

Explainable Depression Detection Using Social Media Data

by

Yuxi Wang

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the degree of
Master of Computer Science

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

© Yuxi Wang, Ottawa, Canada, 2024

Examining Committee

The following served on the Examining Committee for this thesis.

Carleton Member: Olga Baysal

Associate Professor, School of Computer Science
Carleton University

Internal Member: Marina Sokolova

Adjunct Professor, School of Electrical Engineering and Computer Science
University of Ottawa

Supervisor: Diana Inkpen

Professor, School of Electrical Engineering and Computer Science
University of Ottawa

Declaration of Authorship

I hereby certify that this thesis is entirely my own original work except where otherwise indicated. I am aware of the University of Ottawa regulations concerning plagiarism, including those regarding consequent disciplinary actions. Any use of the works of any other author, in any form, is properly acknowledged at their point of use.

Abstract

With the advances in Machine Learning (ML) techniques and the wide availability of social media data, early intervention for mental health issues becomes more and more practicable. An important problem for Natural Language Processing (NLP) practitioners is the automatic detection of mental disorders such as depression, on a large scale. Social media analysis is commonly used to tackle this problem. Due to the rapid growth of user interaction on different social media platforms, publicly available social media data has increased substantially. The sheer amount of data and level of personal information being shared on such platforms has made analyzing textual information to predict mental disorders such as depression a reliable preliminary step when it comes to psychometrics. However, it remains a challenge for computing systems to deal with the huge amount of textual information, and to understand the relationships between the content of the texts and the writers' actual mental health conditions.

In this study, we first proposed a system to search for texts that are related to depression symptoms from the Beck's Depression Inventory (BDI) questionnaire, to extract relevant textual data from huge collections and provide a ranking for further investigation. For each of the 21 symptoms on the BDI questionnaire, 21 queries (for the 21 symptoms) were constructed based on the corresponding questions and possible answers on the questionnaire. Several methods focusing on extracting relevant sentences from Reddit social media posts and comments were introduced. To rank the texts (sentences) based on their relevance, neural embedding vectors were computed as representations; then their cosine similarity to each symptom-query embedding was calculated. With a dataset having only texts and no labels that could be used for training, our system obtained competitive results on 4 metrics with efficient computing, particularly for the metric "precision at 10", which measures how many relevant results there are in the top 10 retrieved sentences. These advantages provide an opportunity for the system to be adopted for the next task.

Then, we address the even more challenging task of automatic depression level detection, using the writings and the voluntary answers provided by users on Reddit. Several explainable machine learning algorithms and several Large Language Models (LLMs) were

applied in our experiments to provide both predictions and explanations for each question. One of our proposed systems is based on glass box models which are interpretable, and another is based on LLMs that could generate explanations for their predictions even if they are considered black boxes. By combining two LLMs for different questions, we achieved better performance on three of four metrics compared to the state-of-the-art and remained competitive on the one remaining metric. In addition, our system is explainable on two levels: first, predicting the answers to the BDI questions provides clues about the possible symptoms that could lead to a clinical diagnosis of depression; second, our system can explain the predicted answer for each question.

Acknowledgements

This thesis is a valuable asset in my life. Since I set foot on Canadian soil, I have to admit that I have suffered a lot and endured many hardships. However, I was also fortunate during this period since I received much help from some noble people, which balanced out a lot of my unluckiness. I want to first thank my supervisor, Professor Diana Inkpen, and all members of the NLP Laboratory under her leadership. I wasn't the smartest student, nor was I the easiest to communicate with. In addition to the complicated research and teaching work, Professor Inkpen gave me enough patience, guidance, and encouragement. It is a great honor to be selected as a graduate student by Professor Inkpen and be able to contribute my modest efforts to the NLP Lab.

I would like to express my sincere gratitude to the members of my examining committee: Professor Olga Baysal and Professor Marina Sokolova. I am grateful for their willingness to take the time to understand my research, critique my thesis, and provide constructive comments and feedback.

I want to thank the Natural Sciences and Engineering Research Council of Canada (NSERC) for supporting my research. I also want to thank the CLEF eRisk organizers, for providing great opportunities for me to explore the task of early risk detection.

Throughout the many difficult days, my thoughts have always been about the people I love. Social connections make me factually exist, and my connections with the people I love and love me make me understand the purpose of my existence. As an ordinary person, my joys and sorrows are really not worth exaggerating in the river of history, but even as an ordinary person, I received much selfless love, which deeply moved me. I don't want to lock myself in my own sentimentality, but I want to always keep my gratitude to them in my heart. Some people may think that we need to have extensive love, and to have great care that even abandons the ego, but for me, I have no shame in my feelings, and I want to do my best to repay the people who care about me. I write these words here to express my deepest emotions.

Table of Contents

List of Tables	xii
List of Figures	xiv
Abbreviations	xvi
1 Introduction	1
1.1 Motivation	1
1.2 Objectives	2
1.3 Contributions	3
1.4 Thesis Organization	4
1.5 Published Papers	5
2 Background	6
2.1 Development of NLP	6
2.1.1 Rule-Based NLP	8
2.1.2 Statistical NLP	8
2.1.3 Traditional Machine Learning	8

2.1.4	Traditional Deep Learning	9
2.1.5	Transfer Learning	9
2.1.6	Prompt Learning	9
2.2	A General Workflow of NLP	10
2.2.1	Data Collection	10
2.2.2	Text Pre-Processing	10
2.2.3	Feature Extraction	10
2.2.4	Modeling	11
2.2.5	Evaluation and Deployment	11
2.2.6	Maintenance	11
2.3	Explainability of NLP	12
2.3.1	Main Categories of Explanations	13
2.3.2	Two Aspects of Explanations	15
2.3.3	Analyzing Explainable Systems	15
2.4	Summary	16
3	Related Work	17
3.1	Lexical Usage and Emotions	17
3.2	Exploring Features Related to Depression	18
3.3	Deep Learning for Depression Detection	18
3.4	Summary	20
4	Datasets	21
4.1	eRisk Shared Tasks	21

4.2	Dataset for Depression Symptoms Search	24
4.3	Dataset for Depression Estimation	25
4.4	Summary	28
5	Search for Depression-Related Writings	29
5.1	Introduction	29
5.2	Methodology	30
5.2.1	Information Retrieval and Search	30
5.2.2	Document Acquisition	31
5.2.3	Data Normalization and Text Processing	31
5.2.4	Searching with Contextual Representations	32
5.2.4.1	DistilBERT with Cosine Similarity	34
5.2.4.2	RoBERTa with Cosine Similarity	34
5.2.4.3	Universal Sentence Encoder with Cosine Similarity	34
5.2.5	Searching with Distributional Word Representations	35
5.3	Evaluation	35
5.4	Results and Discussion	37
5.4.1	Comparison to Related Works	39
5.5	Using USESim for Writing Selection	40
5.6	Summary	41
6	Estimating the Level of Depression	44
6.1	Introduction	44
6.2	Methodology	45

6.2.1	Data Preparation	47
6.2.1.1	Stimuli for Feature Extraction	51
6.2.2	Classification Using Glass Box Models	54
6.2.2.1	Text Pre-Processing	54
6.2.2.2	Text Representation	54
6.2.2.3	Correlation Analysis	55
6.2.2.4	Balancing the Training Data	58
6.2.2.5	Glass Box Classifiers	58
6.2.2.6	Order of Questions	59
6.2.3	Classification Using Large Language Models	61
6.2.3.1	Selection of the LLMs	61
6.2.3.2	Design of Prompts	62
6.2.3.3	Using LLMs	65
6.3	Evaluation	66
6.4	Results and Discussion	67
6.4.1	Glass Box-Based Systems	67
6.4.1.1	Model Explanation	68
6.4.2	LLM-Based Systems	71
6.4.2.1	Local Explanations of LLMs	73
6.4.3	Performance on Individual Questions	75
6.4.4	Evaluation of Explanations	76
6.4.5	Comparison to Related Work	77
6.5	Summary	78

7 Conclusion and Future Work	80
7.1 Conclusion	80
7.2 Research Questions Revisited	81
7.3 Limitations	82
7.4 Future Work	83
7.5 Ethics Statement	85
References	86
References	86
A BDI Questionnaire	96
B Responsible NLP Checklist	102

List of Tables

2.1	Overview of main categories of explanations	14
4.1	Statistics of the dataset for depression symptoms search	25
4.2	Statistics of labels in the training data for depression estimating	26
4.3	Statistics of depression categories in the training data	28
5.1	Queries and keywords for each question	32
5.2	Results for submitted 5 runs (majority voting)	38
5.3	Results for submitted 5 runs (unanimity)	38
5.4	Our results compared to the best results in the shared task (majority voting)	39
5.5	Our results compared to the best results in the shared task (unanimity) . .	39
6.1	Statistics of text length of the cleaned data (Top-5)	47
6.2	Statistics of text length of the cleaned data (Top-1)	49
6.3	Sets of stimuli for feature extraction	51
6.4	Questions that have strong correlations	57
6.5	Order and configurations of glass box models for each question	59
6.6	Rephrased symptoms in the BDI questionnaire	64
6.7	Results of glass box-based systems	68

6.8	Results of LLM-based systems	72
6.9	Question-wise performance of neural-chat+SUS-Chat_top5	76
6.10	Our results compared to the state-of-the-art	78

List of Figures

2.1	Relationship between AI, NLP, ML and DL	7
2.2	A general workflow of NLP development	12
2.3	A black box machine learning model	13
4.1	Structure of resulting file for depression searching	24
4.2	Format of documents to be searched	25
4.3	Depression categories associated with depression levels	27
5.1	Guidelines for labelling sentences related to depression symptoms (Parapar, Martín-Rodilla, Losada, & Crestani, 2023)	42
5.1	Guidelines for labelling sentences related to depression symptoms (Parapar et al., 2023) (cont.)	43
6.1	Flowchart of development pipeline of two model types	46
6.2	Correlation matrix of answers	56
6.3	Some options that might be confusing to models	63
6.4	Designed prompt template for symptom estimation	65
6.5	An example of global variable importance of EBM	69
6.6	An example of local prediction explanation of EBM	70

6.7	An example of an explanation of logistic regression	70
6.8	An example of using a decision tree for classification	71
6.9	An example of concatenated writings. Some words were highlighted for readability. In fact, we rephrased the sentences in the example, because the data sharing agreement that we signed when we obtained the dataset does not allow us to use the real example directly.	74
6.10	Classification and explanation by Llama-2-13b-chat	74
6.11	Classification and explanation by neural-chat-7b-v3-1	75
6.12	Classification and explanation by SUS-Chat-34B	75

Abbreviations

ACL Association for Computational Linguistics 85

ACR Average Closeness Rate 66, 75, 77

ADODL Average Difference between Overall Depression Levels 66, 68, 72, 75, 77, 78

AGI Artificial General Intelligence 6

AHR Average Hit Rate 66, 68, 72, 75

AI Artificial Intelligence 4, 6, 7, 12, 13, 15, 45

BDI Beck’s Depression Inventory 3–5, 19, 22–24, 28–31, 44, 45, 47, 55, 59, 63, 66, 80, 81, 84, 85

CLEF Conference and Labs of the Evaluation Forum 18–20, 22, 28

CNN Convolutional Neural Network 9, 19

DAN Deep Averaging Network 19

DCHR Depression Category Hit Rate 67, 68, 72, 75, 77, 78

DL Deep Learning 7, 9, 11, 29

DPO Direct Preference Optimization 62, 72

DS Data Science 10

DT Decision Tree 58–60, 67, 71, 81

EBM Explainable Boosting Machine 58–60, 67–69, 81

ERDE Early Risk Detection Error 22

ETM Embedded Topic Model 19

GAM Generalized Additive Model 58

HAN Hierarchical Attention Network 19, 77

HMM Hidden Markov Model 8

IAT Implicit Association Test 51

IR Information Retrieval 23, 30, 81

LIWC Linguistic Inquiry and Word Count 17

LLM Large Language Model 9, 14–16, 29, 45, 47, 61, 62, 65, 71, 78, 80–82, 84

LR Logistic Regression 58, 60, 67, 70, 81

LSTM Long Short-Term Memory 77

MC Markov Chain 8

ML Machine Learning 7, 8, 10, 11, 15, 23, 29, 81

NLG Natural Language Generation 61

NLP Natural Language Processing 2, 4, 6–13, 17, 20, 23, 34, 44, 83

PTB Penn Treebank Tokenizer 34

RF Random Forest 8, 13

RLHF Reinforcement Learning from Human Feedback 62

RNN Recurrent Neural Network 9, 19

SVM Support Vector Machine 8, 13

TF-IDF Term Frequency–Inverse Document Frequency 18

TM Text Mining 22, 23

USE Universal Sentence Encoder 19, 37

UX User Experience 15

WHO World Health Organization 1

XAI Explainable Artificial Intelligence 6, 13, 16, 44

Chapter 1

Introduction

In this chapter, the reason and the importance of developing understandable depression detection systems are discussed, and a set of research questions that guided this study are presented. The contributions are presented, followed by the structure of the thesis as well as the list of published papers.

1.1 Motivation

Since the invention of computing machines, the purpose of computers, or their main functionality - computing, was to help human beings solve problems that were unable or less efficient to be solved by humans. Many general and specific applications have been developed through the years, and computer scientists are striving for this purpose.

One of the demanding and realistic problems we are facing is mental health issues. According to World Health Organization (WHO), more than 700,000 persons die by suicide every year globally¹, even though suicide should not be directly considered a mental health issue. However, due to various social and personal factors, psychological problems are rarely exposed or discussed openly, so their seriousness is less noticed and fewer resources

¹<https://www.who.int/health-topics/suicide>

are used to deal with mental health issues. In addition, mental disorder is not well fitted in contemporary psychiatry, because it is often misunderstood, and could be variously interpreted (Radden, 2023). Fortunately, since online social media provide the public with an open and indirect (compared to real life) platform for interaction, they have become an essential part of everyone’s daily life, people use it as a platform to express their feelings about almost everything, thus more information about what individuals are thinking is disclosed. Information on social media contains a variety of valuable content for dealing with mental disorders.

Being one of the leading global public health issues, depression is common, costly, debilitating, and associated with an increased risk of suicide (Marwaha et al., 2023). Since depression has become a prevalent mental health issue, early detection of symptoms could greatly improve the chances of proper treatment. Traditional methods of detection, usually human-led, are expensive to conduct and might be individually biased. There is a high demand for automatic systems to provide efficient depression detection on larger scales, with reliable explanations.

1.2 Objectives

The application of Natural Language Processing (NLP) systems on mental health issues is an active research area, but there is less focus on the explainability of such domain-specific systems. In this research, our main objective is to develop explainable NLP systems that could automatically detect social media users’ levels of depression. Around this goal, the following research questions are aimed to be answered:

- **Research Question 1:** How could smaller and cleaner datasets of user writings be formed from huge social media text collections?
- **Research Question 2:** How to apply domain-knowledge-based features and interpretable machine learning algorithms for automatic depression detection?

- **Research Question 3:** How could large language models be used for classifying depression-related symptoms, even without training?
- **Research Question 4:** How could large language models provide local explanations for their predictions?
- **Research Question 5:** What is the quality of the generated local explanations?

1.3 Contributions

In this study, we propose a method to analyze and select social media writings to help identify potential symptoms of depression. Then, we introduce explainable methods to automatically fill in the Beck’s Depression Inventory (BDI) questionnaire ([Beck, Ward, Mendelson, Mock, & Erbaugh, 1961](#)) for the social media user (see Appendix A for the full questionnaire), using the information contained in the selected user writings. The questionnaire then provides the level of depression of the user based on all the answers and could support early intervention of depression.

This task is considered challenging for several reasons. The user writings on social media could include a large amount of unrelated information, but lack details for specific symptoms; for symptoms that relevant information is available, timeliness issues exist; and even if there is sufficient and valid information to solve individual problems, the task would need the generalization from 21 individual questions on the questionnaire to an overall depression level; and the final estimation, based on the summation of answers to the 21 questions, is into 4 levels of severity of depression (unlike most related works that classify a user as having depression or not).

The key contributions of this thesis are as follows:

- Designed explainable depression detection systems with diverse explainability techniques, spreading from traditional machine learning models to deep learning architectures.

- Extended the applicability of using large language models by proposing a prompt-based method to predict the level of depression for social media users.
- Improved the performance on the task of automatically filling in the BDI questionnaire using social media data through manually designed prompts and without further training the LLMs.
- Explored the practicality of LLMs for generating both the predictions and explanations for the predictions.

1.4 Thesis Organization

This thesis is organized as follows:

- **Chapter 1: Introduction** This chapter introduces the purpose of conducting research on depression detection, the objectives the study aims to achieve, and the contributions the research thesis has made.
- **Chapter 2: Background** The chapter presents the background of Artificial Intelligence (AI) and NLP, the development and application of NLP, and the necessity and status of explainability of NLP.
- **Chapter 3: Related Work** This chapter discusses the previous research around the thesis topic - detecting depression using NLP. Starting from using domain-specific features, more and more complex methods such as deep learning were applied in the field, and the usage of features became less and less explainable.
- **Chapter 4: Datasets** This chapter describes the two datasets that were used in the study, especially how they were shared between the two tasks, how they are formatted, and the findings of statistical explorations.
- **Chapter 5: Search for Depression-Related Writings** This chapter presents the systems that we developed and the experiments we ran for the task of selecting

depression-relevant user writings from a large collection of writings on social media Reddit.

- **Chapter 6: Estimating the Level of Depression** This chapter presents the task of evaluating Reddit users' depression levels using their writings and the BDI questionnaire, and how the proposed explainable systems work.
- **Chapter 7: Conclusion and Future Work** In this chapter, the research findings, limitations as well as potential enhancements that could be achieved in future work are discussed. Ethical considerations are also stated.

1.5 Published Papers

1. Wang, Y., & Inkpen, D. (2023). uOttawa at eRisk 2023: Search for Symptoms of Depression. In Proceedings of ERisk 2023@CLEF 2023, CEUR-WS, pp. 805-812.
2. Wang, Y., Inkpen, D., & Gamaarachchige, P. K. (2024). Explainable Depression Detection Using Large Language Models on Social Media Data. In Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024) (pp. 108-126).

Chapter 2

Background

In this chapter, the development of NLP as a booming research field of AI is presented, main techniques and paradigms in each of its evolution phases are identified, and a general workflow of developing NLP systems is discussed. Next, the growing yet important field of Explainable Artificial Intelligence (XAI) is described, with emphasis on types of explanation techniques.

2.1 Development of NLP

Many people are hoping that through developing AI techniques, multiple facets of human lives would be enhanced; for example, a more efficient food supply could be implemented, or climate changes could be better investigated. However, AI is a broad field with a great number of aspects to work on; thus researchers usually work on some of its sub-fields. Even though there are many concerns and doubts about this approach ([Russell & Norvig, 2020](#)), people with various interests are utilizing their wisdom to help multiple sub-fields flourish, and the current high level of AI development proves that this works better than smaller interested groups that work only in the field of Artificial General Intelligence (AGI).

NLP could be considered as one of the sub-fields of AI, with some non-AI techniques that are being used. Figure 2.1 depicts the relationship between AI and NLP, as well

as other sub-fields of AI such as Machine Learning (ML) and Deep Learning (DL). We could define NLP as designing and deploying computational systems that could acquire information conveyed in human language, and autonomously perform reasonable actions based on pre-designed requirements.

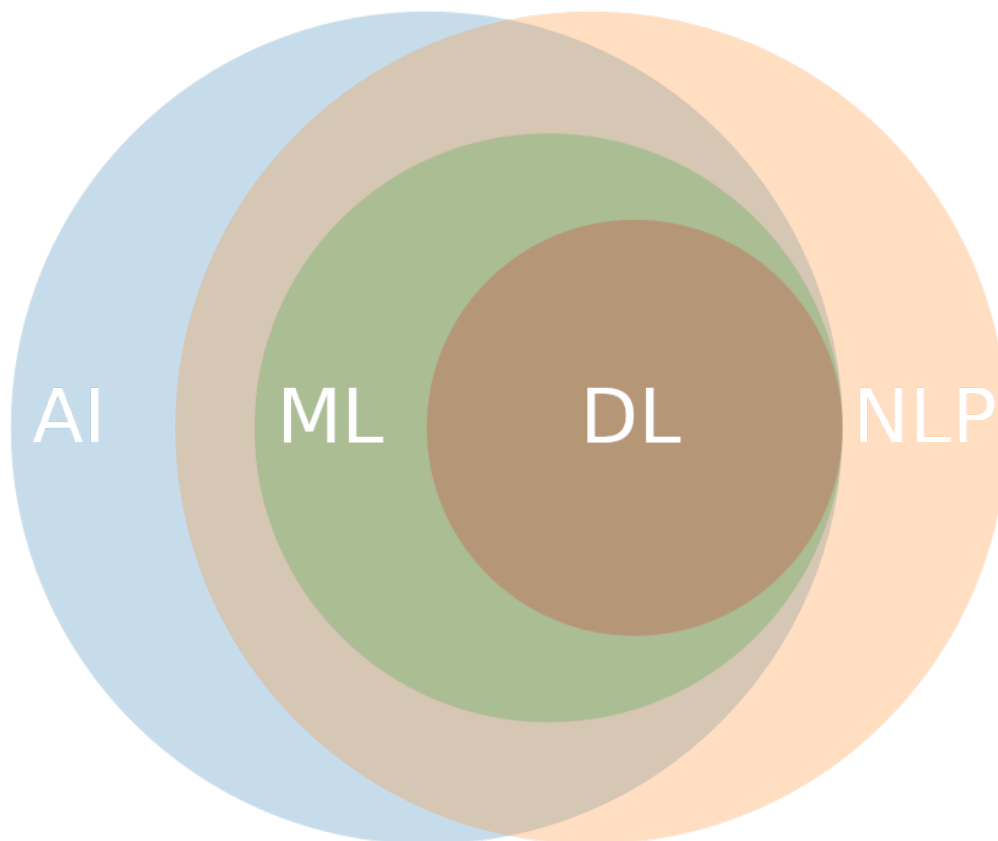


Figure 2.1: Relationship between AI, NLP, ML and DL

NLP systems have become much more powerful in recent decades. The field of NLP has gone through many phases and evolution; some significant phases are described below.

2.1.1 Rule-Based NLP

In this stage of NLP methods, the language structures were considered stable and strictly followed; each part of a sentence (such as subject, verb, and object) was specified and a number of logic rules were designed by researchers for processing systems to make inferences. However, as the complexity of language usage is always surging, this method is considered inflexible and not robust. For these reasons, this type of method is not commonly used as the main approach, but is still being used for some sub-tasks such as feature engineering (Yao, Mao, & Luo, 2019).

2.1.2 Statistical NLP

This is a critical forward step for NLP since it adopted statistical methods to tackle the ever-existing issue of randomness. The key in this stage is the use of corpora. With examples from corpora or a specific corpus, knowledge such as collocations could be extracted and could be further used. For example, n-grams (or most commonly bigrams) could be found and the probability of an occurrence of a word could be calculated. Another example is the Hidden Markov Model (HMM), which is a temporal probabilistic model using hidden states and it played a huge role in speech technology (Russell & Norvig, 2020). Even though this method still struggles with generalization to new or unincluded data, this phase moved way beyond rule-based methods and worked as a foundation of later methods.

2.1.3 Traditional Machine Learning

From a general perspective, ML refers to techniques and procedures of designing and deploying computer systems that could "learn" by observing specific data, then take desired actions on unseen data with presumably the same distribution. In this phase of NLP methods, ML means applying traditional algorithms such as Support Vector Machine (SVM) and ensemble methods such as Random Forest (RF). Many algorithms, for example, Markov Chain (MC) (Chung, 1960), are still being widely used for their relatively simpler structure and faster processing speed compared with more advanced and complex algorithms.

2.1.4 Traditional Deep Learning

This phase is around the resurgence of artificial neural networks. The term "deep learning" refers to learning deep (multiple layers of) representations of information. With the advent of models such as Convolutional Neural Network (CNN) (Kim, 2014) and Recurrent Neural Network (RNN) (Ma et al., 2016), ground-breaking performance was achieved in many tasks, such as text classification. The models could learn features and representations from texts, and traditional feature engineering was greatly reduced (if not completely replaced). However, with models getting much more complicated, it is almost impossible for them to be fully understood by researchers and practitioners.

2.1.5 Transfer Learning

Transfer learning techniques are built upon the development of large DL architectures. Researchers design language models, train them with different training strategies and a very large amount of data, then share the resulting models as "pre-trained models". These generally pre-trained language models, such as BERT (Devlin, Chang, Lee, & Toutanova, 2019) and RoBERTa (Y. Liu et al., 2019) could be adapted into specific tasks with or without further training.

2.1.6 Prompt Learning

Prompt learning (P. Liu et al., 2021) is an emerging and promising paradigm that focuses on guiding the pre-trained Large Language Model (LLM) to generate desired responses, through "prompts". The LLMs would be trained with gigantic and all-encompassing data, and learn to form sentences and paragraphs with "reasonable" knowledge. Since the LLM would gain knowledge from training, the prompt learning techniques try to leverage its ability to generate responses based on carefully designed prompts. Since no or only a few data is needed for specific tasks, this paradigm leads the field of NLP to a new era, such as training versatile language models with comprehensive data rather than model adaption with domain-specific datasets.

2.2 A General Workflow of NLP

NLP applications are amazing and are becoming ubiquitous. To zoom in, NLP systems process language information following common procedures.

2.2.1 Data Collection

Every modern NLP or ML system works with data. Due to the central role that data has been playing, many people consider ML as a sub-field of Data Science (DS). A large NLP system could scrape data from various sources (such as the Web), or receive datasets being publicly shared or from upstream components in a larger system.

2.2.2 Text Pre-Processing

Datasets could contain texts, as well as other information such as metadata or non-textual information. In this step, a NLP system would extract raw texts from the dataset.

Further processing could be applied to the extracted raw texts, if needed. Traditional techniques include stemming, normalization, removing URLs, punctuation, special symbols, special characters, and stopwords. Text segmentation or tokenization is commonly applied as well, especially when using more advanced architectures.

2.2.3 Feature Extraction

Most machine learning algorithms work on numbers, not texts. The goal of feature extraction is to represent data using numbers, especially vectors. In NLP, this step is also called text representation, since it deals with transforming textual data into vectors. Features reside in the data; when properly represented in vectors, the ML algorithm would be able to solve problems using knowledge learned with the help of these features.

In traditional machine learning, a process named "feature engineering" is necessary to improve the quality of the features representing the text. When the age of deep learning

comes, features are learned by DL models with less effort needed from researchers or engineers.

2.2.4 Modeling

When the raw data are represented by numbers (vectors in most cases), multiple machine learning algorithms could be experimented with. A baseline approach could be set first, and researchers or engineers would improve it or propose other methods to compare with. There is no single ML algorithm that would work well on all cases; many experiments and trials are expected in this step.

2.2.5 Evaluation and Deployment

If a well-performing model is built, an evaluation process is necessary to evaluate how good the model actually is. This process would check whether the model generalizes well on unseen data. Another perspective to evaluate the model is inspecting the model's performance according to business objectives. Usually, these evaluations require experts from business teams.

After the model satisfies the requirements, the model could be deployed in the production environment to provide services to end-users.

2.2.6 Maintenance

Similar to any other software system, continuous monitoring is needed to make sure the deployed NLP system is working as expected. Issues need to be investigated and fixed, and hardware usages need to be regularized. It is also common to upgrade the deployed system to let models able to adapt in the ever-changing world.

Figure 2.2 displays the discussed procedures of developing NLP applications. Note that the steps are not necessarily linear and seamless as shown, multiple attempts, errors, or backtracking are common during the development.

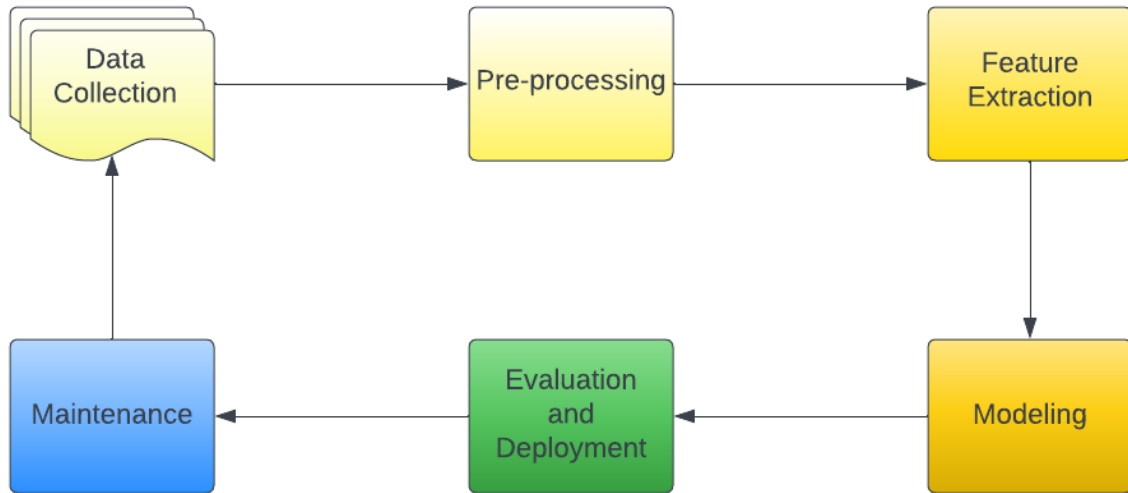


Figure 2.2: A general workflow of NLP development

2.3 Explainability of NLP

The interpretability and explainability of models is an open issue restricting the adoption of AI techniques, especially on NLP systems. It is clear that AI and NLP systems are performing progressively well and they are playing more and more crucial roles in the human society. However, with AI systems getting more controls and their effectiveness being experimentally proven, people are also concerned about their safety and reliability from a human perspective. Explainable models are especially needed for areas such as loan processing and medical diagnosis (Lakkaraju, Kamar, Caruana, & Leskovec, 2019). It is doubtful that critical areas would fully deploy AI systems before they are considered both useful and trustworthy. Furthermore, through a better transparency of AI models, we could investigate bias or discrimination, and improve model performance by identifying issues.

What is undeniable is that there are many systems functioning well without humans' full understanding. An example is our brain: we can process the environmental information

we receive through listening, seeing, reasoning, etc., but without knowing exactly how our brains actually work. Complex machine learning models used for natural language processing, such as SVMs, RFs, and deep neural networks, are in a similar situation: they could take the input and perform an action according to it, but we (humans) are unaware how and why certain actions, such as predictions, are performed. We usually call these non-transparent models "black boxes", as shown in Figure 2.3.

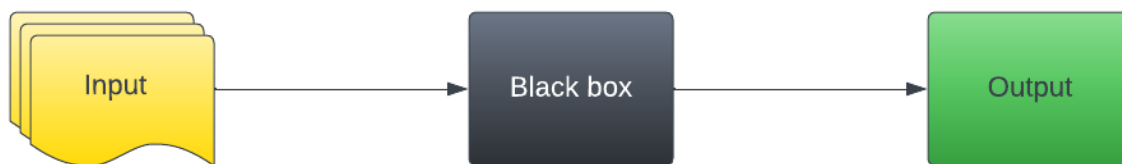


Figure 2.3: A black box machine learning model

2.3.1 Main Categories of Explanations

A sub-field of AI called XAI focuses on explaining all forms of artificial intelligence systems. Specifically for NLP, there are two ways to categorize explanations:

1. Global or Local

- Global explanations explain the comprehensive working mechanism of a model to take actions based on input, such as how a model makes predictions.
- Local explanations explain how a specific sample data was mapped to its corresponding output, usually through features in the data or a surrogate model.

2. Data, Intrinsic, or Post-Hoc

- Data explanations are the information researchers can find in the dataset, such as via statistical analysis or feature engineering. Data visualization is an example of data explanation.

- Intrinsic explanations, sometimes called self-explaining, are from the self-explanatory machine learning models, that have a relatively simple structure and could be naturally understood by researchers. Some common self-explanatory machine learning models are decision trees, K-means, and linear models.
- Post-hoc explanations are usually added to black box models, and do not look into a model’s internal structure, but rather use a surrogate model (a simpler and explainable model) as a proxy and focus on the relationships between input and output data samples. One famous post-hoc technique is LIME (Ribeiro, Singh, & Guestrin, 2016).

Table 2.1 shows an overview of the main categories of explanations discussed above.

Table 2.1: Overview of main categories of explanations

Category	Description
Global & Data	Having a general and high-level description about the dataset
Global & Intrinsic	Explaining the model’s internal reasoning on taking actions
Global & Post-Hoc	Using techniques focusing on data relationships to explain the model
Local & Data	Explaining the sample or data point through its specific characteristics
Local & Intrinsic	Explaining the sample data using the model
Local & Post-Hoc	Using additional techniques to explain a single action (such as prediction) performed by the model

Our studies focus on the local explanations. The glass box-based systems provide both global and local intrinsic explanations, while the LLM-based systems generate local post-hoc explanations.

2.3.2 Two Aspects of Explanations

[Danilevsky et al. \(2020\)](#) introduced two aspects to cover other important characteristics of explainability:

1. **Techniques for Deriving Explanations**

The techniques and operations used by AI or ML professionals (such as researchers and engineers), to derive "raw explanations" that focus more on theoretical justifications of models' actions.

2. **Explanation to End Users**

Explaining the model or model's actions to end users of the AI system, through visual presentations. Ideally, these explanations are designed by User Experience (UX) engineers who are aware of the users' needs.

Our work covered both aspects, with a focus on the latter one. The glass box-based systems present the explanations to the end users through interactive charts, whereas the LLM-based systems generate explanations in textual form.

2.3.3 Analyzing Explainable Systems

[Sokol and Flach \(2020\)](#) proposed a framework for systematically assessing, comparing and discussing explainable approaches: the Explainability Fact Sheets. According to the framework, the explanatory systems could be evaluated along the following five dimensions:

1. **Functional Requirements** consider algorithmic requirements, such as problem types, feature types, scope of explanations, or computational complexity.
2. **Operational Requirements** focus on the type of supported interactions, such as explanatory medium, data and model transparency, and explanation audience.
3. **Usability Requirements** take user-centered perspectives on the properties of explanations, such as completeness, interactiveness, and coherence of the explanations.

4. **Safety Requirements** discuss the robustness and security of the system, on aspects such as information leakage and explanation misuse.
5. **Validation Requirements** evaluate the effectiveness of an explainability approach, such as user studies and synthetic experiments.

As mentioned, in this study, we designed two types of explainable depression estimation systems, with the intent to explain reasons for classifying users' intensity or severity on each of the 21 symptoms on the questionnaire (Function of the Explanation). The systems are designed with different techniques. The first type used glass box models that need training data and the classifiers are transparent. The second one is based on LLMs that do not require (further) training but they are black box models (Data and Model Transparency). The systems could explain how features and/or texts in the users' writings affected the classification on the symptoms (Actionability and Provenance). The explanations, delivered through figures or pieces of text (Explanatory Medium), should be understandable by educated people (Explanation Audience), and interactive (System Interaction).

2.4 Summary

This chapter introduced the field of natural language processing and reviewed its main workflow, stages, and representative techniques. The common types of XAI techniques were also introduced.

Chapter 3

Related Work

Texts written by users could contain much valuable information about mental disorders. Due to the importance of depression treatment and flourishing NLP techniques, researchers attempted to develop computational methods for depression detection, using textual information. In this chapter, relevant literature for depression detection is discussed.

3.1 Lexical Usage and Emotions

To develop computational methods for detecting signs of depression using textual information, analyzing word usage became a natural starting point. Through statistical investigation, researchers found that words that indicate negative emotions, causes, sensory perception, and the first person singular pronouns were more commonly used when students described negative activities such as breakups ([Boals & Klein, 2005](#)).

Linguistic Inquiry and Word Count (LIWC), a computerized text analysis tool was developed to assess word usage in psychologically meaningful categories ([Tausczik & Pennebaker, 2010](#)). This tool was built by creating dictionaries from domain knowledge, in which the words were categorized into different groups. With the help of the LIWC software, [Al-Mosaiwi and Johnstone \(2018\)](#) found that elevated use of absolutist words could

indicate anxiety, depression, and suicidal ideation, and are more faithful in tracking the severity of affective disorder than negative emotion words.

3.2 Exploring Features Related to Depression

Many features other than word counts were designed and applied for this task. Social engagement, emotion, social network, and linguistic style were used by researchers to predict postpartum changes including postpartum depression (De Choudhury, Counts, & Horvitz, 2013). In addition to social and semantic features, linguistic n-gram features were implemented by Tsugawa et al. (2015); the task was using users' social media data for estimating the degree of depression.

Mowery, Park, Bryan, and Conway (2016) further considered using demographic data such as age and gender as features, for classifying depressive symptoms based on social media (Twitter) on a population level. Chen, Sykora, Jackson, and Elayan (2018) proposed an approach for identifying users with or at risk of depression, using eight basic emotions, as well as their temporal evolution as features.

Term Frequency–Inverse Document Frequency (TF-IDF) was used to prepare features for predicting mental illness from social media Reddit (Thorstad & Wolff, 2019); TF-IDF is a classic method for weighting words and is still being widely applied because it gives higher weight to more discriminative words that appear in fewer documents.

3.3 Deep Learning for Depression Detection

Deep learning methods also attracted researchers working on the subject. Yates, Cohan, and Goharian (2017) proposed a method using a neural network model to identify the risk of self-harm or depression, using data from social media Twitter and Reddit.

Researchers who participated in Conference and Labs of the Evaluation Forum (CLEF) eRisk 2017 (Losada, Crestani, & Parapar, 2017) focused on classifying users into binary

targets: at risk or non-risk of depression. The dataset consisted of text examples collected from Reddit users, and bag-of-words, word embedding, and bigram models were examined. [Husseini Orabi, Buddhitha, Husseini Orabi, and Inkpen \(2018\)](#) explored the effectiveness of deep neural architectures, such as CNNs and RNNs, on detecting signs of mental disorder depression using unstructured text data extracted from the Twitter data released earlier, for the shared task on Computational Linguistics and Clinical Psychology (CLPsych) 2015 ([Coppersmith, Dredze, Harman, Hollingshead, & Mitchell, 2015](#)).

As part of CLEF eRisk 2019, there was a proposed task about using the BDI questionnaire to automatically predict the depression level of social media users based on their social media writings, and the task continued in 2020 and 2021 ([Parapar, Martín-Rodilla, Losada, & Crestani, 2021](#)). Importantly, annotated datasets were provided for the tasks. This led to somewhat explainable depression predictors, by indicating possible symptoms (such as lack of sleep, loss of appetite, and reduced physical activity). See Appendix A for the full set of questions from the BDI questionnaire and their possible answers. In general, the performance of the systems that participated in this shared task was poor. Deep learning was used by a few of the participants. For example, while participating in the shared task Task 3 at CLEF eRisk 2021 ([Parapar et al., 2021](#)), [Inkpen, Skaik, Buddhitha, Angelov, and Fredenburgh \(2021\)](#) conducted experiments with Transformer-based models Sentence-BERT (SBERT) and Sentence-RoBERTa (SRoBERTa) ([Reimers & Gurevych, 2019](#)), a Deep Averaging Network (DAN) model Universal Sentence Encoder (USE) ([Y. Yang et al., 2019](#)), as well as a Hierarchical Attention Network (HAN) for text classification tasks inspired by [Z. Yang et al. \(2016\)](#). On the same task, [Maupomé, Armstrong, Rancourt, and Soulas \(2021\)](#) proposed a system that applied topic modeling using Embedded Topic Model (ETM) ([Dieng, Ruiz, & Blei, 2020](#)) which was trained on a depression detection dataset issued from Reddit, and a regression approach with nearest-neighbors on the values of the answers. This system achieved the highest score on one metric and also performed well on the other three metrics from the shared task. In 2022, [Skaik and Inkpen \(2022\)](#) continued working on the task and proposed a method that combined multiple deep learning models to answer different questions. Through all these efforts, a better performance was achieved on some of the metrics. In this thesis, we propose new methods to solve the task, with

improved performance and with added explainability for the predicted answers.

3.4 Summary

This chapter discussed the practices of applying NLP methods for detecting depression through analyzing textual (social media) data, including the CLEF eRisk shared tasks in which we actively participated.

Chapter 4

Datasets

As discussed in Chapter 2, to analyze users' mental conditions, textual data are a suitable source for this purpose. In this study, two datasets were used, both were collected from the social media Reddit. We first developed a sentence (or writings including posts and comments) selector, which would filter out unrelated texts. This selection system was then applied to another dataset to produce a "polished" dataset for our experiments. This chapter will introduce the two shared tasks the datasets were sourced from, and provide details about the two datasets.

4.1 eRisk Shared Tasks

The eRisk is a lab aiming to facilitate the development of early risk detection systems, as well as to explore corresponding evaluation and challenges, specifically in the domain of psychological issues such as depression, addiction, self-harm, and eating disorders. A timely identification of mental disorders could be extremely helpful; for example, alerts or interventions could be conducted when a person starts showing signs of a mental disorder, when a sexual predator starts interacting with a minor or youth, or when a potential offender starts publishing antisocial speeches or comments on the Internet (Losada, Crestani, & Parapar, 2018). Based on test collection and evaluation methods in (Losada & Crestani,

2016), eRisk introduced an exploratory task on the early detection of depression in 2017. This was the starting page of eRisk at CLEF; however, the performance of almost all participated teams' proposed systems was relatively low, with the highest F1-score of 64% (Losada et al., 2017).

In eRisk 2018 (Losada et al., 2018), two tasks were held in the lab, one for detecting depression and the other one for anorexia (an eating disorder). The task on depression detection was a continuation of the shared task in 2017, the data provided in 2017 could be used as a training dataset, and new test cases (depression or non-depression) were provided during the testing stage. The evaluation used standard metrics for classification, such as precision, recall, and F1-score, and an early risk detection measure proposed in (Losada & Crestani, 2016). The Early Risk Detection Error (ERDE) (Losada & Crestani, 2016), which was a measure specifically designed for early risk detection was applied for rewarding early alerts, because standard measures (such as F1-score) are time-unaware and do not penalize late decisions (Losada et al., 2018). With more institutions and teams participating, the highest F1-score of 64%, the highest precision of 67%, and still-modest performance were reached and systems were still struggling to differentiate depressed and non-depressed users (a binary classification problem). Until then, the solutions were mainly based on Text Mining (TM) techniques.

In 2019, eRisk held a new task of analyzing a user's history of social media posts and extracting evidence for estimating the user's depression level (Losada, Crestani, & Parapar, 2019). This task requires participating teams to develop systems to process the user (subject) posts, then estimate the user's depression level by filling a standard depression questionnaire - the BDI questionnaire, which assesses conditions about sadness, loss of interest, change of sleeping patterns, etc., then calculates the overall depression level. The dataset provided to participants consists of 20 users, and the participating systems had to produce a file with the structure shown in Figure 4.1. Each line represents a user's id, as well as the system's prediction for his/her 21 answers to the BDI questionnaire. Notably, the possible values for the answers are 0, 1, 2, and 3 for questions except questions 16 and 18, which have a different answer set of 0, 1a, 1b, 2a, 2b, 3a, 3b. Multiple evaluation metrics were introduced and will be discussed in detail in Chapter 6. This newly introduced

task is even more challenging in depression detection than the modest-performing task in eRisk 2018.

At eRisk 2020 (Losada, Crestani, & Parapar, 2020), a continuation of the depression assessment task from 2019 was held, with the difference that the participants this year could use the dataset shared in 2019 as the training data (that is, a training dataset with 20 examples). The task still used the BDI questionnaire and aimed to explore methods to automatically estimate the severity of social media users' depression level, by assessing each of the 21 symptoms related to depression. The task required participants to automatically fill the BDI questionnaire for 70 test users, with the help of their history of social media posts. The evaluation metrics were the same as the four used in 2019. Even though the participants had a training set that year, the performance scores were even lower (Losada et al., 2020).

The third and last edition of the depression severity estimation task was held at eRisk 2021 (Parapar et al., 2021). In this year, the participants could use the data from the last two years as the training dataset (that is, 90 examples), and predicted answers to the BDI questionnaire for 80 new users in the test set. The evaluation metrics were the same as the four used in 2019 and 2020, and the output format was also the same as described in Figure 4.1. Despite improvements in scores of some metrics, the performance was still modest, and it is still important to continue improving the generalization processes.

Over these years, the eRisk lab employed and compared multiple technologies other than TM, including NLP, ML, or Information Retrieval (IR) techniques.

In 2022, the eRisk lab held a successor edition of the depression detection task from eRisk 2017 and 2018 (Barrón-Cedeño et al., 2022). The dataset contained multiple writings and the users could be classified as depressed or non-depressed users. Data from the 2017 and 2018 editions were used to create the training data in 2022, and users who had clearly mentioned that they were depressed were indicated. Still, no participating systems achieved a good performance.

A new task on searching for sentences that are relevant to symptoms in the BDI questionnaire was proposed at eRisk 2023 (Parapar et al., 2023). The shared dataset contains

the derived sentences from social media users, and participating teams should try to extract and rank relevant writings from millions of sentences. This dataset is described in detail in the Section 4.2.

```

username1 answer1 answer2 answer3 ... answer21

username2 answer1 answer2 answer3 ... answer21

username3 answer1 answer2 answer3 ... answer21

...

```

Figure 4.1: Structure of resulting file for depression searching

4.2 Dataset for Depression Symptoms Search

This dataset was shared for Task 1 of the eRisk 2023 (Search for symptoms of depression) (Parapar et al., 2023). Participants of the shared task were given files in the TREC format containing documents (sentences) of each subject (user). Each document had a document ID as well as the text of the document. The corpus provided to the participants was a sentence-tagged dataset based on eRisk’s past data. Some synthetic examples are given in Figure 4.2 to show the format of stored documents in each TREC file (for privacy reasons, no real examples are shown).

The dataset contains only the derived sentences from social media, with no labels included (to indicate if a sentence is relevant or not; therefore supervised learning methods were not applicable). Languages other than English were not filtered out. The systems participating in the task had to extract the top 1000 relevant sentences for each of 21 symptoms in the BDI questionnaire and provide rankings for the extracted documents. Some statistics of this dataset are shown in Table 4.1.

```

<DOC>
  <DOCNO>s_0_0_0</DOCNO>
  <TEXT> İşte sadece rastgele bir örnek. onları ekledim </TEXT>
</DOC>

<DOC>
  <DOCNO>s_0_1_0</DOCNO>
  <TEXT>I just saw everyone cries ;) ... </TEXT>
</DOC>

<DOC>
  <DOCNO>s_0_2_0</DOCNO>
  <TEXT> an example. </TEXT>
</DOC>

```

Figure 4.2: Format of documents to be searched

4.3 Dataset for Depression Estimation

This dataset was shared at eRisk 2021 Task 3 (Measuring the severity of the signs of depression) (Parapar et al., 2021) and was built upon data shared at eRisk 2020 and eRisk 2019 for the same task. As mentioned above, the dataset contains a training dataset and a test dataset.

The training dataset contains 90 examples, which consist of 43,514 user writings (Reddit

Table 4.1: Statistics of the dataset for depression symptoms search

	Quantity
Number of TREC files	3,107
Number of subjects	3,107
Number of sentences	4,264,693

Table 4.2: Statistics of labels in the training data for depression estimating

Question	Answer 0 ¹	Answer 1	Answer 2	Answer 3
Q1	27 (30%)	47 (52%)	11 (12%)	5 (5%)
Q2	22 (24%)	34 (37%)	20 (22%)	14 (15%)
Q3	22 (24%)	35 (38%)	18 (20%)	15 (16%)
Q4	28 (31%)	33 (36%)	23 (25%)	6 (6%)
Q5	34 (37%)	32 (35%)	12 (13%)	12 (13%)
Q6	60 (66%)	13 (14%)	11 (12%)	6 (6%)
Q7	28 (31%)	17 (18%)	23 (25%)	22 (24%)
Q8	28 (31%)	27 (30%)	23 (25%)	12 (13%)
Q9	41 (45%)	37 (41%)	7 (7%)	5 (5%)
Q10	42 (46%)	23 (25%)	8 (8%)	17 (18%)
Q11	37 (41%)	31 (34%)	14 (15%)	8 (8%)
Q12	28 (31%)	32 (35%)	8 (8%)	22 (24%)
Q13	38 (42%)	21 (23%)	16 (17%)	15 (16%)
Q14	38 (42%)	21 (23%)	20 (22%)	11 (12%)
Q15	17 (18%)	32 (35%)	28 (31%)	13 (14%)
Q16	17 (18%)	36 (40%)	24 (26%)	13 (14%)
Q17	38 (42%)	31 (34%)	16 (17%)	5 (5%)
Q18	32 (35%)	30 (33%)	15 (16%)	13 (14%)
Q19	29 (32%)	25 (27%)	25 (27%)	11 (12%)
Q20	21 (23%)	34 (37%)	21 (23%)	14 (15%)
Q21	51 (56%)	18 (20%)	11 (12%)	10 (11%)
Total	678 (35%)	609 (32%)	354 (19%)	249 (13%)

posts and comments) written by 90 users in the past two years at the time, as well as their answers to 21 questions of the BDI Questionnaire. The test dataset consists of 19,803 posts and comments written by 80 users and their answers to the questionnaire. As discussed in Section 4.1, the labels for questions 16 and 18, which have a different answer set, are revised so that the answers with letters are merged into a single answer (for example, 1a and 1b are merged into 1). Each of the remaining questions has four answers: 0, 1, 2, and 3. Statistics for the user answers (labels) are shown in Table 4.2.

¹Percentage numbers were rounded.

minimal depression	depression levels 0-9
mild depression	depression levels 10-18
moderate depression	depression levels 19-29
severe depression	depression levels 30-63

Figure 4.3: Depression categories associated with depression levels

Through investigations on the training data, it can be seen that most of the symptoms users have are minor, as about 68% of users answered 0 or 1 (with about 35% answered 0 and about 32% answered 1), and a few users reported severe symptoms. Specific questions may have a different distribution, for example, around 40 percent of users answered 1 on question 16 with answer 2 following it (about 26%). As label frequency distributes among multiple labels, and varies by question, we can see that simply choosing a label as the default value in case the classifier cannot reach a decision is not practical.

To calculate the overall level of depression for a user, depression categories introduced in Figure 4.3 (Losada et al., 2019) were considered. Based on this categorization, the calculated levels of depression in the training data are shown in Table 4.3. We could conclude that users are distributed in all four categories, with most users having mild (30%) or severe (30%) depression. It is worth noting that this finding does not fully comply with the findings we had while investigating label distributions, which indicated that most of the users answered having mild or no symptoms on the 21 questions.

Table 4.3: Statistics of depression categories in the training data

Category	# of subjects ²
minimal depression	14 (15%)
mild depression	27 (30%)
moderate depression	22 (24%)
severe depression	27 (30%)

4.4 Summary

In this chapter, we briefly described the shared tasks focusing on depression held at CLEF eRisk. Then we discussed the two datasets we are using in this thesis. The first one is the dataset for searching depression-related documents/sentences, shared at eRisk 2023. The second one is the dataset from eRisk 2021 for estimating the depression level through the BDI questionnaire was introduced and analyzed.

²Percentage numbers were rounded.

Chapter 5

Search for Depression-Related Writings

This chapter¹ discusses the massive selection of depression-related writings conducted on the involved datasets. With these methods, user writings would be less noisy and the length of texts to be further processed could be greatly reduced so that a more transparent depression detection system could be developed. In this study, a text collecting system was built to provide ranked sentences based on their relevance to each of the 21 symptoms of depression derived from the BDI questionnaire ([Beck et al., 1961](#)) was introduced.

5.1 Introduction

Social media users share their thoughts through writing. The amount of writing on social media such as Reddit, including posts and comments, is huge. Massive information resides in these writings, but it is usually sparse and unrelated texts would even distract ML models. To have a denser dataset and make it fit to more advanced architectures such as DL models and LLMs, a selection process is necessary.

A writing (sentence in this context) would be considered relevant to a symptom if it provides information, ideally explicit, about the user’s status of that particular symptom.

¹Parts of this chapter were published in ([Wang & Inkpen, 2023](#)).

The information needs not be negative for the symptom, since neutral or positive writings would also help indicate the user’s status.

The task was considered as an information retrieval task, where user-written sentences are stored as documents. The 21 questions from the BDI questionnaire were transformed into 21 queries. The aim of the task is to retrieve the top 1000 relevant documents for each query and compute their rankings with rank 1 being the most relevant document.

Several text embedding methods were used for transfer learning (directly using knowledge from pre-trained models), including contextual text embedding methods such as DistilBERT (Sanh, Debut, Chaumond, & Wolf, 2020), and the distributional word embedding method GloVe (Pennington, Socher, & Manning, 2014). Combined with a semantic similarity measure (the cosine similarity), the developed system could extract relevant sentences for each query and provide a rank for the top 1000 sentences for each of 21 questions/symptoms in the BDI questionnaire.

5.2 Methodology

5.2.1 Information Retrieval and Search

Information retrieval is the task that, given a query from a system user, the IR system searches and returns a ranked list of documents that match or are related to the specified query. Information retrieval techniques were employed for this task since they were appropriate for the scenario.

In a typical information retrieval system, an "indexer" is used to parse then store for each term a list of documents containing the term. The index is first constructed and then it is used for ranking based on some metrics (similarity formula). A usual search engine collects the documents before the information retrieval step and needs frequent updates. The index needs to be updated too.

In this task, the collection of documents is provided and it is static. Since we do not need to update our collection and the system is not designed for frequent searches, the

inverted index was not constructed. At this stage, the search space was restricted to a set of documents at hand. To accelerate the calculation of contextual representations, we selected many keywords from the questions in the BDI questionnaire, in order to filter out unrelated documents.

5.2.2 Document Acquisition

All the documents, which add up to more than 4 million sentences collected from social media and were provided in the CLEF eRisk 2023 shared task (Parapar et al., 2023), were downloaded and stored on an online storage platform Google Drive (Gallaway & Starkey, 2013). Google Drive is a platform that provides various functionalities, including online storage. Files securely stored could then be accessed through Google Colaboratory, a serverless Jupyter Notebook environment for interactive development (Bisong, 2019). Before the pre-processing steps, the system on the Google Colab connected to Google Drive and extracted all the documents from the files in the dataset.

5.2.3 Data Normalization and Text Processing

As discussed in Section 4.2, the sentences are stored in TREC files containing documents with DOCNO (document number) and TEXT (textual content). After loading the files, sentences were extracted, and additional information related to the data format was discarded.

Depending on the model to be used, different pre-processing steps were applied to the texts:

- For obtaining word embeddings through GloVe (Pennington et al., 2014), tokenization, lowercasing, stopwords and punctuation removal were applied;
- When getting vector representations using contextual representation models, we filtered out the documents (sentences) that did not contain symptom-related keywords,

then we used the transfer learning strategy by employing the representations from the models directly to build the semantic representations. Traditional pre-processing methods were not applied, but the specific tokenization used by each contextual embedding model was used.

The normalization and processing steps were applied to both the documents and the queries.

5.2.4 Searching with Contextual Representations

Various transformer-based models were experimented with to obtain contextual representations of documents and queries. They are text embedding methods that look at all the words in a sentence at the same time (Vaswani et al., 2017). In the study, pretrained models with default configurations from Transformers² are used. We filtered out the documents not containing certain keywords, to reduce the size of the dataset in order to accelerate computation. These keywords were manually selected by us from the 21 questions on the BDI questionnaire. All of the documents (4,264,693 sentences) were loaded for processing; 111,982 sentences were kept after filtering, and 4,152,711 sentences were filtered out using these keywords. The queries we built for each question using both the text of the questions and the text of the possible answers. More information about the keywords we used and the queries we designed is shown in Table 5.1.

Table 5.1: Queries and keywords for each question

Question	Keywords	Query
Q1	sadness, sad, unhappy	Sadness. I feel sad unhappy cannot stand it.
Q2	pessimism, discouraged, hopeless	Pessimism. I feel discouraged about my future is hopeless and will get worse.
Q3	failure, fail	Past Failure. I have failed.

²<https://huggingface.co/docs/transformers/index>

Q4	pleasure, enjoy	Loss of Pleasure. I don't enjoy things.
Q5	guilty	Guilty Feelings. I feel guilty.
Q6	punishment, punish	Punishment Feelings. I am being punished.
Q7	confidence, disappointed	Self-Dislike. I have lost confidence. I am disappointed in myself.
Q8	criticalness, criticize, blame, fault	Self-Criticalness. I criticize myself blame myself for my faults.
Q9	suicidal, suicide, kill	Suicidal Thoughts or Wishes. I kill myself.
Q10	crying, cry	Crying. I cry.
Q11	agitation, restless	Agitation. I am restless or agitated keep moving.
Q12	interest, interested	Loss of Interest. It's hard to get interested.
Q13	indecisiveness, decision, decide	Indecisiveness. I find it difficult to make decisions.
Q14	worthlessness, worthless, worthwhile, useful	Worthlessness. I feel worthless not useful.
Q15	energy, energetic	Loss of Energy. I don't have enough energy.
Q16	sleep, sleeping	Changes in Sleeping Pattern. I sleep more or less than usual.
Q17	irritability, angry	Irritability. I am irritable.
Q18	appetite, food, eat	Changes in Appetite. My appetite is greater or less.
Q19	concentration, concentrate	Concentration Difficulty. It's hard to keep my mind. I can't concentrate.
Q20	tiredness, tired	Tiredness or Fatigue. I am tired or fatigued.
Q21	sex	Loss of Interest in Sex. I am less interested in sex.

5.2.4.1 DistilBERT with Cosine Similarity

The DistilBERT (Sanh et al., 2020), a distilled version of BERT with a smaller model and competitive performance was experimented with for the task. It is faster to train, and lighter to load. After the vector representations of sentences and queries were collected, the cosine similarity was used for calculating semantic similarity between the query and the document, which is defined as follows:

$$sim(v1, v2) = \frac{v1 \cdot v2}{|v1||v2|} = \frac{\sum_{i=1}^n v1_i v2_i}{\sqrt{\sum_{i=1}^n (v1_i)^2} \sqrt{\sum_{i=1}^n (v2_i)^2}}$$

where $v1$ and $v2$ represent vector representations of documents being compared.

The calculated similarity was considered as the relevance of documents. Rankings of document relevance were then saved.

5.2.4.2 RoBERTa with Cosine Similarity

RoBERTa is an improved version of BERT, with a more carefully designed pretraining (Y. Liu et al., 2019). Similar to the method using DistilBERT, we used the cosine similarity (Section 5.2.4.1) to compute the text similarity and the recorded ranks for 21 questions or queries.

5.2.4.3 Universal Sentence Encoder with Cosine Similarity

The Universal Sentence Encoder is a text encoder that directly encodes sentences into vectors. It is specifically designed for transfer learning of various types of NLP tasks. The encoder based on the transformer architecture was trained in the following way: the word representations acquired through the transformer were converted to a fixed-length encoding vector by summing the element-wise representations at each word position, and then the vector was divided by the square root of the length of the sentence to reduce sentence length effects. The inputs to the encoder are lowercased strings that are tokenized using the Penn

Treebank Tokenizer (PTB), and the outputs are 512-dimensional vector representations. Since the model was designed to be of general purpose, multi-task learning was conducted (Cer et al., 2018).

We used the model to obtain embeddings of queries and sentences, and calculated the similarity between them to obtain rankings using the cosine similarity (Section 5.2.4.1).

5.2.5 Searching with Distributional Word Representations

We also used GloVe to get distributional embeddings of sentences and queries. Unlike transformers, GloVe creates co-occurrence matrices of texts and then applies matrix factorization on the global matrix to shapes with various dimensionalities. As described in previous sections, traditional pre-processing steps such as tokenization, lowercasing, stop-word and punctuation removal were conducted on sentences. After data were cleaned, 2 versions of GloVe were experimented with: density of 50-dimension and 100-dimension. The GloVe embeddings were acquired for both documents and queries and then used for cosine similarity calculations (Section 5.2.4.1). Ranks based on similarities are recorded as relevant, the same as other methods.

5.3 Evaluation

We submitted 5 runs to the eRisk 2023 (Parapar et al., 2023); each run included 21 TREC-style formatted rankings of sentences, as described in Section 4.2. For each symptom in the BDI questionnaire, up to 1000 resulting documents (sentences) sorted by estimated relevance were submitted for performance assessment.

At eRisk 2023, expert assessors annotated a pool of sentences associated with each symptom. To label a document as relevant, it should pertain to the symptom and provide explicit information about the author’s state in relation to the symptom.

Since both the topic and the individual’s state were considered for relevance, this introduced a higher complexity level than conventional relevance assessments. The organizers

believe that a robust annotation methodology and formal assessment guidelines to ensure consistency and accuracy were developed (Parapar et al., 2023).

More than 210 hours of assessors’ time were spent on annotation. Among the three annotators, one annotator was a psychologist, and the other two annotators were computer scientists with expertise in early-risk technologies: one postdoctoral fellow and one PhD student. The organizers conducted sessions with the assessors and developed a guideline for annotation, which is shown in Figure 5.1.

Two aggregation criteria were applied to determine sentence relevance, according to the provided relevance judgments (qrels): majority (using majority voting among the available human judgments) and unanimity. Four standard ranking metrics were used for performance assessment:

- Mean Precision at 10

The mean of P@10 values among all queries, with P@10 being the precision for the top 10 retrieved documents. The precision is the fraction of retrieved documents that are considered relevant:

$$Precision = \frac{\# \text{ of retrieved relevant items}}{\# \text{ of retrieved items}}$$

- Mean R-Precision

The mean of R-precision values among all queries, where R-precision is the precision of the top *Rel* documents returned, given a set of known relevant documents *Rel* (Manning, Raghavan, & Schütze, 2008).

- Mean Average Precision

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk})$$

where the set of relevant documents for an information need $q_j \in Q$ is $\{d_1, \dots, d_{m_j}\}$

and R_{jk} is the set of ranked retrieval results from the top result until you get to document d_k (Manning et al., 2008).

- Mean Normalized Discounted Cumulative Gain (NDCG) at 1000

The mean of NDCG@1000 values among all queries, with NDCG being defined as:

$$NDCG(Q, k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_k \sum_{m=1}^k \frac{2^{R(j,m)} - 1}{\log(1 + m)}$$

where $R(j, d)$ being the relevance score assessors gave to document d for query j , and Z_k is a normalization factor calculated so that a perfect ranking's NDCG at k to be 1 (Manning et al., 2008).

5.4 Results and Discussion

As discussed above, the expected criteria for sentence relevance assessments were introduced through examples. In Figure 5.1, some examples from the lab overview paper (Parapar et al., 2023) are given, to illustrate topic relevance.

The systems were evaluated using the majority-based qrels and unanimity-based qrels. The results of the majority voting evaluation for the 5 runs submitted by us are presented in Table 5.2, and the unanimity evaluation for the runs are shown in Table 5.3. Our best performance was achieved by the method that employed the Universal Sentence Encoder with Cosine Similarity (USESim), discussed on Section 5.2.4.3.

The results show that the universal text representation USE performed better overall than the other contextual representation techniques such as DistilBERT, for this task. Also, the contextual representation methods performed better on the metrics P@10 and NDCG@1000, compared to the distributed representation methods based on GloVe. A much larger search space was applied when using the method with GloVe (GloveSim and Glove100Sim), since sentences were not filtered out in this method as the computation was fast enough, and all the sentences were checked for similarity. Lower scores of the

Table 5.2: Results for submitted 5 runs (majority voting)

Run	AP	R-PREC	P@10	NDCG@1000
USESim	0.160	0.248	0.600	0.382
Glove100Sim	0.017	0.052	0.195	0.105
RobertaSim	0.033	0.080	0.329	0.150
GloveSim	0.011	0.038	0.162	0.075
BertSim	0.084	0.150	0.505	0.271

Table 5.3: Results for submitted 5 runs (unanimity)

Run	AP	R-PREC	P@10	NDCG@1000
USESim	0.139	0.232	0.438	0.380
GloveSim	0.008	0.028	0.110	0.063
Glove100Sim	0.011	0.042	0.110	0.092
RobertaSim	0.025	0.068	0.190	0.140
BertSim	0.070	0.130	0.357	0.260

GloVe-based methods could be due to the removal of stopwords, for example, pronouns that refer to the participant in the discourse (the agent) were removed, but they could contain relevant information.

The performance of Glove100Sim is better than GloveSim which provides embeddings with densities of 100 and 50 dimensions separately. This could demonstrate again the value of having more information and features being encoded for documents, provided that the vectors are not sparse.

Similar to situations met by many other participating teams at eRisk 2023, our methods performed generally worse on unanimity-based evaluations than on majority-based evaluations. The reason could be attributed to the stricter (but more convincing) nature of evaluation with unanimity-based qrels on relevance judgements of sentences.

The proposed systems did not perform well according to metrics mean average precision (AP in the tables) and Mean R-Precision (R-PREC in the tables). The main reason could be the inherent difficulty of the task, as well as the unknowable nature of the dataset since

Table 5.4: Our results compared to the best results in the shared task (majority voting)

Metric	Our system	Our best	Best system	Best in shared task
AP	USESim	0.160	SentenceTransformers_0.25	0.319
R-PREC	USESim	0.248	SentenceTransformers_0.25	0.375
P@10	USESim	0.600	SentenceTransformers_0.25	0.861
NDCG@1000	USESim	0.382	SentenceTransformers_0.25	0.596

Table 5.5: Our results compared to the best results in the shared task (unanimity)

Metric	Our system	Our best	Best system	Best in shared task
AP	USESim	0.139	SentenceTransformers_0.1	0.293
R-PREC	USESim	0.232	SentenceTransformers_0.25	0.360
P@10	USESim	0.438	SentenceTransformers_0.25	0.709
NDCG@1000	USESim	0.380	SentenceTransformers_0.25	0.615

no extra data were provided other than the merest user sentences. Despite the relatively low score on AP and R-PREC, the performance of the other two metrics mean precision at 10 (P@10 in the tables) and mean NDCG at 1000 (NDCG@1000 in the tables) could be considered decent. With 0.600 on P@10 (0.438 with unanimity), one could conclude that the top documents retrieved by the proposed system USESim could provide considerable confidence, that the retrieved top documents would contain relevant information regarding the users’ status on depression symptoms.

USESim’s performance on the metric P@10 leads to the adoption of the system to the dataset for estimating depression levels from the next chapter.

5.4.1 Comparison to Related Works

Our team (uOttawa) ranked 3rd among the 10 participating teams of task 1 at eRisk 2023 (Parapar et al., 2023) on both aggregation criteria. In Table 5.4 and Table 5.5, we compare our results with the best results from the shared task at eRisk 2023, on the four metrics.

Models based on sentence transformers developed by the Formula-ML team (Recharla,

[Bolimera, Gupta, & Madasamy, 2023](#)), achieved the best performance across all 4 metrics and judgment types. Different from our best system, their method pre-processed the texts using the usual techniques, including lowercasing, removing punctuation, unnecessary spaces, numerals, emojis, and contractions, then tokenization. The para-MiniLM-L3-v2 sentence transformer model was applied, and then the cosine similarity was used to measure similarities (but not as the final score). The significant difference is that, unlike the methods we used to design queries for each symptom, the systems proposed by Formula-ML utilized all the options of symptoms and formed specific weights. The summation of the weighted similarity scores was used as the final similarity score. This method is slightly more inefficient than ours, but it utilized the questions on the questionnaire more extensively and achieved a better performance.

The systems proposed by the BLUE team ([Bucur, 2023](#)) experimented with two methods: a variant of MPNet ([Song, Tan, Qin, Lu, & Liu, 2020](#)) and MentalRoBERTa ([Ji et al., 2021](#)), and used two types of queries: the 90 options on the BDI questionnaire as well as 2,700 synthetic Reddit texts generated by ChatGPT. The generated Reddit texts were believed to have more diversity than the original responses on the questionnaire, and be more specific to Reddit. The system using only the original options as queries and encoded with the variant of MPNet performed best among their runs, achieving a score of 0.781 on P@10 which is higher than ours, but obtained lower scores on the remaining three metrics. The MPNet variant they used was explicitly designed for semantic search, which could lead to a high score on P@10, but only using the initial options on the questionnaire as the queries could be the reason for having lower performance on the three other metrics. The generated texts, according to the team, might be too specific and provide too many details, which could be unhelpful for searching.

5.5 Using USESim for Writing Selection

As described in Section 4.3, for the dataset for estimating depression, 90 users’ 43,514 writings, including Reddit posts and comments, are in the training data and 19,803 writings

from 80 users are in the test data. It is good to have more data as more information could reside in it, but too much text for a user could also introduce more noise, and it is difficult for machine learning models to process all of them.

The document retrieval systems proposed for the task of depression symptoms search could provide great value in pre-processing the dataset for depression estimation. In the next chapter, we will discuss in more detail about how we applied our best-performing searching system USESim to generate a smaller and more relevant dataset.

5.6 Summary

In this chapter, we discussed the task of searching for documents (user writings) relevant to depression symptoms. Then, 4 methods and 5 developed systems were introduced to tackle the task. Finally, we discussed the performance of our 5 runs submitted to the shared task. We concluded that the system USESim has value in selecting relevant user writings for the symptoms in the BDI questionnaire, therefore we adopted it for the next task.

Assume you are given a BDI item, e.g.:

15. Loss of Energy

- I have as much energy as ever.
- I have less energy than I used to have.
- I don't have enough energy to do very much.
- I don't have enough energy to do anything.

The task consists of annotating sentences in the collection that are topically-relevant to the item (related to the question and/or to the answers).

Note: A relevant sentence should provide some information about the state of the own writer related to the topic of the BDI item. But it is not necessary that the exact same words are used.

Your job is to assess sentences on how topically-relevant they are for a concrete BDI item. The relevance grades are:

1. Relevant: A relevant sentence should be topically-related to the BDI-item (regardless of the wording) and, additionally, it should refer to the state of the writer about the BDI-item.

0. Non-relevant: A non-relevant sentence does not address any topic related to the question and/or the answers of the BDI-item (or it is related to the topic but does not represent the writer's state about the BDI-item). For example, for BDI-item 15, a sentence that does not talk about the individual's level of energy (regardless of the wording), then is a non-relevant sentence.

Figure 5.1: Guidelines for labelling sentences related to depression symptoms ([Parapar et al., 2023](#))

Examples (assessment of sentences ranked for BDI-item number 15):

I cannot control my energy these days	Relevant (1)
My sister has no energy at all	Non-relevant (0) (because it does not refer to the writer who posted this sentence)
The book was about a highly energetic man	Non-relevant (0) (because it does not refer to the writer who posted this sentence)
I feel more tired than usual	Relevant (1)
The football team is named Top Energy	Non-relevant (0)
I am totally lonely	Non-relevant (0) (because it does not mention energy)
I have just recharged my batteries	Relevant (1)
I am lost	Non-relevant (0)

We advise you to not stop the assessment session in the middle of one BDI-item (this helps to maintain consistency in the judgments). To measure the assessment effort, we ask you to record the time spent on fully evaluating the sentences presented for each BDI-item.

Figure 5.1: Guidelines for labelling sentences related to depression symptoms ([Parapar et al., 2023](#)) (cont.)

Chapter 6

Estimating the Level of Depression

In this chapter¹, we present a machine learning-based system using glass box models in XAI, and a prompt learning-based system that can generate explanations, to predict users' responses in the BDI questionnaire, by analyzing their social media writings on Reddit in the past, and ultimately predict the severity of users' depression.

6.1 Introduction

A large number of harmful or even fatal acts, such as excessive usage of alcohol and suicide, are associated with depression. Depression is also considered a factor and cause of several kinds of disabilities. In recent years, applying NLP techniques to the task of massive early depression detection has become one of the important topics in the field of psychology and computer science.

Various machine learning techniques are being developed to predict depression ([Zulfiker, Kabir, Biswas, Nazneen, & Uddin, 2021](#)). At the same time, it is interesting and important to understand the working mechanisms of machine learning algorithms, which are used to

¹Parts of this chapter were published in ([Wang, Inkpen, & Kirinde Gamaarachchige, 2024](#)).

predict mental health conditions such as depression. This could greatly improve the reliability of AI models, and lead us to improve our understanding of mental health problems from multiple points of view.

This study tries to automatically fill in a standard depression questionnaire: the BDI questionnaire, by using machine learning models to do multi-class predictions of answers to questions in the questionnaire. As mentioned, the questionnaire has 21 questions in total, which can be used to evaluate conditions of feelings about sadness, sleeping, etc. Each question has 4 answers, except 2 questions (question 16 about sleep patterns and question 18 about appetite) have more than 4 answers of which answers were transformed into 4 classes.

As discussed in Section 4.3, the dataset contains a training dataset and a test dataset. The training dataset has 90 users' 43,514 writings on Reddit, and their answers to the BDI questionnaire. The test dataset consists of 80 users' 19,803 posts and comments as well as their responses to the questionnaire.

6.2 Methodology

Text classification models were developed for this task. Based on different scenarios, two types of systems were designed: systems using glass box models which are interpretable classifiers, and systems using LLMs which could generate predictions as well as explanations. From a general perspective, three steps are involved in the process:

1. Dataset preparation
2. Pre-processing and transforming textual data
3. Developing and testing

Figure 6.1 depicts the two proposed systems and their development procedures.

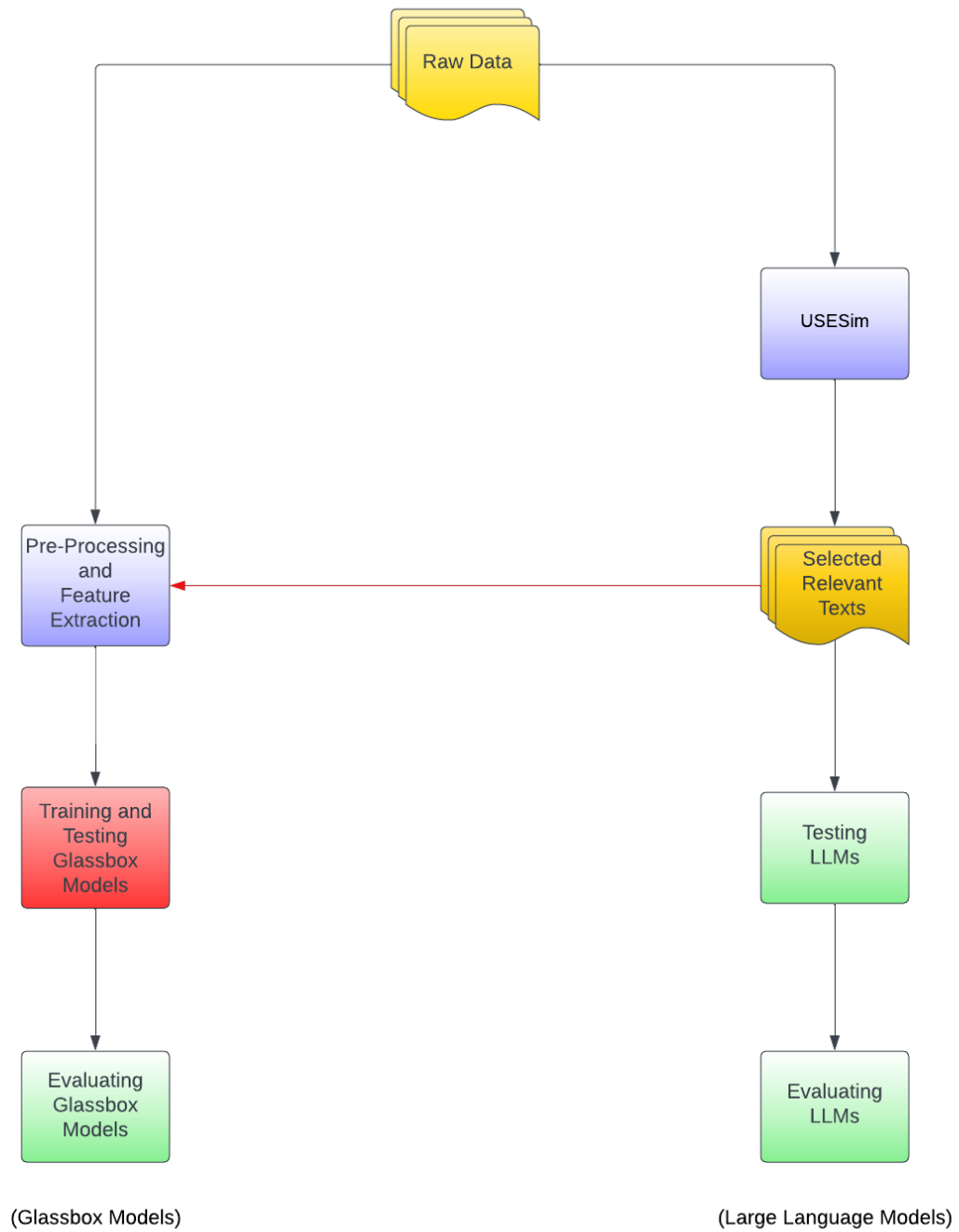


Figure 6.1: Flowchart of development pipeline of two model types

6.2.1 Data Preparation

The original datasets were used to extract features, i.e., stimuli for each depression-related symptom. In addition, as mentioned in Section 5.5, the writing-selection system USESim is adopted for cleaning the dataset for estimating depression levels, by selecting only symptom-relevant user writings and forming a more useful dataset. Based on our settings, two datasets were generated using USESim:

1. The Top-5 Dataset

Collected by applying the USESim to collect the top 5 relevant writings for each symptom in the BDI questionnaire. The statistics for text length of this type of training and test data are shown in Table 6.1.

2. The Top-1 Dataset

Formed by applying the USESim to collect the top 1 relevant writings (the most relevant post or comment) for each of the 21 symptoms. The statistics for the text length of this dataset are shown in Table 6.2.

The top 5 and top 1 relevant writings were selected with the consideration of the maximum length: as many LLMs have a short context length which refers to the maximum number of tokens that the model can process, shorter input texts are desired. As an example, the Llama 2 models have a context length of 4,096 tokens.

Table 6.1: Statistics of text length of the cleaned data (Top-5)

	Mean	Min	Max
Training-Q1	183.78	31.00	890.00
Training-Q2	276.74	37.00	958.00
Training-Q3	177.38	32.00	743.00
Training-Q4	206.89	39.00	1505.00
Training-Q5	171.72	32.00	994.00

Training-Q6	177.67	27.00	1171.00
Training-Q7	197.49	41.00	923.00
Training-Q8	161.58	44.00	783.00
Training-Q9	245.66	27.00	962.00
Training-Q10	112.19	27.00	445.00
Training-Q11	268.56	33.00	1502.00
Training-Q12	159.20	43.00	567.00
Training-Q13	211.44	25.00	851.00
Training-Q14	201.67	39.00	759.00
Training-Q15	248.41	35.00	716.00
Training-Q16	207.50	50.00	744.00
Training-Q17	217.79	31.00	1406.00
Training-Q18	177.93	32.00	630.00
Training-Q19	223.30	45.00	866.00
Training-Q20	228.73	49.00	960.00
Training-Q21	276.11	61.00	811.00
Training-All	2122.86	287.00	11039.00
<hr/>			
Test-Q1	304.71	36.00	1020.00
Test-Q2	413.99	43.00	2312.00
Test-Q3	237.18	29.00	1261.00
Test-Q4	274.61	33.00	945.00
Test-Q5	246.57	28.00	955.00
Test-Q6	250.69	26.00	1154.00
Test-Q7	314.34	42.00	1091.00
Test-Q8	232.59	33.00	693.00
Test-Q9	329.64	26.00	1703.00
Test-Q10	206.88	21.00	978.00
Test-Q11	382.46	37.00	1579.00
Test-Q12	225.05	28.00	960.00
Test-Q13	295.59	22.00	1571.00

Test-Q14	283.14	30.00	1083.00
Test-Q15	357.49	34.00	1115.00
Test-Q16	253.56	42.00	768.00
Test-Q17	322.65	25.00	1080.00
Test-Q18	230.90	40.00	857.00
Test-Q19	321.32	38.00	1022.00
Test-Q20	345.68	41.00	1665.00
Test-Q21	367.70	52.00	1467.00
Test-All	2561.28	202.00	11424.00

Table 6.2: Statistics of text length of the cleaned data (Top-1)

	Mean	Min	Max
Training-Q1	29.69	5.00	181.00
Training-Q2	51.88	5.00	234.00
Training-Q3	26.08	4.00	245.00
Training-Q4	36.30	4.00	255.00
Training-Q5	28.13	4.00	266.00
Training-Q6	31.40	4.00	693.00
Training-Q7	31.10	5.00	314.00
Training-Q8	31.19	4.00	331.00
Training-Q9	43.63	5.00	305.00
Training-Q10	25.19	4.00	213.00
Training-Q11	50.84	6.00	260.00
Training-Q12	27.90	5.00	146.00
Training-Q13	34.90	4.00	159.00
Training-Q14	35.27	5.00	255.00
Training-Q15	47.36	5.00	260.00
Training-Q16	44.61	5.00	260.00

Training-Q17	37.88	4.00	304.00
Training-Q18	35.23	4.00	260.00
Training-Q19	40.43	5.00	245.00
Training-Q20	59.67	4.00	382.00
Training-Q21	47.08	4.00	204.00
Training-All	524.31	77.00	2261.00
<hr/>			
Test-Q1	56.11	4.00	438.00
Test-Q2	84.61	6.00	438.00
Test-Q3	37.94	4.00	321.00
Test-Q4	44.86	5.00	438.00
Test-Q5	40.58	4.00	405.00
Test-Q6	42.67	5.00	532.00
Test-Q7	60.31	4.00	438.00
Test-Q8	46.88	4.00	509.00
Test-Q9	76.28	4.00	438.00
Test-Q10	38.52	4.00	326.00
Test-Q11	77.05	6.00	887.00
Test-Q12	43.34	4.00	429.00
Test-Q13	59.45	5.00	752.00
Test-Q14	57.23	6.00	438.00
Test-Q15	78.53	6.00	438.00
Test-Q16	59.17	6.00	398.00
Test-Q17	76.19	4.00	446.00
Test-Q18	38.81	4.00	242.00
Test-Q19	65.39	7.00	367.00
Test-Q20	63.86	5.00	398.00
Test-Q21	90.78	6.00	445.00
Test-All	742.86	61.00	2777.00
<hr/>			

6.2.1.1 Stimuli for Feature Extraction

Inspired by the Implicit Association Test (IAT) and the depression IAT (Meites, Deveney, Steele, Holmes, & Pizzagalli, 2008), 23 sets of depression stimuli are developed: 21 sets for each of 21 symptoms in the BDI questionnaire were built using the keywords from the questionnaire and their synonyms acquired from the WordNet (Miller, 1995; Oram, 2001); one set named "depression stimuli" adopted from the "stimuli for depression IAT" in (Meites et al., 2008); one set that was a union of the other 22 stimuli sets. To fit the results of the text pre-processing, these stimuli are also pre-processed (for example, by applying a stemmer that removes frequent suffixes).

Stimuli are counted on the cleaned texts and recorded as numeric features. In total 23 features were based on depression stimuli, including 21 features for 21 symptoms in the questionnaire, one feature with counts on all stimuli, and one feature with the adopted stimuli for depression IAT. The sets of stimuli are listed in Table 6.3.

Table 6.3: Sets of stimuli for feature extraction

Feature	Stimuli
Q1	gloomi, sorri, deplor, dysphor, distress, lament, unhappi, infelicit, sorrow, lugubri, sad, piti
Q2	demor, deter, admonish, warn, discourag, pessim, demoralis, dishearten, hopeless, monish
Q3	break, bankruptci, flush it, miscarri, bomb, give out, fail, go wrong, die, neglect, conk out, go bad, flunk, go, run out, loser, betray, break down, give way, nonstart, unsuccessful person, failur
Q4	pleasanc, bask, revel, enjoy, joy, pleasur, relish, savor, savour, delight, love

Q5	shame, shamefac, guilti, hangdog
Q6	penalis, punish, penal, penalti
Q7	frustrat, defeat, self-assur, author, confid, sure, trust, self-confid, let down, discomfit, foil, thwart, disappoint, assur
Q8	break, blast, error, decis, charg, criticis, darn, rap, fault, pick apart, damn, mistak, demerit, goddam, knock, flaw, pick, find fault, bless, geological fault, critic, shift, deuc, incrimin, blame, defect, infern, goddamn, vital, crucial, fractur, inculp
Q9	defeat, suicid, bolt down, wipe out, toss off, felo-de-s, stamp out, self-annihil, kill, pour down, drink down, obliter, belt down, putting to death, pop, vote down, self-destruct, vote out, shoot down, down
Q10	vocifer, cry out, squall, insist, call, shout, outcri, rallying cri, battle cri, war cri, blazon out, rank, flagrant, egregi, exig, weep, tear, cri, yell, shout out, gross, hollo, glare, exclaim, call out, scream, holler, instant, watchword, clamant
Q11	commov, ungratifi, stir, excit, unrest, charg, hullabaloo, vex, charge up, turn on, turmoil, press, push, unsatisfi, upheav, foment, raise up, shift, tempestu, shake up, rous, shake, fight, uneasi, restless, stir up, disturb, budg, agit, campaign, crusad, ferment
Q12	concern, sake, stake, interest group, occupi, pursuit, pastim, interest, matter to, worri, involv, interesting

Q13	settl, adjud, decis, resolv, decid, indecis, irresolut, determin, make up ones mind, conclus
Q14	utilitarian, unworthi, slimi, ugli, ineptitud, worthless, despic, worth-whil, wretch, util, vile, use
Q15	gumptiou, industri, up-and-com, energy depart, free energi, energet, energi, vital, push, vigor, doe, zip, get-up-and-go, muscular, vim, vigour, department of energi
Q16	rest, quiescenc, eternal sleep, dormant, sleep, dormanc, log z, slumber, sopor, eternal rest, catch some z, kip, nap, quietu
Q17	excit, fret, angri, nettlesom, peevish, choler, petul, scratchi, furiou, fractiou, temper, testi, fussi, cross, irrit, techi, peckish, pettish, tempestu, cranki, bilious, tetchi, wild, rage, surli, snappish
Q18	food, wipe out, use up, eat on, appet, eat, solid food, nutrient, feed, food for thought, rust, appetit, eat up, consum, run through, deplet, corrode, intellectual nourish, exhaust
Q19	contract, centralis, concentr, immers, rivet, centr, decoct, compact, reduc, digest, focu, central, condens, center, pore, assidu, tight, densiti, dens, boil down, absorpt, engross, dressed or
Q20	banal, well-worn, wear upon, sap, timeworn, trite, wear, tire, thread-bar, bore, fag out, play out, stock, tire out, wear out, fatigue duti, shopworn, old-hat, fatigu, tired, weari, fag, exhaust, run down, outwear, commonplac, pall, wear down, jade, hackney

Q21	sex, arous, turn on, excit, gender, wind up, sexual act, sex act, sexual, sexual urg, sexual practic
Depression IAT	smile, i, mine, gloomi, glad, them, delight, myself, their, depress, cheer, me, other, joy, self, withdrawn, they, hopeless, helpless
All	Union of all above stimuli sets

Note that some stimuli gathered from the WordNet might not seem to have sensible relationships with depression (e.g., department of energy). These terms were kept as the usage of metaphors is possible among users on social media.

6.2.2 Classification Using Glass Box Models

6.2.2.1 Text Pre-Processing

For each user (subject) in the datasets, their writing texts, including posts and comments are extracted. For the extracted texts, all digits and punctuation are removed, and lower-cased. Then, stopwords in English are cleaned from the texts, and the remaining tokens are stemmed using the Porter Stemmer ([Porter, 1980](#)). Finally, all the pre-processed writings are concatenated into a single paragraph for each user.

6.2.2.2 Text Representation

In this step, the pre-processed textual data will be represented as vectors of numeric values. Two vectorization approaches are applied for glass box models:

- **Term Frequency–Inverse Document Frequency (TF-IDF)**

This method quantifies the importance of a term relative to other terms in the corpus.

The term frequency (TF) measures how often a term appears in a document:

$$TF(t, d) = \frac{\# \text{ of occurrences of } t \text{ in } d}{\# \text{ of terms in } d}$$

where t and d represent the term and the document separately.

The inverse document frequency (IDF) measures the importance of a term across documents in the corpus. It is calculated as follows:

$$IDF(t) = \log_e \frac{\# \text{ of documents in the corpus}}{\# \text{ of documents having } t}$$

where t represents the term.

Then, the TF-IDF is calculated as follows:

$$TF-IDF = TF * IDF$$

- **Bag of Words (BoW)**

This method considers the texts as a collection of words, without other information such as sentence structure or order of the words. The BoW converts the representing document as vectors with the number of unique words as its dimension and then assigns the number of occurrences for each unique word to its unique component of the representation vector.

6.2.2.3 Correlation Analysis

Based on an assumption that there could be relationships between symptoms, we explored the correlations between questions in the BDI questionnaire, in the training data. The calculated correlations are shown in Figure 6.2.

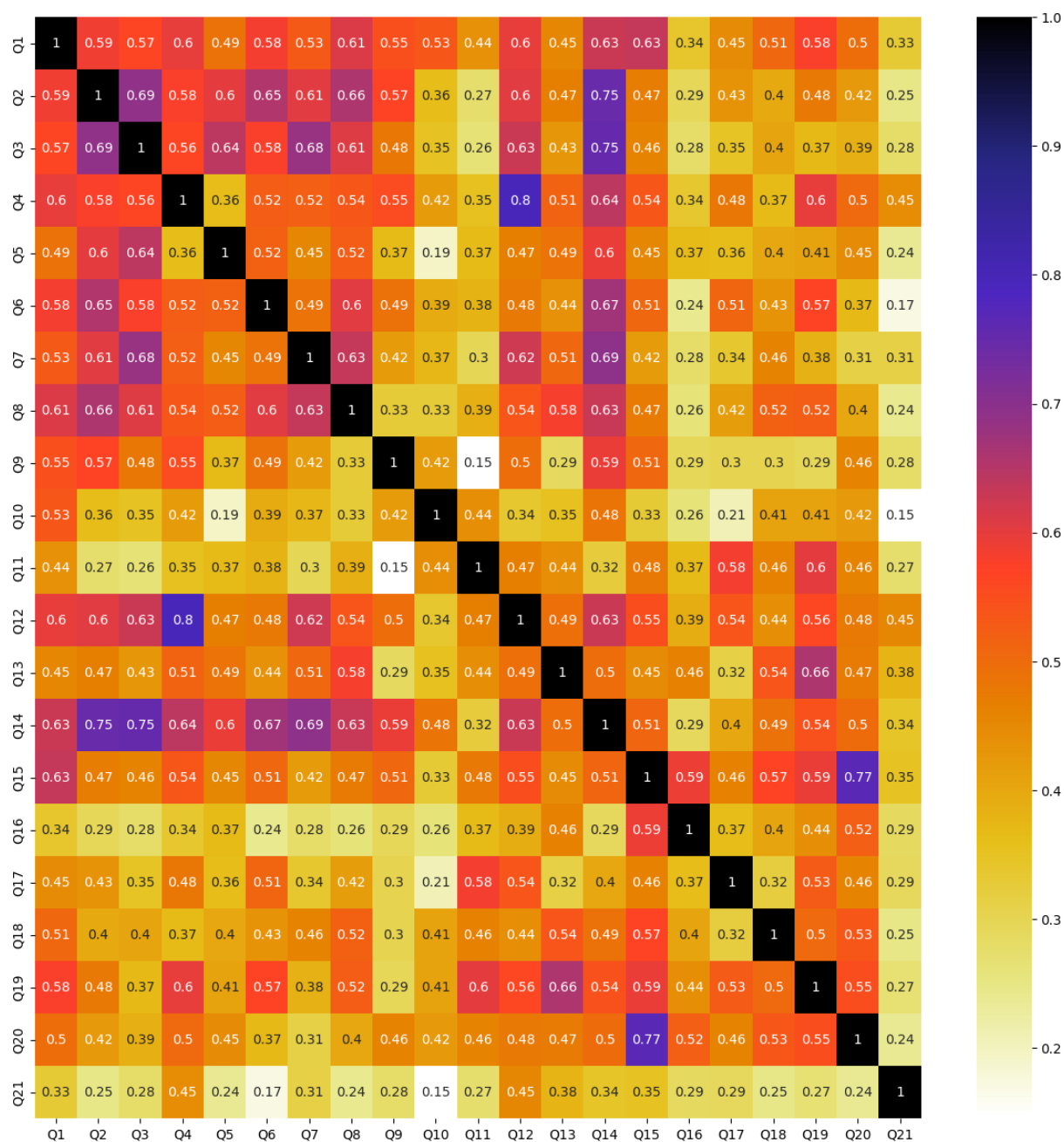


Figure 6.2: Correlation matrix of answers

For correlations between question answers, we picked out those that are strongly correlated (i.e., correlation greater than 0.6) and summarized in Table 6.4.

Table 6.4: Questions that have strong correlations

Q	Related Q	r.	Related Q	r.	Related Q	r.	Related Q	r.
Q1	Q15	0.61						
Q2	Q14	0.72	Q8	0.64	Q3	0.62		
Q3	Q14	0.70	Q2	0.62	Q7	0.61	Q12	0.60
Q4	Q12	0.78						
Q6	Q14	0.62	Q8	0.60				
Q7	Q8	0.62	Q3	0.61	Q14	0.61		
Q8	Q2	0.64	Q7	0.62	Q14	0.61	Q6	0.60
Q9	Q14	0.60						
Q12	Q4	0.78	Q3	0.60				
Q13	Q19	0.61						
Q14	Q2	0.72	Q3	0.70	Q6	0.62	Q7	0.61
Q15	Q20	0.76	Q1	0.61				
Q19	Q20	0.61	Q13	0.61				
Q20	Q15	0.76	Q19	0.61				

6.2.2.4 Balancing the Training Data

As shown in Section 4.3, the labels are imbalanced for many questions. Since classification algorithms tend to be skewed due to the existence of dominant classes, random over-sampling, which is a technique to over-sample the instances belonging to the minority classes (answers) by randomly selecting samples with replacement, was performed for the questions.

6.2.2.5 Glass Box Classifiers

Glass box models are machine learning models designed to be interpretable (Nori, Jenkins, Koch, & Caruana, 2019). In this proposed system, three glass box models are applied: Explainable Boosting Machine (EBM), Decision Tree (DT), and Logistic Regression (LR).

- The explainable boosting machine is an interpretable machine learning algorithm that was designed to have accuracy comparable to black box methods while maintaining its explainability. It is an extended Generalized Additive Model (GAM) of the form:

$$g(E[y]) = \beta_0 + f_1(x_1) + f_2(x_2) + \dots + f_j(x_j)$$

where g is the link function that adapts the GAM to different settings such as regression or classification, and f_j is the learnt feature function.

- The decision trees are a famous interpretable machine learning algorithm for classification. When making predictions, the built tree would split the input data several times at decision points, according to the values of the feature. As more splits are made, more subsets of data are created. At last, the leaf nodes in the tree would be the final subsets for the data. To predict on a leaf node, the model uses an average outcome learned from the training data.
- The logistic regression models apply a transformation on the output of a linear combination to values between 0 and 1. The logistic function is generally used for the

transformation. For classification tasks, 0 and 1 are used as probabilities for predicting labels.

Through experiments, the three glass box models are applied to different questions in the BDI questionnaire based on their training performance. Details about which models are applied to questions are summarised in Table 6.5.

6.2.2.6 Order of Questions

With the knowledge of the correlation between questions summarised in Table 6.4, the predictions of questions could then be used as features for correlated questions.

By analyzing the initial training results, we designed the order of questions that the models will answer, to better utilize the correlations. The questions having a higher accuracy score on the training data will be answered earlier, and then the predictions will be used as features for answering correlated questions with a lower accuracy. Stimuli of correlated questions are also considered.

Feature engineering, which is a technique of combining existing features to determine new sets of features in order to create an effective dataset for the model, was conducted. The details about the designed order of questions to be answered, applied vectorizers and other collected features of each question are shown in Table 6.5.

Table 6.5: Order and configurations of glass box models for each question

	Question	Vectorizer	Features besides vectorization	Model
1	Q1	TF-IDF	Q1 stimuli, depression stimuli	EBM
2	Q15	BoW	Q1 answer, Q15 stimuli, Q20 stimuli, depression stimuli	EBM
3	Q6	-	Q6 stimuli, Q14 stimuli, Q8 stimuli, Q2 stimuli, Q3 stimuli, depression stimuli	EBM
4	Q4	BoW	Q4 stimuli, Q12 stimuli, Q1 stimuli, depression stimuli	DT

5	Q14	BoW	Q6 answer, Q14 stimuli, Q2 stimuli, depression stimuli	DT
6	Q8	BoW	Q6 answer, Q8 stimuli, Q2 stimuli, Q14 stimuli, depression stimuli	EBM
7	Q7	BoW	Q8 answer, Q7 stimuli, Q12 stimuli, depression stimuli	EBM
8	Q12	BoW	Q4 answer, Q12 stimuli, Q4 stimuli, Q3 stimuli, depression stimuli	DT
9	Q2	BoW	Q14 answer, Q2 stimuli, Q14 stimuli, Q8 stimuli, depression stimuli	EBM
10	Q3	-	Q14 answer, Q3 stimuli, Q14 stimuli, Q2 stimuli, Q7 stimuli, Q12 stimuli, depression stimuli	LR
11	Q19	TF-IDF	Q19 stimuli, Q13 stimuli, Q15 stimuli, Q1 stimuli, depression stimuli	LR
12	Q13	TF-IDF	Q19 answer, Q13 stimuli, Q19 stimuli, Q8 stimuli	EBM
13	Q20	-	Q15 answer, Q20 stimuli, Q15 stimuli, Q19 stimuli, depression stimuli	LR
14	Q9	BoW	Q9 stimuli, Q14 stimuli, Q4 stimuli, depression stimuli	LR
15	Q5	BoW	Q5 stimuli, Q3 stimuli, Q8 stimuli, depression stimuli	EBM
16	Q10	BoW	Q10 stimuli, Q1 stimuli, Q14 stimuli, depression stimuli	LR
17	Q11	-	Q11 stimuli, Q17 stimuli, Q10 stimuli, Q18 stimuli, depression stimuli	LR
18	Q16	TF-IDF	Q16 stimuli, depression stimuli	EBM
19	Q17	-	Q17 stimuli, Q12 stimuli, Q1 stimuli, depression stimuli	LR
20	Q18	-	Q18 stimuli, Q8 stimuli, Q15 stimuli, Q20 stimuli, depression stimuli	DT
21	Q21	BoW	Q21 stimuli, Q4 stimuli, Q15 stimuli, Q12 stimuli	EBM

6.2.3 Classification Using Large Language Models

As introduced in Chapter 2, prompt learning is a new paradigm that is showing promising results. The large language models are trained to estimate the probability $P(x; \theta)$ for text x . Prompt learning techniques utilize the probability P , to predict the output y . As an example, the output y can be the label in a classification task, and it can be extracted or transformed from the text generated by the LLM.

6.2.3.1 Selection of the LLMs

A wide variety of pre-trained language models are available. In this study, the following open LLMs were applied in experiments:

- **Llama-2-13b-chat**

Meta’s Llama 2 models (Touvron et al., 2023) are LLMs that are well-supported and powerful. As auto-regressive language models, they are particularly useful for Natural Language Generation (NLG) tasks, which means that not only the output label for our classification task could be generated, but also the LLMs’ explanations for predictions.

Llama-2-13b-chat², which is optimized for dialogue use cases was applied in this study.

- **SUS-Chat-34B**

Released by the Southern University of Science and Technology (SUSTech) and IDEA-CCNL, SUS-Chat-34B³ is a bilingual (Chinese-English) dialogue model. It showed significant improvements on many benchmarks of evaluation; it achieved the highest scores among open source models of similar size (34 billion parameters), and it is one of the best models with a size below 70B.

²<https://huggingface.co/meta-llama/Llama-2-13b-chat-hf>

³<https://huggingface.co/SUSTech/SUS-Chat-34B>

The model which was based on Yi-34B⁴ was trained with 1.4 billion tokens of complex instruction data, including multi-turn dialogues, mathematics, reasoning, and others, thus the model is capable of focusing on long-text dialogue and imitating human thought processes.

- **Neural-chat-7b-v3**

Based on Mistral-7B-v0.1⁵, the Intel neural-chat-7b-v3-1⁶ is a LLM that was fine-tuned on the SlimOrca⁷ dataset and with the Direct Preference Optimization (DPO) algorithm to align with human preferences. The DPO derives the probability of human preference data for the optimal policy so that the reward model which the Reinforcement Learning from Human Feedback (RLHF) needs is replaced (Lv, Zhang, & Shen, 2023).

The model performed exceptionally well among 7-billion-parameter models.

6.2.3.2 Design of Prompts

In this step, a prompting function f_{prompt} is designed (P. Liu et al., 2021), which could then be applied to transform the input text x into the prompt $x' = f_{prompt}(x)$.

Prompt engineering is the process of designing an effective prompting function f_{prompt} . Through experiments, we noticed that if only user writings relevant to the symptom are given, the model tends to exaggerate the user's condition and give severe evaluations. With the knowledge we learned from Section 4.3 that most of the symptoms users have are minor, we added "suppose you are a psychiatrist who prefers to give minor diagnoses rather than serious ones" in the instruction, to alleviate the models' tendency of exaggeration.

Instead of giving the exact meaning of each question's answers, such as "0. I do not feel sad.", "1. I feel sad much of the time.", "2. I am sad all the time." and "3. I am so sad or unhappy that I can't stand it." for the symptom of sadness, we give only the

⁴<https://huggingface.co/01-ai/Yi-34B>

⁵<https://huggingface.co/mistralai/Mistral-7B-v0.1>

⁶<https://huggingface.co/Intel/neural-chat-7b-v3-1>

⁷<https://huggingface.co/datasets/Open-Orca/SlimOrca>

possible labels to the model. In this way, the labels 0, 1, 2, and 3 are considered levels of intensity or the magnitude of the change, rather than solely markers of classes. The first consideration is the length of the prompt: it is inefficient to reserve a diverse space for this information, and models may not have a large enough context window (e.g. Llama 2 supports up to 4,096 tokens). Another consideration is that some label meanings might confuse the model. Examples of options that could be confusing are shown in Figure 6.3.

9. Suicidal Thoughts or Wishes

- 0. I don't have any thoughts of killing myself.
- 1. I have thoughts of killing myself, but I would not carry them out.
- 2. I would like to kill myself.
- 3. I would kill myself if I had the chance.

10. Crying

- 0. I don't cry anymore than I used to.
 - 1. I cry more than I used to.
 - 2. I cry over every little thing.
 - 3. I feel like crying, but I can't.
-

Figure 6.3: Some options that might be confusing to models

To clarify the symptom that the model should work on, we rephrased the 21 questions in the BDI questionnaire into 21 sentences. All the rephrased sentences talk about levels of the symptom, to help the model's understanding. The rephrased symptoms are summarized in Table 6.6.

Table 6.6: Rephrased symptoms in the BDI questionnaire

Question	Rephrased symptom
Q1	how sad the user feels
Q2	how discouraged the user is about future
Q3	how much the user feels like a failure
Q4	how much the user loses pleasure from things
Q5	how often the user feels guilty
Q6	how much the user feels punished
Q7	how much the user feels disappointed about him/herself
Q8	how often the user criticizes or blames him/herself
Q9	how much the user thinks about killing him/herself
Q10	how often the user cries
Q11	how much the user feels restless or agitated
Q12	how much the user loses interest in things
Q13	how difficult the user to make decisions
Q14	how much the user feels worthless
Q15	how much the user loses energy
Q16	how much the user experienced changes in sleeping
Q17	how much the user feels irritable
Q18	how much the user experienced changes in appetite
Q19	how difficult the user to concentrate
Q20	how much the user feels tired or fatigued
Q21	how much the user loses interest in sex

Our designed prompt template is shown in Figure 6.4. During experiments, "[*rephrased symptom*]" will be filled by using rephrased symptoms in Table 6.6, and "[*input text*]" will be substituted with the concatenated user writings from the cleaned dataset: texts for each symptom when using the Top-5 sentences or all texts when using the Top-1 sentences.

Instruction: The following paragraph was concatenated from a user's posts on social media. Suppose you are a psychiatrist who prefers to give minor diagnoses rather than serious ones, read the posts as a whole, determine the level of "[*rephrased symptom*]" and give a number in 0, 1, 2 or 3, then explain why.

User's posts: [*input text*]

Level (0, 1, 2 or 3):

Figure 6.4: Designed prompt template for symptom estimation

6.2.3.3 Using LLMs

The same configurations are applied to all the experimented LLMs.

The tokenizers for each model were used to encode prompts and decode outputs. The maximum length for the models was set to 4,096, and only the first sentence of a pair is truncated if it is longer.

To reduce the costs of utilizing LLMs, QLoRA ([Dettmers, Pagnoni, Holtzman, & Zettlemoyer, 2023](#)) was used: models are run in 4-bit precision, using the NF4 (Normal Float 4) data type, double quantization, and the computational type of torch.bfloat16.

While generating texts, the models are set to use multinomial sampling, keeping the top 10 highest probability vocabulary tokens and the smallest set of most probable tokens with probabilities that add up to 0.7 or higher. The maximum length is 4,096, and the temperature (the value used to modulate probabilities of tokens) was set to be 0.1.

We experimented with a single LLM at a time, as well as combining two LLMs working on separate questions based on the results from experiments using the training data. The experiments were conducted using the Top-1 and Top-5 datasets.

The outputs generated by the LLMs contain labels for the predicted answers, as well as explanations for the predictions. We extracted the labels using regular expression (regex)

and recorded them as the automated responses to the BDI questionnaire for each user in the test set.

6.3 Evaluation

The same evaluation metrics were applied for the task in eRisk 2019, 2020, and 2021([Parapar et al., 2021](#)). The four metrics used for evaluation are:

- **Average Hit Rate (AHR)**

The AHR is the hit rate averaged over all the users. The hit rate measures the number of answers systems automatically fill in that are exactly the same as the actual answers provided by the users.

- **Average Closeness Rate (ACR)**

The ACR is the Closeness Rate averaged over all the users. It takes into account that the answers represent an ordinal scale, rather than merely separate options. To get the closeness rate, first compute the absolute difference between the automatically filled answer and the actual answer, then transform the calculated absolute difference into an effectiveness score as follows:

$$CR = \frac{maxad - ad}{maxad}$$

where *maxad* stands for the maximum absolute difference, and *ad* is the absolute difference.

- **Average Difference between Overall Depression Levels (ADODL)**

The ADODL is the difference between the system’s and actual overall depression levels averaged over all users. The Difference between Overall Depression Levels (DODL) is obtained by first calculating the overall depression levels for the system-filled and actual questionnaire, then computing the absolute difference between the

two overall scores.

The DODL is normalized as follows:

$$DODL = \frac{63 - ad}{63}$$

where ad is the absolute difference between the automated and actual overall score.

- **Depression Category Hit Rate (DCHR)**

The DCHR measures the closeness of the depression estimation achieved over all users according to the established depression categories introduced in Figure 4.3. It calculates the fraction of cases where the automated questionnaire led to a category that is identical to the user’s actual depression category.

6.4 Results and Discussion

6.4.1 Glass Box-Based Systems

While conducting experiments with systems based on glass box models EBM, DT and LR, through InterpretML⁸, four types of settings are experimented:

- Use the system with all cleaned texts, no stimuli features described in Table 6.5 are used. The run is named "glassbox_all_text" in the Table 6.7.
- Use the system with all cleaned texts and the stimuli features. The run is named "glassbox_feature+all_text" in the Table 6.7.
- Use the system with the stimuli features as well as texts from the Top-1 dataset by concatenating all the most relevant writings for each symptom into a single paragraph. This run is named "glassbox_feature+top1_text" in the Table 6.7.

⁸<https://interpret.ml/>

- Use the system with texts for each of the 21 symptoms from the Top-5 dataset, as well as the stimuli features. This run is named "glassbox_feature+top5_text" in the Table 6.7.

Table 6.7: Results of glass box-based systems

Run	AHR	ACR	ADODL	DCHR
glassbox_all_text	29.64	64.56	82.14	33.75
glassbox_feature+all_text	31.01	64.62	82.83	37.5
glassbox_feature+top1_text	30.23	64.64	83.69	40.0
glassbox_feature+top5_text	31.19	65.47	84.12	45.0

As we can see from the results, the application of USESim for selecting relevant user writings is helpful, especially for metrics ADODL and DCHR. The score on AHR dropped while using the Top-1 dataset compared with the run using all texts and features; this could be due to the lack of information since only one writing was left for each symptom. The fact that the run with the Top-5 dataset had the highest scores on all metrics proves this, since more information is kept and features in the texts are more likely to be relevant when used to classify the new data.

The stimuli features are shown to be valuable. The system runs using the stimuli features scored higher than the runs without them. By comparing the results of the two runs using all texts, it can be seen that AHR and DCHR are higher when stimuli features are used.

6.4.1.1 Model Explanation

Intrinsic explanations are available for the three experimented glass box models. The generated explanations focus on justifying models' predictions, and could also be helpful for debiasing, error detection, and debugging.

- **Interpretation of EBM**

The EBM could provide global explanations in terms of variable importance for

features, and also local explanations for instances with explanations on feature contribution.

An example of a global explanation for the EBM of Q15 (Loss of Energy) is shown in Figure 6.5. The mean absolute contribution score each feature (or interaction) makes to predictions averaged across the training dataset determines the feature importance. Contributions are weighted by the number of samples in each bin and the sample weights (if any). The figure shows the 15 most important features for Q15.

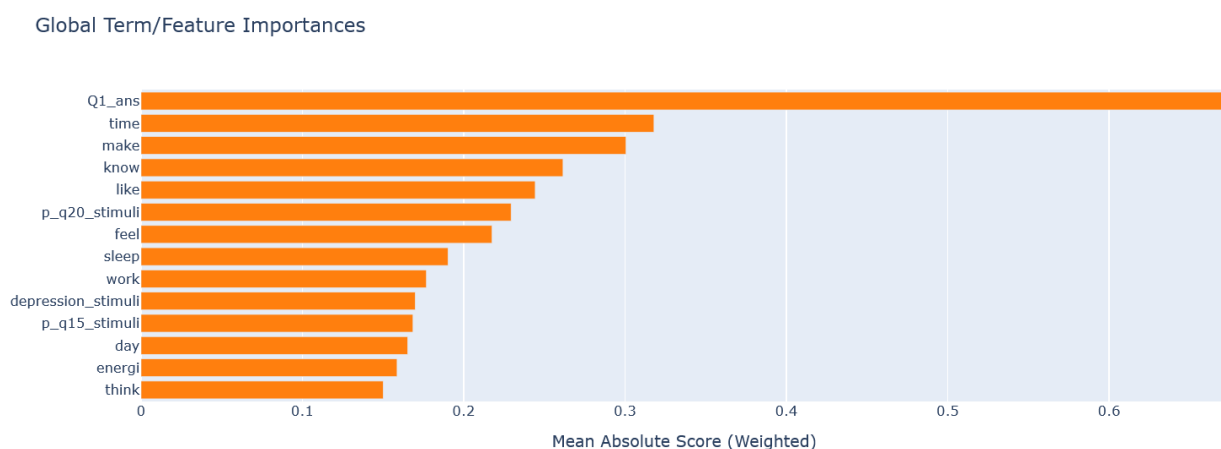


Figure 6.5: An example of global variable importance of EBM

In Figure 6.6, an example of a local explanation of a prediction is shown. A local explanation shows the breakdown of how much each feature contributed to the prediction for a single sample. The intercept reflects the average case. The figure shows the 15 most important features for this prediction on Q15. In this example, the possible labels are represented by 4 colors, and feature contributions (positive and negative) are shown as bars.

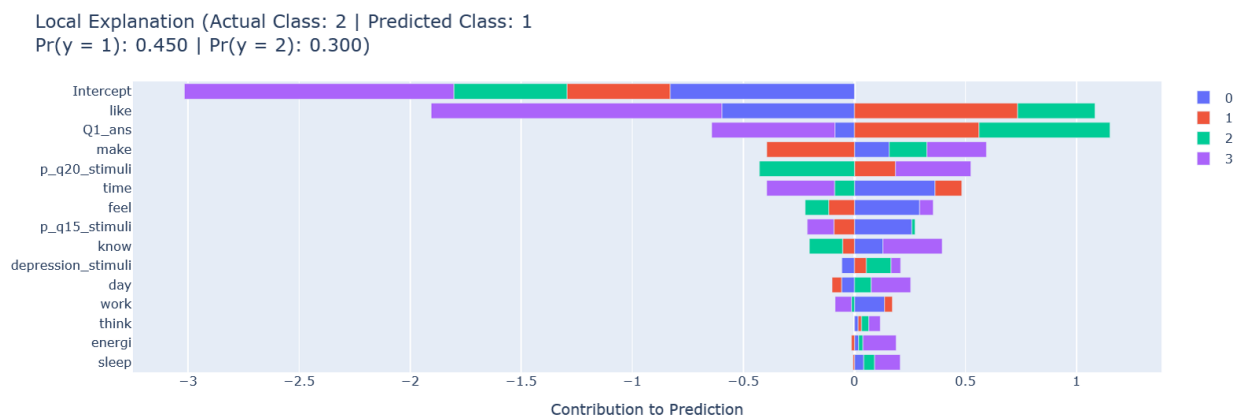


Figure 6.6: An example of local prediction explanation of EBM

• Interpretation of LR

To interpret a LR classifier, the weighted sum in logistic regression is transformed by the logistic function to a probability. The trained weights show the importance of features. If we change the value of the feature by some units, the estimated probability will change accordingly.

An example of a global explanation for the LR for Q10 (Crying) is shown in Figure 6.7.

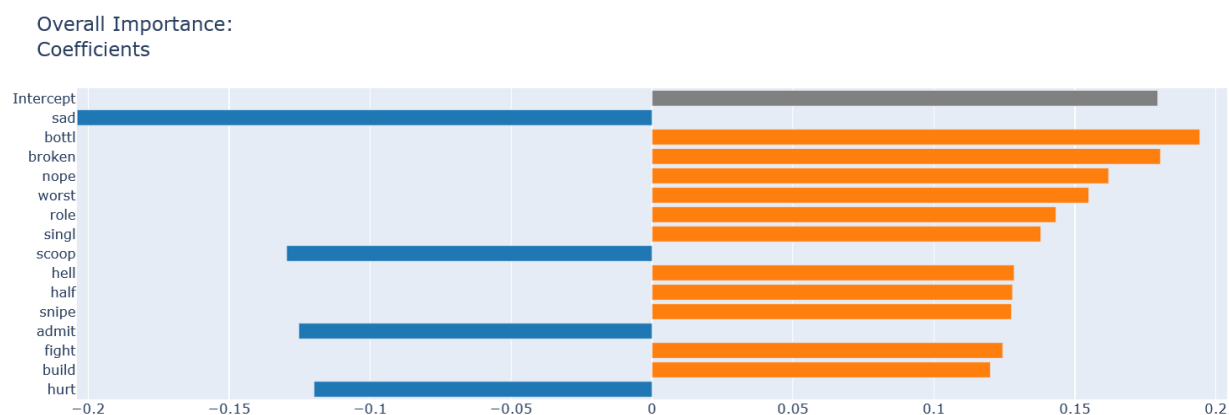


Figure 6.7: An example of an explanation of logistic regression

- **Interpretation of DT**

Each node on a DT has the splitting feature and threshold, as well as the population in each class. To interpret the built tree-based model, we start by comparing the input data's feature from the root node, and then continually move to the child node which is the subset the next feature's value is in. The prediction is given by the finally reached leaf node.

An example of a constructed DT for Q4 (Loss of Pleasure) and a prediction is shown in Figure 6.8.

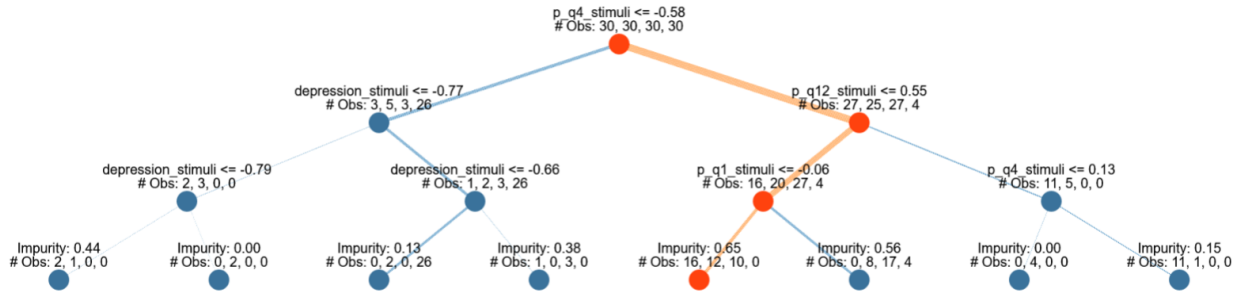


Figure 6.8: An example of using a decision tree for classification

6.4.2 LLM-Based Systems

Experimental results of the systems using LLMs are shown in Table 6.8.

Table 6.8: Results of LLM-based systems

Run	AHR	ACR	ADODL	DCHR
Llama-2-13b-chat_top1	21.90	63.29	72.22	42.5
Llama-2-13b-chat_top5	22.32	63.51	72.16	42.5
neural-chat-7b-v3-1_top1	31.96	71.82	84.12	48.75
neural-chat-7b-v3-1_top5	33.63	70.83	85.87	52.5
SUS-Chat_top1	32.61	72.02	84.64	50.0
SUS-Chat_top5	33.51	72.57	83.53	52.5
neural-chat+SUS-Chat_top1	34.70	72.91	85.41	48.75
neural-chat+SUS-Chat_top5	37.32	73.25	85.63	50.0

Similar to what we learned from glass box-based systems, it is generally better to have more writings kept so that the model could have more information about the user, and the writings would be more focused on the specific question. In our experiments, the usage of the Top-5 dataset leads to a better performance than using the Top-1 dataset.

When both using the Top-5 dataset, the model neural-chat-7b-v3-1 performed better than SUS-Chat-34B on metrics AHR, ADODL, and DCHR. This is surprising since the neural-chat-7b-v3-1 is much smaller than SUS-Chat-34B in terms of size/number of parameters. The reason could be the language focus and the application of the DPO algorithm.

Through experimenting on the training data, the neural-chat-7b-v3-1’s answers on questions 4, 8, 9, 11, 12, 16, 18, 19, 20 and 21, are combined with SUS-Chat-34B’s answers on questions 1, 2, 3, 5, 6, 7, 10, 13, 14, 15 and 17 as the combined system. The combined runs performed well on all metrics and achieved the highest scores on two of the metrics.

Although the performance on some metrics is still not outstanding, our systems scored over 85% on ADODL, which is a good improvement considering ADODL is the most critical metric for measuring depression at the population level (Skaik & Inkpen, 2022). Many runs scored over 50 on DCHR which means they were correct for more than half of the test subjects when predicting their depression category.

Our experiments proved that LLMs have learned knowledge about various depression-related symptoms, and they can make better inferences than supervised deep learning

techniques, in the convenient zero-shot learning setting (no training) if we use properly-designed prompts.

6.4.2.1 Local Explanations of LLMs

Through prompts, the LLMs were asked to provide explanations for their predictions. Even though these explanations are not necessarily factual, they provide insights about the important information in the given user writings. In Figure 6.9, an example of user writings is given, which is answered by LLMs for Q18 as a change in appetite.

Figure 6.10 shows the prediction and explanation from Llama-2-13b-chat for the text in Figure 6.9. In the explanation, the model mentioned several physical and mental issues described in the user’s writings, such as inflammation and mental health issues. The model predicted 2 as the answer.

In Figure 6.11, the classification and explanation for the given example generated by neural-chat-7b-v3-1 are shown. The model mentioned that the user needed to set alarms to eat, having stomach flu and fluctuating weights, which could affect the user’s appetite. An answer label of 2 is given by the model.

SUS-Chat-34B’s prediction and explanation are shown in Figure 6.12. The model presumed that the user had a higher level of change in appetite since the user had to set alarms to eat at some points, but the model also mentioned that no significant weight changes were presented. The model generated 2 as the answer to the question.

User's posts: I would say coriander. Grab a package and eat it. **Great snacks.** I mean the psychosis is still untreated I'm thinking about what to do next. **My weight fluctuates** because, in addition to severe inflammation, I also have some mental health issues that cause loss of memory. I actually **have to set alerts to eat** for certain times. I had a huge success that day and was approved to work from home. Now I **got stomach flu**. I don't feel accomplished enough. I'm **vomiting for days** and my mental state is getting worse and I hear constant noise from the church. Oh my goodness, my back pain, my **constant digestive issues**, plus my mental illness, have all gotten worse. I think all I really need is a happy message. Usually good news upsets me, but I really need it today.

Figure 6.9: An example of concatenated writings. Some words were highlighted for readability. In fact, we rephrased the sentences in the example, because the data sharing agreement that we signed when we obtained the dataset does not allow us to use the real example directly.

Level (0, 1, 2 or 3): 2

Explanation:

The user's posts suggest that they have experienced significant changes in appetite, as they mention "weight fluctuations" and setting alerts to eat. They also mention "severe inflammation" and "mental health issues" which could be contributing to their appetite changes. Additionally, they mention "stomach flu" and "constant digestive issues" which could also be affecting their appetite. Based on this information, I would rate the user's experience of changes in appetite as a level 2.

Figure 6.10: Classification and explanation by Llama-2-13b-chat

Level (0, 1, 2 or 3): 2

The user experienced changes in appetite at a level of 2. This is because they mentioned having mental health issues that cause loss of memory, which can affect appetite. They also mentioned setting alerts to eat for certain times, indicating that they have to be more mindful of their eating habits. Additionally, they mentioned having stomach flu, which can also affect appetite. The user's fluctuating weight and other health issues, such as back pain, digestive issues, and mental illness, further contribute to the complexity of their situation.

Figure 6.11: Classification and explanation by neural-chat-7b-v3-1

Level (0, 1, 2 or 3): 2

Explanation: The user mentions having to set alerts to eat for certain times, which indicates that they have experienced changes in appetite. However, they do not mention any significant weight loss or gain, which would indicate a higher level of change in appetite. Therefore, a level of 2 is appropriate.

Figure 6.12: Classification and explanation by SUS-Chat-34B

6.4.3 Performance on Individual Questions

The question-wise performance of neural-chat+SUS-Chat_top5, which is the system that achieved the highest scores on AHR and ACR are summarized in Table 6.9. The 2 remaining metrics (ADODL and DCHR) are not discussed here since they focus more on overall depression levels.

It could be seen that the system performed better on some questions, such as questions 4 (Loss of Pleasure), 18 (Changes in Appetite), and 19 (Concentration Difficulty). However, for questions like questions 6 (Punishment Feelings), 17 (Irritability), and 21 (Loss of Interest in Sex), the performance is poor (but still higher than 25%). One possible reason for this difference could be different amounts of information, for example, some topics

might be less discussed on social media compared to others. Performance on individual questions could be further investigated to have more focused fine-tuning or prompting.

Table 6.9: Question-wise performance of neural-chat+SUS-Chat_top5

Question	AHR	ACR
Q1	35.0	71.25
Q2	38.75	75.41
Q3	43.75	74.99
Q4	48.75	81.24
Q5	41.25	75.83
Q6	27.5	66.66
Q7	35.0	72.08
Q8	37.5	72.49
Q9	43.75	77.91
Q10	30.0	65.41
Q11	38.75	74.16
Q12	40.0	75.83
Q13	30.0	68.33
Q14	32.5	71.24
Q15	41.25	77.49
Q16	37.5	74.58
Q17	27.5	67.49
Q18	48.75	77.91
Q19	48.75	79.58
Q20	31.25	72.49
Q21	26.25	65.83

6.4.4 Evaluation of Explanations

One important yet difficult issue with the generated (instead of human-written) explanations is to quantitatively evaluate them. In (Rajagopal, Balachandran, Hovy, & Tsvetkov, 2021), several criteria were introduced for evaluating the explanations, including sufficiency (via BERT-score), plausibility, and trustability. Due to limitations on the amount

of time available for conducting evaluations, we only evaluated explanations for the best-performing system on one metric: sufficiency – to automatically evaluate how well the explanations reflect the system’s predictions. Manual evaluation with experts to calculate the other measures is left as future work.

Rajagopal et al. (2021) used the "Faithfulness-by-construction" (FRESH) pipeline (Jain, Wiegrefe, Pinter, & Wallace, 2020) to evaluate sufficiency: a BERT-based classifier is trained to perform a task using only the extracted explanations. A high accuracy would indicate a high sufficiency of the explanations, which means that the explanations are indicative of the predicted labels. Based on this method, we implemented a BERT classifier, trained it with the generated explanations for the training dataset as the training data, and evaluated the classifier with the explanations for the test data.

The sufficiency score for the explanations generated by our best system was 80.77%. This indicates the effectiveness of generated explanations of the system. The configurations are as follows: a fast Rust-based pretrained tokenizer of "bert-base-cased", with default settings on the rest; the pretrained "bert-base-cased" model for the classification into the 4 classes/labels; and with default values for the hyperparameters except for the total number of training epochs performed that was set to 8.

6.4.5 Comparison to Related Work

In this section, we compare our results for estimating depression levels with the state-of-the-art systems (not in terms of explainability because they did not address this aspect). Skaik and Inkpen (2022) proposed a method for this task that first applies topic-based filtering to clean the irrelevant posts, then conducts multiple iterations to select and assign different models to different questions. The final model named "BDI_Multi_Model" was formed using 5 HANs, 3 Long Short-Term Memory (LSTM) models and 2 transformers. Their proposed system set new records on several metrics. The comparison between our best systems and the state-of-the-art is concluded in Table 6.10.

Through comparisons, we can see that our systems achieved the best scores on three out of four metrics: ACR, ADODL, and DCHR. It is worth noting that the run neural-

chat+SUS-Chat_top5 achieved 85.63 on ADODL and 50.0 on DCHR, which indicates that the system neural-chat+SUS-Chat_top5 performed better on 3 metrics (that were achieved by different systems in previous work).

Our best results are achieved by LLM-based systems, not the glass box-based systems. However, in addition to answering questions without further training, our LLM-based systems could generate local explanations for all their answers, and these explanations are more readable and straightforward compared with the features-based model explanations. More importantly, no examples with labels are needed while using these LLMs, which means that the methods would fit the tasks that lack labeled data, and could be adapted to other tasks and domains.

We believe that with the availability of more computing power, if more user writings (not just selected ones) can be included in the prompts and a larger model can be used, higher scores could be achieved using our method.

Table 6.10: Our results compared to the state-of-the-art

Metric	Our system	Our best	Best system	State-of-the-art
AHR	neural-chat+SUS-Chat_top5	37.32	BDI_Multi_Model	41.25
ACR	neural-chat+SUS-Chat_top5	73.25	UPV-Symanto 4_symanto_upv_lingfeat_cors	73.17
ADODL	neural-chat-7b-v3-1_top5	85.87	BDI_Multi_Model	83.79
DCHR	neural-chat-7b-v3-1_top5	52.5	BDI_Multi_Model	48.75

6.5 Summary

In this chapter, we presented in detail the two types of explainable systems we developed for estimating levels of depression for social media users. The LLM-based systems out-

performed the glass box-based systems. We evaluated the systems' performance on the task of estimating the level of depression, scores achieved on each question as well as the explanation sufficiency of the best system. Finally, a comparison between our proposed systems and the state-of-the-art is presented.

Chapter 7

Conclusion and Future Work

This chapter presents an overall summary of the work, and reviews the research questions of the study. In addition, we discussed limitations and the possible future improvements that could be made to enhance the application.

7.1 Conclusion

With the further popularity of the Internet, more and more people are used to expressing themselves on social media platforms, thus it has become practical to detect people's depression beforehand by analyzing social media data. The BDI questionnaire is a very useful tool for this purpose. By automatically answering the questions on the BDI questionnaire, the depression status of social media users can be intuitively assessed. In the study, we designed a system that searches for relevant sentences in numerous user writings, and applied it to provide cleaner data for depression estimation systems based on glass box models or LLMs that automatically answer the questions from the BDI questionnaire. As the recommended method, the LLM-based systems have a better performance on several metrics and could sufficiently explain their answers to every question on the questionnaire for every user, without training on labeled data.

7.2 Research Questions Revisited

In Section 1.2, we presented several research questions to guide our study. As a part of the conclusion, each question is discussed below:

Research Question 1: How could smaller and cleaner datasets of user writings be formed from huge social media text collections?

In this study, we considered the task of cleaning/reducing the datasets as an IR task of selecting relevant user writings to form smaller and cleaner datasets. Through multiple experiments, we developed a document-selection system that uses keywords to filter out irrelevant posts and the Universal Sentence Encoder with Cosine Similarity (USESim) to search and rank writings that are relevant to symptoms in the BDI questionnaire.

Using this method, cleaner datasets with controllable size (number of top relevant writings) could be generated, with beneficial effects on the downstream tasks of estimating the level of depression.

Research Question 2: How to apply domain-knowledge-based features and interpretable machine learning algorithms for automatic depression detection?

In our glass box-based systems for estimating depression levels, the domain-knowledge-based features were generated by counting occurrences of depression stimuli from 23 stimuli sets, in users’ texts. These numerical features can then be used as inputs to the interpretable ML models such as EBM, LR, and DT.

Research Question 3: How could large language models be used for classifying depression-related symptoms, even without training?

The LLMs have gained knowledge through training, but the challenge is to present the tasks to the model so that they can understand the tasks well and use their knowledge to solve them.

The first technique we proposed in the study is to rephrase the questions in the questionnaire into sentences asking about levels of intensity of changes, in this way, the labels of questions (0, 1, 2, and 3) are transformed from solely markers of labels into meaningful and scalable levels. In addition, a role that LLMs should play could be established

through prompts. Most importantly, the expectation of responses from the models should be explicitly stated in the prompts to make sure the models generate texts as desired.

Research Question 4: How could large language models provide local explanations for their predictions?

The chat models, which are models trained or optimized for dialogue use cases could work well in generating explanations. In this study, the LLMs were clearly asked through prompts that they should first give answers, and then explain why. The models could understand this requirement and provide local explanations for each of their predictions.

Research Question 5: What is the quality of the generated local explanations?

Many metrics can be used to evaluate explanations. In this study, we evaluated the sufficiency of the generated local explanations and achieved a score of 80.77%, which indicated the effectiveness of the generated explanations of our best-performing system. In the future, human-involved evaluations could be conducted to further estimate the quality of explanations.

7.3 Limitations

There are limitations of this study that should be addressed or considered. Some of the limitations are:

- **Limited Datasets**

The datasets used in this study are collected from Reddit, which could lead to sampling biases. Also, as discussed in Section 4.3, the dataset for estimating depression levels contains only 90 users and 80 users for training and testing. The performance of the proposed systems applied to data gathered from other social media platforms is not guaranteed.

- **English Focus**

The proposed systems on user writing selection would result in datasets mostly in English, thus the systems are limited to English-written texts. More investigation will be needed in multilingual settings.

- **Preliminary Evaluation of Explanations**

We applied the automatic evaluation to our best-performing system, on only one metric (sufficiency). More metrics should be considered if resources are available.

- **Lack of Validation in Clinical Settings**

All evaluations are conducted without human health practitioners. It is better to have mental health practitioners review system predictions and explanations. Careful testing of the system in medical settings will also be needed.

7.4 Future Work

Despite active progress of the application of NLP techniques, due to the randomness of cyberspace, users' speeches on social media platforms cannot fully, objectively, accurately, and consistently describe their status of various depression-related symptoms. Based on the limitations of the work, some of the future work could be:

- **High-Quality Datasets**

It would be good for the community to produce and share larger high-quality datasets so that we can run more experiments to calibrate our systems and verify their effectiveness.

- **Multilanguage Support**

Pretrained models supporting more languages, and machine translation techniques could be applied and experimented with, so that systems could support languages other than English.

- **Further Prompt Engineering**

The prompts for LLMs are manually designed. Other prompts, that could achieve better results, likely exist. With more resources and properly designed experiments, better prompts could be found. For example, more roles could be played by the LLMs, or more symptom-specific details could be included through instructions.

- **Application of Domain Knowledge**

Few domain-specific information was used in the proposed LLM-based systems. A better performance might be achieved if domain knowledge could be automatically applied in addition to the user’s input texts.

- **Clinical Evaluations**

We evaluated the systems’ performance based on 80 users’ answers to the BDI questionnaire, and automatically evaluated the explanations generated by the best LLM-based system on sufficiency. Having healthcare practitioners qualitatively evaluate the systems’ predictions and explanations would be beneficial.

- **Better Usage of LLMs**

We set many restrictions on context length, sampling, and model size, due to the high requirements of computing resources. These restrictions could affect the performance but can be removed if more resources are available.

- **Adaptation to Other Questionnaires**

The questions (symptoms) in the BDI questionnaire were rephrased into single sentences, to ease LLMs’ understanding. When applying the method to other types of mental health questionnaires, more works need to be done on transforming their questions.

7.5 Ethics Statement

This study complies with all the ethical guidelines and principles of the University of Ottawa and the Ottawa-Carleton Institute of Computer Science, as well as the Association for Computational Linguistics (ACL) Ethics Policy¹.

The ethical aspects were always under consideration during this study. Since the datasets are collected from Reddit and are anonymized, privacy is respected and no bias is introduced. Grammarly was used when writing this thesis, but no AI assistance was involved in writing or programming.

The datasets and models in the study were only used for research purposes. The proposed systems should be used within research contexts, the automatically filled BDI questionnaires are meant to be for initial information and to be used as references by professionals, not for self-diagnosis.

To guide and evaluate the ethical implications of this study, we have voluntarily used the ACL Responsible NLP Checklist². The answered checklist can be found in Appendix B.

¹<https://www.aclweb.org/portal/content/acl-code-ethics>

²<https://aclrollingreview.org/responsibleNLPresearch/>

References

- Al-Mosaiwi, M., & Johnstone, T. (2018, July). In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation. *Clinical Psychological Science*, 6(4), 529–542. Retrieved 2023-10-25, from <https://doi.org/10.1177/2167702617747074> (Publisher: SAGE Publications Inc) doi: doi: 10.1177/2167702617747074
- Barrón-Cedeño, A., Martino, G. D. S., Esposti, M. D., Sebastiani, F., Macdonald, C., Pasi, G., ... Ferro, N. (2022). *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings*. Springer Nature. (Google-Books-ID: LzaFEAAQBAJ)
- Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961, June). An Inventory for Measuring Depression. *Archives of General Psychiatry*, 4(6), 561–571. Retrieved 2023-11-01, from <https://doi.org/10.1001/archpsyc.1961.01710120031004> doi: doi: 10.1001/archpsyc.1961.01710120031004
- Bisong, E. (2019). Google Colaboratory. In E. Bisong (Ed.), *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners* (pp. 59–64). Berkeley, CA: Apress. Retrieved 2023-11-21, from https://doi.org/10.1007/978-1-4842-4470-8_7 doi: doi: 10.1007/978-1-4842-4470-8_7
- Boals, A., & Klein, K. (2005, September). Word Use in Emotional Narratives about Failed Romantic Relationships and Subsequent Mental Health. *Journal of Language and Social Psychology*, 24(3), 252–268. Retrieved 2023-10-25, from <https://doi.org/10.1177/0261927X05278386> (Publisher: SAGE Publications Inc) doi: doi: 10.1177/0261927X05278386
- Bucur, A.-M. (2023, July). *Utilizing ChatGPT Generated Data to Retrieve Depression Symptoms from Social Media*. arXiv. Retrieved 2024-03-29, from <http://arxiv.org/abs/2307.02313> (arXiv:2307.02313 [cs]) doi: doi: 10.48550/arXiv.2307.02313
- Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., ... Kurzweil, R. (2018,

- April). *Universal Sentence Encoder*. arXiv. Retrieved 2023-05-26, from <http://arxiv.org/abs/1803.11175> (arXiv:1803.11175 [cs]) doi: doi: 10.48550/arXiv.1803.11175
- Chen, X., Sykora, M. D., Jackson, T. W., & Elayan, S. (2018, April). What about Mood Swings: Identifying Depression on Twitter with Temporal Measures of Emotions. In *Companion Proceedings of the The Web Conference 2018* (pp. 1653–1660). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. Retrieved 2023-10-24, from <https://dl.acm.org/doi/10.1145/3184558.3191624> doi: doi: 10.1145/3184558.3191624
- Chung, K. L. (1960). *Markov Chains with Stationary Transition Probabilities*. Berlin, Heidelberg: Springer. Retrieved 2023-10-20, from <http://link.springer.com/10.1007/978-3-642-49686-8> doi: doi: 10.1007/978-3-642-49686-8
- Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015, June). CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 31–39). Denver, Colorado: Association for Computational Linguistics. Retrieved 2023-10-25, from <https://aclanthology.org/W15-1204> doi: doi: 10.3115/v1/W15-1204
- Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020, October). *A Survey of the State of Explainable AI for Natural Language Processing*. arXiv. Retrieved 2023-10-27, from <http://arxiv.org/abs/2010.00711> (arXiv:2010.00711 [cs]) doi: doi: 10.48550/arXiv.2010.00711
- De Choudhury, M., Counts, S., & Horvitz, E. (2013, April). Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3267–3276). New York, NY, USA: Association for Computing Machinery. Retrieved 2023-10-24, from <https://doi.org/10.1145/2470654.2466447> doi: doi: 10.1145/2470654.2466447
- Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023, May). *QLoRA: Efficient Finetuning of Quantized LLMs*. arXiv. Retrieved 2023-09-26, from <http://arxiv.org/abs/2305.14314> (arXiv:2305.14314 [cs]) doi: doi: 10.48550/arXiv.2305

.14314

- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv. Retrieved 2023-10-24, from <http://arxiv.org/abs/1810.04805> (arXiv:1810.04805 [cs]) doi: doi: 10.48550/arXiv.1810.04805
- Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020, July). Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, 8, 439–453. Retrieved 2023-12-18, from https://doi.org/10.1162/tacl_a_00325 doi: doi: 10.1162/tacl_a_00325
- Gallaway, T. O., & Starkey, J. (2013, January). *Google Drive* [Text]. Retrieved 2023-11-21, from <https://annurev.publisher.ingentaconnect.com/content/annurev/tca/2013/00000014/00000003/art00008> (Publisher: Annual Reviews) doi: doi: 10.5260/chara.14.3.16
- Husseini Orabi, A., Buddhitha, P., Husseini Orabi, M., & Inkpen, D. (2018, June). Deep Learning for Depression Detection of Twitter Users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (pp. 88–97). New Orleans, LA: Association for Computational Linguistics. Retrieved 2023-10-25, from <https://aclanthology.org/W18-0609> doi: doi: 10.18653/v1/W18-0609
- Inkpen, D., Skaik, R., Buddhitha, P., Angelov, D., & Fredenburgh, M. T. (2021). uOttawa at eRisk 2021: Automatic Filling of the Beck’s Depression Inventory Questionnaire using Deep Learning.
- Jain, S., Wiegrefe, S., Pinter, Y., & Wallace, B. C. (2020, April). *Learning to Faithfully Rationalize by Construction*. arXiv. Retrieved 2024-01-14, from <http://arxiv.org/abs/2005.00115> (arXiv:2005.00115 [cs]) doi: doi: 10.48550/arXiv.2005.00115
- Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2021, October). *Mental-BERT: Publicly Available Pretrained Language Models for Mental Healthcare*. arXiv. Retrieved 2024-03-31, from <http://arxiv.org/abs/2110.15621> (arXiv:2110.15621 [cs]) doi: doi: 10.48550/arXiv.2110.15621
- Kim, Y. (2014, September). *Convolutional Neural Networks for Sentence Classifi-*

- cation. arXiv. Retrieved 2023-10-24, from <http://arxiv.org/abs/1408.5882> (arXiv:1408.5882 [cs]) doi: doi: 10.48550/arXiv.1408.5882
- Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019, January). Faithful and Customizable Explanations of Black Box Models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 131–138). New York, NY, USA: Association for Computing Machinery. Retrieved 2023-10-26, from <https://dl.acm.org/doi/10.1145/3306618.3314229> doi: doi: 10.1145/3306618.3314229
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021, July). *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. arXiv. Retrieved 2023-06-26, from <http://arxiv.org/abs/2107.13586> (arXiv:2107.13586 [cs] version: 1) doi: doi: 10.48550/arXiv.2107.13586
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019, July). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv. Retrieved 2023-05-26, from <http://arxiv.org/abs/1907.11692> (arXiv:1907.11692 [cs]) doi: doi: 10.48550/arXiv.1907.11692
- Losada, D. E., & Crestani, F. (2016). A Test Collection for Research on Depression and Language Use. In N. Fuhr et al. (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (pp. 28–39). Cham: Springer International Publishing. doi: doi: 10.1007/978-3-319-44564-9_3
- Losada, D. E., Crestani, F., & Parapar, J. (2017). eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Foundations. In G. J. Jones et al. (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (pp. 346–360). Cham: Springer International Publishing. doi: doi: 10.1007/978-3-319-65813-1_30
- Losada, D. E., Crestani, F., & Parapar, J. (2018). Overview of eRisk 2018: Early Risk Prediction on the Internet (extended lab overview).
- Losada, D. E., Crestani, F., & Parapar, J. (2019). Overview of eRisk 2019 Early Risk Prediction on the Internet. In F. Crestani et al. (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (pp. 340–357). Cham: Springer International Publishing. doi: doi: 10.1007/978-3-030-28577-7_27
- Losada, D. E., Crestani, F., & Parapar, J. (2020). Overview of eRisk at CLEF 2020: Early

Risk Prediction on the Internet (Extended Overview).

- Ly, K., Zhang, W., & Shen, H. (2023, November). *Supervised Fine-Tuning and Direct Preference Optimization on Intel Gaudi2*. Retrieved 2023-12-10, from <https://medium.com/intel-analytics-software/the-practice-of-supervised-finetuning-and-direct-preference-optimization-on-habana-gaudi2-a1197d8a3cd3>
- Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., & Cha, M. (2016, July). Detecting rumors from microblogs with recurrent neural networks. *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)*, 3818–3824. Retrieved from https://ink.library.smu.edu.sg/sis_research/4630
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Retrieved 2024-01-23, from <https://www.cambridge.org/highereducation/books/introduction-to-information-retrieval/669D108D20F556C5C30957D63B5AB65C> (ISBN: 9780511809071 Publisher: Cambridge University Press) doi: doi: 10.1017/CBO9780511809071
- Marwaha, S., Palmer, E., Suppes, T., Cons, E., Young, A. H., & Upthegrove, R. (2023, January). Novel and emerging treatments for major depression. *The Lancet*, 401(10371), 141–153. Retrieved 2023-11-26, from [https://www.thelancet.com/article/S0140-6736\(22\)02080-3/fulltext](https://www.thelancet.com/article/S0140-6736(22)02080-3/fulltext) (Publisher: Elsevier) doi: doi: 10.1016/S0140-6736(22)02080-3
- Maupomé, D., Armstrong, M. D., Rancourt, F., & Soulas, T. (2021). Early Detection of Signs of Pathological Gambling, Self-Harm and Depression through Topic Extraction and Neural Networks.
- Meites, T. M., Deveney, C. M., Steele, K. T., Holmes, A. J., & Pizzagalli, D. A. (2008, September). Implicit depression and hopelessness in remitted depressed individuals. *Behaviour Research and Therapy*, 46(9), 1078–1084. Retrieved 2023-11-28, from <https://www.sciencedirect.com/science/article/pii/S0005796708001332> doi: doi: 10.1016/j.brat.2008.05.008
- Miller, G. A. (1995, November). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39–41. Retrieved 2023-11-28, from <https://dl.acm.org/doi/10.1145/219717.219748> doi: doi: 10.1145/219717.219748

- Mowery, D. L., Park, A., Bryan, C., & Conway, M. (2016, December). Towards Automatically Classifying Depressive Symptoms from Twitter Data for Population Health. In *Proceedings of the Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media (PEOPLES)* (pp. 182–191). Osaka, Japan: The COLING 2016 Organizing Committee. Retrieved 2023-10-25, from <https://aclanthology.org/W16-4320>
- Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019, September). *InterpretML: A Unified Framework for Machine Learning Interpretability*. arXiv. Retrieved 2023-11-26, from <http://arxiv.org/abs/1909.09223> (arXiv:1909.09223 [cs, stat])
- Oram, P. (2001, March). WordNet: An electronic lexical database. Christiane Fellbaum (Ed.). Cambridge, MA: MIT Press, 1998. Pp. 423. *Applied Psycholinguistics*, 22(1), 131–134. Retrieved 2023-11-28, from <https://www.cambridge.org/core/journals/applied-psycholinguistics/article/abs/wordnet-an-electronic-lexical-database-christiane-fellbaum-ed-cambridge-ma-mit-press-1998-pp-423/8A9F540FB453B327C1AF0AC74E2F7D4D#> (Publisher: Cambridge University Press) doi: doi: 10.1017/S0142716401221079
- Parapar, J., Martín-Rodilla, P., Losada, D. E., & Crestani, F. (2021, September). Overview of eRisk 2021: Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings* (pp. 324–344). Berlin, Heidelberg: Springer-Verlag. Retrieved 2023-06-23, from https://doi.org/10.1007/978-3-030-85251-1_22 doi: doi: 10.1007/978-3-030-85251-1_22
- Parapar, J., Martín-Rodilla, P., Losada, D. E., & Crestani, F. (2023). Overview of eRisk 2023: Early Risk Prediction on the Internet. In A. Arampatzis et al. (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (pp. 294–315). Cham: Springer Nature Switzerland. doi: doi: 10.1007/978-3-031-42448-9_22
- Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association

- for Computational Linguistics. Retrieved 2023-05-26, from <https://aclanthology.org/D14-1162> doi: doi: 10.3115/v1/D14-1162
- Porter, M. (1980, January). An algorithm for suffix stripping. *Program*, 14(3), 130–137. Retrieved 2023-12-04, from <https://doi.org/10.1108/eb046814> (Publisher: MCB UP Ltd) doi: doi: 10.1108/eb046814
- Radden, J. (2023). Mental Disorder (Illness). In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2023 ed.). Metaphysics Research Lab, Stanford University. Retrieved 2023-12-15, from <https://plato.stanford.edu/archives/fall2023/entries/mental-disorder/>
- Rajagopal, D., Balachandran, V., Hovy, E., & Tsvetkov, Y. (2021, September). *SelfExplain: A Self-Explaining Architecture for Neural Text Classifiers*. arXiv. Retrieved 2024-01-09, from <http://arxiv.org/abs/2103.12279> (arXiv:2103.12279 [cs]) doi: doi: 10.48550/arXiv.2103.12279
- Recharla, N., Bolimera, P., Gupta, Y., & Madasamy, A. K. (2023). Exploring Depression Symptoms through Similarity Methods in Social Media Posts.
- Reimers, N., & Gurevych, I. (2019, August). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv. Retrieved 2023-10-25, from <http://arxiv.org/abs/1908.10084> (arXiv:1908.10084 [cs]) doi: doi: 10.48550/arXiv.1908.10084
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). New York, NY, USA: Association for Computing Machinery. Retrieved 2023-10-27, from <https://dl.acm.org/doi/10.1145/2939672.2939778> doi: doi: 10.1145/2939672.2939778
- Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson. Retrieved 2023-10-16, from <http://aima.cs.berkeley.edu/>
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020, February). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv. Retrieved 2023-05-26, from <http://arxiv.org/abs/1910.01108> (arXiv:1910.01108 [cs]) doi: doi: 10.48550/arXiv.1910.01108

- Skaik, R. S., & Inkpen, D. (2022). Predicting Depression in Canada by Automatic Filling of Beck's Depression Inventory Questionnaire. *IEEE Access*, 10, 102033–102047. Retrieved 2023-11-27, from <https://ieeexplore.ieee.org/abstract/document/9903477> (Conference Name: IEEE Access) doi: doi: 10.1109/ACCESS.2022.3208470
- Sokol, K., & Flach, P. (2020, January). Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 56–67). New York, NY, USA: Association for Computing Machinery. Retrieved 2024-03-20, from <https://doi.org/10.1145/3351095.3372870> doi: doi: 10.1145/3351095.3372870
- Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and Permuted Pre-training for Language Understanding. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 16857–16867). Curran Associates, Inc. Retrieved 2024-03-31, from https://proceedings.neurips.cc/paper_files/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html
- Tausczik, Y. R., & Pennebaker, J. W. (2010, March). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1), 24–54. Retrieved 2023-10-25, from <https://doi.org/10.1177/0261927X09351676> (Publisher: SAGE Publications Inc) doi: doi: 10.1177/0261927X09351676
- Thorstad, R., & Wolff, P. (2019, August). Predicting future mental illness from social media: A big-data approach. *Behavior Research Methods*, 51(4), 1586–1600. Retrieved 2023-10-25, from <https://doi.org/10.3758/s13428-019-01235-z> doi: doi: 10.3758/s13428-019-01235-z
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... Scialom, T. (2023, July). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv. Retrieved 2023-12-07, from <http://arxiv.org/abs/2307.09288> (arXiv:2307.09288 [cs]) doi: doi: 10.48550/arXiv.2307.09288
- Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., & Ohsaki, H. (2015, April). Recognizing Depression from Twitter Activity. In *Proceedings of the 33rd Annual*

- ACM Conference on Human Factors in Computing Systems* (pp. 3187–3196). New York, NY, USA: Association for Computing Machinery. Retrieved 2023-10-24, from <https://doi.org/10.1145/2702123.2702280> doi: doi: 10.1145/2702123.2702280
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. Retrieved 2023-05-26, from https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
- Wang, Y., & Inkpen, D. (2023). uOttawa at eRisk 2023: Search for Symptoms of Depression.
- Wang, Y., Inkpen, D., & Kirinde Gamaarachchige, P. (2024, March). Explainable Depression Detection Using Large Language Models on Social Media Data. In A. Yates et al. (Eds.), *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)* (pp. 108–126). St. Julians, Malta: Association for Computational Linguistics. Retrieved 2024-03-20, from <https://aclanthology.org/2024.clpsych-1.8>
- Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., ... Kurzweil, R. (2019, July). *Multilingual Universal Sentence Encoder for Semantic Retrieval*. arXiv. Retrieved 2023-10-25, from <http://arxiv.org/abs/1907.04307> (arXiv:1907.04307 [cs]) doi: doi: 10.48550/arXiv.1907.04307
- Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016, June). Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1480–1489). San Diego, California: Association for Computational Linguistics. Retrieved 2023-10-26, from <https://aclanthology.org/N16-1174> doi: doi: 10.18653/v1/N16-1174
- Yao, L., Mao, C., & Luo, Y. (2019, April). Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Medical Informatics and Decision Making*, 19(3), 71. Retrieved 2023-10-20, from <https://doi.org/10.1186/s12911-019-0781-4> doi: doi: 10.1186/s12911-019-0781-4

- Yates, A., Cohan, A., & Goharian, N. (2017, September). *Depression and Self-Harm Risk Assessment in Online Forums*. arXiv. Retrieved 2023-10-25, from <http://arxiv.org/abs/1709.01848> (arXiv:1709.01848 [cs]) doi: doi: 10.48550/arXiv.1709.01848
- Zulfiker, M. S., Kabir, N., Biswas, A. A., Nazneen, T., & Uddin, M. S. (2021, November). An in-depth analysis of machine learning approaches to predict depression. *Current Research in Behavioral Sciences*, 2, 100044. Retrieved 2023-11-26, from <https://www.sciencedirect.com/science/article/pii/S2666518221000310> doi: doi: 10.1016/j.crbeha.2021.100044

Appendix A

BDI Questionnaire

Instructions:

This questionnaire consists of 21 groups of statements. Please read each group of statements carefully, and then pick out the one statement in each group that best describes the way you feel.

If several statements in the group seem to apply equally well, choose the highest number for that group.

1. 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.

2. Pessimism

0. I am not discouraged about my future.

1. I feel more discouraged about my future than I used to be.

2. I do not expect things to work out for me.

3. I feel my future is hopeless and will only get worse.

3. Past Failure

- 0. I do not feel like a failure.
- 1. I have failed more than I should have.
- 2. As I look back, I see a lot of failures.
- 3. I feel I am a total failure as a person.

4. Loss of Pleasure

- 0. I get as much pleasure as I ever did from the things I enjoy.
- 1. I don't enjoy things as much as I used to.
- 2. I get very little pleasure from the things I used to enjoy.
- 3. I can't get any pleasure from the things I used to enjoy.

5. Guilty Feelings

- 0. I don't feel particularly guilty.
- 1. I feel guilty over many things I have done or should have done.
- 2. I feel quite guilty most of the time.
- 3. I feel guilty all of the time.

6. Punishment Feelings

- 0. I don't feel I am being punished.
- 1. I feel I may be punished.
- 2. I expect to be punished.
- 3. I feel I am being punished.

7. Self-Dislike

- 0. I feel the same about myself as ever.
- 1. I have lost confidence in myself.
- 2. I am disappointed in myself.
- 3. I dislike myself.

8. Self-Criticalness

- 0. I don't criticize or blame myself more than usual.
- 1. I am more critical of myself than I used to be.
- 2. I criticize myself for all of my faults.
- 3. I blame myself for everything bad that happens.

9. Suicidal Thoughts or Wishes

- 0. I don't have any thoughts of killing myself.
- 1. I have thoughts of killing myself, but I would not carry them out.
- 2. I would like to kill myself.
- 3. I would kill myself if I had the chance.

10. Crying

- 0. I don't cry anymore than I used to.
- 1. I cry more than I used to.
- 2. I cry over every little thing.
- 3. I feel like crying, but I can't.

11. Agitation

- 0. I am no more restless or wound up than usual.
- 1. I feel more restless or wound up than usual.
- 2. I am so restless or agitated that it's hard to stay still.
- 3. I am so restless or agitated that I have to keep moving or doing something.

12. Loss of Interest

- 0. I have not lost interest in other people or activities.
- 1. I am less interested in other people or things than before.
- 2. I have lost most of my interest in other people or things.
- 3. It's hard to get interested in anything.

13. Indecisiveness

- 0. I make decisions about as well as ever.
- 1. I find it more difficult to make decisions than usual.
- 2. I have much greater difficulty in making decisions than I used to.
- 3. I have trouble making any decisions.

14. Worthlessness

- 0. I do not feel I am worthless.
- 1. I don't consider myself as worthwhile and useful as I used to.
- 2. I feel more worthless as compared to other people.
- 3. I feel utterly worthless.

15. Loss of Energy

- 0. 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.

16. Changes in Sleeping Pattern

- 0. I have not experienced any change in my sleeping pattern.
- 1a. I sleep somewhat more than usual.
- 1b. I sleep somewhat less than usual.
- 2a. I sleep a lot more than usual.
- 2b. I sleep a lot less than usual.
- 3a. I sleep most of the day.
- 3b. I wake up 1-2 hours early and can't get back to sleep.

17. Irritability

- 0. I am no more irritable than usual.
- 1. I am more irritable than usual.
- 2. I am much more irritable than usual.
- 3. I am irritable all the time.

18. Changes in Appetite

- 0. I have not experienced any change in my appetite.
- 1a. My appetite is somewhat less than usual.
- 1b. My appetite is somewhat greater than usual.
- 2a. My appetite is much less than before.
- 2b. My appetite is much greater than usual.
- 3a. I have no appetite at all.
- 3b. I crave food all the time.

19. Concentration Difficulty

- 0. I can concentrate as well as ever.
- 1. I can't concentrate as well as usual.
- 2. It's hard to keep my mind on anything for very long.
- 3. I find I can't concentrate on anything.

20. Tiredness or Fatigue

- 0. I am no more tired or fatigued than usual.
- 1. I get more tired or fatigued more easily than usual.
- 2. I am too tired or fatigued to do a lot of the things I used to do.
- 3. I am too tired or fatigued to do most of the things I used to do.

21. Loss of Interest in Sex

- 0. I have not noticed any recent change in my interest in sex.
- 1. I am less interested in sex than I used to be.
- 2. I am much less interested in sex now.
- 3. I have lost interest in sex completely.

Appendix B

Responsible NLP Checklist

To reinforce the ethical considerations of our work, the Responsible NLP Checklist¹ was answered and displayed below.

A. For every submission

A1. Did you describe the limitations of your work?

Yes. In Section 7.3.

A2. Did you discuss any potential risks of your work?

Yes. In Section 7.5.

A3. Do the abstract and introduction summarize the paper’s main claims?

Yes.

¹<https://aclrollingreview.org/responsibleNLPresearch/>

B. Did you use or create scientific artifacts?

Yes. We wrote codes to design application systems and conduct experiments. We used many existing scientific artifacts, such as modules, frameworks, and shared datasets.

B1. Did you cite the creators of artifacts you used?

Yes.

B2. Did you discuss the license or terms for use and / or distribution of any artifacts?

Not applicable.

B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?

Yes. In Section 7.5.

B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?

Not applicable.

B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?

Yes. In Section 4.2, Section 6.2, and Section 7.3.

B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created?

Yes. In Section 4.2 and Section 4.3.

C. Did you run computational experiments?

Yes.

C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?

Yes. In Section 5.2 and Section 6.2.

C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Yes. In Section 5.2 and Section 6.2.

C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

Yes. In Section 5.4 and Section 6.4.

C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

Yes. In Section 5.2 and Section 6.2.

D. Did you use human annotators (e.g., crowdworkers) or research with human participants?

Not applicable.

D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

Not applicable.

D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

Not applicable.

D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating?

Not applicable.

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

Not applicable.

D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

Not applicable.