denoted as *selection threshold*.

[4] We want to note that the tuning of hyperparameters in our method did not result in significant changes to its performance. We thoroughly analysed the impact of hyperparameters on our results and found that the changes were not significant enough to include another section in the chapter.

Depending on the selection strategy (BM25, SBERT or Aug Dataset), we used the next sentence selection hyperparameters:

3) **The number of silver sentences to generate the silver dataset**, Ag^s. Using BM25, we explored with two different top k values, $k \in \{50, 100\}$ for retrieving the sentences of each training user. In the case of semantic retrieval (SBERT), we explored with the same semantic similarity thresholds as in the semantic ranking: $[0.45, 0.50, 0.55, 0.60, 0.65]$. Higher cosine similarity implies more restrictions, so the number of silver sentences generated will be lower. Finally, the semantic threshold values explored with the augmented dataset were the same.

With respect to the sentence transformers models Reimers and Gurevych 2019, we experimented with different pre-trained models: *msmarco-bert-base-dot-v5*[5], *msmarco-distilbert-cos-v5*[6], *all-roberta-large-v1*[7] *and stsb-roberta-large*[8] via the huggingface transformers library. All these models were fine-tuned on diverse semantic similarity datasets. In pilot experiments, the best results were obtained with the model *all-roberta-large-v1*. Thus, all our reported results correspond to the use of that model.

Table 5.2: Best hyperparameter values for all the variants considered in our methods. These values were obtained by performing leave-one-out cross-validation in the training set by maximizing the DCHR metric.

Training Set	Method	k Test Queries	Semantic Ranking Threshold	Silver Sentence Selection
eRisk2019	Acc. Voting-BM25	25	0.55	Top k = 100
	Acc. Recall-BM25	25	0.60	Top k = 100
	Acc. Voting-SBERT	25	0.50	0.45
	Acc. Recall-SBERT	25	0.50	0.35
	Acc. Voting-Aug Dataset	40	0.55	0.50
	Acc. Recall-Aug Dataset	35	0.55	0.50
eRisk2020 eRisk2021	Acc. Voting-BM25	25	0.55	Top k = 100
	Acc. Recall-BM25	25	0.55	Top k = 100
	Acc. Voting-SBERT	25	0.50	0.50
	Acc. Recall-SBERT	30	0.55	0.40
	Acc. Voting-Aug Dataset	25	0.55	0.50
	Acc. Recall-Aug Dataset	25	0.55	0.50

Metrics. Similar to the previous chapter, we have adhered to the official metrics proposed by organizers (Losada et al. 2020). In this chapter, we have also included one additional error metric, RMSE. We use RMSE to compare the models predictions of the total BDI-II score. Thus, the lower

[5] https://huggingface.co/sentence-transformers/msmarco-bert-base-dot-v5

[6] https://huggingface.co/sentence-transformers/msmarco-distilbert-cos-v5

[7] https://huggingface.co/sentence-transformers/all-roberta-large-v1

[8] https://huggingface.co/cross-encoder/stsb-roberta-large

Table 5.3: Results on eRisk 2020 and 2021 collections(questionnaire metrics). The numbers of the official metrics are in percentage. Best values are bolded. Methods using external datasets for training the model are marked. Statistical significant differences in the severity level category assignment according to the Stuart-Maxwell marginal homogeneity test w.r.t to the baselines are super-scripted (p-values < 0.05).

			Questionnaire Metrics		
Collection	Model	External Dataset	DCHR (↑)	ADODL (↑)	RMSE (↓)
eRisk 2020	(Trifan et al. 2020) (a)	✓	30.00	76.01	18.78
	(Martínez-Castaño et al. 2020) (b)	✓	27.14	81.70	14.89
	(Maupomé et al. 2020) (c)	✓	34.29	83.15	14.37
	(Uban and Rosso 2020b) (d)	✗	35.71	80.63	15.40
	(Pérez et al. 2022) (e)	✗	37.14	82.61	12.40
	Our methods				
	Acc. Voting-BM25	✗	$38.57^{a,b,c,d,e}$	85.19	12.37
	Acc. Recall-BM25	✗	$40.00^{a,b,c,d,e}$	84.65	12.13
	Acc. Voting-SBERT	✗	$42.86^{a,c,d}$	83.08	14.25
	Acc. Recall-SBERT	✗	$42.86^{a,b,c,e}$	84.51	12.37
	Acc. Voting-Aug Dataset	✗	$47.14^{a,b,c,d,e}$	**85.33**	**11.87**
	Acc. Recall-Aug Dataset	✗	$\mathbf{50.00}^{a,b,c,d}$	85.24	12.09
eRisk 2021	(Spartalis et al. 2021) (a)	✗	15.00	73.97	19.60
	(Basile et al. 2021) (b)	✓	32.50	82.42	14.46
	(Shih-Hung and Qiu 2021) (c)	✗	41.25	**83.59**	**12.78**
	Our methods				
	Acc. Voting-BM25	✗	$45.00^{a,b,c}$	82.16	14.11
	Acc. Recall-BM25	✗	$42.50^{a,b,c}$	80.62	15.13
	Acc. Voting-SBERT	✗	$45.00^{a,b,c}$	81.92	14.15
	Acc. Recall-SBERT	✗	$41.25^{a,b,c}$	81.86	14.2
	Acc. Voting-Aug Dataset	✗	$46.25^{a,b,c}$	81.72	14.83
	Acc. Recall-Aug Dataset	✗	$\mathbf{51.25}^{a,b,c}$	81.65	14.96

the value reported by RMSE, the lower the difference between predictions and real scores are. We refer to the reader to the Section 3.2 for a detailed description of all the metrics.

5.3.2 *Results*

Tables 5.3 and 5.4 present a comparison of the results achieved by all variants of our approach with competing methods, focusing on questionnaire and symptom-level metrics[9]. Table 5.3 showcases the outcomes related to the questionnaire metrics, while Table 5.4 displays the results pertaining to symptom-level metrics. Our variants are the combination of our two aggregation methods (*Accum Voting* and *Accum Recall*) and the sentences selection strategies (BM25, SBERT and the augmented dataset, *Aug Dataset*). Next, we will comment our results obtained for both metrics:

Questionnaire level: Looking at Table 5.3. Our approach achieves the best DCHR, which considers the percentage of times that the system estimates the severity level of the users correctly. Most of our variants outperform all prior work in this metric, with the *Accum Recall-Aug Dataset* correctly estimating at least 50% of the depression levels for both collections. In more detail, it improves 13 and 10 points over the best previous results for eRisk2020 and 2021, respectively. A similar phenomenon occurs in the rest of the questionnaire metrics. In the error metric, RMSE, our results also show less estimation error in the BDI-II score.

Symptom level: Although in eRisk2020, our AHR figures are close to the best baselines, that is not the case in 2021. AHR computes the ratio of option responses estimated correctly. The explanation is that we tuned the model hyperparameters for the DCHR metric since clinicians believe that assessing overall depression levels is more valuable than focusing on specific symptoms (Richter et al. 1998). Tuning for AHR may produce worse overall results because the model could be failing to a greater amount in the non-correct answers, resulting in higher overall error. To illustrate that effect, we produced an oracle to obtain the best hyperparameters for each symptom-classifier, maximizing AHR using the *Accum Voting-SBERT* variant. With this oracle, we achieved an AHR of 41.77 and 37.32 for eRisk2020 and 2021, which improves all baselines. However, the oracle obtained worse results in DCHR (24.29 and 36.25). This is because tuning each individual symptom-classifier would require much more training data. We may improve the results for some symptoms with enough data but produce predictions with higher errors (e.g., 0 vs 3) for symptoms with few training samples.

Finally, with respect to the sentence selection strategies, we can observe that using the options descriptions as queries (BM25 and SBERT) performs worse than the augmented dataset (*Aug Dataset*). This emphasizes the importance of a precise candidate selection. Moreover, despite the distribution of depression levels varies in both collections (see Table 3.1),

[9] To see a detailed description of the metrics, we refer to the reader to the Evaluation Section (§ 3.2).

Table 5.4: Results on eRisk 2020 and 2021 collections (symptom metrics). The numbers of the official metrics are in percentage. Best values are bolded. Methods using external datasets for training the model are marked. We found no statistically significant different with respect to the best prior work and our methods proposed.

			Symptom Metrics	
Collection	Model	External Dataset	AHR (↑)	ACR (↑)
eRisk 2020	BioInfo (Trifan et al. 2020) (a)	✓	38.30	69.21
	ILab (Martínez-Castaño et al. 2020) (b)	✓	37.07	69.41
	Relai (Maupomé et al. 2020) (c)	✓	36.39	68.32
	UPV (Uban and Rosso 2020b) (d)	✗	34.56	67.44
	Sense2vec (Pérez et al. 2022) (e)	✗	**38.97**	**70.10**
	Our methods			
	Acc. Voting-BM25	✗	35.24	68.37
	Acc. Recall-BM25	✗	35.71	67.60
	Acc. Voting-SBERT	✗	34.83	65.90
	Acc. Recall-SBERT	✗	33.33	66.05
	Acc. Voting-Aug Dataset	✗	35.24	67.41
	Acc. Recall-Aug Dataset	✗	35.44	67.23
eRisk 2021	DUTH (Spartalis et al. 2021) (a)	✗	**35.36**	67.18
	Symanto (Basile et al. 2021) (b)	✓	34.17	**73.17**
	CYUT (Shih-Hung and Qiu 2021) (c)	✗	32.62	69.46
	Our methods			
	Acc. Voting-BM25	✗	30.97	64.54
	Acc. Recall-BM25	✗	28.03	62.92
	Acc. Voting-SBERT	✗	29.67	64.27
	Acc. Recall-SBERT	✗	27.47	62.89
	Acc. Voting-Aug Dataset	✗	27.95	62.40
	Acc. Recall-Aug Dataset	✗	27.66	61.72

our methods show robustness as we keep achieving good performance in DCHR.

5.3.3 *Effect of Data Augmentation Strategy*

To better understand the performance of the data augmentation, we report the number of augmented silver sentences along with the F1 metric for each depression level. For this reason, we divide the eRisk users in the four severity categories (minimal, mild, moderate, severe), and compute the F1 of our models per each category. Table 5.5 shows the F1 results of our best variant using the augmented dataset, *Accum Recall-Aug Dataset*, in eRisk2020 and 2021. Looking at the statistics, we see more presence in golden sentences of high-risk levels (moderate and severe). In addition, the number of silver sentences augmented for each of them is also higher. For example, using eRisk2019 as the training set, an average of three silver sentences were augmented from each golden one in the minimal level ($\frac{310}{98} \approx 3$). In contrast, the average of silver sentences augmented from the severe category is 7 ($\frac{2414}{354} \approx 7$). This suggests that users with higher depressive levels tend to manifest more explicit thoughts related to the symptoms. As a result, our augmentation method finds pieces of evidence in these levels easier.

Table 5.5: Number of golden and the augmented silver sentences for each severity level and their F1 using *Accum Recall-Aug Dataset*.

Training	Level	Golden sentences	Silver sentences	F1
eRisk 2019	Minimal	98	310	0.42
	Mild	49	171	0.37
	Moderate	237	2298	0.46
	Severe	354	2414	0.74
eRisk2019 eRisk 2020	Minimal	98	442	0.24
	Mild	49	614	0.42
	Moderate	237	1633	0.51
	Severe	354	1207	0.63

Table 5.6: F1 results in a binary classification scenario using the *Accum Recall-Aug Dataset* variant and the best prior model.

		F1		
Test	Level	Ours	Best prior model	
eRisk2020	Low risk	0.72	Sense2vec	0.64
	High risk	0.74		0.62
eRisk2021	Low risk	0.52	CYUT	0.00
	High risk	0.82		0.85

If we observe the F1 results in Table 5.5, we also see considerable variability among depressive levels. In both collections, we achieve better results for higher risk categories. This seems to be related to the number of golden sentences. Therefore, if we obtain more samples belonging to the lower risk levels, there may be an improvement in these categories. Finally, we examine our results with a binary classification setting. For this purpose, we categorize the four depression levels into only two: 1) *low risk* (minimal + mild levels) and 2) *high risk* (moderate + severe levels). Table 5.6 shows the results for the *Accum Recall-Aug Dataset* variant along with the best prior work under this setting. Our results suggest the effectiveness of our method, which distinguishes with fair accuracy between higher and lower risks.

5.4 MANUAL DATASET AND ANNOTATION PROCESS

This section describes the construction and annotation schema of our manual dataset. As commented in the method Section, the main idea of this dataset is to obtain a few representative samples that indicate the presence of BDI-II depressive symptoms. For this reason, we develop an annotation schema based on the BDI-II questionnaire (Lasa et al. 2000) to collect a different set of golden sentences belonging to each BDI-II symptom. For each symptom s, and the corresponding options o, where $o \in \{0, 1, 2, 3\}$, we collect a different set of golden sentences, denoted as G_o^s.

To annotate the golden sentences, we used as data source the training users from the eRisk2019 collection of depression severity (Losada et al. 2019). However, the large size of the eRisk collection requires an exhaustive filter for reasonable annotation efforts. For this purpose, we leveraged the data selection strategy of using the option descriptions as queries (§ 5.2.2). In particular, we applied the semantic retrieval variant (SBERT). Using this strategy, we selected candidate sentences for annotating each BDI-II symptom. We have considered this strategy following a recent study that has shown great results in identifying diverse expressions of symptoms for candidate retrieval annotation (Zhang et al. 2022b). Previous studies on symptom annotation (Zhang et al. 2022b) demonstrated a high variance in the distribution of each symptom. For some of them, it is much easier to find representative sentences than for others. To keep the number of annotations per symptom stable, we fixed a similarity threshold of 0.6 to filter out sentences. However, this similarity threshold still produced too many candidate sentences for some symptoms. For this reason, we further restricted the annotator's work to the first 750 sentences in the symptoms with too many candidates.

More specifically, 17% of the candidate sentences have been labelled positive following the semantic retrieval strategy from the total of 5004 candidates. From the same labelled sentences, using keyword matching with BM25 reduced this percentage to 4%. With a random retrieval strategy, it dropped to 0.01% due to the small number of relevant sentences compared to the size of the entire pool. These findings align with previous research indicating that pattern matching is not effective in retrieving diverse sentences relevant to depressive symptoms (Mowery et al. 2017). Instead, a semantic similarity-based strategy is better suited to retrieve representative sentences without relying on specific keywords covered in the clinical questionnaires.

Following the above candidate annotation schema, we constructed a small dataset for all the BDI-II symptoms. The annotation task was carried out by two psychologists and two PhD students with knowledge in the field. Before the annotation process, we removed all supplementary metadata to avoid bias in the annotators, such as the severity option label ($0 - 3$) of the user who wrote the sentence. We followed the same annotation procedure as Karisani and Agichtein (2018a) to validate the annotation outcomes. This procedure consisted of two phases:

1) First, an initial annotator answered the following question in a binary setting (Positive/Negative): *Does the sentence refer to the symptom, and the user talks about himself/herself (first person)?*. This first annotator labelled a total of 738 positive sentences from the candidate sentences. We considered all the sentences annotated as positive for each symptom to obtain our final labels corresponding to the option levels ($0 - 3$). Subsequently, we label these positive sentences with the severity option reported by the user who wrote them. Therefore, for each option o and symptom s, we obtained a different set of golden labels, G_o^s, where the sentences come from the eRisk users that answered the BDI-II symptoms.

2) Once we had the previous initial annotated sentences, the rest of the annotators validated them. For this purpose, they were provided with a subset containing a random sample of the 20% of the sentences of each symptom for re-annotation. Since in our pilot experiments, we found much more disagreement with positive labels, the 20% random sample only contained positive ones. The re-annotation process obtained an 82.44% agreement among the three annotators, which is an acceptable number considering the sensitivity of this topic (Coppersmith et al. 2018).

Table 5.7 and 5.8 show the main statistics of our manual dataset. Visualizing these tables, we can extract several findings. We note that, for all the symptoms, the number of sentences associated with the option 0 is very low. In some symptoms, even none of the sentences corresponded to option 0. This suggests retrieving sentences representing positive feelings towards the symptom is more complicated. We attribute this fact to two main reasons, (i) the descriptions of BDI-II options 0 are not entirely appropriate for the candidate retrieval process (most of them are just negations of a negative feeling), and (ii) users are not as likely to talk about positive as they do with negative feelings. To address this, for the symptoms that lacked sentences with option 0, we manually included between 1 and 3 sentences that provide a positive description of the symptom and labelled them with option 0.

Finally, the statistics also show that, despite our efforts, there is a clear imbalance in the number of sentences for each symptom and their options. Further details on the dataset will be described with its public release. The dataset will be made available under a research data agreement in accordance with eRisk policies.

Table 5.7: Annotations statistics of the first ten BDI-II symptoms.

	Sadness	***Pessimism***	***Sense of Failure***	***Loss of Pleasure***	***Guilty Feelings***	***Punishment***	***Self-dislike***	***Self-incrimination***	***Suicidal Ideas***	***Crying***
Option 0	2	2	3	35	2	2	3	1	1	2
Option 1	97	4	6	51	7	0	8	1	29	8
Option 2	8	9	2	18	0	15	42	3	17	32
Option 3	0	44	45	0	23	1	6	5	0	3
Total Labels	108	59	56	104	32	18	59	10	47	45

Table 5.8: Annotations statistics of the last eleven BDI-II symptoms.

	Agitation	***Social withdrawal***	***Indecision***	***Worthlesness***	***Loss of energy***	***Sleep changes***	***Irritability***	***Changes in appetite***	***Concentration difficulty***	***Tiredness/Fatigue***	***Low libido***
Option 0	2	1	3	1	4	1	2	1	2	3	2
Option 1	7	12	2	1	3	1	26	3	4	15	4
Option 2	4	4	6	3	7	3	4	1	10	4	1
Option 3	13	7	3	19	0	4	0	2	1	3	1
Total Labels	26	24	14	24	14	9	32	7	17	25	8

Table 5.9: Example of the top query sentences from a test user for two symptoms along with the golden option response of that user and the predicted option of our method.

Symptom	Golden label	Predicted label	Test queries with more retrieved silver sentences from the predicted label
Sleep problems	1	2	*"My sleep cycle consists of staying awake for 48 hours until I can't keep my eyes open."* *"Same as you, I usually can't go back to sleep once I'm awake."* *"I went through a phase where I slept for up to 16 hours (usually partially waking up)."*
Loss of pleasure	3	3	*"Look, no matter how hard you try, things don't get any better from here."* *"I don't even enjoy simple things like food that I used to enjoy; there are just foods that I dislike less."* *"Why am I not supposed to enjoy life?"*

Table 5.10: Examples of retrieved silver sentences with their assigned label from two test queries from a user.

Test query	Silver sentences retrieved
"My sleep cycle consists of staying awake for 48 hours until I can't keep my eyes open."	**(Option 2)** *"Always had trouble sleeping, no big deal but it's gotten worse in the last two months."* **(Option 2)** *"I have to get up early to get to university, and I've recently been getting no more than 3-4 hours of sleep."*
"Look, no matter how hard you try, things dont get any better from here."	**(Option 3)** *"Hoping for a "better thing" never makes me feel better unless it comes from this sub because I know people get it."* **(Option 3)** *"Things stop being enjoyable, and everything becomes a chore."*

5.5 CASE STUDY – INTERPRETABILITY

The lack of reliable clinical markers is one of the barriers to the practical use of mental health prediction models (Amini and Kosseim 2020; Walsh et al. 2020b). By considering a more refined grain in the symptom presence, we provide valuable information that may be strong clinical markers. Table 5.9 showcases how our approach offers interpretability of the symptom decisions, showing three query sentences from an anonymized test user. The symptoms in the Table are *Sleep problems* and *Loss of pleasure*, and the user declared the option 1 and 3 for them, respectively. We can see that these test queries are robust indicators of symptom concerns. Following this approach, clinicians may inspect sentences as a first step towards further diagnosis or monitoring methods during treatment.

In addition, Table 5.10 displays some of the silver sentences retrieved for the same test queries selected from the anonymized user. The silver sentences are related to the content of the query, and clinicians may evaluate the justifications for every symptom decision by reviewing their labels. Moreover, in our method, false positive/negative predictions can still be helpful for future inspection. For example, for the symptom *Sleep problems*, the test user reported the option 1, but our method retrieved more silver sentences with the option 2. While the prediction may be incorrect (golden label (1) ≠ predicted label (2)), the risk may still be present.

5.6 CONCLUSIONS

In this chapter, we present an effective semantic pipeline to estimate depression severity in individuals from their social media data. We address this challenge as a multi-class classification task, where we distinguish between depression severity levels. The proposed methods base their decisions on the presence of clinical symptoms collected by the BDI-II questionnaire. With this aim, we introduce two data selection strategies to screen out candidate sentences, both unsupervised and semi-supervised. For the latter, we also propose an annotation schema to obtain relevant training samples. Our approaches achieve state-of-the-art performance in two different Reddit benchmark collections in terms of measuring the depression level of individuals. Additionally, we illustrate how our semantic retrieval pipeline provides strong interpretability of the symptom decisions, highlighting the most relevant sentences by semantic similarities.

Part III

RESOURCES AND APPLICATIONS FOR DEPRESSION DETECTION BASED ON SYMPTOM MARKERS

6

BDI-SEN: A SENTENCE DATASET FOR CLINICAL SYMPTOMS OF DEPRESSION

In the preceding two chapters, we have presented a range of solutions for estimating the severity of depression based on symptom markers, which involved the development of various symptom classifiers. As we discussed earlier, constructing solutions based on symptom identification enhances generalization and interpretation, particularly in clinical settings. However, most of the existing datasets on depression detection only provide binary labels at the user level (depressive vs control users).

In the previous chapter, we presented a methodology to obtain a small dataset related to depressive symptoms. Following and extending that idea, in this chapter, we introduce *BDI-Sen*, a symptom-annotated sentence dataset for depressive disorder. *BDI-Sen* covers all the symptoms present in the BDI-II, a reliable questionnaire used for detecting and measuring depression. The annotations in the collection reflect whether a statement about the specific symptom is informative (i.e., exposes traces about the individual's state regarding that symptom). We thoroughly analyze this resource and explore linguistic style, emotional attribution, and other psycholinguistic markers.

Additionally, we conducted a series of experiments investigating the utility of BDI-Sen for various tasks, including the detection and severity classification of symptoms. We also examine their generalization when considering symptoms from other mental diseases. BDI-Sen may aid the development of future models that consider trustworthy and valuable depression markers. The contributions presented in this chapter have been previously published in the 46th International ACM SIGIR Conference (Perez et al. 2023).

6.1 INTRODUCTION AND MOTIVATION

Symptom-based prediction models showed the importance of presenting reliable depression markers to aid health professionals in their diagnosis (Coppersmith et al. 2018). Most of them leveraged the use of Large Language Models (LLMs) to design classifiers, like the ones presented in previous chapters (Pérez et al. 2023b; Pérez et al. 2022). Recent similar works also explored the use of symptoms to estimate depression (Nguyen et al. 2022a; Zhang et al. 2022a). For instance, Zhang et al. (2022a) developed a BERT-based model that aggregates markers from different clinical inventories to calculate the risk of symptoms at the post level. To improve the efficiency of their approach, they designed templates based on standard questionnaires to pre-filter only representative posts.

Nguyen et al. (2022a) also explored BERT-based methods using symptom classifiers and compared them against a standard depression classifier. They used the nine symptoms from the 9-Question Patient Health Questionnaire (PHQ-9) (Kroenke et al. 2001) to design the symptom classifiers. Covering three different datasets, the authors found that these classifiers performed well compared to the standard depression classifier while generalizing better to other collections. Moreover, the authors found that when leveraging the weights from the attention architecture, these symptom classifiers provide a model that can highlight specific posts based on relevant symptoms, improving their interpretability. In addition to this, the works proposed to solve the *eRisk depression estimation shared task* were pioneering contributions to the development of symptom detection models, where their approaches predicted the BDI-II symptom responses of Reddit users. For a detailed analysis of these works, we refer the reader to the corresponding shared task overviews (Losada et al. 2019, 2020; Parapar et al. 2021b).

Following the works mentioned above, we can find collections that go beyond providing binary labels for depression detection. Some datasets focused on more diverse aspects, such as including temporality in their annotations (MacAvaney et al. 2018), or the combination of different modalities of data, including text and images (Shen et al. 2017). When considering the time factor, the collections released by the shared tasks of early risk detection add further challenges over the classical binary classification problem (Losada et al. 2019, 2020; Parapar et al. 2021a). Additionally, the eRisk collections on depression severity estimation adopt a human-in-the-loop approach, requiring self-reported information directly from individuals. These collections consist of users' social media posts and the real users' responses to the 21 BDI-II symptoms.

In another exciting contribution to this new trend of symptom-based models, Zhang et al. (2022b) released the PsySym dataset. This dataset is the first annotated symptom sentence dataset that covers multiple mental disorders. PsySym includes annotations of 38 symptoms from 7 mental disorders. The authors established the symptom classes according to the DSM-V, accompanied by the descriptions of diverse inventories. While our work shares similar motivations, we differ in our approach by adhering directly to the clinical questionnaire of the BDI-II and providing the actual responses of the writers to the analysed symptoms' questions. The present study represents a step towards considering reliable symptoms as depression markers to design more robust mental health detection models.

In this chapter, we introduce *BDI-Sen* to promote further the development of models based on symptom markers to identify depressive signs. BDI-Sen is a dataset comprising 4973 annotated sentences covering depressive symptoms and 41 200 control sentences. Following a similar approach to PsySym (Zhang et al. 2022b), which includes symptoms of different diseases based on the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) (Nuckols and Nuckols 2013), we identify relevant sentences to depressive symptoms. However, in our case, the sentences are associated with users' responses to the 21 BDI-II symptoms. BDI-Sen is also a valid resource for ranking representative sentences of depressive symptoms, following the recent CLEF eRisk task (Parapar et al. 2023). Adhering to established clinical schemas for diagnosing depression, such as BDI-II, is crucial for facilitating the integration of more effective and consistent diagnostic support tools.

For building our dataset, we first semantically ranked the whole sentences from the annotated users for relevance to a symptom. To do so, we estimated semantic similarities using sentence transformers embeddings (Reimers and Gurevych 2019) and relying on the descriptions of elements provided by the BDI-II as queries. In the second phase, we follow a manual annotation schema as in similar works (Karisani and Agichtein 2018b; MacAvaney et al. 2018; Mowery et al. 2017) where experts decided the actual relevance of the filtered candidates.

Our study includes a symptom-by-symptom analysis of the language and emotional characteristics of the annotated sentences. Additionally, we perform experiments to validate the usefulness of BDI-Sen for various tasks, including the detection and severity estimation of symptoms. Using a wide range of classification models, we find that the methods can effectively detect sentences representative of depressive symptoms. However, when considering different severity risk levels, we observe a significant decrease in performance. Further examination via error analysis reveals the challenge of distinguishing between closely related severity levels. Finally, we investigated the generalization of our models to symptoms from other mental diseases, showing that the models trained on our dataset can generalize well. BDI-Sen dataset and the code implemented is available under the eRisk dataset research license[1].

6.2 BDI-SEN DATASET

This section describes the construction and annotation schema of the *BDI-Sen* dataset. We create a symptom-based dataset with relevant sentences that trace the presence of clinical symptoms. For this reason, we develop an annotation schema based on the BDI-II (Beck et al. 1996a). As commented in previous chapters, the BDI-II is a highly reliable tool to diagnose depression in clinical settings (Lasa et al. 2000). The BDI-II covers 21 recognized symptoms, including emotional, cognitive and physical markers. To create the *BDI-Sen* dataset, we used as data source the eRisk2019 depression severity collection (Losada et al. 2019). We used Reddit as the target platform due to its wide acceptance in previous studies (Cohan et al. 2018; MacAvaney et al. 2018; Parapar et al. 2021a; Zhang et al. 2022b).

6.2.1 *Dataset Construction*

Candidate Sentences Selection. The large volume of publications from eRisk2019 training users requires an exhaustive filter for reasonable annotation efforts. For this reason, we design an initial retrieval stage based on filtering candidate sentences that may be relevant to each symptom. The retrieval phase uses the options' descriptions (severity descriptions) as queries to select the candidate sentences. We generate four queries (one per severity level) and search the entire set of sentences. For this purpose, we produce semantic rankings using cosine-similarity with sentence transformers (Reimers and Gurevych 2019) leveraging a pre-trained model based on RoBERTa (Liu et al. 2019). To obtain a reasonable balance between the amount and quality of candidates, we conducted pilot experiments involving expert annotators[2]. We presented them with candidate sentences from different similarity thresholds. This process resulted in

[1] https://erisk.irlab.org/BDISen.html

[2] Prior research showed a high variance in symptoms distributions, since for some of them is easier to retrieve relevant sentences (Mowery et al. 2017; Zhang et al. 2022b).

a minimum value of 0.6 to filter out candidates. We further restricted the assessor's work to the first 750 ranked sentences in those symptoms where this threshold produced too many candidates.

Annotation schema. After selecting the candidate sentences, a team of expert annotators consisting of a psychologist, a speech therapist, and a PhD student with knowledge in the field were responsible for annotating BDI-Sen. The three annotators individually examined the whole set of candidates, with all supplementary metadata removed beforehand to avoid potential bias. To ensure the quality of the labels, we conducted training sessions with the annotators. We discussed the labelling rules with all of them, providing examples of positive and negative cases for each symptom. We instructed the annotators on the goal of our study, and explained the concept of relevance: a sentence is deemed relevant if it offers information about the specific symptom for the individual. Specifically, each annotator answered the following question in a binary setting (Positive/Negative): *Does the sentence offers information about the symptom, and the user talks in first person?* If in doubt, annotators could leave a sentence unlabelled, and there was no time limit on their annotations. We presented the sentences for each symptom in a different ranking, and the same sentence can appear in the rankings for different symptoms. Each sentence was considered positive following a majority voting approach among the annotators' decisions. Finally, we obtained a total of 4973 annotated sentences. The inter-annotation agreement among the three annotators was 84.93%, which is a substantial agreement compared to similar works (MacAvaney et al. 2018; Mowery et al. 2017; Zhang et al. 2022b).

Overall Annotation Results. The first row of Table 6.1 shows the main statistics of our dataset. The first row display the average annotator agreement per symptom (in percentage). The analysis of the agreement is conducted in the next subsection (§ 6.2.2). Next, we can see the number of positive, negative and control sentences obtained: (1) *Positive sentences* are those identified as relevant to the BDI-II symptoms, with a total of 853 sentences. (2) *Negative sentences* represent the highest percentage of annotations, totalling 4120. Despite being semantically related to the symptom, the negative sentences are not relevant to it. However, they can still be valuable for developing efficient depression detection models, being examples of false positives, one of the main challenges in detecting risks in social media (Loveys et al. 2018c). (3) We include a set of *Control sentences*. For each symptom, we obtain ten sets of control sentences, each set having the same number of sentences as the negative group. The control sentences were randomly obtained from the rest of the sentences not selected for annotation. The experts annotated the 17% sentences from the pool of candidate ones as relevant. Among the BDI-II symptoms, *Loss of pleasure* has the most annotations (739), while *Low libido* has the least (24). Comparing the positive and negative groups, we can see that the number of sentences annotated as negative is always higher than the number of positive ones.

Severity Weak Labels. In addition to the relevance labels provided by our annotators, using the eRisk2019 users as data source allowed us to include severity labels (0-3) for each BDI-II symptom. The severity labels correspond with the responses from users who authored the sentences to the BDI-II. We leveraged this additional information using a weakly-supervised approach to generate weak labels for each sentence. Specifically, we assigned the severity label corresponding to the user's response to each sentence related to the symptom. Table 6.2 shows examples of sentences from our dataset, along with their binary relevance labels and weak severity labels for the symptom *Sleep issues*. For instance, looking at the relevant sentence *"I just have energy to eat and sleep"*, its author responded 3 for that symptom. Therefore, the weak severity label is 3. On the other hand, if we observe the sentence *"I'm lying in my bed, and I'm still feeling it"*, despite being topically related to sleep, it is not relevant to the symptom. In this case, the severity label is 1. The severity labels allowed us to study the relationship between language and symptom severity at the sentence level, despite not having severity labels annotated by experts for each sentence.

6.2.2 *Dataset Analysis*

Next, we present an analysis of the constructed dataset. This section aims to determine if there are any differences among the three groups (positive, negative, and control) and among the positive group along the different symptoms. Following Ríssola et al. (2020) approach, we analyze psycholinguistic and emotional features that characterize the writing style from the groups (Cohan et al. 2018; Ríssola et al. 2022; Yates et al. 2017). While the previous works studied the overall language of positive individuals vs control ones, we present the analysis at the symptom level in this case. First, Table 6.1 shows the main statistics and vocabulary comparison of the three groups of sentences for each symptom. The first block (first four rows) corresponds to the average annotation agreement of the symptom, along with the number of sentences per group. We note the high agreement among the symptoms, with only five having an agreement of less than 80%. While we considered including the Cohen's

Table 6.1: Main statistics and vocabulary comparison of the three sentences groups. The first block shows the average annotator agreement (in percentage) and the number of sentences per group (positive, negative and control). The second and third blocks display the Jaccard's Index and the KLD comparison of the groups, respectively.

	Sadness	*Pessimism*	*Sense of Failure*	*Loss of Pleasure*	*Guiltiness*	*Punishment*	*Self-dislike*	*Self-incrimination*	*Suicidal Ideas*	*Crying*	*Agitation*
Agreement (%)	81.1	75.4	85.7	79.1	93.7	91.6	93.5	83.6	89.0	87.1	78.9
# Pos. sent. (P)	154	72	62	140	27	18	60	10	44	34	26
# Neg. sent. (N)	490	202	237	599	290	269	445	108	186	244	135
# Con. sent. (C)	4900	2020	2370	5990	2900	2690	4450	1080	1860	2440	1350
Jaccard (P vs N)	0.21	0.21	0.18	0.19	0.15	0.09	0.15	0.13	0.21	0.14	0.13
Jaccard (P vs C)	0.10	0.12	0.10	0.11	0.07	0.04	0.06	0.07	0.11	0.08	0.09
Jaccard (N vs C)	0.15	0.16	0.15	0.17	0.14	0.14	0.15	0.16	0.16	0.15	0.17
Jaccard (C vs C)	0.25	0.21	0.20	0.27	0.21	0.20	0.24	0.18	0.19	0.21	0.18
KLD (P||N)	1.01	1.40	1.54	0.91	1.50	1.69	1.41	1.95	1.20	1.79	1.98
KLD (P||C)	1.42	1.67	1.92	1.20	1.94	2.29	1.87	2.36	1.67	2.12	2.37
KLD (N||C)	1.24	1.49	1.57	0.96	1.51	1.51	1.28	1.56	1.40	1.37	1.69
KLD (C||C)	0.74	1.20	1.12	0.63	1.00	1.04	0.78	1.55	1.24	1.09	1.45

	Social issues	*Indecision*	*Worthlesness*	*Low energy*	*Sleep issues*	*Irritability*	*Appetite issues*	*Concentration*	*Fatigue*	*Low libido*
Agreement (%)	85.6	76.7	90.5	81.2	87.1	86.5	85.2	86.3	85.5	80.6
# Pos. sent. (P)	25	22	28	16	9	55	8	13	27	3
# Neg. sent. (N)	118	58	119	108	22	220	19	128	102	21
# Con. sent. (C)	1180	580	1190	1080	220	2200	190	1280	1020	210
Jaccard (P vs N)	0.14	0.19	0.15	0.14	0.15	0.16	0.15	0.12	0.16	0.11
Jaccard (P vs C)	0.11	0.11	0.09	0.07	0.09	0.09	0.09	0.07	0.09	0.05
Jaccard (N vs C)	0.17	0.15	0.14	0.14	0.13	0.16	0.13	0.15	0.16	0.14
Jaccard (C vs C)	0.18	0.15	0.18	0.17	0.12	0.20	0.12	0.18	0.18	0.13
KLD (P||N)	1.62	1.64	1.92	1.96	2.54	1.24	2.45	1.97	1.61	2.61
KLD (P||C)	1.94	2.24	2.30	2.53	2.95	1.95	2.92	2.50	2.21	3.50
KLD (N||C)	1.45	1.89	2.00	1.85	2.60	1.41	2.61	1.45	1.68	2.62
KLD (C||C)	1.50	1.99	1.43	1.60	2.65	1.07	2.71	1.44	1.58	2.59

Table 6.2: Examples of paraphrased sentences for the symptom *Sleep issues*.

(Relevance, Severity)	**Sentence**
(N,1)	*"I'm lying in my bed, and I'm still feeling it."*
(N,2)	*"You might be having trouble sleeping or anything."*
(N,3)	*"If it persists, consult a sleep expert."*
(P,1)	*"I have sleeping issues, that's why I miss school."*
(P,2)	*"Even when I'm exhausted, I can't sleep."*
(P,3)	*"I just have energy to eat and sleep."*

kappa coefficient, we decided against it since our dataset labels were highly unbalanced. In scenarios where labels are very unbalanced (e.g., our positive sentences represent a small percentage of the candidates), kappa can be a misleading measure of agreement. In particular, for rare classes, very low kappa values may not necessarily reflect low rates of overall agreement (Viera, Garrett et al. 2005). Therefore, using the average annotator agreement may be a more appropriate measure.

Words Usage. The second block of Table 6.1 corresponds to the Jaccard index between the sentence groups. This index is a statistic used to quantify the diversity of sample sets (Fletcher, Islam et al. 2018). Therefore, the higher the Jaccard value, the more similar the use of words from the groups[3]. Visualizing these results, we see that positive vs control are the groups with the least common vocabulary. For example, in some symptoms like *Punishment feelings*, they only share the 4% of the vocabulary. On the other hand, the most similar groups are positive vs negative (average Jaccard index of all symptoms of 15.60%). This makes distinguishing between negative and positively labelled sentences hard when only considering bag of words models (e.g. 'you might be having trouble sleeping.' is a challenging negative sentence).Control vs control groups also obtained similar numbers in terms of word usage (average of 18.81%). To calculate the numbers for the control vs control groups, we used the ten different control sets and computed the average over all the possible pairs.

Words Distribution. We analyzed the differences in word probability distributions among groups. The third block of Table 6.1 reports the difference in word probability distributions among groups. We measured how the probability distributions (i.e., the language models) differ using

[3] Please note that the comparison with the control group is always the averaged value over the ten sampled control sets.

Kullback-Leibler Divergence (KLD). If the two distributions are identical, the KLD value is 0. Visualizing the numbers, we observe that the word distributions for most symptoms have more KLD when comparing positive vs control groups. Again, we observe lower similarities between positive and negative groups.

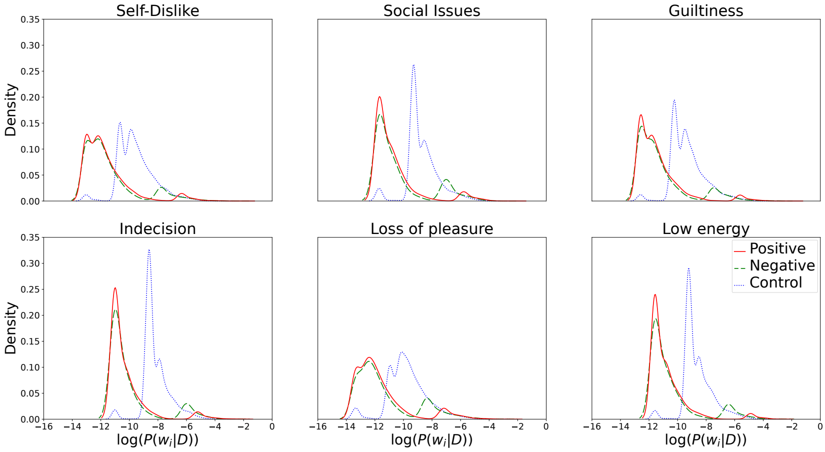


Figure 6.1: Density plot comparing word distributions of the three sentence groups for different symptoms.

Finally, Figure 6.1 illustrates the kernel density estimation (KDE) of the word probabilities of six BDI-II symptoms. KDE represents the words distribution using a continuous probability density curve. More specifically, we used a Gaussian kernel to smooth the observations. The x-axis represents the logarithm of the word probabilities. Thus, the right side of this axis corresponds to the words with higher probabilities (i.e., frequent words). The y-axis corresponds to the kernel density estimations. We compare the word distributions of the LMs from the three groups considered. We may observe apparent differences between the control vs positive/negative groups. The word probabilities in the control groups result in a high density of words with high frequencies (i.e., the control group uses common words more frequently). However, that is not the case in the positive and negative groups, where many used words correspond to less probable terms (i.e., they use uncommon terms more frequently). Moreover, the distributions of positive vs negative groups show more differences on the right side of the x-axis (associated with high probability words), where the positive group uses more common words than the negative. These differences may correspond with first-person pronoun use (more popular and more used by depressed individuals (Ortega-Mendoza et al. 2022)) versus second-person pronoun usage.

Emotions and Sentiments Association: Similar to prior works that revealed significant differences in emotional expressions between depressive and control groups (De Choudhury et al. 2013a; Ríssola et al. 2020), we investigated to extend this type of analysis at the symptom level. We used the Plutchik set of emotions (Plutchik 1980), which considers: 1) eight primary *emotions*: anger, fear, sadness, disgust, surprise, anticipation, trust, and joy and 2) two basic *sentiments*: positive (SP) and negative (SN). To quantify the emotion levels, we relied on the NRC emotion lexicon (Mohammad and Turney 2013), which includes a set of words associated with the Plutchik emotions. In our analysis, we calculated the percentage of sentences from each group (positive, negative, control) that contain at least one word associated with the primary emotions and sentiments.

When comparing these results among the groups, we identified two different patterns among the symptoms. We illustrate both in Figure 6.2. The symptoms in the first row show marked differences between the positive and negative/control groups. For example, for the symptom *Social Issues*, the percentage is always the highest in the positive sentences, with terms associated with *fear* or *SN* being in more than 30% of the positive sentences. Interestingly, the percentage of words referring to *SP* is also higher. This aligns with previous studies that demonstrated individuals with depressive conditions tend to be more emotional in social media (Ríssola et al. 2022). However, in the second row, we observe a different pattern. The differences are much lower in this second type of symptom, with a high degree of overlapping.

6.3 EXPERIMENTS

In this section, we provide an experimental analysis to evaluate the impact of the BDI-Sen dataset. As previously discussed, integrating clinical symptoms for developing mental health detection models has significant practical implications. For this reason, we divided our experiments into two tasks: 1) *Symptom Detection* and 2) *Symptom Severity Classification*. In the symptom detection task, we explored models identifying sentences relevant to BDI-II symptoms. On the other hand, the severity classification task leverages the four levels of severity of the BDI-II (corresponding with the four possible responses to each symptom) to classify the sentences based on them (0-3). In addition, to evaluate the generalization

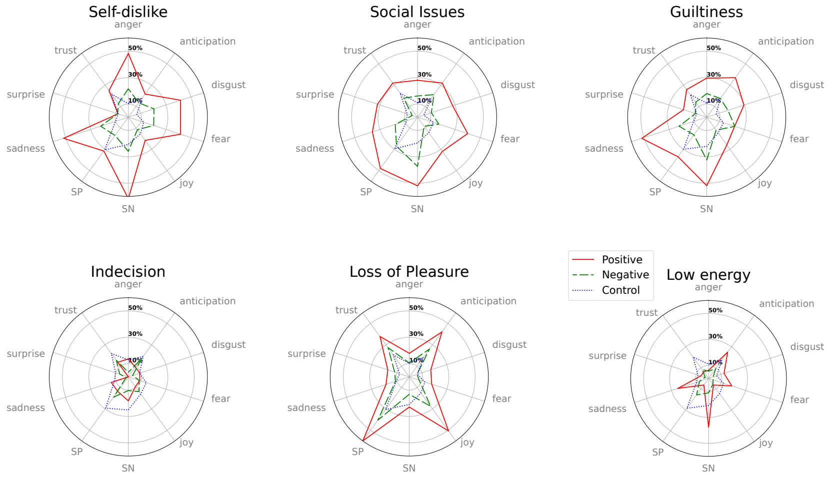


Figure 6.2: Radar plots illustrating the percentage of sentences that contain a word associated to the Plutchik emotions for each group (positive, negative, control).

ability of our classification models, we also explored how the models trained on BDI-Sen behave on sentences from symptoms related to other mental diseases from the PsySym dataset (Zhang et al. 2022b).

6.3.1 *Models*

Similar to recent literature (Nguyen et al. 2022a; Zhang et al. 2022b), we considered different types of LLMs formulated as classifiers. First, we used BERT-based models (Devlin et al. 2018) for text classification. We finetuned the pre-trained BERT base uncased model, which represents a strong baseline. We also finetuned MentalBERT (Ji et al. 2022) (**MBERT**) [4], a masked language model explicitly trained for the mental health domain. MBERT is pretrained with a corpus coming from subreddits associated with various mental diseases. As the last BERT variant, we included **BERT-mini**[5], a cost-effective alternative to BERT with fewer parameters, to explore the performance of a more lightweight model. Finally, we included Text-to-Text Transfer Transformer (T5) (Raffel et al. 2020) in our experiments, which we finetuned to generate labels in textual form.

[4] https://huggingface.co/mental/mental-bert-base-uncased

[5] https://huggingface.co/prajjwal1/bert-mini

In addition to these deep learning models, we included two traditional classification approaches based on textual features. We used Term Frequency-Inverse Document Frequency (TF-IDF) features with a linear classifier based on Logistic Regression (LR) to predict the labels (**TF-IDF+LR**). We also explored text features derived from LIWC categories. LIWC (Pennebaker et al. 2001) provides a set of linguistic categories that can extract psychological features from the text, such as the presence of words related to positive or negative emotions. We extracted the LIWC features for each sentence and employed those with an Support Vector Machine SVM classifier (**LIWC+SVM**). These two traditional approaches are good baselines for examining the improvements of complex deep learning models.

6.3.2 *Experimental Settings*

In all our experiments, we used three splits of our dataset corresponding with training/validation/testing in a ratio of 7:1:2. For the training and validation sets, we included the sentences annotated as positive, and we randomly selected the same number of control sentences to balance the labels. That resulted in 1194 sentences in the training set and 172 in the validation set. We included more control sentences for the testing split to simulate a more realistic scenario. When processing user data in social networks, most sentences are not about depressive symptomatology. In a real setting, there is a high unbalanced towards the control class. For this reason, the number of control sentences in the test set is always five times greater than the number of positive sentences (resulting in 1026 sentences in the test split).

It is worth mentioning that, in our experimentation, we have not considered the sentences labeled as negative, focusing solely on the positive and control sentences. During the initial stages of our experiments, we explored the inclusion of negative sentences. However, we observed that due to their semantic similarity to positive sentences, the models struggled to effectively distinguishing and accurately classifying them. Therefore, we have left the inclusion of negative sentences as an area for future investigation and potential improvement.

Regarding model choices and hyperparameters, in case of the TF-IDF+LR model, we removed stopwords and used 5-fold-cross-validation with the regularization strength (C) as hyperparameter in the next ranges: $[0.1, 0.5, 1, 2, 5, 10, 100, 1000]$.

In LIWC+SVM, we incorporated all the 64 categories from LIWC, used 5-fold-cross-validation with linear and RBF kernels, and the penalty

Table 6.3: Symptom detection results of our classification models on BDI-Sen.

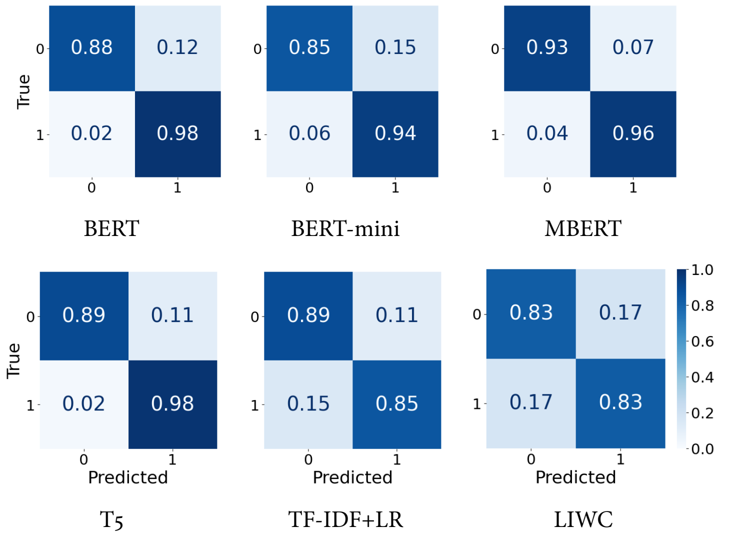
Method	AUC	P	R	F1
TF-IDF+LR	0.87	0.61	0.85	0.71
LIWC+SVM	0.83	0.49	0.83	0.62
MBERT	**0.95**	**0.74**	0.96	**0.83**
BERT	0.93	0.63	0.98	0.77
BERT-mini	0.90	0.57	0.94	0.70
T5	0.94	0.65	**0.98**	0.78

parameter C in the ranges: $[0.1, 1, 10, 50, 100]$. We follow the same procedure for all the transformer-based models by using existing implementations from the HuggingFace library. Thus, we did not include any additional hyperparameter tuning. Specifically, for MBERT, BERT, and BERT-mini, we used a learning rate of $2e^{-5}$, the maximum sequence length of 128 during 20 epochs and a batch size of 32. For T5, we used a learning rate of $1e^{-3}$, a maximum sequence length of 256 during 10 epochs with a batch size of 16.

6.3.3 *Symptom Detection*

Identifying symptoms is crucial for diagnosing and researching mental health diseases (Walsh et al. 2020a). Therefore, detecting depressive symptoms may be highly beneficial for early detection from social media data. In the symptom detection task, our goal is to determine if a sentence is relevant to a depressive symptom or not. We formulate this task as a binary classification problem, where the models detect if the sentence is related to a depressive symptom (1) or not (0). For the T5 model, we finetuned it by training T5 to generate *"true"* or *"false"* tokens. Table 6.3 shows the results of all our methods considered for symptom detection.

The results in Table 6.3 show that all the methods have relatively high F1 and AUC, with AUC scores ranging from 0.83 to 0.95, and F1 scores from 0.62 to 0.83. We can see that the transformer-based models, except BERT-mini, performed better than methods based on textual features (TF+IDF and LIWC). The results also show that the standard BERT model and T5 perform similarly despite T5 being pre-trained on a larger corpus of data. In line with prior research (Ji et al. 2022; Zhang et al. 2022b), the model pre-trained on mental health-related corpora (MBERT) achieved



BERT BERT-mini MBERT

T5 TF-IDF+LR LIWC

Figure 6.3: Confusion matrices showing the predictions accuracy of our symptom detection methods.

higher scores in almost every metric. The models seem to perform worse in terms of precision, with MBERT obtaining the best value (0.74) and LIWC with the worst (0.49).

To better understand these results, Figure 6.3 provides a visual representation of the distribution of true and false predictions made by each of our classification models. These numbers show that the transformer-based models have a high ratio of true positives, from 0.94 (BERT-mini) to 0.98 (T5 and BERT). The percentage of false negatives for these models is small, with less than 0.06 in all cases. On the other hand, the number of false positive errors is higher. In the mental health domain, missing an individual at risk of being reviewed by professionals, is much more worrying than therapists examining a healthy person. For this reason, a good prediction performance for false negatives is crucial. Finally, we can also observe that the prediction errors of the methods using the textual features have the lowest accuracy overall.

6.3.4 *Symptom Detection - Generalization*

Recent studies have demonstrated the low generalizability of mental disease detection models (Harrigian et al. 2020). In this experiment, we want to analyze whether models trained on the BDI-Sen data can generalize to detect symptoms from other mental illenesess. We evaluated this premise using the *PsySym* dataset (Zhang et al. 2022b). The mental disorders covered by PsySym are *depression, anxiety, attention deficit hyperactivity disorder (ADHD), bipolar disorder, obsessive-compulsive disorder (OCD), post-traumatic stress disorder (PTSD) and eating disorder*. We aimed to test the ability of our models to generalize across these conditions, given the potential overlap in symptom expression between different mental disorders. For this purpose, we used the same models in the *symptom detection* task[6] trained in BDI-Sen and tested them over the symptoms of the seven mental disorders of PsySym.

Table 6.4 shows the results of our models on the PsySym data. The number of positive test sentences for each disease is indicated between brackets. We also display their number of common symptoms with the BDI-Sen symptoms (third row). We only considered positive sentences of each illness and reported the precision that the models achieved. Based on the figures, deep learning methods exhibited good generalization capabilities to other mental diseases. However, a significant performance gap exists between the models using textual features (TF-IDF and LIWC+SVM) and transformer-based ones. Specifically, the best-performing model, T5, achieved an average accuracy of 0.81 across the symptoms for all diseases. Meanwhile, the worst (TF-IDF+LR) had a precision of only 0.44.

These results suggest that the models trained on our dataset can generalize well to symptoms from other mental diseases. However, as shown in the *Disease Average* row 6.4, the performance varies among illnesses, indicating that some disorders may be more challenging to detect than others. Unsurprisingly, the models have their best accuracy when evaluated in depression, with an average of 0.81. For the other diseases, the results suggest that the more symptoms they share with BDI-Sen, the better the model performs. Specifically, we achieved at least 0.70 accuracy in anxiety, bipolar disorder, OCD, and PTSD. In contrast, the worst results correspond to ADHD and eating disorders, with accuracy numbers of 0.59 and 0.51, respectively. Conducting these types of multi-disease analyses may provide valuable insights into the similarities and differences between different mental health conditions, potentially leading to new avenues of research.

6.3.5 *Symptom Severity Classification*

In this experiment, we aim to classify the sentences from BDI-Sen based not only on whether they are relevant to the symptom but according to the

[6] The experimental settings remains the same as the described in § 6.3.2.

Table 6.4: Generalization ability results of our models with other mental diseases. The Table shows the precision of the proposed sentence classification models when confronted with the positive sentences for the different disorders from the PsySym dataset Zhang et al. 2022b. Second row displays the number of test sentences from that disease. In the third row, c.s. (*common symptoms*) refers to the number of symptoms that are in common between the disease and BDI-Sen.

Method	**Depression**	**Anxiety**	**ADHD**	**Bipolar Disorder**	**OCD**	**PTSD**	**Eating Disorder**	**Method Average**
	(1433 sent.)	(2822 sent.)	(528 sent.)	(1131 sent.)	(449 sent.)	(1284 sent.)	(907 sent.)	
	(14 c.s.)	(19 c.s.)	(4 c.s)	(14 c.s)	(2 c.s)	(5 c.s)	(4 c.s)	
TF-IDF+LR	0.60	0.46	0.30	0.55	0.43	0.43	0.33	*0.44*
LIWC+SVM	0.69	0.73	0.46	0.53	0.64	0.62	0.49	*0.59*
MBERT	0.88	0.80	0.61	0.82	0.83	0.82	0.48	*0.75*
BERT	0.89	0.80	0.71	0.84	0.79	0.82	0.58	*0.78*
BERT-mini	0.85	0.84	0.74	0.79	0.62	0.69	0.58	*0.73*
T5	0.93	0.87	0.72	0.87	0.87	0.82	0.58	***0.81***
Disease Average	***0.81***	*0.75*	*0.59*	*0.73*	*0.70*	*0.70*	*0.51*	*0.68*

Table 6.5: Symptom severity classification results of our methods considering all the BDI-II severity levels and control sentences.

Method		Severity 0			Severity 1			Severity 2			Severity 3			Control		
	Micro F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
TF-IDF+LR	0.79	0.15	0.30	0.20	0.55	0.27	0.36	0.21	0.23	0.22	0.41	0.37	0.39	0.87	0.96	0.92
LIWC+SVM	0.76	0.00	0.00	0.00	0.80	0.24	0.37	0.00	0.00	0.00	0.03	1.00	0.06	0.83	0.95	0.89
MBERT	0.80	0.35	0.08	0.13	0.52	0.37	0.43	0.32	0.22	0.26	0.44	0.39	0.42	0.86	0.99	0.92
BERT	0.77	0.24	0.01	0.10	0.56	0.32	0.40	0.34	0.19	0.24	0.25	0.20	0.22	0.82	0.99	0.90
BERT-mini	0.70	0.05	0.50	0.09	0.89	0.25	0.39	0.03	0.33	0.05	0.00	0.00	0.00	0.83	0.98	0.90
T5	0.66	0.12	0.05	0.07	0.09	0.09	0.09	0.06	0.06	0.06	0.09	0.07	0.08	0.76	0.83	0.80

declared severity level. This task represents a step forward from our previous experiments enabled by the weak labels we provide in the BDI-Sen from the users' response to the BDI-II. By identifying the severity of each symptom, mental health detection models may provide a more nuanced and accurate diagnosis of an individual's situation. We formulate the task as a multi-classification problem. The models classify each sentence severity according to the BDI-II schema, with the levels ranging from 0 to 3. We refer the reader to Table 6.2 to see descriptions and example sentences from our datasets of the severity levels. In this experiment, we used the same text classification models trained in a multi-class setting and considered two experimental variants:

1. The first experiment considers all severity levels, which includes a separate category for control sentences that were randomly selected (i.e., unrelated to any symptom). The aim was also to investigate whether the multi-class classification models may distinguish sentences talking about the symptom in a non-negative way (severity level 0) from those unrelated to the symptoms (control). Table 6.5 presents the results of our classification methods under this setting. We can observe a significant decrease in performance compared to the symptom detection experi-

ments, where only two classes were considered. Although all methods achieved a reasonably good Micro F1 score due to the large number of control sentences in the test set, there was poor performance in sentences in non-control classes. Furthermore, the gap in performance between the transformer-based and textual feature models is reduced, with T5 being the worst-performing method. MBERT remains the top-performing model across all severity levels.

To further analyze these results and examine prediction errors between categories, Figure 6.4 (a) presents the confusion matrices for the best-performing method (MBERT). The matrix shows very few misclassification errors between severities that are far apart. For instance, for the *True* sentences with severity label 3, none of them were labelled as 0 or control sentences. Similarly, for the sentences with severity level 2, only 6% of them were misclassified with the level 0, and none of them were misclassified as control. Overall, most prediction errors occurred between severity levels 1, 2 and 3, indicating that the models find it challenging to correctly distinguish between categories with subtle differences. The obtained results highlight the challenge of accurately classifying sentences of symptoms based on their graded severity. To enhance the performance of our models in this regard, it would be beneficial to employ a labeling methodology that involves a substantial number of sentences manually labeled by domain experts.

2. After analysing the above results, we performed an additional experiment, combining in one class the control and labelled sentences with severity level 0. The rationale behind this is that, when using severity detection approaches, the main practical interest would be to detect high-risk sentences. Both severity level 0 (i.e., no risk) and control sentences may not provide much value to support the diagnosis (they would sum up zero to the BDI-II final score). The more severe and negative symptoms expressions are more likely to require attention. Therefore, we grouped them to investigate this more practical scenario.

Table 6.6 shows higher Micro F1 values than the previous experiment. The accuracy of this new class is higher for all models than the one of the control class from previous results. However, even on most occasions, their F1 values are improved, the models still struggle to distinguish between severity levels with risk. MBERT is still the top-performing model, and its F1 scores for severity classes were 0.46, 0.25, and 0.41. T5 continues to be the worst model in this multi-class scenario. Finally, we also included the confusion matrices of the MBERT model in Figure 6.4 (b).

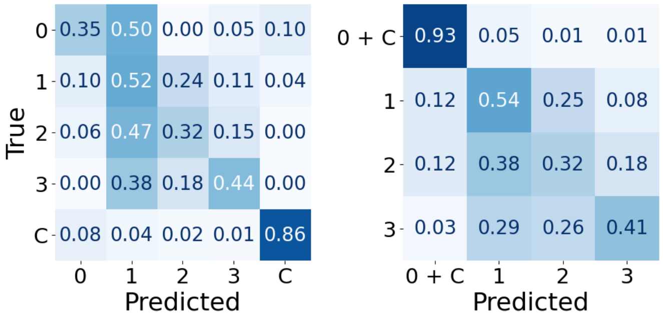
Table 6.6: Symptom severity classification results considering grouping the BDI-II severity level 0 and control sentences.

Method		Sev. 0 + Control			Severity 1			Severity 2			Severity 3		
	Micro F1	P	R	F1	P	R	F1	F1	P	R	P	R	F1
TF-IDF+LR	0.83	0.91	0.96	0.93	0.45	0.32	0.37	0.26	0.22	0.24	0.38	0.35	0.37
LIWC+SVM	0.84	0.94	0.93	0.93	0.54	0.31	0.40	0.00	0.00	0.00	0.00	0.00	0.00
MBERT	0.86	0.93	0.98	0.95	0.54	0.40	0.46	0.32	0.20	0.25	0.41	0.40	0.41
BERT	0.85	0.92	0.99	0.95	0.58	0.34	0.43	0.18	0.17	0.17	0.35	0.41	0.38
T5	0.75	0.85	0.88	0.86	0.15	0.16	0.15	0.00	0.00	0.00	0.20	0.14	0.16

As in the previous experiment, the matrices reveal that most misclassifications occur between adjacent severity levels. Specifically, only 3% of the sentences with severity level 3 were mislabeled as severity level 0.

6.4 CONCLUSIONS

In this chapter, we presented BDI-Sen, a symptom-annotated dataset for depression that includes manually labelled sentences addressing the 21 BDI-II symptoms. By leveraging the eRisk2019 collections as data source, our dataset provides binary relevance labels for the BDI-II symptoms and weak labels regarding their severity level. We designed a retrieval phase to filter-out candidate sentences based on the descriptions of the BDI-II elements, and three experts decided the actual relevance of the candidates. We explored this resource, revealing linguistic and emotional differences among the symptoms. Moreover, we performed two main experiments with state-of-the-art models trained solely on BDI-Sen: symptom detection and symptom severity classification, including an extensive error analysis for both tasks. The good generalization ability of our models



(a) Control sentences separated from 0 level.

(b) Control sentences grouped with 0 level.

Figure 6.4: Confusion matrices of our best method (MBERT) classifying different symptom severity levels.

further underlines the usefulness of BDI-Sen as a resource for developing robust mental health detection models.

7

THE ROLE OF LLMS AS ASSESSORS OF PSYCHOLOGICAL MARKERS

In the preceding chapter, we presented *BDI-Sen*, a dataset consisting of sentences annotated with symptoms related to depressive disorder. This comprehensive dataset encompasses all the symptoms present in the BDI-II. In addition to releasing and analyzing this dataset, we conducted a series of experiments to explore its applicability in various tasks. The aim was to advance models that leverage symptom markers and enhance the robustness of mental health detection models. Aligned with this purpose, the eRisk 2023 initiative fosters research on this area and has recently proposed a new ranking task focused on developing search methods to find sentences related to depressive symptoms. This search challenge relies on the symptoms specified by the BDI-II. In line with our efforts to improve and expand such collections, in this chapter, we continue this line of work by releasing DepreSym. The creation of this dataset is based on the participant systems' results, consisting of 21 580 sentences annotated according to their relevance to the 21 BDI-II symptoms.

In this resource, the labelled sentences come from a pool of diverse ranking methods, and the final dataset serves as a valuable resource for advancing the development of models that incorporate depressive markers such as clinical symptoms. Due to the complex nature of this relevance annotation, we designed a robust assessment methodology carried out by three expert assessors (including an expert psychologist). Additionally, we explore here the feasibility of employing recent Large Language Models (ChatGPT and Generative Pre-trained Transformer (GPT-4)) as potential assessors in this complex task. We undertake a comprehensive examination of their performance, determine their main limitations and analyze their role as a complement or replacement for human annotators. The contributions and dataset presented in this chapter are currently available on arxiv (Pérez et al. 2023) and have been previously submitted to the 46th European Conference on Information Retrieval (ECIR), and it is currently under review.

7.1 INTRODUCTION AND MOTIVATION

Inspired by clinical practice, there has been a growing interest in designing predictive models that focus on identifying depressive symptoms. These approaches diverge from traditional depression screening models that rely on the presence of general markers, which are based on the use of engineered features (e.g., word counts, emotion levels, posting hours). However, these features offer less personalized and interpretable solutions (Harrigian et al. 2020; Walsh et al. 2020a). Based on this idea, recent studies, like the ones commented in previous chapters, have shown the potential of symptom-based detection models (Nguyen et al. 2022b; Pérez et al. 2023b; Pérez et al. 2022; Zhang et al. 2022a).

Two recent studies have attempted to fill this void by constructing fine-grained datasets that label depressive symptoms at sentence level (Perez et al. 2023; Zhang et al. 2022a). The first paper contributes to this line of research by introducing *DepreSym*, a dataset to encourage the development of models that rely on symptom-level screening of depression. *DepreSym* consists of 21 580 sentences that are labeled in terms of their relevance to the BDI-II symptoms[1]. This resource comes from a shared-data ranking task introduced in the CLEF 2023 eRisk Lab[2]. To construct our dataset, three expert assessors annotated a pool of sentences associated with each symptom. The candidate sentences were obtained using top-k pooling from the relevance rankings designed by the participants in the task, with a total of 37 different ranking methods presented. Pooling over the participants' results helps increase the diversity of the candidate sentences.

The assessors were instructed to consider the candidate sentences as relevant if they were on-topic but also provide explicit information about the individual state related to the symptom. This two-side notion of relevance is more complex compared to previous works, requiring us to develop a robust annotation methodology with formal assessment guidelines. To validate the effectiveness of our methodology, we calculate the inter-rater agreement and conduct further analysis of the resulting set of judgements.

[1] https://erisk.irlab.org/depresym_dataset.html

[2] https://early.irlab.org/

High-quality assessments are essential to obtain accurate and reliable results (Büttcher et al. 2007), and annotations need to be consistent, unbiased, and representative of the task at hand. Low-quality assessments potentially lead to inaccurate evaluations and unreliable conclusions (Scholer et al. 2011). Manually annotating test collections requires significant human effort, frequently requiring domain experts. Consequently, several steps have been taken to reduce the cost and biases of the labelling process (Moghadasi et al. 2013; Sakai 2009). With the incredible development of LLMs, a potential application of these models is to assist in tasks such as relevance labelling. This represents a natural advance, as was the replacement of TREC annotators by crowdsourcing (Alonso and Mizzaro 2009).

Initial steps were taken by Gilardi and his colleagues (Gilardi et al. 2023) demonstrating that ChatGPT outperforms crowd-workers for a tweet annotation task. Other researchers focused their efforts on improving their performance as annotators through prompt engineering (He et al. 2023). Faggioli et al. (2023) tested the accuracy of LLMs for annotating two TREC test collections. In this chapter, we intend to go one step further by evaluating the most recent LLMs for a highly demanding annotation task. Specifically, we put them under scrutiny for assessing the relevance of sentences given specific BDI-II symptoms. Thus, we are considering a scenario where the two-side relevance notion is more complex (i.e., on-topic and providing explicit information about the individual). Moreover, the context is much shorter (i.e., only judging short sentences). To study this effect, we analyse the agreement between human annotations, including those coming from experts in the field, and machine annotations.

In this part of the thesis, we explore the ability of recent state-of-the-art LLMs to annotate the dataset. Specifically, we employ the latest versions of GPT conversational applications (ChatGPT (Forbes 2022) and GPT-4 (OpenAI 2023)) as complex relevance assessors. One of the main advantages of LLMs is their ability to accurately process large amounts of data, which can significantly reduce the time and effort required for manual assessment. Comparing the performance of LLMs with human assessors provides insights into the strengths and limitations of both approaches. Human assessors are considered the gold standard for relevance assessment, but they are also subject to biases and errors that can affect their performance. Examining the performance of LLMs in relation to humans can help us to understand how well these models can replicate human behaviour.

7.2 RESOURCE

This section describes the construction of *DepreSym*, a resource derived from Task 1 of the eRisk 2023 Lab. This is a novel task that consists of identifying sentences that are indicative of the presence of clinical symptoms from the individuals who wrote these sentences.

As in the rest of this thesis, we follow the BDI-II, a well-studied clinical questionnaire. It includes emotional (*Pessimism or Sadness*), cognitive (*Indecision*) and physical (*Fatigue*) symptoms (Beck et al. 1996a). The sentences come from a large corpus of users' posts that were written by multiple social media users, coming from the Reddit platform. The users' posts were segmented into sentences and a TREC-style collection was created (3 807 115 sentences from 3107 unique users). All extracted sentences were public and Reddit terms allow the use of its contents for research purposes.

Table 7.1: Examples of sentences for the symptom *Loss of Energy*. Sentences are paraphrased for anonymity purposes.

Relevance	Sentence
0	"*Learn new ideas consumes energy, but builds neural connections.*" "*Low electrolytes can cause a person to feel low on energy.*"
1	"*Even brushing my teeth is too exhausting for me right now.*" "*I became constantly lethargic, drowsy, and unable to concentrate.*"

The eRisk participants were given the full collection of sentences and were asked to submit 21 rankings of sentences (one for each BDI-II symptom) ordered by decreasing relevance to the symptom. Each participant team could submit up to 5 variants (runs) and each ranking had up to 1000 sentences. Prior to annotation, we obtained candidate sentences by following a top-k ($k = 50$) pooling approach on the submitted runs (37 runs from 10 different teams). Table 7.1 provides two examples of candidate sentences annotated as relevant and non-relevant for the symptom *Loss of energy*. Note that all sentences in the Table are somehow on-topic, but only those in the lower block were labelled as relevant. These two relevant sentences offer insights into the individuals' state related to the BDI-II symptom. This stringent notion of relevance adds complexity to the labelling process. The first block of Table 7.2 reports the total number of annotated and relevant sentences. Here, the number of relevant sentences corresponds to the ones unanimously agreed upon by all human assessors. We can see that the number of relevant sentences is substantially low, with the 11% of the sentences annotated as relevant from the pool of candidates (pool sizes ranging from 829 to 1150 sentences). The number of relevant sentences ranges from 21 to 260. The rest of the blocks correspond to the annotations agreement among the symptoms, explained in Subsection 7.4.1.

7.3 MANUAL ANNOTATION PROCESSS

A sentence should be considered relevant only if it provides "information about the individual state related to the BDI-II symptom". To that end, we designed a set of instructions that guide the assessment process [3]. The guidelines were given to the human annotators and, additionally, these textual instructions were used to prompt the LLMs in our evaluation of automatic judgements.

We selected three human assessors with different backgrounds: a field expert (background in Psychology), a PhD student and a Postdoc (both with backgrounds in Computer Science). First, we asked them to label the pools of the first three BDI-II topics. At this point, judges were allowed to mark sentences as “undecided”. Next, we calculated pairwise, Cohen’s Kappa to assess the agreement between single raters, and Krippendorff’s Alpha for ordinal scales to evaluate the agreement between all raters [4]. Kappa values ranged between 0.18 and 0.51, with a median of 0.38 for the three initial symptoms. Mean Krippendorff’s α was 0.32. These Kappa values indicate low average agreement, and α falls below the desirable limit of $\alpha \geq 0.667$ for reliable annotations (Krippendorff 2018; McHugh 2012).

Next, we had a briefing with the three annotators to resolve ambiguities and make the assessments more consistent. After this meeting, they were asked to relabel again the three initial symptoms. This time, “undecided” labels were not allowed. Cohen’s Kappa ranged between 0.30 and 0.68 with a median of 0.55. Mean Krippendorff’s α increased up to 0.51, but still below the recommendable limit. The agreement analysis for the overall assessments (over the 21 BDI-II symptoms) led to Cohen’s Kappa between 0.58 and 0.65 (median of 0.58) and Krippendorff’s α of 0.60. These figures are much higher than those obtained before the briefing. This suggests that the annotation process was solid but, still, the agreement scores are moderate, reflecting the difficulty of the task. In any case, we produced two types of relevance assessments, consensus and

[3] https://erisk.irlab.org/guidelines_erisk23_task1.html
[4] The inter-rater agreement metrics are described in the evaluation Section 3.2.

Table 7.2: Number of sentences and annotations agreement (in percentage) statistics per symptom.

	Sadness	*Pessimism*	*Sense of Failure*	*Loss of Pleasure*	*Guiltiness*	*Self-Punishment*	*Self-dislike*	*Self-incrimination*	*Suicidal Ideas*	*Crying*	*Agitation*
# Sentences	1110	1150	973	1013	829	1079	1005	1072	953	983	1080
# Rel. sent.	179	104	160	97	83	21	158	76	260	230	69
GPT-4 vs Cons.	72.6	75.7	73.6	74.6	75.6	88.2	72.0	75.6	77.2	79.8	80.7
GPT-4 vs Maj.	76.1	77.1	76.0	82.0	81.2	89.3	81.6	81.3	85.9	87.7	80.7
PhD s. vs Rest	80.0	70.5	84.2	88.5	90.8	97.0	81.4	86.5	87.4	89.3	88.9
GPT-4 vs Rest	73.0	75.6	75.3	79.6	78.9	89.6	73.3	76.6	77.4	82.8	81.4
Psy. vs Rest	83.0	76.8	74.1	82.7	90.0	95.8	85.3	86.7	89.8	88.1	88.4
GPT-4 vs Rest	73.6	75.6	73.9	74.5	76.2	88.2	74.2	77.4	81.7	82.0	80.0
Postdoc vs Rest	85.4	78.3	84.5	82.1	86.9	92.4	84.2	86.2	89.4	88.9	87.6
GPT-4 vs Rest	74.8	77.3	74.0	77.2	77.3	88.4	78.1	78.5	81.2	82.5	80.7

	Social issues	*Indecision*	*Worthlessness*	*Low energy*	*Sleep issues*	*Irritability*	*Appetite issues*	*Concentration*	*Fatigue*	*Low libido*	**Avg.**
# Sentences	1077	1110	1067	1082	938	1047	984	1024	1033	971	-
# Rel. sent.	70	61	71	129	203	94	103	83	123	97	
GPT-4 vs Cons.	75.9	79.2	80.9	78.3	62.1	84.1	71.7	83.9	74.8	81.6	-
GPT-4 vs Maj.	81.2	83.3	84.7	84.5	76.4	88.0	80.5	88.2	83.7	85.6	
PhD s. vs Rest	87.7	90.8	90.3	88.1	80.7	86.1	84.6	90.9	88.2	92.4	86.9
GPT-4 vs Rest	78.4	80.9	81.4	79.5	65.1	85.0	75.3	84.4	77.2	82.2	78.7
Psych. vs Rest	85.7	90.6	91.7	91.9	83.1	93.6	86.8	94.0	89.2	93.3	87.6
GPT-4 vs Rest	76.6	80.5	82.9	80.8	70.0	85.6	75.2	86.8	80.0	84.8	79.1
Postdoc vs Rest	86.4	88.3	89.6	90.5	79.2	92.5	80.9	91.7	86.5	86.0	86.5
GPT-4 vs Rest	78.0	80.3	82.3	80.8	65.4	85.5	73.3	84.8	76.3	81.8	79.0

majority, and consider the first as a high-quality container of sentences that are unambiguously relevant. We compare both assessments in the next section.

7.4 LLMS AS AUTOMATIC ANNOTATORS

We present the results of the agreement between the human annotators and the two LLMs regarding sentences' relevance to depressive symptoms in Table 7.3. The analysis is conducted for the two classes of ground truth annotations: (*i*) *Consensus*, where relevant sentences are identified by all human assessors, and (*ii*) *Majority*, where relevant sentences are identified by at least two human annotators.

Under the consensus ground truth, ChatGPT accurately identified 95% of the relevant sentences, compared to 93% for GPT-4. However, both models struggled to correctly identify sentences marked as non-relevant (accuracies of 51% and 75% for ChatGPT and GPT-4, respectively). This trend persists for the majority ground truth, but the correlation with human judgements shows a significant improvement. The Cohen's κ level of agreement increases from 0.18 to 0.38 for ChatGPT and from 0.38 to 0.57 for GPT-4. These figures indicate a considerable increase in the performance of GPT-4 compared to ChatGPT. Another finding is that the "non-relevant" predictions of the models tend to be trustworthy. For instance, ChatGPT identified correctly 9832 out of 9945 non-relevant sentences. However, the predictions of relevance are much noisier. This suggests that LLMs could be the basis of a hybrid annotation approach that we will further discuss in Section 7.5.

Table 7.3: Agreement between each LLM and two types of ground truth (consensus and majority).

LLM	Prediction	Consensus			Majority		
		Rel.	Not Rel.	κ	Rel.	Not Rel.	κ
ChatGPT	Rel.	**2358**	9277	0.18	**4241**	7394	0.38
	Not Rel.	113	**9832**		290	**9655**	
GPT-4	Rel.	**2296**	4755	0.38	**3916**	3135	0.57
	Not Rel.	175	**14354**		615	**13914**	

7.4.1 *Symptom-based Agreement*

Table 7.2 also displays the agreement statistics for each symptom. In the second block, we report the percentage of agreement between GPT-4 and the two classes of ground truth. The agreement percentages are generally stable across all symptoms. These percentages are higher for majority (82.63% mean overall) compared to consensus (77.04%). Moreover, ChatGPT achieved substantially lower agreement values (65% and 56%, respectively)[5].

We further conducted a comparison between the annotations provided by GPT-4 and each human annotator, reported in the last three blocks of Table 7.2. We made pairwise comparisons between GPT-4 and each human annotator. To that end, the reference ground truth was obtained from the consensus of the two remaining human annotators. For example, to compare the PhD student vs GPT-4 we ran them against the ground truth of relevant sentences obtained from the postdoc and the psychologist. In all cases, each human annotator achieved a higher percentage of agreement than GPT-4. Only in one symptom (*Pessimism*) against the PhD student, GPT-4 achieved a higher percentage of agreement (75.57% vs 70.52%, respectively). The humans led to mean scores (last column of the lower table) that were substantially higher than those achieved by GPT-4 (greater than 85% while GPT-4 was always lower than 80%). By a narrow margin, the psychologist was the human who produced superior agreement scores.

7.4.2 *Inter-rater Agreement and Correlation of Systems Rankings*

7.4.2.1 *Inter-rater Agreement*

The inter-rater agreement, measured using Cohen's κ between ChatGPT and the human annotators, ranged from 0.29 to 0.32, with a median value of 0.31. For GPT-4, the Cohen's κ scores ranged from 0.52 to 0.54, with a median of 0.53. Additionally, Krippendorff's α for the combination of the three human annotators and ChatGPT was 0.40, while Krippendorff's α with GPT-4 was 0.56. These results confirm the previous findings that GPT-4 is a more reliable annotator compared with ChatGPT.

7.4.2.2 *Systems Rankings Correlation*

We compared the ranking of the 37 participating search systems obtained with the official assessments (consensus of the three human assessors) against a hypothetical ranking based on assessments from a single annotator. To that end, we ranked the systems by decreasing Mean Average Precision (MAP) and compared the rankings with Kendall's τ and AP Correlation (τ_{ap}[6]) (Yilmaz et al. 2008). This analysis allows us to explore to

[5] We did not include ChatGPT's results in Table 7.2 due to page limitations.

[6] τ_{ap} assigns greater weight to errors made to the systems positioned higher in the ranking.

what extent the use of a single annotator alters the system's rankings. We refer to the reader to the Evaluation Section 3.2 for a detailed description of both metrics.

Table 7.4: Correlations between the official ranking of systems (consensus qrels) and the ranking of systems obtained from the qrels of a single annotator.

	Annotators			
	GPT-4	Psychologist	Postdoc	PhD student
Kendall τ	0.86	**0.98**	0.95	0.94
τ_{ap}	0.81	**0.97**	0.91	0.88

Looking at the results in Table 7.4, we can observe that GPT-4 yields a high correlation (0.86 and 0.81), although lower than the correlation levels achieved by the human annotators. Note that the human assessors were involved in the construction of the official qrels, while GPT-4 was not part of the official evaluation process. The results also suggest that the assessment effort could have been reduced by involving a single human assessor. Notably, the psychologist correlates nearly perfectly with the official consensus-based ranking (0.98). The correlations suggest a relative order among human annotators, Psych > Postdoc > PhD student, which is a natural consequence of their domain knowledge and level of experience. Lastly, AP correlation and Kendall's τ show similar trends and, thus, the rankings from individual judges do not seem to induce major swaps at the top-ranked positions.

7.5 DISCUSSION

Our results suggest that LLMs are significantly better at identifying sentences marked as relevant in the ground truth compared to non-relevant ones. This finding deviates from the tendencies observed in prior research (Faggioli et al. 2023), wherein varying patterns emerged based on the specific dataset. We believe our results give grounds to propose a new efficient hybrid labelling strategy, where LLMs act as filters that automatically remove non-relevant sentences from the pools. As shown in Table 7.3, the "non-relevance" predictions of LLMs are quite accurate and, thus, the human annotation effort could be reduced to review those sentences estimated as relevant by the LLM. Thus, GPT-4 would reduce the human workload by approximately 68%, eliminating the need to annotate around

15 000 sentences. Considering that the average human effort per assessor was 70 hours (21 580 sentences), this reduction would save around 49 hours of work per human. Furthermore, reducing the burden on human annotators could potentially lead to improved annotation quality and allow for an increase in the size of the annotation pool, allowing for more documents to be reviewed.

7.6 CONCLUSIONS

In this chapter, we presented *DepreSym*, a novel resource to foster research on new depression screening models that rely on symptom markers at sentence level. The annotated sentences were obtained from a pooling approach that utilised multiple search systems and a thorough assessment method involving domain experts. We also reported here our endeavours to evaluate the capabilities of LLMs as relevant sentence annotators. We found that these models, particularly GPT-4, are promising but still make many false positive errors. Related to this, we further intend to explore the capabilities of other models, such as LLaMA (Touvron et al. 2023), and implement hybrid annotation approaches where the LLMs act as filters of non-relevant sentences.

8

PSYPROF: A PLATFORM FOR ASSISTED SCREENING OF DEPRESSION IN SOCIAL MEDIA

The field of mental health detection models has witnessed significant advancements, with numerous approaches developed for detecting and assessing various mental health conditions. However, effectively integrating these models into clinical settings is a persistent challenge. As discussed in previous chapters, health professionals emphasize the importance of models that rely on interpretable features, which can be easily inspected and validated. Consequently, a new line of work has emerged, focusing on developing solutions that integrate symptoms from different clinical questionnaires as reliable markers.

The translation of symptom-based models from research articles into practical tools holds great potential for advancing the integration of recent models by health professionals. This transition may be crucial for facilitating widespread adoption and maximizing the impact of these models. In this chapter, we introduce PsyProf, an innovative monitoring platform designed to screen depression in social media. The primary objective of PsyProf is to provide a comprehensive solution for estimating the severity level of depression in individuals by leveraging models that detect the presence of recognized symptoms.

PsyProf serves a dual purpose: firstly, as a demonstrative platform showcasing models for the task of severity estimation of depression. Secondly, as a tool that can be employed by healthcare professionals to automate user screening and validate the results. By offering automated assistance in the screening process, PsyProf aims to improve efficiency and accuracy in the identification and assessment of depression. To provide a broader context for each individual, we have augmented our tool with user profiling capabilities. This enhancement allows professionals to gain insights into user data labelling, leveraging both depression estimators and profiling models. The contributions presented in this chapter have been previously published in the 45th ECIR conference (Pérez et al. 2023a).

Throughout this chapter, we will explore the architecture and functionality of PsyProf and the underlying models used for depression severity estimation. Furthermore, we will discuss the potential implications and benefits of integrating such a platform into clinical practice, addressing the challenges and opportunities it presents. By highlighting the significance of PsyProf as a valuable tool for both researchers and healthcare professionals in the field of mental health assessment, we aim to contribute to the advancement and practical implementation of symptom-based models in clinical settings.

8.1 INTRODUCTION AND MOTIVATION

Social media platforms are channels people tend to consider comfortable for expressing their honest feelings and concerns (Kauer et al. 2014), where factors such as the anonymity status may influence people on a sincere manifestation of their thoughts (Chancellor and De Choudhury 2020). As exposed during this thesis, computational methods have obtained promising results in detecting mental health states by exploiting this user-generated data. There is a large body of prior work in assessing users at risk from different mental disorders, such as suicidal ideation (Ramírez-Cifuentes et al. 2020), eating disorders (Losada et al. 2019) or pathological gambling (Parapar et al. 2021a). In this context, Major Depressive Disorder (MDD), also known as depression, attracted the attention of many researchers, as it is one of the most common and debilitating mental illnesses (Hollon et al. 2002). We can find rich bodies of work identifying indicators that characterize depression based on user texts from different social platforms, such as Twitter, Reddit and Facebook (Couto et al. 2022b; De Choudhury et al. 2013a; Ríssola et al. 2021; Trotzek et al. 2018).

The solutions mentioned above obtained remarkable results in a variety of datasets (Cohan et al. 2018; Yates et al. 2017) and benchmark evaluations (Losada et al. 2019, 2020; Parapar et al. 2021a) that consider depression and control groups. However, the integration of these methods in clinical settings faces several challenges. Health professionals favour models that base their decisions on interpretable features, as they need to be inspected and validated (Walsh et al. 2020a). However, the exclusive use of engineered features does not provide enough context to be interpretable indicators (Coppersmith et al. 2018). Diverse studies have also shown that the performance of mental health models is not stable across different social media platforms (Ernala et al. 2019; Harrigian et al. 2020). To overcome these limitations, a new line of work focused on developing solutions integrating symptoms from different clinical questionnaires as reliable markers. In this regard, recent works demonstrated the potential of symptom-based models in terms of performance, interpretability and generalisation (Nguyen et al. 2022a; Pérez et al. 2022; Zhang et al. 2022a).

In this Chapter, we present PsyProf, a monitoring platform for assisted screening of depression in social media. To measure the depression severity level of the individuals, we use models that estimate the presence of recognized symptoms. For this purpose, and similar to previous chapters, we use the symptoms of a validated clinical questionnaire, the BDI-II. PsyProf also reports the BDI-II score, which is the sum of the option responses to the 21 symptoms. The BDI-II score is associated with four depression levels. Moreover, we have complemented our tool with user profiling methods that can bring wider context when measuring at-risk users. We use Reddit as the target platform due to its wide acceptance in previous studies (Cohan et al. 2018; Parapar et al. 2021a; Yates et al. 2017), and our tool provides scalability to process the large amount of data coming from this platform. Finally, the data from the social media users can be downloaded to CSV format and can help create symptom-based datasets with the inspection and labels coming from health professionals.

While several efforts have been made in the mental health detection in social media field, few works have presented platforms that directly integrate depression detection models from such media. There are, indeed, several platforms that have been developed to provide help and support, facilitating interaction with individuals suffering from depression, allowing for valuable conversational engagements (Graham et al. 2020; Morris et al. 2015). To the best of our knowledge, only one other platform with a similar orientation exists. Martínez-Castaño et al. (2020) presented *Catenae*, a Python library that facilitates the construction of scalable real-time streaming applications, exemplifying its potential in social media screening tasks. In this work, the authors demonstrated how *Catenae* can analyse Reddit publications to support the detection of early signs of depression. Their work differs from ours in that they use simple feature-based models to perform a binary classification of depression. Consequently, their application offers limited insight into their predictions, providing just a likelihood that a user might be suffering from depression. Furthermore, their application is not explicitly oriented towards the professional healthcare community for direct interaction and validation of model outcomes.

8.2 PSYPROF

We conceived PsyProf both as a demonstration platform for models for the task of severity estimation of depression and for being used by professionals for doing automated user screening and validation of the results. This dual motivation is the reason for designing the platform's use cases that show the results to the professionals for validation and correction. In this way, PsyProf is not only a proof-of-concept of the utility of the automated models for massive screening but also a tool for obtaining insights from the corrections or validations that the professionals make based on the provided evidence.

Architecture and Implementation. Figure 8.1 illustrates PsyProf's overall architecture, which consists of (1) a web-based front-end and back-end built with the web framework Django. (2) A scalable system for processing user publications from Reddit, built with Celery and Redis. Our Application Programming Interface (API) calls to the Reddit API in a asynchronously way. Therefore, the models can infer user estimations without the need of waiting to process all the remanining data. As a result, the clinicians and potential administrators of the platform can analyze the inferences from the models in real time. (3) Two different REST APIs regarding to the depression and profiling models that consume the calls from the Celery workers. Furthermore, PsyProf is portable since all the components run in a different Docker container orchestrated with Docker compose. This containerization approach ensures easy deployment and scalability, making PsyProf a flexible and adaptable solution for different environments and infrastructures.

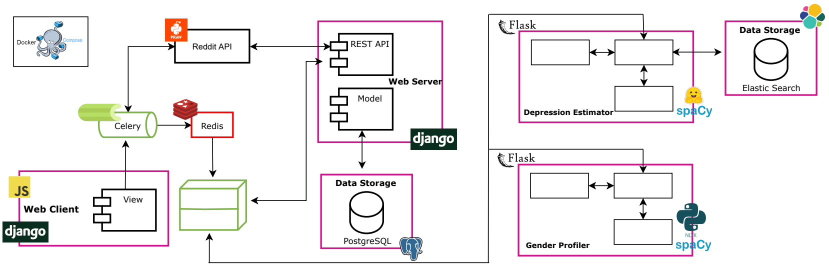
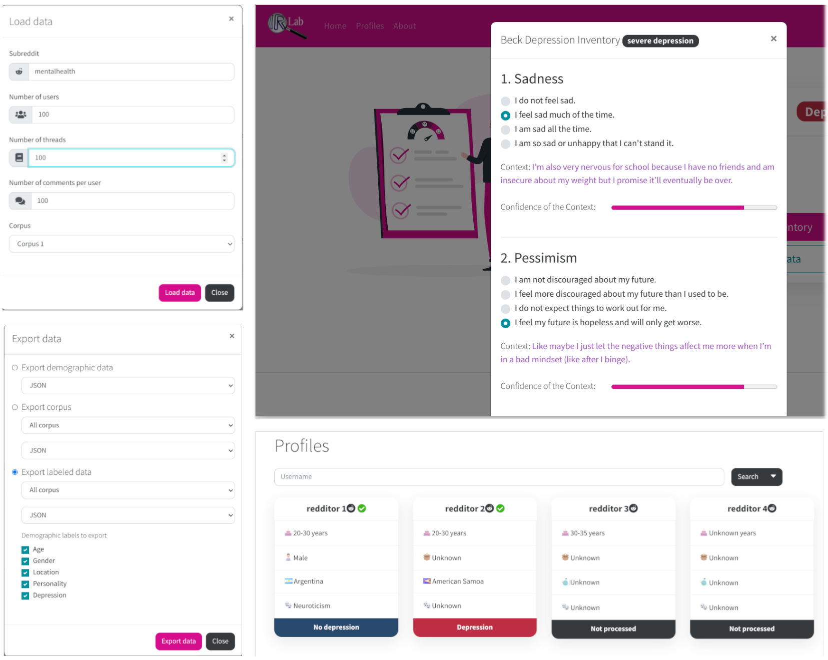


Figure 8.1: PsyProf architecture overview.

User interface and interaction. PsyProf is a web application intended to be used exclusively by clinical professionals where administrators can run experiments related to profiling and estimating depression severity levels from Reddit users. Figure 8.2 shows screenshots regarding the platform's main functionalities. The clinicians have two ways of monitoring: (1) In the upper left corner, we can see the form to obtain a pool of Reddit users estimations. It contains five fields: (*i*) the subreddit from which to obtain users. A subreddit is a specific Reddit community that is focused on a certain topic. For example, there are subreddits related to mental health problems or specific mental disorders, such as depression, eating disorders or anxiety. (*ii*) The number of users to process, (*iii*) the number of threads per user and (*iv*) comments per user to inspect. Finally, (*v*) the platform also allows you to determine the corpus, as it let you structure the users processed via different corpuses. After filling in these fields, the application will obtain estimates of users that meet these characteristics. (2) Moreover, PsyProf includes the feature to obtain the estimations of specific users by introducing the Reddit username. The number of fields are the same, but instead of selecting the subreddit, in this case, we only have to introduce the specific nickname. The application also allows the clinician to export all the data. This can be seen as a tool for creating new unsupervised or supervised collections, and it can be very promising for different research purposes. It includes the functionality to export the data to JSON and CSV formats, and including as labels the depression estimation and the profiling attributes of each user.



Load data

Subreddit: mentalhealth

Number of users: 100

Number of threads: 100

Number of comments per user: 100

Corpus: Corpus 1

Export data

○ Export demographic data: JSON

○ Export corpus: All corpus, JSON

● Export labeled data: All corpus, JSON

Demographic labels to export

☑ Age
☑ Gender
☑ Location
☑ Personality
☑ Depression

Beck Depression Inventory severe depression

1. Sadness

○ I do not feel sad.
● I feel sad much of the time.
○ I am sad all the time.
○ I am so sad or unhappy that I can't stand it.

Context: I'm also very nervous for school because I have no friends and am insecure about my weight but I promise it'll eventually be over.

Confidence of the Context:

2. Pessimism

○ I am not discouraged about my future.
○ I feel more discouraged about my future than I used to be.
○ I do not expect things to work out for me.
● I feel my future is hopeless and will only get worse.

Context: Like maybe I just let the negative things affect me more when I'm in a bad mindset (like after I binge).

Confidence of the Context:

Profiles

Username

redditor 1 ✓	redditor 2 ✓	redditor 3	redditor 4
20-30 years	20-30 years	30-35 years	Unknown years
Male	Unknown	Unknown	Unknown
Argentina	American Samoa	Unknown	Unknown
Neuroticism	Unknown	Unknown	Unknown
No depression	Depression	Not processed	Not processed

Figure 8.2: User interfaces related to the main functionalities of PsyProf.

Data Inspection and Validation. When a clinician requests to process users, PsyProf showcases the model estimations under the "Profiles" view, as seen in the bottom screenshot of Figure 8.2. Within this view, we have two categories of users: (1) Users not yet validated, labelled as *Not processed*. (2) Validated users, labelled either *Depression* or *No depression*. A *validated* label indicates that a clinician has reviewed the model's predictions and has confirmed the depression risk assessment.

Both user categories are profiled based on attributes such as *age, gender, country, and personality*. Currently, PsyProf has an API designed specifically for gender profiling. However, its flexible platform design allows for seamless integration of other profiling tools linked to different attributes. The gender profiling tool displays gender information only when the model's confidence level exceeds a preset threshold. If not, the label *Unknown* appears.

A primary functionality of PsyProf is its capability to assess symptoms as clinical markers, which can be viewed at the top of Figure 8.2. Thus, the integrated models within PsyProf can automatically predict responses to the BDI-II symptoms, allowing clinicians to view the predicted responses for all symptoms. Additionally, the platform computes the total BDI-II score, indicating the user's associated severity level. These models also emphasize the user's most pertinent comment regarding the symptoms to provide decision context. It is pivotal to recognize that these model-generated responses are not absolute. The platform empowers clinicians to manually modify any decision, if clinicians consider it neccesary.

Models. We utilize two distinct models for user estimations:

1) The depression estimator, which predicts responses to the 21 symptoms of the BDI-II questionnaire, effectively filling it out automatically (refer to the right component in Figure 8.1). For this purpose, we use the Sense2Vec model from Pérez et al. (2022), which is based on the use of word embeddings to capture the semantics of the symptom options. The predict these options, Sense2Vec compares the similarity of each option with the user embedding representation[1]. To do so, we index the embeddings from the training texts associated to each symptom option in an Elasticsearch index and perform a vector search using the test user embedding as query. Following this approach, we can compare test users with the options of the BDI-II. Sense2vec training data corresponds to the eRisk2019 collections (Losada et al. 2019), obtained from Reddit.

2) The gender profiler, which includes a set of models for author profiling based on the work of Piot-Perez-Abadin et al. (2021) using as training

[1] We refer the reader to the Chapter 4 for more information about the depression detection model

data the collections from *PAN Author Profiling* shared task. This task belong to the CLEF campaign in 2019 (Rangel and Rosso 2019). The authors used feature engineering to calculate the most relevant text features from the corpus. Specifically, they experimented with three main types of features (sociolingusitic, sentiment analysis and topic modelling). Within the sociolinguistic features, they included repeated alphabets, emojis, the use of personal pronouns or Part-of-Speech (PoS) information. Regarding the sentiment analysis, the authors used the NLTK sentiment analysis analyzer to extract the sentiment scores for each user. Finally, with respect to the topic modelling part, they applied Latent Dirichlet Allocation (LDA) to obtain the twenty most significant topics per user.

8.3 CONCLUSION

In this chapter, we introduce PsyProf, a web platform for assisted examination and monitoring depressive symptoms in social media users. PsyProf is conceived as a demonstrative platform to produce effective depression screening tools. To improve the interpretability of the decisions, the platform also includes a gender profiler model, which allows to improve the context.

PsyProf does not intend to replace health professionals but rather to complement their work. Due to the sensitive nature of the mental health domain, we do not provide public access to the platform. Instead, we provide a demonstration video[2] and the source code of the platform and the profiler models are publicly available[3][4]. Finally, and following eRisk policies, the depression models will be available under research data agreement.

[2] https://irlab.org/psyprof.mp4

[3] https://github.com/palomapiot/early

[4] https://github.com/palomapiot/profiler-buddy

Part IV

CLOSING

9

CONCLUSIONS AND FUTURE WORK

This concluding chapter summarizes the primary conclusions of this doctoral thesis. We also delve into the ethical nuances associated with detecting mental health markers in social media. Finally, we provide future work suggestions.

9.1 CONCLUSIONS

In this thesis, our research aim was to investigate the development of models based on clinical symptoms to identify depressive signs. We adhered to an accepted clinical questionnaire by considering the symptoms covered in the BDI-II. Using this questionnaire, we explored classification frameworks for depression severity estimation of social media users. We also explored the construction of new resources for helping the development of new symptom-based models. Finally, and with the aim of a practical integration of depression detection models, we incorporate our contributions into a demonstrative platform to be used by health professionals. In the following, we present in more detail the findings of this work.

In PART II, we proposed different classification frameworks to automatically estimate the 21 symptoms of the BDI-II questionnaire from online user-generated content. To this aim, in Chapter 4 we used word embeddings to explore the presence of BDI-II symptoms depending on their sensitivity. Sensitivity refers to users' inclination to discuss the symptoms openly. To validate this idea, we proposed two methods: 1) General symptom-classifiers designed to capture general language patterns from social media users and 2) Direct-symptom classifiers aimed at identifying explicit mentions of symptoms. Our experimental findings showed the effectiveness of our approaches in estimating depressive states. At the same time, the methods are also flexible and easy to interpret.

Furthermore, outcomes from our symptom-by-symptom analysis indicated that some symptoms are more challenging to capture than others. As a result, individuals might be more predisposed to discuss certain symptoms publicly on social platforms. Such tendencies highlight the importance that the stigma and other external factors play in symptom manifestation.

In Chapter 5, we built upon the work presented in the preceding chapter, introducing a classification pipeline that estimates depression severity through semantic similarities. Again, we designed symptom-classifiers aligned with the BDI-II symptoms. In this work, we focus on selecting users' posts related to depressive symptoms by exploring different data selection strategies. Once we selected the most risky posts from the test user, we produce a semantic ranking that gives training labelled sentences that we know they are associated with depressive symptoms. Subsequently, we utilize the training sentences derived from these rankings as evidence for predicting symptoms severity in users. We rely on pre-trained models with SBERT to calculate the semantic similarities. This new approach surpassed our previous results, achieving state-of-the-art performance in two different collections in terms of measuring the depression level of individuals. More specifically, our best method correctly estimates at least 50% of the depression levels for both collections. To construct the training sentences, we also proposed an annotation schema to obtain them. Finally, we illustrated how our semantic pipeline provides interpretability of the symptom decisions.

The development of depression detection models based on symptom markers is important. Yet, the foundation of such models (i.e., the data they rely upon) is equally significant. Moreover, this is even more important considering that the use of symptoms is a recent line of research, and there was only one dataset in the literature dedicated to symptom markers on social media (Zhang et al. 2022b). Constructing robust datasets allows models to observe different symptom manifestations and can help them to have better performance and generalisation. In Chapter 6, we presented *BDI-Sen*, a symptom-annotated sentence dataset for depressive disorder. BDI-Sen contains 4973 annotated sentences covering the 21 symptoms from the BDI-II. For the dataset construction, we designed an initial retrieval stage by filtering candidate sentences that may be relevant to each symptom. After the candidate selection, we followed a manual annotation schema carried out by three assessors. Outcomes from the overall annotation results revealed that the number of sentences annotated as negative is always higher than positive. Only the 17% of candidates were annotated as positive, which reflects the challenge of retrieving sentences associated with depressive symptoms.

Furthermore, we conducted an depth symptom-by-symptom analysis, examining the linguistic and emotional features of both positive and negative sentences. In terms of word distribution, we observed lower differences when comparing positive and negative groups for most of the symptoms. Nonetheless, for certain symptoms, there were notable variances in the eight primary emotions between the depressive and control groups, using the NRC lexicon of Mohammad and Turney (2013). For instance, for the symptom *Social Issues*, we found many more emotional terms linked to emotions of fear and sadness.

We further carried out an experimental analysis to assess the impact of this dataset in two different classification tasks: 1) Symptom Detection and 2) Symptom Severity Classification. To do so, we employed various types of LLMs formulated as classifiers. In the symptom detection task, our trained models showed a great prediction performance for detecting relevant sentences to depressive symptoms. Moreover, in alternative experiments, we showcased their good generalization ability when considering symptoms from other diseases. However, when these models are formulated to do multi-classification based on different levels of symptom severity, there was a notable drop in their efficacy. These results may point out the need for more severity-labelled sentences to train models accurately with this level of granularity.

In light of the insightful analysis and the performance of our models under BDI-Sen, the next part of this thesis explores deeper into the creation of resources centred on depressive symptoms. Specifically, we designed a robust methodology by predefined guidelines [1] and studied how modern conversational LLMs perform compared to human assessors. In Chapter 7, we introduced *DepreSym*, a collection of 21 580 sentences annotated for their relevance to the BDI-II symptoms. Distinct from *BDI-Sen*, this resource is derived from the shared-data ranking task presented at the CLEF 2023 eRisk Lab. As a result, the candidate sentences were sourced using top-k pooling based on the relevance rankings derived by task participants, which resulted in a diverse set of 37 ranking methods.

In Chapter 7, we also investigate the efficacy of conversational LLMs (ChatGPT and GPT-4) in scenarios of complex relevance annotation. This involves ensuring sentences are not only symptom topics but also provide information about their author. Our research yielded insights into the capabilities and limitations of these models. To study this effect, we had three human assessors who annotated an entire set of candidate

[1] https://erisk.irlab.org/guidelines_erisk23_task1.html

sentences, of which only 11% were annotated as relevant. With this golden truth established, we carried out a series of experiments leveraging LLMs as automated annotators. Studying their agreement with the human annotations, we observed several findings. Both LLMs displayed modest agreement with human judgments in the sentences associated with depression, as evidenced by ChatGPT and GPT-4, with Cohen Kappa scores of 0.18 and 0.38, respectively. Interestingly, in discerning non-relevant sentences, the models displayed remarkable precision. ChatGPT, for instance, accurately identified 9832 out of 9945 non-relevant sentences.

We further examined the agreement of the LLMs with the human annotators per each symptom, which remained consistent with an average agreement rate of 82.63% across all symptoms. Additionally, we also made a comparison between each assessor. As expected, the highest agreement was observed among human evaluators. Intriguingly, the assessor with a clinical psychology background showcased the strongest alignment with the LLMs assessments. The outcomes of our experiments support the idea of efficient, hybrid labelling strategies that exploit the capabilities of LLMs for constructing resources. In our proposed scenario, LLMs would serve as preliminary filters, automatically discarding non-relevant sentences. This would allow human annotators to concentrate solely on sentences deemed “relevant" by the LLMs. Applied to our dataset, this method could alleviate human annotators’ workload by roughly 68%, obviating the need to review about 15 000 sentences. Given that each assessor, on average, spent 70 hours annotating 21 580 sentences, this approach translates to a time-saving of approximately 49 hours for each individual. Such efficiency enhances the quality of annotations due to reduced fatigue and paves the way for expanding the volume of annotations, enabling a broader review of sentences.

In our last work (Chapter 8), we introduced PsyProf, a demonstrative platform designed to assess depression severity. We aimed to integrate our previous models on BDI-II estimation to demonstrate a practical integration of their depression screening capabilities. Designed to be used by health professionals, PsyProf is a proof-of-concept of the potential of automated models in large-scale screening but also a tool for obtaining insights from the corrections or validations that the professionals make based on the provided evidence. PsyProf is a web-based application built with the web framework Django. It processes Reddit user publications asynchronously, with its core functionality being the prediction of responses to the BDI-II symptoms. This scenario allows clinicians to validate the predictions and modify any decision if necessary. It is also accompanied by profiling models to provide as much context as possible.

In addition to its screening functionalities, PsyProf allows administrators to export all the data, which can be seen as a tool for creating new unsupervised and supervised collections for different research purposes.

Throught this doctoral thesis, a systematic and rigorous methodology was consistently employed. In the first part of the dissertation, we explored the existing literature to understand the current state of depression detection techniques, particularly those using data from social media platforms. A key component of the methodology used in this thesis was the aligment with established standards in the domain of online risk detection. It follows the eRisk initiative protocols, which stand as a recognized benchmark in this area. All the classification models were evaluated following the eRisk schemas, providing a structure framework for testing grounded in real-word scenarios. As a result, the results from these works were compared against several eRisk participants ensuring a fair comparison. Regarding the datasets construction, BDI-Sen and DepreSym, all the process that involved data mining from social media platforms ensured privacy measures and rigorously annotating data. Given the sensitive nature of this topic, strict ethical guidelines were adhered during data collections, which are detailed in the next section.

9.2 ETHICAL STATEMENT AND DISCUSSION

Ethical considerations play a crucial role in the rapidly evolving domain of mental health detection on the Internet. Previous research has delved into the ethical aspects of exploiting vast amounts of online information for predictive modeling (Chancellor et al. 2019; Walsh et al. 2020a). Key areas of concern highlighted by studies include handling sensitive personal data and the vital importance of anonymizing user information (Saha et al. 2017). There is also an emphasis on the crucial involvement of domain experts during data analysis and creation (Chancellor et al. 2016), as well as the need for models that yield interpretable and high-quality outcomes for healthcare professionals (Walsh et al. 2020a). This chapter discusses the ethical considerations that have guided our research. We have prioritised ethical considerations at every step of this dissertation, ensuring that our work will serve as a positive advancement in the field.

In this thesis, we released two datasets related to sentences with depressive symptoms coming from the Reddit platform. First, we have taken meticulous steps to ensure data privacy and integrity standards. Reddit is a publicly available source, and the sentences were collected in such a way that they rely on the exempt status under title 45 CFR §46.104.

We adhered to the corresponding data usage policies. We ensured that personal information could not be identified from the data. The essence of our datasets lies in the content of the messages rather than the identity of users. Thus, anonymization is crucial. Furthermore, to prevent misuse, we have adhered to strict permissions and licensing for the datasets, ensuring that they are only used for research purposes. Beyond data gathering, we have received active feedback from domain experts to validate the relevance of the resources. This not only guarantees the quality of the data but also ensures it is being used appropriately in the broader context of mental health research. Regarding the annotation methodology, the annotators did not report any adverse effects after their work.

Another fundamental part of this dissertation was the presentation of different classification frameworks that automatically estimate the 21 symptoms of the BDI-II questionnaire. Focusing on symptoms offers several advantages that enhance the transparency of model outcomes. First, our methods move away from making oversimplified claims about a user's mental health. This not only provides a more detailed insights into their health state but also reduces the potential for misclassification. For instance, stating that a user exhibits signs consistent with 'diminished interest in activities" or 'feelings of worthlessness" is more tangible and understandable than stating they have a 70% likelihood of depression. This granularity provides clearer insight into the rationale behind the model's decision, promoting trust and understanding. By aligning our methods to the BDI-II clinical criteria, our methods invite scrutiny and dialogue. This open-door approach allows for continuous feedback, ensuring our methods evolve with ethical and clinical considerations.

Finally, in terms of impact in real-world settings, there is still work to be done to produce practical depression screening tools. The development of such technologies should be approached with caution to ensure that their use is ethical and respects patient privacy and autonomy. Our work aims to supplement the efforts of health professionals rather than replace them. We acknowledge the validation gap between mental health detection models and their clinical applicability. We aim to develop automated technologies that complement current online screening approaches. The final decision must always be supported by the validation of a health professional. Our study highlights the potential of NLP-based approaches in assisting clinicians with diagnosis, but further research and testing are needed before it can be considered for clinical deployment.

9.3 FUTURE DIRECTIONS

The findings and outputs of this thesis have paved the way for numerous research opportunities. These not only encompass improvements to our existing methods and resources but also explore novel applications. Every chapter provided new insights accompanied by future directions to investigate. Next, we propose future lines of work to continue the research presented in this thesis.

- In PART II, our attention was primarily centred on developing classification frameworks to estimate the 21 symptoms covered by the BDI-II. An interesting avenue for future exploration lies in adapting our approaches to other depression assessment tools, such as the PHQ-9 (Cameron et al. 2011) or the Hamilton Rating Scale for Depression (Hamilton 1980). This would not only allow a comparison of the symptoms covered by each questionnaire but also to study the different performance and correlation obtained among the questionnaires. Parallel to this approach, there is potential in translating our models to diagnose related disorders. For instance, assessing pathological gambling tendencies could be pursued with tools such as the DSM-V. Similarly, eating disorders might be examined with tools like the Eating Disorder Inventory-III (Espelage et al. 2003). However, venturing into these new applications mandates the prerequisite of appropriate training data to explore the precision and validity of our methods.

- A significant challenge for mental health detection models on social media is their limited capability to explain their predictions clearly. For healthcare professionals to trust and rely on these models, they must understand how the decisions are made (Chancellor et al. 2019). To address this, we plan to develop models that offer trustworthy and comprehensive explanations to detect the presence of depressive symptoms. A promising direction is the utilization of generative language models designed to provide a clear rationale for each prediction they make. Specifically, we will explore text-to-text approaches that accomplish two main objectives: classifying the relevance of social media publications to depressive symptoms and explaining the classification decisions. By prioritizing explainability, we will aim to reduce the gap between automated predictions and human understanding, facilitating more informed clinical decision-making.

- In PART III, we explored how well our models, explicitly trained for BDI-Sen, could recognize symptoms of other mental disorders. This was based on the premise that many mental conditions have overlapping symptoms. Encouraged by the strong generalization ability shown by our models, we are interested in further investigating how models designed for depression can be applied to other mental diseases. Another intriguing aspect we aim to delve into is the significance of the presence of certain symptoms in the diagnosis of one disease compared to another. For instance, a symptom common to two diseases might be more critical for diagnosing one condition over the other. Pursuing such multi-disease studies can offer profound insights into the similarities and differences between mental disorders.

- Another promising avenue for exploration involves optimizing the balance between the speed and accuracy of mental health detection models. The sooner a potential risk is identified, the earlier healthcare professionals can take appropriate actions, which can be life-saving in critical cases. For this reason, eRisk organized the first shared task on early risk detection of depression in 2017 (Losada et al. 2017). To tackle the time-aware nature of this framework, the organizers proposed specific metrics that consider both the accuracy and delay of the predictions. Our preliminary efforts in this area showed potential (Couto et al. 2022a). In future work, we aim to adapt our symptom-based models further to consider time-aware metrics. Our goal will be to explore an optimal balance between rapid classification and maintaining high-quality, reliable outcomes.

- In the continuously evolving domain of digital mental health, the potential to exploit symptom detection models with content recommendation presents a promising potential for improving personalized support. As we have discussed in earlier parts of this thesis, we can implement models that detect depressive symptoms from user-generated content on platforms like social media. Building upon this foundation, there is an exciting opportunity to integrate these detections with recommender systems. Traditional mental health resources often offer generic advice. By leveraging the precision of symptom classification, we can exploit recommender systems that suggest highly personalized content. For instance, is a user's posts strongly align with symptoms of isolation or loneliness, the system could recommend resources specifically addressing these feelings, ensuring the user feels seen and understood. Moreover, instead of waiting for a user to seek help, proactive recommendations can be made based on detected symptoms. By continually monitoring user content and providing timely recommendations, these platforms can act as first-line responders, guiding users towards the help they need.

APPENDICES

A

PUBLICATIONS

In this appendix, we list all the articles published during the doctoral period. For the conferences, we provide their rank according to CORE 2023[1]. For each journal, we detail its Journal Citation Reports Impact Factor[2] and its quartile.

A.1 CONFERENCE ARTICLES

Manuel Couto, Anxo Pérez and Javier Parapar. “Temporal word embeddings for early detection of signs of depression.” In: *Proceedings of the CIRCLE (Joint Conference of The Information Retrieval Communities in Europe)*. Toulouse, France, 2022.

Jorge Gabín, Anxo Pérez and Javier Parapar. “Multiple-Choice Question Answering Models for Automatic Depression Severity Estimation.” In: vol. 7. 1. MDPI, 2021, p. 23.

Anxo Pérez, Marcos Fernández-Pichel, Javier Parapar and David E Losada. “DepreSym: A Depression Symptom Annotated Corpus and the Role of LLMs as Assessors of Psychological Markers.” In: 2023.

Anxo Perez, Javier Parapar, Alvaro Barreiro and Silvia Lopez-Larrosa. “BDI-Sen: A Sentence Dataset for Clinical Symptoms of Depression.” In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Taipei, Taiwan, 2023. DOI: 10.1145/3539618.3591905. CORE 2023: A*.

Anxo Pérez, Paloma Piot-Pérez-Abadín, Javier Parapar and Álvaro Barreiro. “PsyProf: A Platform for Assisted Screening of Depression in Social Media.” In: *European Conference on Information Retrieval*. Springer. Dublin, Ireland, 2023, pp. 300–306. CORE 2023: A.

[1] The CORE 2023 Conference ranking is available at: http://portal.core.edu.au/conf-ranks.

[2] The JCR Impact Factor can be consulted at: https://jcr.clarivate.com/jcr/home.

Anxo Pérez, Neha Warikoo, Kexin Wang, Javier Parapar and Iryna Gurevych. "Semantic Similarity Models for Depression Severity Estimation." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapeur, 2023, "To be published". CORE 2023: A*.

A.2 JOURNAL ARTICLES

Anxo Pérez, Javier Parapar and Álvaro Barreiro. "Automatic depression score estimation with word embedding models." In: *Artificial Intelligence in Medicine* 132 (2022), p. 102380. DOI: https://doi.org/10.1016/j.artmed.2022.102380. IF 2021: 7.5 (Q1).

B

EXTENDED SUMMARY IN SPANISH

In accordance with the current Regulations of the PhD Studies of the Universidade da Coruña, we present in this appendix an extended summary of this doctoral thesis in Spanish.

B.1 RESUMEN

Por un lado, existe una extensa evidencia proveniente de los campos de la medicina y la psicolingüística sobre los cambios en el uso del lenguaje de las personas que sufren problemas de salud mental. Se ha observado que tanto el lenguaje como nuestro comportamiento con los demás puede cambiar cuando enfrentamos problemas de salud mental. Estas manifestaciones lingüísticas ofrecen pistas cruciales para identificar posibles trastornos. Paralelamente, las redes sociales, en su explosivo auge, se han convertido en un gran repositorio de datos lingüísticos. En este contexto, muchas personas, buscando un espacio de apoyo o simplemente un lugar para expresarse, recurran a plataformas como Twitter, Reddit o Facebook para discutir sobre sus problemas de salud mental.

Dada esta conjunción de factores, no es de extrañar que la lingüística computacional haya visto una tendencia emergente hacia el aprovechamiento de estas vastas cantidades de datos para fines diagnósticos, especialmente en la detección temprana de trastornos como la depresión, que, según las estadísticas actuales, está liderando como uno de los trastornos mentales más comunes a nivel global.

Aunque este campo de investigación es prometedor, no estamos pisando terreno inexplorado. Ya existen estudios que han abordado la detección de la depresión utilizando datos de redes sociales y han obtenido cifras alentadoras usando conjuntos de datos provenientes de Twitter, Reddit o Facebook. Sin embargo, un área de mejora identificada es la interpretabilidad de los modelos. Si bien estos modelos pueden clasificar con precisión, a menudo carecen de transparencia en sus decisiones, lo que podría generar escepticismo en la comunidad médica. Si no se entienden las razones subyacentes de la clasificación de un modelo, es mucho más complicado que los profesionales sanitarios puedan actuar en función de ello.

Abordando esta preocupación, en esta tesis, hemos decidido explorar un enfoque centrado en síntomas depresivos. Nuestro objetivo es utilizar indicadores clínicamente validados, como los síntomas presentes en el Inventario de Depresión de Beck (BDI-II). Esta herramienta no solo detecta la presencia de depresión, sino que también proporciona una métrica de su gravedad. Al incorporar este enfoque basado en síntomas, esperamos ofrecer resultados más transparentes y, por lo tanto, más confiables para los profesionales de la salud. Para ello, nuestros trabajos se encuentran en la intersección de los campos de Recuperación de la Información (IR), Procesamiento de Lenguaje Natural (NLP) y Aprenzidaje Automático o *Machine Learning* (ML).

Las contribuciones de esta tesis tienen tres enfoques diferentes: i): nuevos modelos para la estimación de la gravedad basados en marcadores de síntomas, ii) la creación de conjuntos de datos para ayudar al desarrolo de métodos basados en síntomas, y iii) la exploración de los recientes modelos masivos de lenguaje para ayudar a escalar la creación de estos datasets. Como último paso, y en nuestra búsqueda de una integración práctica de los modelos de detección de la depresión, incorporamos nuestras aportaciones anteriores a una plataforma demostrativa para su uso por parte de profesionales sanitarios. Finalmente, presentamos nuestras conclusiones y discutimos los resultados obtenidos a través de nuestra investigación. Esta tesis contribuye a avanzar en la comprensión y detección de la depresión a través de marcadores de síntomas, y sienta las bases para futuras investigaciones en esta área crítica de la detección de la depresión en las redes sociales.

B.2 INTRODUCCIÓN

Los trastornos mentales, incluida la depresión, están entre los problemas de salud pública más prevalentes. Según la Organización Mundial de la Salud (OMS), aproximadamente 332 millones de personas en todo el mundo padecen un trastorno depresivo [1]. La salud mental juega un papel fundamental en fomentar la felicidad, promover la interacción social y contribuir a la salud individual y de la población. Una buena salud mental es un requisito fundamental para el éxito en todos los as-

[1] https://apps.who.int/iris/handle/10665/254610

pectos de la vida. Además, impacta significativamente en la producción nacional y la productividad laboral [2]. Se sabe bien que la intervención temprana en trastornos depresivos es esencial para mitigar su impacto y consecuencias (Picardi et al. 2016). Sin embargo, debido al estigma que rodea a los trastornos mentales, más del 60% de las personas afectadas no buscan apoyo profesional (Gulliver et al. 2010), lo cual es particularmente preocupante considerando el creciente número de casos entre los jóvenes (Thapar et al. 2022). Para ayudar con este problema, los gobiernos y las agencias han lanzado programas para concienciar sobre la importancia de la salud mental en sus ciudadanos. Sin embargo, los recursos limitados de los sistemas de salud pública restringen gravemente su capacidad para detectar y diagnosticar casos (Arango et al. 2018).

Como alternativa a los sistemas de salud pública, las plataformas sociales son un canal prometedor para evaluar riesgos de manera no intrusiva (Coppersmith et al. 2015). La proliferación de las redes sociales constituye un recurso valioso para detectar signos tempranos de depresión. Las personas que experimentan depresión a menudo encuentran consuelo al expresar sus pensamientos y emociones en estas plataformas, motivadas por factores como la privacidad y el anonimato (Callahan and Inckle 2012; Kauer et al. 2014). En consecuencia, las redes sociales ofrecen una oportunidad única para acceder a información valiosa sobre los riesgos de salud de los individuos que de otro modo sería imposible de obtener. Investigadores en los campos de Recuperación de Información (IR), Procesamiento de Lenguaje Natural (NLP) y Aprendizaje Automático (ML) han aprovechado los vastos recursos de las redes sociales para obtener avances considerables en la detección de signos de depresión (Ríssola et al. 2021). Sin embargo, una limitación en los esfuerzos actuales de investigación es la necesidad de una mayor interpretabilidad en las decisiones de los modelos (Walsh et al. 2020a). En el dominio de la detección de salud mental, donde una interpretación fiable de los resultados es crucial para los clínicos, se vuelve esencial que los modelos produzcan resultados de clasificación confiables e interpretables (Ernala et al. 2019).

En línea con este objetivo, esta tesis doctoral se centra en el desarrollo de modelos basados en síntomas clínicos validados para identificar signos depresivos en redes sociales. Incorporando marcadores clínicos en las decisiones de los modelos, nuestro objetivo es mejorar la interpretabilidad de sus resultados por parte de profesionales de la salud. Por esta razón, nos adherimos a protocolos clínicos establecidos, considerando los 21 síntomas incluidos en el Inventario de Depresión de Beck (BDI-II), un

[2] https://www.who.int/europe/health-topics/mental-health

cuestionario ampliamente utilizado para medir la depresión. El BDI-II abarca una serie de síntomas depresivos como irritabilidad, pesimismo o problemas de sueño. El BDI-II no solo sirve como herramienta para detectar la depresión, sino también como un instrumento de calificación para estimar la severidad. Al adherirnos a protocolos clínicos, nuestra aspiración es construir modelos predictivos que no solo detecten sino que también proporcionen estimaciones detalladas de la severidad, dotando así a los profesionales de la salud con herramientas robustas para un diagnóstico completo.

B.3 MOTIVACIÓN

Los trastornos depresivos tienen numerosos efectos perjudiciales. Sin embargo, existen tratamientos que han sido validados y efectivos, y pueden ser potenciados con terapias y programas de intervención (Duarte et al. 2009). Como se subrayó en la Introducción, una detección temprana y precisa reduce significativamente el impacto negativo del trastorno (Halfin 2007; Picardi et al. 2016). En la práctica clínica, el diagnóstico y la gravedad de la depresión se basan en pruebas psicométricas validadas. Estos cuestionarios tienen un desempeño satisfactorio al diagnosticar a individuos (Smarr and Keefer 2011). Ejemplos relevantes son el Cuestionario de Salud del Paciente 9 (PHQ-9)(Kroenke et al. 2001), la Escala de Depresión del Centro de Estudios Epidemiológicos(Eaton et al. 2004) o la Escala de Evaluación de la Depresión de Hamilton (Hamilton 1980). Entre estos, el BDI-II es uno de los instrumentos más reconocidos y fiables, existiendo amplias evidencias empíricas que respaldan su eficacia (Dozois et al. 1998).

Sin embargo, la autoevaluación y la notificación por parte de la familia a menudo sirven como los principales métodos para detectar casos de enfermedades depresivas (Sanchez-Villegas et al. 2008). El análisis a nivel de población a través de métodos tradicionales a menudo requiere recursos sustanciales. Por ejemplo, las encuestas telefónicas son un enfoque común que puede llevar a retrasos significativos en la obtención de resultados prácticos [3]. Por esta razón, tanto las organizaciones de salud públicas como privadas han puesto estos cuestionarios a disposición de los usuarios para su auto-completado. En ciertos casos, los exámenes en línea basados en estos cuestionarios incluso ofrecen recomendaciones para que las personas busquen ayuda médica profesional según sus puntuaciones. En contraparte, cuando se busca un diagnóstico global preciso,

[3] https://www.cdc.gov/brfss/about/index.html

los procedimientos convencionales tienen ciertas limitaciones. Más allá del estigma social asociado con los problemas de salud mental, que puede influir en la voluntad de los individuos para proporcionar respuestas precisas al cuestionario, los estudios han examinado cómo estas respuestas pueden fluctuar drásticamente en función de factores variables (Cameron et al. 2011). Las puntuaciones finales pueden ser fácilmente manipuladas, ya que pueden ser minimizadas o exageradas. Bowling (2005) estudió las variaciones en la calidad de los resultados en función de la administración de estas pruebas. Las expectativas sociales, como hacer una prueba frente a un médico, cambiarían drásticamente los resultados en comparación con hacerlo en un entorno amigable como tu habitación.

El desarrollo de estos instrumentos proviene de un extenso trabajo previo para entender las causas subyacentes de la depresión. Se han realizado estudios sobre temas relacionados con condiciones depresivas en los campos de la medicina y la psicolingüística (Campbell and Pennebaker 2003; Rude et al. 2004). Todos ellos han intentado identificar la presencia de síntomas, causas y cómo realizar un diagnóstico preciso. Gran parte de esta investigación se ha centrado en comprender la conexión entre el lenguaje y la salud mental. Estos trabajos subrayan el impacto que las palabras pueden tener en nuestro estado emocional y cognitivo. El trabajo pionero de Pennebaker exploró las sutiles matices del uso del lenguaje en la vida diaria, demostrando que ciertos patrones de lenguaje, como el uso frecuente de pronombres en primera persona, pueden servir como indicadores del bienestar mental de un individuo (Pennebaker et al. 2003).

Consecuentemente, las redes sociales ofrecen una oportunidad complementaria para obtener información valiosa sobre los estados mentales de las personas, complementando la terapia profesional tradicional. La combinación de la lingüística computacional con los extensos datos derivados de las redes sociales ha producido un progreso significativo en la detección de indicadores de depresión (Garg 2023; Ríssola et al. 2021). Reconociendo la gran importancia de este dominio, se han dedicado esfuerzos sustanciales para crear puntos de referencia experimentales cuidadosamente organizados (Parapar et al. 2023; Zirikly et al. 2022). Estos recursos han facilitado el desarrollo y evaluación de numerosos modelos predictivos.

Aunque los investigadores en este campo no buscan reemplazar a los profesionales de la salud mental, sí buscan apoyar su trabajo. Los clínicos desempeñan un papel indispensable en validar las predicciones hechas por modelos computacionales y tomar medidas adecuadas con las personas cuando es necesario. Sin embargo, la mayoría de los modelos actuales presentan varias limitaciones para lograr este objetivo (Walsh et al. 2020a). Una barrera significativa es su capacidad limitada para explicar sus predicciones, lo que a menudo resulta en escepticismo entre los profesionales (Hauser et al. 2022). Una forma de abordar esto implica diseñar nuevos modelos que incorporen explicaciones confiables (Ernala et al. 2019). Siguiendo ese camino, la investigación emergente ha explorado el uso de síntomas obtenidos de cuestionarios clínicos validados. La mayoría de estas propuestas, particularmente en el campo de la depresión, aprovechan los marcadores de síntomas del BDI-II (Beck et al. 1996b) o los inventarios PHQ-9 (Kroenke et al. 2001), que cubren una variedad de síntomas depresivos como irritabilidad, pesimismo y trastornos del sueño. Se ha demostrado que la aplicación de tales marcadores de síntomas mejora la explicabilidad, la generalización y el rendimiento general de los modelos de detección de depresión (Nguyen et al. 2022a; Zhang et al. 2022a,b).

B.4 OBJETIVOS Y ALCANCE

Nuestro principal objetivo es explotar el lenguaje utilizado en las redes sociales para construir modelos computacionales que detecten y estimen la gravedad de la depresión. Un requisito previo de los modelos presentados en esta tesis es que sigan esquemas clínicos con el fin de proporcionar resultados interpretables y prácticos. Por esta razón, nuestras soluciones aprovechan el contenido generado por los usuarios para desarrollar modelos que puedan predecir eficazmente la evidencia de síntomas depresivos. Para hacerlo, nos centramos en la aplicabilidad de técnicas dentro de los campos de IR, NLP y ML.

Usar un cuestionario establecido para diagnosticar la depresión, como el BDI-II, es vital para asegurar que nuestros modelos ofrezcan un apoyo diagnóstico en el que confiar. Viendo los recursos limitados disponibles para identificar síntomas depresivos, nuestro segundo objetivo se centra en construir conjuntos de datos centrados en marcadores de síntomas. Para lograr esto, aprovechamos las descripciones del BDI-II para emplear varias técnicas de minería de texto con el fin de filtrar expresiones lingüísticas candidatas que puedan estar asociadas con síntomas depresivos. Además, incluimos análisis síntoma por síntoma de nuestros recursos y realizamos experimentos para validar su utilidad práctica en diferentes tareas de clasificación. Paralelamente a nuestros esfuerzos de creación de recursos, también analizamos la importancia de una metodología de anotación robusta para construir recursos en este dominio complejo.

Siguiendo esta idea, exploramos las capacidades de los recientes modelos de lenguaje conversacionales (LLM) en la creación y ampliación de conjuntos de datos. Reuniendo todos estos esfuerzos, la parte final de la tesis introduce estas contribuciones en forma de una plataforma demostrativa diseñada para profesionales de la salud.

La evaluación juega un papel crucial en ciencias experimentales como IR y NLP. En el dominio de detección de salud mental, una evaluación segura es especialmente crítica, ya que los resultados del modelo pueden influir directamente en las decisiones clínicas y en las evaluaciones generales de salud mental. En esta tesis, nuestros enfoques y recursos se construyen sobre la base establecida por un conocido *workshop* referencia experimental. Específicamente, la Predicción Temprana de Riesgos en Internet (eRisk [4]) (Losada et al. 2017, 2018, 2019, 2020; Parapar et al. 2021b, 2022). Al evaluar nuestros modelos en las colecciones eRisk, no solo aseguramos una base consistente de comparación con otros investigadores líderes, sino que también alineamos nuestro trabajo con un marco establecido conocido por su fiabilidad y relevancia clínica.

B.5 METODOLOGÍA

A lo largo de esta tesis doctoral, se empleó de una metodología rigurosa de manera sistemática. En la primera parte de la tesis, se realiza una investigación completa de la literatura existente para comprender el estado actual de las técnicas de detección de depresión, particularmente aquellas que utilizan datos de plataformas de redes sociales. Un componente clave de la metodología utilizada en esta tesis fue la alineación con estándares establecidos en el dominio de la detección de riesgos en Internet. Se siguen los protocolos de la iniciativa eRisk, que se consideran un punto de referencia reconocido en esta área. Todos los modelos de clasificación se evaluaron siguiendo los esquemas de eRisk, proporcionando un marco estructurado para las pruebas basado en escenarios del mundo real. Como resultado, los resultados de estos trabajos se compararon con varios participantes de eRisk, asegurando una comparación justa.

En cuanto a la construcción de los conjuntos de datos presentados aquí, BDI-Sen y DepreSym, todo el proceso que involucró la minería de datos de las plataformas de redes sociales garantizó medidas de privacidad y una anotación rigurosa de los datos. Dada la naturaleza sensible de este tema, se siguieron estrictas pautas éticas durante la recopilación de datos. Además, los datos obtenidos pasaron por un robusto proceso de ano-

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

tación. Evaluadores expertos revisaron y etiquetaron todo el contenido. En resumen, las consideraciones éticas subrayaron todo el proceso de investigación, reconociendo las responsabilidades que conlleva investigar un área tan sensible y con un impacto tan grande.

B.6 ESTRUCTURA Y CONTENIDOS

Esta tesis doctoral está dividida en siete partes con nueve capítulos. Los capítulos de contribución están diseñados para ser lo más auto-contenido posible. A continuación, presentamos la organización de esta tesis con mayor detalle:

PART I La parte inicial contiene tres capítulos base: primero, la introducción a esta tesis (Capítulo 1), que presenta el trabajo relacionado con los principales temas cubiertos en nuestro estudio. Se encuentra estructurada en tres partes: *i*) el contexto y la motivación, *ii*) el objetivo y alcance, y *iii*) la estructura y principales contribuciones de nuestro trabajo. En el trabajo relacionado, proporcionamos una revisión de los avances más relevantes en el campo de la detección de la depresión en Internet y su evolución hasta el día de hoy. Empezamos dando una primera visión sobre estudios pioneros que estudiaban la conexión entre el lenguage y la manifestación de enfermedades mentales, y como toda esta evidencia ha transicionado hacia el estudio del lenguaje empleado por las personas en las redes sociales. El último capítulo (Capítulo 3) presenta los métodos de investigación y las pautas experimentales seguidas en nuestro trabajo. Inicialmente, contextualizamos el marco experimental eRisk y cómo se alinea con nuestras propuestas, ya que todas nuestras contribuciones están relacionadas con el marco eRisk. A continuación, proporcionamos una visión general de las tareas y colecciones que guían nuestros enfoques, explicando cómo nuestra investigación se alinea con ellas. Concluyendo esta parte, presentamos las principales métricas adoptadas en nuestras contribuciones, elaborando cómo evaluamos la eficacia de nuestros modelos.

PART II Aquí presentamos dos diferentes marcos de clasificación que estiman automáticamente los 21 síntomas del cuestionario BDI-II. Por un lado, el Capítulo 4 utiliza incrustaciones de palabras para explorar la presencia de estos síntomas según su sensibilidad.

Nos referimos a la sensibilidad de los síntomas como la inclinación de los usuarios a discutir abiertamente sobre ellos (es decir, hay síntomas que son más íntimos y los usuarios evitan hablar explícitamente sobre ellos). Por esta razón, analizamos la sensibilidad de cada síntoma y diseñamos dos métodos diferentes para capturar mejor las principales características de cada uno. Por otro lado, el Capítulo 5 utiliza transformadores de frases para seleccionar oraciones de los usuarios de test que producen rankings semánticos basados en su asociación con los síntomas. Posteriormente, utilizamos las frases derivadas de estos rankings como evidencia para predecir la gravedad de los síntomas. Para construir el ranking, indexamos frases de entrenamiento representativas que están asociadas con síntomas depresivos, y exploramos algoritmos de selección de frases para obtenerlas.

PART III En primer lugar, esta sección contiene el trabajo relacionado con la construcción de recursos basados en síntomas para el trastorno depresivo. Presentamos dos recursos principales: *BDI-Sen*, discutido en el Capítulo 6, y *DepreSym*, detallado en el Capítulo 7. Ambos recursos consisten en conjuntos de datos de oraciones anotadas con síntomas de depresión, proporcionando anotaciones manuales relacionadas con los 21 síntomas incluidos en el BDI-II. En el Capítulo 6 (BDI-Sen), comenzamos describiendo la estrategia de recuperación que utilizamos para obtener las oraciones candidatas para anotación. Posteriormente, tres evaluadores decidieron sobre la relevancia real de los candidatos. Profundizando más en esta sección, ofrecemos un análisis detallado de este recurso, estudiando el estilo lingüístico, los atributos emocionales y otros marcadores psicolingüísticos de las oraciones. Además, llevamos a cabo una serie de experimentos investigando la utilidad de BDI-Sen para diversas tareas, incluida la detección y clasificación de la severidad de los síntomas. Finalmente, también examinamos su generalización al considerar síntomas de otras enfermedades mentales.

En el Capítulo 7 (DepreSym), estudiamos formas alternativas de etiquetado de síntomas. Para ello, aprovechamos las tareas de clasificación de eRisk 2023, que se centran en desarrollar métodos de clasificación para encontrar oraciones asociadas con síntomas depresivos. La construcción de DepreSym se basa en los métodos de clasificación de los participantes de la tarea. En este caso, las oraciones etiquetadas provienen de un conjunto de diversos métodos de clasificación, y las oraciones candidatas finales se obtuvieron utilizando agrupación top-k a partir de ellas. Debido a la naturaleza compleja de la anotación de relevancia, diseñamos una metodología de evaluación robusta llevada a cabo por tres evaluadores expertos. Para validar la efectividad de esta metodología, calculamos el acuerdo entre evaluadores y realizamos un análisis adicional del conjunto resultante de juicios. Además, también exploramos la viabilidad de emplear los recientes modelos de lenguaje conversacional LLM (ChatGPT y GPT-4) para ayudar en esta tarea compleja. Llevamos a cabo un examen exhaustivo de su rendimiento, determinamos sus principales limitaciones y analizamos su papel como complemento o sustituto de los anotadores humanos.

Finalmente, en el Capítulo 8, presentamos PsyProf, que es una plataforma demostrativa diseñada para la tarea de evaluar la gravedad de la depresión. La plataforma está pensada para ser utilizada por profesionales de la salud, y se ha desarrollado para demostrar las capacidades efectivas de nuestros modelos de detección de la depresión. Por ello, integramos en esta plataforma los modelos presentados anteriormente, que estiman la presencia de síntomas del BDI-II. Además, hemos complementado nuestra herramienta con métodos de perfilado de usuarios para aportar contexto al medir a usuarios en riesgo. Esto incluiría un modelo que permita predecir el género de los usuarios. Finalmente, también incluimos la funcionalidad de recopilar datos de usuarios de redes sociales, lo que puede ayudar a crear conjuntos de datos basados en síntomas con la inspección proveniente de profesionales de la salud.

PART IV Concluyendo esta tesis, presentamos las conclusiones principales de nuestra investigación y discutimos la dirección potencial para trabajos futuros. Además, consideramos las consideraciones éticas y los desafíos en torno a la detección de indicadores de salud mental en redes sociales.

B.7 RESULTADOS PRINCIPALES Y CONCLUSIONES

Esta tesis doctoral propone varios trabajos que contribuyen en conferencias y revistas de alto prestigio, como ECIR, SIGIR, EMNLP o la revista

Artificial Intelligence in Medicine (AIM). El tema central profundiza en un tema muy importante: explorar la salud mental a través de nuestras actividades en línea. La novedad presentada es alta, aprovechando prácticas reales al centrarse en los síntomas del BDI-II. Todas las contribuciones vienen acompañadas de una evaluación experimental y metodología exhaustiva que muestra los avances sobre el estado del arte. Además, las diferentes contribuciones están vinculadas en un marco que combina modelos predictivos, conjuntos de datos y aplicaciones prácticas.

En esta tesis, se proponen diversas contribuciones para la estimación de la severidad de la depresión a través del uso del lenguaje en las redes sociales. Primero, al introducir dos marcos de clasificación distintos destinados a estimar automáticamente los síntomas del BDI-II (Capítulo 4 y 5). Aprovechando técnicas novedosas relacionadas con la minería de texto y similitudes semánticas, estos modelos han demostrado potencial para detectar con precisión la severidad de la depresión superando al estado del arte en diferentes conjuntos de datos de eRisk. Más allá de estos modelos, se realiza un análisis sobre los patrones lingüísticos y marcadores de síntomas asociados con la depresión, como se discute en el capítulo relacionado con BDI-Sen (Capítulo 6). Además, este trabajo de investigación ofrece integraciones novedosas de LLMs conversacionales para una construcción escalable de conjuntos de datos, un componente crítico en el panorama impulsado por el escenario de big-data en el que nos movemos hoy en día (Capítulo 7). En la sección de conclusiones, ofrecemos un buen resumen que encapsula toda la investigación realizada, destacando la importancia de integrar protocolos clínicamente validados con modelos computacionales. También se proponen vías interesantes para futuras investigaciones, marcadas por la importancia de las consideraciones éticas y la aplicabilidad en el mundo real. Estas direcciones no solo muestran la naturaleza expansiva del trabajo ya completado, sino que también resaltan el gran potencial que se vislumbra en el futuro.

BIBLIOGRAPHY

Pegah Abed-Esfahani, Derek Howard, Marta Maslej, Sejal Patel, Vamika Mann, Sarah Goegan and Leon French. "Transfer Learning for Depression: Early Detection and Severity Prediction from Social Media Postings." In: *CLEF (working notes)* 1 (2019), pp. 1–6.

Mohammed Al-Mosaiwi and Tom Johnstone. "In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation." In: *Clinical Psychological Science* 6.4 (2018), pp. 529–542.

Hayda Almeida, Antoine Briand and Marie-Jean Meurs. "Detecting Early Risk of Depression from Social Media User-generated Content." In: *CLEF (working notes)*. 2017.

Omar Alonso and Stefano Mizzaro. "Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment." In: *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*. Vol. 15. 2009, p. 16.

Hessam Amini and Leila Kosseim. "Towards Explainability in Using Deep Learning for the Detection of Anorexia in Social Media." In: *Natural Language Processing and Information Systems* 12089 (2020), pp. 225 –235.

Nazanin Andalibi, Pinar Ozturk and Andrea Forte. "Sensitive Self-Disclosures, Responses, and Social Support on Instagram: The Case of #Depression." In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17. Portland, Oregon, USA: Association for Computing Machinery, 2017, 1485–1500. DOI: 10.1145/2998181.2998243.

Mario Ezra Aragón, Adrián Pastor López Monroy, Luis Carlos González-Gurrola and Manuel Montes. "Detecting depression in social media using fine-grained emotions." In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 1481–1486.

Celso Arango, Covadonga M Díaz-Caneja, Patrick D McGorry, Judith Rapoport, Iris E Sommer, Jacob A Vorstman, David McDaid, Oscar Marín, Elena Serrano-Drozdowskyj, Robert Freedman et al. "Preventive strategies for mental health." In: *The Lancet Psychiatry* 5.7 (2018), pp. 591–604.

Lisa J Barney, Kathleen M Griffiths, Anthony F Jorm and Helen Christensen. “Stigma about depression and its impact on help-seeking intentions.” In: *Australian & New Zealand Journal of Psychiatry* 40.1 (2006), pp. 51–54.

Angelo Basile, Mara Chinea-Rios, Ana Sabina Uban, Thomas Müller, Luise Rössler, Seren Yenikent, Berta Chulvi, Paolo Rosso and Marc Franco-Salvador. “UPV-Symanto at eRisk 2021: Mental Health Author Profiling for Early Risk Prediction on the Internet.” In: *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*. Ed. by Guglielmo Faggioli, Nicola Ferro, Alexis Joly, Maria Maistro and Florina Piroi. Vol. 2936. CEUR Workshop Proceedings. CEUR-WS.org, 2021, pp. 908–927. URL: http://ceur-ws.org/Vol-2936/paper-75.pdf.

Aaron T Beck, Robert A Steer, Roberta Ball and William F Ranieri. “Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients.” In: *Journal of personality assessment* 67.3 (1996), pp. 588–597.

Aaron T Beck, Robert A Steer and Gregory Brown. “Beck depression inventory–II.” In: *Psychological Assessment* (1996).

Daniel Berrar. “Bayes’ theorem and naive Bayes classifier.” In: *Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics* 403 (2018), p. 412.

Ann Bowling. “Mode of questionnaire administration can have serious effects on data quality.” In: *Journal of public health* 27.3 (2005), pp. 281–291.

Margaret M Bradley and Peter J Lang. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Tech. rep. Technical report C-1, the center for research in psychophysiology . . ., 1999.

Stefan Büttcher, Charles LA Clarke, Peter CK Yeung and Ian Soboroff. “Reliable information retrieval evaluation with incomplete and biased judgements.” In: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 2007, pp. 63–70.

Fidel Cacheda, Diego Fernandez, Francisco J Novoa and Victor Carneiro. “Early detection of depression: social network analysis and random forest techniques.” In: *Journal of medical Internet research* 21.6 (2019), e12554.

Fidel Cacheda, Diego Fernández Iglesias, Francisco Javier Nóvoa and Victor Carneiro. “Analysis and Experiments on Early Detection of Depression.” In: *CLEF (Working Notes)* 2125 (2018), p. 43.

Amy Callahan and Kay Inckle. "Cybertherapy or psychobabble? A mixed methods study of online emotional support." In: *British Journal of Guidance & Counselling* 40.3 (2012), pp. 261–278.

Isobel M Cameron, Amanda Cardy, John R Crawford, Schalk W du Toit, Steven Hay, Kenneth Lawton, Kenneth Mitchell, Sumit Sharma, Shilpa Shivaprasad, Sally Winning et al. "Measuring depression severity in general practice: discriminatory performance of the PHQ-9, HADS-D, and BDI-II." In: *British Journal of General Practice* 61.588 (2011), e419–e426.

R Sherlock Campbell and James W Pennebaker. "The secret life of pronouns: Flexibility in writing style and physical health." In: *Psychological science* 14.1 (2003), pp. 60–65.

Tianfeng Chai and Roland R Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)." In: *Geoscientific Model Development Discussions* 7.1 (2014), pp. 1525–1534.

Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio and Munmun De Choudhury. "A taxonomy of ethical tensions in inferring mental health states from social media." In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 79–88.

Stevie Chancellor and Munmun De Choudhury. "Methods in predictive techniques for mental health status on social media: a critical review." In: *NPJ digital medicine* 3.1 (2020), pp. 1–11.

Stevie Chancellor, Zhiyuan Lin, Erica L Goodman, Stephanie Zerwas and Munmun De Choudhury. "Quantifying and predicting mental illness severity in online pro-eating disorder communities." In: *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 2016, pp. 1171–1184.

Xuetong Chen, Martin D Sykora, Thomas W Jackson and Suzanne Elayan. "What about mood swings: Identifying depression on twitter with temporal measures of emotions." In: *Companion Proceedings of the The Web Conference 2018*. 2018, pp. 1653–1660.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney and Nazli Goharian. "SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions." In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1485–1497.

Glen Coppersmith, Mark Dredze and Craig Harman. "Quantifying Mental Health Signals in Twitter." In: *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to*

Clinical Reality. Baltimore, Maryland, USA: Association for Computational Linguistics, June 2014, pp. 51–60. DOI: 10.3115/v1/W14-3207.

Glen Coppersmith, Mark Dredze, Craig Harman and Kristy Hollingshead. "From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses." In: *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Denver, Colorado: Association for Computational Linguistics, 2015, pp. 1–10. DOI: 10.3115/v1/W15-1201.

Glen Coppersmith, Craig Harman and Mark Dredze. "Measuring post traumatic stress disorder in Twitter." In: *Proceedings of the international AAAI conference on web and social media*. Vol. 8. 1. 2014, pp. 579–582.

Glen Coppersmith, Ryan Leary, Patrick Crutchley and Alex Fine. "Natural language processing of social media as screening for suicide risk." In: *Biomedical informatics insights* 10 (2018), p. 1178222618792860.

Manuel Couto, Anxo Pérez and Javier Parapar. "Temporal word embeddings for early detection of signs of depression." In: *Proceedings of the CIRCLE (Joint Conference of The Information Retrieval Communities in Europe)*. 2022.

Manuel Couto, Anxo Pérez and Javier Parapar. "TemporalWord Embeddings for Early Detection of Signs of Depression." In: *Proceedings of the 2nd Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2022), Samatan, Gers, France, July 4-7, 2022*. Vol. 3178. CEUR Workshop Proceedings. CEUR-WS.org, 2022.

Munmun De Choudhury, Scott Counts and Eric Horvitz. "Social Media as a Measurement Tool of Depression in Populations." In: WebSci '13. Paris, France: Association for Computing Machinery, 2013, 47–56. DOI: 10.1145/2464464.2464480.

Munmun De Choudhury, Scott Counts and Eric Horvitz. "Social media as a measurement tool of depression in populations." In: *Proceedings of the 5th annual ACM web science conference*. 2013, pp. 47–56.

Munmun De Choudhury, Michael Gamon, Scott Counts and Eric Horvitz. "Predicting depression via social media." In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 7. 1. 2013.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith and Mrinal Kumar. "Discovering shifts to suicidal ideation from mental health content in social media." In: *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2016, pp. 2098–2110.

Munmun De Choudhury, Sanket S Sharma, Tomaz Logar, Wouter Eekhout and René Clausen Nielsen. "Gender and cross-cultural differences in social media disclosures of mental illness." In: *Proceedings of the 2017*

ACM conference on computer supported cooperative work and social computing. 2017, pp. 353–369.

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

David JA Dozois, Keith S Dobson and Jamie L Ahnberg. "A psychometric evaluation of the Beck Depression Inventory–II." In: *Psychological assessment* 10.2 (1998), p. 83.

Priscila Silveira Duarte, Maria Cristina Miyazaki, Sergio Luís Blay and Ricardo Sesso. "Cognitive–behavioral group therapy is an effective treatment for major depression in hemodialysis patients." In: *Kidney international* 76.4 (2009), pp. 414–421.

William W Eaton, C Muntaner, C Smith, A Tien and M Ybarra. "Center for epidemiologic studies depression scale: Review and revision." In: *The use of psychological testing for treatment planning and outcomes assessment* (2004).

Sindhu Kiranmai Ernala, Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane and Munmun De Choudhury. "Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, 1–16. DOI: 10.1145/3290605.3300364.

Dorothy L Espelage, Suzanne E Mazzeo, Steven H Aggen, Alexandra L Quittner, Roberta Sherman and Ron Thompson. "Examining the construct validity of the Eating Disorder Inventory." In: *Psychological Assessment* 15.1 (2003), p. 71.

Guglielmo Faggioli, Laura Dietz, Charles Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein et al. "Perspectives on Large Language Models for Relevance Judgment." In: *arXiv preprint arXiv:2304.09161* (2023).

Sam Fletcher, Md Zahidul Islam et al. "Comparing sets of patterns with the Jaccard index." In: *Australasian Journal of Information Systems* 22 (2018).

Forbes. *Introducing ChatGPT*. Ed. by openai.com/blog. [Accessed April 4, 2023]. 2022. URL: https://openai.com/blog/chatgpt.

Muskan Garg. "Mental health analysis in social media posts: a survey." In: *Archives of Computational Methods in Engineering* 30.3 (2023), pp. 1819–1842.

Fabrizio Gilardi, Meysam Alizadeh and Maël Kubli. "Chatgpt outperforms crowd-workers for text-annotation tasks." In: *arXiv preprint arXiv:2303.15056* (2023).

Nazli Goharian, Philip Resnik, Andrew Yates, Molly Ireland, Kate Niederhoffer and Rebecca Resnik, eds. *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. Online: Association for Computational Linguistics, June 2021. URL: https://aclanthology.org/2021.clpsych-1.0.

Andrea K Graham, Carolyn J Greene, Mary J Kwasny, Susan M Kaiser, Paul Lieponis, Thomas Powell and David C Mohr. "Coached mobile app platform for the treatment of depression and anxiety among primary care patients: a randomized clinical trial." In: *JAMA psychiatry* 77.9 (2020), pp. 906–914.

Amelia Gulliver, Kathleen M Griffiths and Helen Christensen. "Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review." en. In: *BMC Psychiatry* 10 (Dec. 2010), p. 113.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar and Johannes C Eichstaedt. "Detecting depression and mental illness on social media: an integrative review." In: *Current Opinion in Behavioral Sciences* 18 (2017). Big data in the behavioural sciences, pp. 43–49. DOI: https://doi.org/10.1016/j.cobeha.2017.07.005.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat and Ming-Wei Chang. "REALM: Retrieval-Augmented Language Model Pre-Training." In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org, 2020.

Aron Halfin. "Depression: the benefits of early and appropriate treatment." In: *American Journal of Managed Care* 13.4 (2007), S92.

Max Hamilton. "Rating depressive patients." In: *The Journal of clinical psychiatry* 41.12 Pt 2 (1980), pp. 21–24.

Keith Harrigian, Carlos Aguirre and Mark Dredze. "Do Models of Mental Health Based on Social Media Data Generalize?" In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3774–3788. DOI: 10.18653/v1/2020.findings-emnlp.337.

Tobias U Hauser, Vasilisa Skvortsova, Munmun De Choudhury and Nikolaos Koutsouleris. "The promise of a model-based psychiatry: building computational models of mental ill health." In: *The Lancet Digital Health* 4.11 (2022), e816–e828.

Rebecca A Hayes, Caleb T Carr and Donghee Yvette Wohn. "One click, many meanings: Interpreting paralinguistic digital affordances in social media." In: *Journal of Broadcasting & Electronic Media* 60.1 (2016), pp. 171–187.

Xingwei He, Zhenghao Lin, Yeyun Gong, A Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen et al. "AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators." In: *arXiv preprint arXiv:2303.16854* (2023).

Steven D Hollon, Michael E Thase and John C Markowitz. "Treatment and prevention of depression." In: *Psychological Science in the public interest* 3.2 (2002), pp. 39–77.

Zunaira Jamil, Diana Inkpen, Prasadith Buddhitha and Kenton White. "Monitoring Tweets for Depression to Detect At-risk Users." In: *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology From Linguistic Signal to Clinical Reality*. Vancouver, BC: Association for Computational Linguistics, Aug. 2017, pp. 32–40. DOI: 10.18653/v1/W17-3104.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari and Erik Cambria. "MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare." In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, June 2022, pp. 7184–7190. URL: https://aclanthology.org/2022.lrec-1.778.

Zheng Ping Jiang, Sarah Ita Levitan, Jonathan Zomick and Julia Hirschberg. "Detection of mental health from reddit via deep contextualized representations." In: *Proceedings of the 11th international workshop on health text mining and information analysis*. 2020, pp. 147–156.

Ankur Joshi, Saket Kale, Satish Chandel and D Kumar Pal. "Likert scale: Explored and explained." In: *Current Journal of Applied Science and Technology* (2015), pp. 396–403.

Payam Karisani and Eugene Agichtein. "Did You Really Just Have a Heart Attack? Towards Robust Detection of Personal Health Mentions in Social Media." In: WWW '18. Lyon, France: International World Wide Web Conferences Steering Committee, 2018, 137–146. DOI: 10.1145/3178876.3186055.

Payam Karisani and Eugene Agichtein. "Did you really just have a heart attack? Towards robust detection of personal health mentions in social media." In: *Proceedings of the 2018 World Wide Web Conference*. 2018, pp. 137–146.

Sylvia Deidre Kauer, Cheryl Mangan and Lena Sanci. "Do online mental health services improve help-seeking for young people? A systematic review." In: *Journal of medical Internet research* 16.3 (2014), e3103.

Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018.

Kurt Kroenke, Robert L Spitzer and Janet BW Williams. "The PHQ-9: validity of a brief depression severity measure." In: *Journal of general internal medicine* 16.9 (2001), pp. 606–613.

Lourdes Lasa, Jose L Ayuso-Mateos, Jose L Vázquez-Barquero, FJ Dıez-Manrique and Christopher F Dowrick. "The use of the Beck Depression Inventory to screen for depression in the general population: a preliminary analysis." In: *Journal of affective disorders* 57.1-3 (2000), pp. 261–265.

Brianna A Lienemann, Jason T Siegel and William D Crano. "Persuading people with depression to seek help: Respect the boomerang." In: *Health Communication* 28.7 (2013), pp. 718–728.

Chung-Ying Lin, Peyman Namdar, Mark D Griffiths and Amir H Pakpour. "Mediated roles of generalized trust and perceived social support in the effects of problematic social media use on mental health: A cross-sectional study." In: *Health Expectations* 24.1 (2021), pp. 165–173.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. DOI: 10.48550/ARXIV.1907.11692.

David E Losada, Fabio Crestani and Javier Parapar. "eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations." In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*. Springer. 2017, pp. 346–360.

David E Losada, Fabio Crestani and Javier Parapar. "Overview of erisk 2018: Early risk prediction on the internet (extended lab overview)." In: *Proceedings of the 9th International Conference of the CLEF Association, CLEF*. 2018, pp. 1–20.

David E Losada, Fabio Crestani and Javier Parapar. "Overview of erisk 2019 early risk prediction on the internet." In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*. Springer. 2019, pp. 340–357.

David E Losada, Fabio Crestani and Javier Parapar. "eRisk 2020: Self-harm and depression challenges." In: *European Conference on Information Retrieval*. Springer. 2020, pp. 557–563.

Kate Loveys, Kate Niederhoffer, Emily Prud'hommeaux, Rebecca Resnik and Philip Resnik, eds. *Proceedings of the Fifth Workshop on Compu-*

tational Linguistics and Clinical Psychology: From Keyboard to Clinic. New Orleans, LA: Association for Computational Linguistics, June 2018. DOI: 10.18653/v1/W18-06.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty and Glen Coppersmith. “Cross-cultural differences in language markers of depression online.” In: *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*. 2018, pp. 78–87.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty and Glen Coppersmith. “Cross-cultural differences in language markers of depression online.” In: *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. New Orleans, LA: Association for Computational Linguistics, June 2018, pp. 78–87. DOI: 10.18653/v1/W18-0608.

Mufan Luo and Jeffrey T Hancock. “Self-disclosure and social media: motivations, mechanisms and psychological well-being.” In: *Current opinion in psychology* 31 (2020), pp. 110–115.

Sean MacAvaney, Bart Desmet, Arman Cohan, Luca Soldaini, Andrew Yates, Ayah Zirikly and Nazli Goharian. “RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses.” In: *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. New Orleans, LA: Association for Computational Linguistics, June 2018, pp. 168–173. DOI: 10.18653/v1/W18-0618.

Alessia Mammone, Marco Turchi and Nello Cristianini. “Support vector machines.” In: *Wiley Interdisciplinary Reviews: Computational Statistics* 1.3 (2009), pp. 283–289.

Rodrigo Martínez-Castaño, Amal Htait, Leif Azzopardi and Yashar Moshfeghi. “Early Risk Detection of Self-Harm and Depression Severity using BERT-based Transformers.” In: *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*. Ed. by Linda Cappellato, Carsten Eickhoff, Nicola Ferro and Aurélie Névéol. Vol. 2696. CEUR Workshop Proceedings. CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2696/paper_50.pdf.

Rodrigo Martínez-Castaño, Juan C Pichel and David E Losada. “A big data platform for real time analysis of signs of depression in social media.” In: *International journal of environmental research and public health* 17.13 (2020), p. 4752.

Diego Maupomé, Maxime D. Armstrong, Raouf Moncef Belbahar, Josselin Alezot, Rhon Balassiano, Marc Queudot, Sébastien Mosser and Marie-Jean Meurs. “Early Mental Health Risk Assessment through

Writing Styles, Topics and Neural Models.” In: *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*. Ed. by Linda Cappellato, Carsten Eickhoff, Nicola Ferro and Aurélie Névéol. Vol. 2696. CEUR Workshop Proceedings. CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2696/paper_53.pdf.

Mary L McHugh. “Interrater reliability: the kappa statistic.” In: *Biochemia medica* 22.3 (2012), pp. 276–282.

Anna Metzler and Herbert Scheithauer. “The long-term benefits of positive self-presentation via profile pictures, number of friends and the initiation of relationships on Facebook for adolescents’ self-esteem and the initiation of offline relationships.” In: *Frontiers in psychology* 8 (2017), p. 1981.

Tomas Mikolov, Kai Chen, G.s Corrado and Jeffrey Dean. “Efficient Estimation of Word Representations in Vector Space.” In: *Proceedings of Workshop at ICLR* 2013 (Jan. 2013).

Shiva Imani Moghadasi, Sri Devi Ravana and Sudharshan N Raman. “Low-cost evaluation techniques for information retrieval systems: A review.” In: *Journal of Informetrics* 7.2 (2013), pp. 301–312.

Saif M Mohammad and Peter D Turney. “Crowdsourcing a word–emotion association lexicon.” In: *Computational intelligence* 29.3 (2013), pp. 436–465.

Robert R Morris, Stephen M Schueller and Rosalind W Picard. “Efficacy of a web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: randomized controlled trial.” In: *Journal of medical Internet research* 17.3 (2015), e72.

Danielle Mowery, Hilary Smith, Tyler Cheney, Greg Stoddard, Glen Coppersmith, Craig Bryan, Mike Conway et al. “Understanding depressive symptoms and psychosocial stressors on Twitter: a corpus-based study.” In: *Journal of medical Internet research* 19.2 (2017), e6895.

Usman Naseem, Adam G Dunn, Jinman Kim and Matloob Khushi. “Early identification of depression severity levels on reddit using ordinal classification.” In: *Proceedings of the ACM Web Conference 2022*. 2022, pp. 2563–2572.

Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh and Michael Berk. “Affective and content analysis of online depression communities.” In: *IEEE transactions on affective computing* 5.3 (2014), pp. 217–226.

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet and Arman Cohan. “Improving the Generalizability of Depression Detection by Leveraging Clinical Questionnaires.” In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022. Ed. by Smaranda Muresan, Preslav Nakov and Aline Villavicencio. Association for Computational Linguistics, 2022, pp. 8446–8459. DOI: 10.18653/v1/2022.acl-long.578.

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet and Arman Cohan. "Improving the Generalizability of Depression Detection by Leveraging Clinical Questionnaires." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*. Ed. by Smaranda Muresan, Preslav Nakov and Aline Villavicencio. Association for Computational Linguistics, 2022, pp. 8446–8459. URL: https://aclanthology.org/2022.acl-long.578.

Kate Niederhoffer, Kristy Hollingshead, Philip Resnik, Rebecca Resnik and Kate Loveys, eds. *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019. URL: https://aclanthology.org/W19-3000.

Cardwell C Nuckols and Cardwell C Nuckols. "The diagnostic and statistical manual of mental disorders,(DSM-5)." In: *Philadelphia: American Psychiatric Association* (2013).

Luıs Oliveira. "BioInfo@ UAVR at eRisk 2020: on the use of psycholinguistics features and machine learning for the classification and quantification of mental diseases." In: *Proceedings of the CEUR Workshop Proceedings, Thessaloniki, Greece*. 2020, pp. 22–25.

OpenAI. "GPT-4 Technical Report." In: *arXiv:submit/4812508* (2023).

Yaakov Ophir, Refael Tikochinski, Christa SC Asterhan, Itay Sisso and Roi Reichart. "Deep neural networks detect from textual facebook posts." In: *Scientific reports* 10.1 (2020), p. 16685.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi and Diana Inkpen. "Deep learning for depression detection of twitter users." In: *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*. 2018, pp. 88–97.

Rosa María Ortega-Mendoza, Delia Irazú Hernández Farías, Manuel Montes-y-Gómez and Luis Villaseñor Pineda. "Revealing traces of depression through personal statements analysis in social media." In: *Artif. Intell. Medicine* 123 (2022), p. 102202. DOI: 10.1016/j.artmed.2021.102202.

Rosa María Ortega-Mendoza, Adrián Pastor López-Monroy, Anilu Franco-Arcega and Manuel Montes-y Gómez. "PEIMEX at eRisk2018: Emphasizing Personal Information for Depression and Anorexia Detection." In: *CLEF (Working Notes)*. 2018.

Rosa María Ortega-Mendoza, Delia Irazú Hernández-Farías, Manuel Montes y Gómez and Luis Villaseñor-Pineda. "Revealing traces of depression through personal statements analysis in social media." In: *Artificial Intelligence in Medicine* 123 (2022), p. 102202. DOI: https://doi.org/10.1016/j.artmed.2021.102202.

Javier Parapar, Patricia Martín-Rodilla, David E Losada and Fabio Crestani. "Overview of eRisk 2021: Early risk prediction on the internet." In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer. 2021, pp. 324–344.

Javier Parapar, Patricia Martín-Rodilla, David E Losada and Fabio Crestani. "erisk 2021: Pathological gambling, self-harm and depression challenges." In: *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*. Springer. 2021, pp. 650–656.

Javier Parapar, Patricia Martín-Rodilla, David E Losada and Fabio Crestani. "Overview of erisk 2022: Early risk prediction on the internet." In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings*. Springer. 2022, pp. 233–256.

Javier Parapar, Patricia Martín-Rodilla, David E Losada and Fabio Crestani. "eRisk 2023: Depression, Pathological Gambling, and Eating Disorder Challenges." In: *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*. Springer. 2023, pp. 585–592.

Sungkyu Park, Inyeop Kim, Sang Won Lee, Jaehyun Yoo, Bumseok Jeong and Meeyoung Cha. "Manifestation of Depression and Loneliness on Social Networks: A Case Study of Young Adults on Facebook." In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '15. Vancouver, BC, Canada: Association for Computing Machinery, 2015, 557–570. DOI: 10.1145/2675133.2675139.

James W Pennebaker, Martha E Francis and Roger J Booth. "Linguistic inquiry and word count: LIWC 2001." In: *Mahway: Lawrence Erlbaum Associates* 71.2001 (2001), p. 2001.

James W Pennebaker, Matthias R Mehl and Kate G Niederhoffer. "Psychological aspects of natural language use: Our words, our selves." In: *Annual review of psychology* 54.1 (2003), pp. 547–577.

Jeffrey Pennington, Richard Socher and Christopher D Manning. "Glove: Global vectors for word representation." In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

Anxo Pérez, Marcos Fernández-Pichel, Javier Parapar and David E Losada. "DepreSym: A Depression Symptom Annotated Corpus and the Role of LLMs as Assessors of Psychological Markers." In: 2023.

Anxo Perez, Javier Parapar, Alvaro Barreiro and Silvia Lopez-Larrosa. "BDI-Sen: A Sentence Dataset for Clinical Symptoms of Depression." In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Taipei, Taiwan, 2023. DOI: 10.1145/3539618.3591905.

Anxo Pérez, Paloma Piot-Pérez-Abadín, Javier Parapar and Álvaro Barreiro. "PsyProf: A Platform for Assisted Screening of Depression in Social Media." In: *European Conference on Information Retrieval*. Springer. Dublin, Ireland, 2023, pp. 300–306.

Anxo Pérez, Neha Warikoo, Kexin Wang, Javier Parapar and Iryna Gurevych. "Semantic Similarity Models for Depression Severity Estimation." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapeur, 2023, "To be published".

A Picardi, I Lega, L Tarsitani, M Caredda, G Matteucci, MP Zerella, R Miglio, A Gigantesco, M Cerbo, A Gaddini et al. "A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care." In: *Journal of affective disorders* 198 (2016), pp. 96–101.

Paloma Piot-Perez-Abadin, Patricia Martín-Rodilla and Javier Parapar. "Experimental Analysis of the Relevance of Features and Effects on Gender Classification Models for Social Media Author Profiling." In: *ENASE*. 2021, pp. 103–113.

Robert Plutchik. "A general psychoevolutionary theory of emotion." In: *Theories of emotion*. Elsevier, 1980, pp. 3–33.

Daniel Preoţiuc-Pietro, Maarten Sap, H Andrew Schwartz and Lyle Ungar. "Mental illness detection at the world well-being project for the clpsych 2015 shared task." In: *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*. 2015, pp. 40–45.

Anxo Pérez, Javier Parapar and Álvaro Barreiro. "Automatic depression score estimation with word embedding models." In: *Artificial Intelligence in Medicine* 132 (2022), p. 102380. DOI: https://doi.org/10.1016/j.artmed.2022.102380.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li and Peter J Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer." In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 5485–5551.

Pranav Rajpurkar, Robin Jia and Percy Liang. “Know What You Don’t Know: Unanswerable Questions for SQuAD.” In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*. Ed. by Iryna Gurevych and Yusuke Miyao. Association for Computational Linguistics, 2018, pp. 784–789. DOI: 10.18653/v1/P18-2124.

Faneva Ramiandrisoa, Josiane Mothe, Farah Benamara and Véronique Moriceau. “IRIT at e-Risk 2018.” In: *9th Conference and Labs of the Evaluation Forum, Living Labs (CLEF 2018)*. 2018, pp. 1–12.

Diana Ramírez-Cifuentes, Ana Freire, Ricardo Baeza-Yates, Joaquim Puntí, Pilar Medina-Bravo, Diego Alejandro Velazquez, Josep Maria Gonfaus, Jordi Gonzàlez et al. “Detection of suicidal ideation on social media: multimodal, relational, and behavioral analysis.” In: *Journal of medical internet research* 22.7 (2020), e17758.

Francisco Rangel and Paolo Rosso. “Overview of the 7th author profiling task at PAN 2019: bots and gender profiling in twitter.” In: *Proceedings of the CEUR Workshop, Lugano, Switzerland*. 2019, pp. 1–36.

Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.” In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019.

Philip Resnik, Anderson Garron and Rebecca Resnik. “Using topic modeling to improve prediction of neuroticism and depression in college students.” In: *Proceedings of the 2013 conference on empirical methods in natural language processing*. 2013, pp. 1348–1353.

Paul Richter, Joachim Werner, Andrés Heerlein, Alfred Kraus and Heinrich Sauer. “On the validity of the Beck Depression Inventory.” In: *Psychopathology* 31.3 (1998), pp. 160–168.

Paul van Rijen, Douglas Teodoro, Nona Naderi, Luc Mottin, Julien Knafou and Patrick Ruch. “Data-driven approach for measuring the severity of the signs of depression using reddit posts: women and men in the orchestra.” In: *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019. 2019.

Esteban A. Ríssola, David E. Losada and Fabio Crestani. “A Survey of Computational Methods for Online Mental State Assessment on Social Media.” In: *ACM Trans. Comput. Healthcare* 2.2 (2021). DOI: 10.1145/3437259.

Esteban A Ríssola, David E Losada and Fabio Crestani. “A survey of computational methods for online mental state assessment on social media.” In: *ACM Transactions on Computing for Healthcare* 2.2 (2021), pp. 1–31.

Esteban Andrés Ríssola, Mohammad Aliannejadi and Fabio Crestani. “Beyond Modelling: Understanding Mental Disorders in Online Social Media.” In: *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I*. Ed. by Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva and Flávio Martins. Vol. 12035. Lecture Notes in Computer Science. Springer, 2020, pp. 296–310. DOI: 10.1007/978-3-030-45439-5_20.

Stephanie Rude, Eva-Maria Gortner and James Pennebaker. “Language use of depressed and depression-vulnerable college students.” In: *Cognition & Emotion* 18.8 (2004), pp. 1121–1133.

Esteban A. Ríssola, Mohammad Aliannejadi and Fabio Crestani. “Mental disorders on online social media through the lens of language and behaviour: Analysis and visualisation.” In: *Information Processing & Management* 59.3 (2022), p. 102890. DOI: 10.1016/j.ipm.2022.102890.

Farig Sadeque, Dongfang Xu and Steven Bethard. “Uarizona at the clef erisk 2017 pilot task: linear and recurrent models for early depression detection.” In: *CEUR workshop proceedings*. Vol. 1866. NIH Public Access. 2017.

Koustuv Saha, Larry Chan, Kaya De Barbaro, Gregory D Abowd and Munmun De Choudhury. “Inferring mood instability on social media by leveraging ecological momentary assessments.” In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.3 (2017), pp. 1–27.

Tetsuya Sakai. “On the robustness of information retrieval metrics to biased relevance assessments.” In: *Journal of Information Processing* 17 (2009), pp. 156–166.

Almudena Sanchez-Villegas, Javier Schlatter, Felipe Ortuno, Francisca Lahortiga, Jorge Pla, Silvia Benito and Miguel A Martinez-Gonzalez. “Validity of a self-reported diagnosis of depression among participants in a cohort study using the Structured Clinical Interview for DSM-IV (SCID-I).” In: *BMC psychiatry* 8.1 (2008), pp. 1–8.

Falk Scholer, Andrew Turpin and Mark Sanderson. “Quantifying test collection quality based on the consistency of relevance judgements.” In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 2011, pp. 1063–1072.

H Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski and Lyle Ungar. “Towards assessing changes in degree of depression through facebook.” In:

Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality. 2014, pp. 118–125.

Mihye Seo, Jinhee Kim and Hyeseung Yang. "Frequent interaction and fast feedback predict perceived social support: Using crawled and self-reported data of Facebook users." In: *Journal of Computer-Mediated Communication* 21.4 (2016), pp. 282–297.

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, Wenwu Zhu et al. "Depression detection via harvesting social media: A multimodal dictionary learning solution." In: *IJCAI*. 2017, pp. 3838–3844.

Shih-Hung and Zhao-Jun Qiu. "A RoBERTa-based model on measuring the severity of the signs of depression." In: *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*. Ed. by Guglielmo Faggioli, Nicola Ferro, Alexis Joly, Maria Maistro and Florina Piroi. Vol. 2936. CEUR Workshop Proceedings. CEUR-WS.org, 2021, pp. 1071–1080. URL: http://ceur-ws.org/Vol-2936/paper-86.pdf.

Karen L Smarr and Autumn L Keefer. "Measures of depression and depressive symptoms: Beck depression Inventory-II (BDI-II), center for epidemiologic studies depression scale (CES-D), geriatric depression scale (GDS), hospital anxiety and depression scale (HADS), and patient health Questionnaire-9 (PHQ-9)." In: *Arthritis care & research* 63.S11 (2011), S454–S466.

Christoforos Spartalis, George Drosatos and Avi Arampatzis. "Transfer Learning for Automated Responses to the BDI Questionnaire." In: *Working Notes of CLEF 2021 – Conference and Labs of the Evaluation Forum*. Ed. by Guglielmo Faggioli, Nicola Ferro, Alexis Joly, Maria Maistro and Florina Piroi. Vol. 2936. Bucharest, Romania, 2021, pp. 1046–1058.

RA Steer, AT Beck and B Garrison. "Applications of the Beck Depression Inventory." In: *Assessment of depression*. Springer, 1986, pp. 123–142.

Anita Thapar, Olga Eyre, Vikram Patel and David Brent. "Depression in young people." In: *The Lancet* 400.10352 (2022), pp. 617–631.

Hugo Touvron et al. "LLaMA: Open and Efficient Foundation Language Models." In: *CoRR* abs/2302.13971 (2023). DOI: 10.48550/arXiv.2302.13971.

Andrew Trask, Phil Michalak and John Liu. "sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings." In: *arXiv preprint arXiv:1511.06388* (2015).

Alina Trifan and José Luís Oliveira. “Bioinfo@ UAVR at erisk 2019: delving into social media texts for the early detection of mental and food disorders.” In: *CLEF (working notes)*. 2019.

Alina Trifan, Pedro Salgado and José Luís Oliveira. “BioInfo@UAVR at eRisk 2020: on the Use of Psycholinguistics Features and Machine Learning for the Classification and Quantification of Mental Diseases.” In: *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*. Ed. by Linda Cappellato, Carsten Eickhoff, Nicola Ferro and Aurélie Névéol. Vol. 2696. CEUR Workshop Proceedings. CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2696/paper_43.pdf.

Marcel Trotzek, Sven Koitka and Christoph Friedrich. “Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences.” In: *IEEE Transactions on Knowledge and Data Engineering* 32 (Apr. 2018), pp. 588–601. DOI: 10.1109/TKDE.2018.2885515.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh and Hiroyuki Ohsaki. “Recognizing depression from twitter activity.” In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 2015, pp. 3187–3196.

Ana Sabina Uban and Paolo Rosso. “Deep Learning Architectures and Strategies for Early Detection of Self-harm and Depression Level Prediction.” In: *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*. Ed. by Linda Cappellato, Carsten Eickhoff, Nicola Ferro and Aurélie Névéol. Vol. 2696. CEUR Workshop Proceedings. CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2696/paper_52.pdf.

Ana-Sabina Uban and Paolo Rosso. “Deep learning architectures and strategies for early detection of self-harm and depression level prediction.” In: *CEUR Workshop Proceedings*. Vol. 2696. Sun SITE Central Europe. 2020, pp. 1–12.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser and Illia Polosukhin. “Attention is all you need.” In: *Advances in neural information processing systems* 30 (2017).

Anthony J Viera, Joanne M Garrett et al. “Understanding interobserver agreement: the kappa statistic.” In: *Fam med* 37.5 (2005), pp. 360–363.

Maria Paula Villegas, Darío Gustavo Funez, Maria José Garciarena Ucelay, Leticia Cecilia Cagnina and Marcelo Luis Errecalde. “LIDIC-UNSL’s Participation at eRisk 2017: Pilot Task on Early Detection of Depression.” In: *CLEF (Working Notes)*. 2017.

Colin G Walsh, Beenish Chaudhry, Prerna Dua, Kenneth W Goodman, Bonnie Kaplan, Ramakanth Kavuluru, Anthony Solomonides and Vignesh Subbian. “Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence.” In: *JAMIA open* 3.1 (2020), pp. 9–15.

Colin G Walsh, Beenish Chaudhry, Prerna Dua, Kenneth W Goodman, Bonnie Kaplan, Ramakanth Kavuluru, Anthony Solomonides and Vignesh Subbian. “Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence.” In: *JAMIA Open* 3.1 (Jan. 2020), pp. 9–15. DOI: 10.1093/jamiaopen/ooz054.

JT Wolohan, Misato Hiraga, Atreyee Mukherjee, Zeeshan Ali Sayyed and Matthew Millard. “Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP.” In: *Proceedings of the first international workshop on language cognition and computational models*. 2018, pp. 11–21.

Chia-chen Yang, Sean M Holden and Mollie DK Carter. “Emerging adults’ social media self-presentation and identity development at college transition: Mindfulness as a moderator.” In: *Journal of Applied Developmental Psychology* 52 (2017), pp. 212–221.

Andrew Yates, Arman Cohan and Nazli Goharian. “Depression and Self-Harm Risk Assessment in Online Forums.” In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2968–2978. DOI: 10.18653/v1/D17-1322.

Emine Yilmaz, Javed A. Aslam and Stephen Robertson. “A new rank correlation coefficient for information retrieval.” In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*. Ed. by Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua and Mun-Kew Leong. ACM, 2008, pp. 587–594. DOI: 10.1145/1390334.1390435.

Zhiling Zhang, Siyuan Chen, Mengyue Wu and Kenny Q. Zhu. “Psychiatric Scale Guided Risky Post Screening for Early Detection of Depression.” In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*. Ed. by Luc De Raedt. ijcai.org, 2022, pp. 5220–5226. DOI: 10.24963/ijcai.2022/725.

Zhiling Zhang, Siyuan Chen, Mengyue Wu and Kenny Q. Zhu. “Symptom Identification for Interpretable Detection of Multiple Mental Disorders on Social Media.” In: *Proceedings of the 2022 Conference on Empirical*

Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022. Ed. by Yoav Goldberg, Zornitsa Kozareva and Yue Zhang. Association for Computational Linguistics, 2022, pp. 9970–9985. URL: https://aclanthology.org/2022.emnlp-main.677.

Zhuosheng Zhang, Junjie Yang and Hai Zhao. "Retrospective Reader for Machine Reading Comprehension." In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021, pp. 14506–14514. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17705.

Ayah Zirikly et al., eds. *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*. Seattle, USA: Association for Computational Linguistics, July 2022. URL: https://aclanthology.org/2022.clpsych-1.0.