NUS-IDS@eRisk2024: Ranking Sentences for Depression Symptoms using Early Maladaptive Schemas and Ensembles

Notebook for the eRisk Lab at CLEF 2024

Beng Heng Ang[1], Sujatha Das Gollapalli[2] and See-Kiong Ng[2]

[1]*Integrative Sciences and Engineering Programme, National University of Singapore*
[2]*Institute of Data Science, National University of Singapore*

Abstract

In this paper, we describe the techniques developed by our team, *NUS-IDS*, for eRisk2024 Task 1. This task focuses on ranking and identifying sentences from social media that express symptoms of depression as outlined in the Beck's Depression Inventory-II (BDI) questionnaire. We developed five different configurations for this task using ensembles of unique sets of base predictors. The base predictors use sentence-transformer models fine-tuned with Contrastive Learning using the task's training data and expressive exemplars relevant to both the BDI symptoms and the features from the well-established taxonomy of Early Maladaptive Schemas (EMS) obtained by prompting GPT-4. Among our five submitted configurations, three of them outperformed competing runs in at least two metrics. Overall, our Configuration 5 is the best performing among all runs in three out of four competition metrics for both the Majority voting and Unanimity evaluations. We further analyzed our configurations and derive insights on the complexities of Ensemble Learning as well as the use of expressive exemplars for this task.

1. Introduction

Depression is one of the most common mental disorders affecting at least 10% of the global population [1, 2, 3]. Worryingly, most people do not address their depression due to limited access to professional mental health services, lack of awareness, and the social stigma from publicly sharing their mental disorder [4, 5, 6]. Rather, with the rising adoption of Social Media and the anonymity that these platforms provide, more people with depression are seeking alleviation of their disorder by posting about their personal experiences online [7, 8, 9, 10]. However, these people form only a subset of the users of these online platforms as there are also people who use Social Media for other purposes like socializing and entertainment [11, 12, 13]. Hence, it is essential to identify which users are more susceptible to depression on these platforms. This crucial step enables timely support and better intervention outcomes for the support seekers [8, 10].

Recent advances in Natural Language Processing (NLP), especially Large Language Models (LLMs) [14], have enabled the prediction of one's susceptibility to depression from their online posts. Most studies used LLMs to encode a user's posts and used the representations to predict whether a user is depressed [15, 16, 17, 18, 19, 20]. However, this approach lacks explainability. Depression is a complex mental disorder with symptoms that can vary significantly from person to person [21, 22]. Consequently, merely predicting whether someone is depressed from their online posts excludes valuable information on the addressable symptoms that could otherwise explain the depression and guide personalized support [23, 24].

To address this limitation, recent studies have approached predicting depression from its symptoms. One of them proposed using the nine symptoms in the Patient Health Questionnaire-9 to substantiate the prediction of depression [25]. The study found that grounding their predictions in the symptoms

not only improved their models' generalizability across out-of-distribution data but also offered better explainability of the predictions. In another study, the authors augmented the explainability of their depression prediction model by building a disease-symptom knowledge graph with the symptoms described in clinical manuals (e.g. DSM-5) and using the graph to predict depression [26]. A later study curated linguistic features corresponding to 13 expert-validated depression symptoms from 43 subreddit communities and used these features to predict depression [27]. Their results highlighted the importance of symptom-specific features in improving depression prediction accuracy and reliability.

Placing similar importance on depression symptoms, the eRisk2024 Task 1 "Search for Symptoms of Depression" aims to retrieve sentences exemplifying depression symptoms from Social Media posts [28, 29]. Unlike previous studies, this task focuses on the 21 symptoms outlined in the Beck's Depression Inventory-II (BDI) [30]. The BDI is a questionnaire commonly used by mental health professionals to screen people for depression through symptoms such as "Sadness", "Pessimism", "Guilty Feelings", and "Changes in Appetite". For this eRisk2024 task, participating teams are required to rank sentences from a given dataset of Social Media posts according to their relevance to each symptom [28, 29]. The subsequent sections of this paper outline the contributions of our team *(NUS-IDS)* to this task, detailing the techniques we developed and the performance of our configurations. We also discuss the insights acquired from analyzing the performance of our configurations, highlighting our challenges and exciting opportunities for future work.

2. Related Work

Recent popularity of Zero-Shot LLMs like ChatGPT[1] have introduced a novel method of generating synthetic data for data augmentation [31, 32, 33]. Models using these synthetic data typically exhibit improved task performance, particularly in low-resource settings. However, for tasks requiring specialized domain knowledge, performance can decline if the synthetic data are not well-aligned with the ground truths [33]. For instance, a study in eRisk2023 Task 1 [34] used ChatGPT to synthesize Reddit posts that exemplify specific BDI symptoms and used these posts as exemplars when scoring a sentence in the task. However, the system from the study underperformed. Upon further investigation, the author attributed this underperformance to the inclusion of extraneous details (e.g., "I just got back from a great vacation ...") in the synthesized exemplars and recommended that subsequent studies minimize introducing irrelevant information during the synthesis. Inspired by this study, we also made use of synthetic data in this competition. Unlike the previous study, we based our synthetic data not only on the BDI symptoms but also on Early Maladaptive Schemas (EMS), a well-established taxonomy in Schema Therapy. EMS are deeply ingrained, dysfunctional belief patterns formed in childhood, which negatively influence behavior and relationships throughout one's life [35]. As existing studies have identified associations between EMS and depression [36, 37], we leveraged EMS to synthesize more diverse and expressive exemplars that are relevant to the task.

We also experimented with Ensemble Learning in this study. Ensemble learning is a powerful machine learning technique that aggregates the predictions of multiple base predictors to produce a more accurate and robust output than any single predictor could achieve independently [38, 39, 40, 38]. Previous studies [41, 42, 43, 34, 44, 45, 46] have focused on using a single pair of a model and exemplars to score and rank the sentences. In contrast, each of our ensembles consists of a set of base predictors, with each base predictor being a unique pair of a model and exemplars. To increase the selection of models available for our ensembles, we also experimented fine-tuning several pretrained sentence-transformer models [47] with the Contrastive Learning approach [48].

---

3. Methodology

3.1. Dataset Description

The organizers provided participants with TREC-formatted sentence-tagged eRisk2024 T1 test and train datasets [28, 29]. For each set, we compiled the sentences along with their respective unique "DOCNO" identifiers. The test dataset comprises ∼15M sentences from about 500k users, with an average sentence length of ∼18 words. The train dataset, comprising competition data for the same task last year, includes ∼4M sentences from about 3k users, with an average sentence length of ∼14 words. Accompanying the train dataset are gold labels for ∼21k sentences, each denoting relevance (0 for irrelevant; 1 for relevant) of a sentence to a BDI symptom. Of these, 4,552 sentences are labeled relevant, while 17,028 are labeled irrelevant to some symptom. The following are some examples that are relevant to their respective symptoms:

- I feel this emotional pain straight to my soul. (Sadness symptom)
- So basically, it feels like I'm stuck in a prison, unable to do anything but work. (Punishment Feelings symptom)
- I'm constantly fidgeting, pacing back and forth across the room, getting up from my computer chair, sitting back down, getting back up, rinse repeat. (Agitation symptom)
- Allowing myself to become addicted to food. (Changes in Appetite symptom)

From the examination of the gold labels, we discovered that most symptom-relevant sentences contain first-person personal pronouns (e.g. I, I'm, my, me, mine, myself). In contrast, symptom-irrelevant sentences generally lack first-person personal pronouns. While also considering the task definition that a sentence is relevant to a symptom if it conveys information about the user's state [28, 29], we excluded sentences that do not contain first-person personal pronouns from the datasets. After filtering the data, we have ∼10M sentences in our test dataset and ∼1M sentences in our train dataset.

3.2. Scoring, Ranking, and Retrieving Symptom-related Sentences

In this study, we provided each sentence a score that indicates their relevance to a BDI symptom. This score can be formalized as follows. Let s be any sentence from the dataset $\mathcal{S}$, x_b an exemplar from a set $\mathcal{X}$ and is related to a symptom $b \in \mathcal{B}$, and $m \in \mathcal{M}$ a candidate model with which we acquire representations. We compute the score $\mathcal{SR}$ for s with respect to b, m, and $\mathcal{X}$ as:

$$\mathcal{SR}(s \mid b, m, \mathcal{X}) = \frac{1}{|\mathcal{X}_b|} \sum_{x_b \in \mathcal{X}_b} sim(m(s), m(x_b))$$

where sim denotes the cosine similarity function, $m(s)$ and $m(x_b)$ represent the representations of the sentence s and exemplar statement x_b from model m.

The score $\mathcal{SR}$ enabled us to rank sentences on their relevance to a symptom b. For each b, we retrieved the top 1,000 ranked sentences and the collection over $\mathcal{B}$ formed our submission set:

$$S_{submission} = \bigcup_{b \in \mathcal{B}} \{s_i \in S \mid rank(\mathcal{SR}(s_i \mid b, m, \mathcal{X})) \leq 1.0 \times 10^3\}$$

3.3. Features based on Early Maladaptive Schemas

EMS are entrenched patterns of thought, emotion, and behavior that often originate in childhood and endure into adulthood, often leading to self-defeating outcomes. Young et al. [35] have delineated 18 distinct types of EMS (see Table A), which frequently underlie prevalent mental disorders like Depression. In our on-going work [49], we curated posts from users of online mental health community forums such as 7Cups,[2] annotated these posts with EMS labels [50], clustered these posts according to

---

[2]https://www.7cups.com/forum/depression/

their EMS annotations, and extracted the characteristic features of each EMS from their cluster with GPT-4.[3] Following previous studies which demonstrated associations between EMS and depression [37, 51, 52, 53], we used EMS-based features relevant to Depression to curate exemplars that are more expressive, to augment those provided in the BDI.

3.4. Curating the Exemplars $\mathcal{X}_{BDI}$, $\mathcal{X}_{EMS\_BDI}$, and $\mathcal{X}_{EMS\_BDI+}$

In this study, we used three sets of exemplars[4] in our configurations to score and rank each sentence. The first set $\mathcal{X}_{BDI}$ consists of the original exemplars for each symptom in the BDI (e.g. "I feel sad much of the time" for "Sadness").

The second set $\mathcal{X}_{EMS\_BDI}$ consists of exemplars synthesized using the features of EMS and the BDI symptoms. Specifically, we computed the semantic similarity scores between each pair of EMS feature and BDI symptom using a sentence-transformer model[5] and retained feature-symptom pairs that have a semantic similarity score of ≥ 0.5. We conducted a manual review of the pairs to verify that they are logically sound. For example, the pair between the feature "Feeling angry or frustrated towards oneself" of the Punitiveness EMS and the symptom "Suicidal Thoughts or Wishes" have been previously reported in human subject research [54]. For each feature-symptom pair, we synthesized the corresponding exemplars using GPT-4[6] and our prompt (see Table 1). As illustrated in the table, the synthesized exemplars are more expressive about the "Suicidal Thoughts or Wishes" symptom than the original exemplar in the BDI.

Table 1
Example of a prompt to synthesize expressive exemplars related to a BDI symptom and EMS feature.

Input Prompt:	Given the Beck's Depression Inventory (BDI) symptom and Early Maladaptive Schema (EMS) feature below, generate a list of example sentences, each sentence expressing the BDI symptom and the EMS feature in first person to reflect the state of the person. BDI symptom: Suicidal Thoughts or Wishes EMS feature: Feeling angry or frustrated towards oneself
Output Exemplars:	I keep wishing I could just disappear, frustrated that I can never live up to my own expectations I can't escape the desire to end my pain, and it's fueled by this relentless anger I have towards myself for not being stronger
Original Exemplar in the BDI:	I would like to kill myself.

For the final set of exemplars $\mathcal{X}_{EMS\_BDI+}$, we adhered to the task definition that a sentence remains relevant even if it suggests that the user is okay with the symptom. Since most exemplars in the preceding sets imply a negative user state, we prompted GPT-4 to generate corresponding variations of the exemplars that could satisfy this task definition. As illustrated in Table 2, the exemplars generated by GPT-4 with our prompt indeed adhere to the task definition.

3.5. Ensemble Learning

Ensemble learning is a machine learning technique that combines multiple base predictors to create a more powerful and robust predictor with improved overall performance, accuracy, and generalization

---

[3]We used the variant "gpt-4-1106-preview"
[4]The resources will be available in https://github.com/NUS-IDS/ems_mentalhealth/tree/eRisk2024T1
[5]https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[6]Following our other study, we also used the variant "gpt-4-1106-preview"

Table 2
Example of a prompt to synthesize exemplars showing that the user is ok with the symptoms.

Input Prompt:	A sentence will be deemed relevant to a Beck's Depression Inventory symptom when it conveys information about the user's state concerning the symptom. That is, it may be relevant even when it indicates that the user is ok with the symptom. For each sentence from the given list, generate a corresponding sentence that indicates that the user is ok with the symptom. For example, given the sentence "I am so sad or unhappy that I can't stand it", the corresponding sentence is "I do not feel sad".
	Given sentences: I just can't shake the feeling that I've let myself down, and it's led to a profound sense of self-dislike that I can't escape. [...]
Output Exemplar:	I've learned to appreciate my efforts and I no longer feel a profound sense of self-dislike that I can't escape. [...]

compared to any individual base predictors [38, 39]. In this study, each base predictor is a unique pair of a model $m_i \in \mathcal{M}_{select}$ and exemplars $\mathcal{X}_i \in \mathcal{X}_{select}$, with a scoring function equivalent to that described in Section 3.2. Formally, an ensemble containing ≥ 2 base predictors can be formally described as:

$$\mathcal{E} = \{(m_i, \mathcal{X}_i) \mid m_i \in \mathcal{M}_{select}, \mathcal{X}_i \in \mathcal{X}_{select}, \text{ and } i = 2, 3, \dots, n\}$$

For ensembles, $\mathcal{SR}_{ensemble}$ is obtained from averaging the $\mathcal{SR}$ scores across all the base predictors:

$$\mathcal{SR}_{ensemble}(s \mid b, \mathcal{E}) = \frac{1}{|\mathcal{E}|} \sum_{(m,\mathcal{X}) \in \mathcal{E}} \left(\frac{1}{|\mathcal{X}_b|} \sum_{x_b \in \mathcal{X}_b} sim(m(s), m(x_b)) \right)$$

For the base predictors of our ensembles, we experimented with various pretrained sentence-transformer models and also their fine-tuned variants.[7] The fine-tuned models were obtained by training the pretrained models using Contrastive Learning. Contrastive learning involves training a model to learn new representations by bringing positive examples closer in the representation space while pushing negative ones further apart [48]. This method enhances the model's ability to differentiate between positive (similar) and negative (dissimilar) pairs of data and has proven effective in information retrieval [55, 56], which is advantageous to the current task of ranking sentences for depression symptoms.

Let $m \in \mathcal{M}_{pre}$ represent a pretrained sentence-transformer model. We fine-tuned m via contrastive learning using a dataset of positive and negative pairs of sentences $\mathcal{S}_{FT}$. The contrastive loss $\mathcal{L}_{contrastive}$ can be formally defined as:

$$\mathcal{L}_{contrastive} = \frac{1}{|\mathcal{S}_{FT}|} \sum_{(s_i, s_j, y) \in \mathcal{S}_{FT}} \begin{cases} \|m(s_i) - m(s_j)\|^2 & \text{if } y = 1 \\ \max(0, t - \|m(s_i) - m(s_j)\|)^2 & \text{if } y = 0 \end{cases}$$

Here, s_i and s_j are pairs of input data points from $\mathcal{S}_{FT}$, y is a binary label that indicates whether the pair is positive (1) or negative (0), and $m(s)$ denotes the representation of s by the model m. The margin t specifies the minimum desired distance between the representations of the negative pairs.

To prepare $\mathcal{S}_{FT}$ for contrastive learning, we made use of the given gold labels in the train dataset. Let s be a sentence in S_{gold}, the gold subset of the train dataset. Each s is either related or unrelated to a symptom b, where $b \in B$ and B is the set of all possible symptoms covered by the BDI. To identify positive pairs of sentences, we considered sentences s_i and s_j that are both related to the same symptom

---

[7]In this study, we use the top three pretrained sentence-transformer models for their performant sentence embeddings [47], "all-mpnet-base-v2" (m_{mpnet}), "all-MiniLM-L12-v2" (m_{minilm}) and "all-distilroberta-v1" (m_{distil}). Their fine-tuned versions are correspondingly labelled $m_{mpnetFT}$, $m_{minilmFT}$ and $m_{distilFT}$

b. Conversely, we identified negative pairs between sentences that are not related to the same symptom. Formally, they can be represented as:

$$\mathcal{S}_{pos} = \{(s_i, s_j, 1) \mid s_i \in S_{gold} \wedge s_j \in S_{gold} \wedge related(s_i, b) \wedge related(s_j, b)\}$$

$$\mathcal{S}_{neg} = \{(s_i, s_j, 0) \mid s_i \in S_{gold} \wedge s_j \in S_{gold} \wedge related(s_i, b) \wedge \neg related(s_j, b)\}$$

Collectively, $\sim$4M pairs of sentences from $\mathcal{S}_{pos}$ and $\mathcal{S}_{neg}$ were obtained from the gold labels in the train dataset to form $\mathcal{S}_{FT}$. With this and $\mathcal{L}_{contrastive}$, m is iteratively fine-tuned to improve its ability in differentiating between semantically similar and dissimilar sentences, which should provide better task performance in this competition.

3.6. Submitted Configurations

This section describes our top five configurations, which we have found to achieve competitive validation performance on the data shared as part of the eRisk 2024 competition.

3.6.1. Configuration 1

This configuration uses only the fine-tuned model $m_{distilFT}$ and the exemplars $\mathcal{X}_{BDI}$ to score the sentences.

$$\mathcal{SR}_1(s \mid b, m_{distilFT}, \mathcal{X}_{BDI}) = \frac{1}{|\mathcal{X}_{BDI,b}|} \sum_{x_b \in \mathcal{X}_{BDI,b}} sim(m_{distilFT}(s), m_{distilFT}(x_b))$$

3.6.2. Configuration 2

Configuration 2 extends from the previous configuration by using an ensemble involving $m_{mpnetFT}$, $m_{minilmFT}$, and $m_{distilFT}$. Following the previous, this configuration uses the exemplars $\mathcal{X}_{BDI}$ when scoring the sentences.

$$\mathcal{E}_2 = \{(m_{mpnetFT}, \mathcal{X}_{BDI}), (m_{minilmFT}, \mathcal{X}_{BDI}), (m_{distilFT}, \mathcal{X}_{BDI})\}$$

$$\mathcal{SR}_2(s \mid b, \mathcal{E}_2) = \frac{1}{|\mathcal{E}_2|} \sum_{(m, \mathcal{X}) \in \mathcal{E}_2} \left(\frac{1}{|\mathcal{X}_b|} \sum_{x_b \in \mathcal{X}_b} sim(m(s), m(x_b))\right)$$

3.6.3. Configuration 3

Configuration 3 is also an ensemble involving the same models as Configuration 2. Unlike the latter, Configuration 3 involved augmenting $\mathcal{X}_{BDI}$ with the other exemplar sets $\mathcal{X}_{EMS\_BDI}$ and $\mathcal{X}_{EMS\_BDI+}$. Formally, this ensemble can be expressed as:

$$\mathcal{E}_{3a} = \{(m_{mpnetFT}, \mathcal{X}_{BDI} \cup \mathcal{X}_{EMS\_BDI}), (m_{minilmFT}, \mathcal{X}_{BDI} \cup \mathcal{X}_{EMS\_BDI}),$$
$$(m_{distilFT}, \mathcal{X}_{BDI} \cup \mathcal{X}_{EMS\_BDI})\}$$

$$\mathcal{E}_{3b} = \{(m_{mpnetFT}, \mathcal{X}_{BDI} \cup \mathcal{X}_{EMS\_BDI} \cup \mathcal{X}_{EMS\_BDI+})$$
$$(m_{minilmFT}, \mathcal{X}_{BDI} \cup \mathcal{X}_{EMS\_BDI} \cup \mathcal{X}_{EMS\_BDI+}),$$
$$(m_{distilFT}, \mathcal{X}_{BDI} \cup \mathcal{X}_{EMS\_BDI} \cup \mathcal{X}_{EMS\_BDI+})\}$$

$$\mathcal{E}_3 = \mathcal{E}_{3a} \cup \mathcal{E}_{3b}$$

3.6.4. Configuration 4

Configuration 4 is an ensemble that combines the base predictors used in the ensembles of both Configuration 2 and 3, specifically:

$$\mathcal{E}_4 = \mathcal{E}_2 \cup \mathcal{E}_3$$

3.6.5. Configuration 5

Configuration 5 is a variant of Configuration 4. Instead of using the fine-tuned models in $\mathcal{E}_2$, we used their pretrained versions.

$$\mathcal{E}_{2\_pretrained} = \{(m_{mpnet}, \mathcal{X}_{BDI}), (m_{minilm}, \mathcal{X}_{BDI}), (m_{distil}, \mathcal{X}_{BDI})\}$$

$$\mathcal{E}_5 = \mathcal{E}_3 \cup \mathcal{E}_{2\_pretrained}$$

4. Results

A total of 29 systems were submitted by all participating teams for this task. Table 3 and 4 present the performance of our configurations. In these tables, we highlighted the best performance for each of the evalution metric in **bold** for the top-2 runs shared by the competition organizers [28, 29]. Performance of our configurations was evaluated along the metrics of Average Precision (AP), R-Precision (R-PREC), Precision at 10 (P@10), and Normalized Discounted Cumulative Gain at 1000 (NDCG). The performance in Table 3 were computed by comparing the output from a configuration against the organizer's majority-voted ground truths of the test dataset. On the other hand, Table 4 reflected the performance with respect to the unanimity ground truths of the test dataset.

Table 3
Performance of our Configurations against the best run from the other participating teams (Majority Voting evaluation). Our best configuration is highlighted using an *.

Team	Run	AP	R-PREC	P@10	NDCG
APB-UC3M	APB-UC3M sentsim-all-MiniLM-L6-v2	0.354	0.391	**0.986**	0.591
NUS-IDS	Configuration 5*	**0.375**	**0.434**	0.924	**0.631**
NUS-IDS	Configuration 2	0.352	0.415	0.881	0.616
NUS-IDS	Configuration 4	0.336	0.401	0.890	0.599
NUS-IDS	Configuration 1	0.312	0.386	0.871	0.576
NUS-IDS	Configuration 3	0.286	0.359	0.857	0.556

Table 4
Performance of our Configurations against the top-2 runs from the other participating teams (Unanimity evaluation). Our best configuration is highlighted using an *.

Team	Run	AP	R-PREC	P@10	NDCG
MeVer-REBECCA	TransformerEmbeddings CosineSimilarity gpt	0.305	0.357	**0.833**	0.551
APB-UC3M	APB-UC3M sentsim-all-MiniLM-L6-v2	0.345	0.407	0.829	0.630
NUS-IDS	Configuration 5*	**0.392**	**0.436**	0.795	**0.692**
NUS-IDS	Configuration 2	0.370	0.431	0.752	0.677
NUS-IDS	Configuration 4	0.358	0.416	0.771	0.662
NUS-IDS	Configuration 1	0.329	0.391	0.786	0.636
NUS-IDS	Configuration 3	0.312	0.375	0.757	0.621

Across all runs, Configuration 5 is the best performing along the metrics AP, R-PREC, and NDCG in both Majority voting and Unanimity evaluations. For these metrics, Configuration 5 yielded 0.375, 0.434, and 0.631 respectively in the Majority voting evaluation, and 0.392, 0.436, and 0.692 in the Unanimity evaluation. Similar to Configuration 5 but only for Unanimity evaluation, Configuration 2 and 4 also performed better than the best run from the other teams in terms of AP, R-PREC, and NDCG. For Majority voting evaluation, however, both configurations only performed better in R-PREC and NDCG. Although Configuration 1 and 3 did not surpass the best run from the other teams in most metrics, the performance of these configurations are still marginally competitive. Likewise, although our configurations didn't attain the best score for P@10, they still achieved P@10 performance that closely approached the best performance from the other teams. Our best performing Configuration 5 yielded P@10 of 0.924 for Majority voting and 0.795 for Unanimity evaluation. In comparison, the best run from the other teams obtained the corresponding scores of 0.986 and 0.833. Overall, the performance of our configurations highlights their effectiveness in ranking and identifying sentences relevant to the symptoms of depression.

We noted that Configuration 5 performed best amongst all our configurations on the competition test dataset, as evaluated by the organizers. Configuration 4 though comprising of an ensemble of fined-tuned models for all base predictors obtained lower performance than Configuration 5 which is a mix of fine-tuned as well as pretrained sentence transformers. We hypothesize that this outcome, contrary to our expectation, signals the underlying complexities among and within each base predictor of an ensemble (e.g. model m_i and exemplars $\mathcal{X}_i$ compatibility) that could have a greater influence on the overall performance than just solely the use of fine-tuned models. Certainly, this is something interesting to investigate further in subsequent studies.

Similarly, though Configuration 3 which supplemented $\mathcal{X}_{BDI}$ with more expressive exemplars from $\mathcal{X}_{EMS\_BDI}$ and $\mathcal{X}_{EMS\_BDI+}$, its performance is lower than Configuration 2 which uses $\mathcal{X}_{BDI}$ alone. We posit that this discrepancy might be due to the task's inherent bias for shorter sentences and those with explicit keywords. As an example, for the BDI symptom "Sadness", compare between the sentence "I am always sad" and a more expressive one "Today I feel more depressed and miserable than I felt in a very long time (I am generally a happy guy)". The first is a short sentence that explicitly mentioned "sad", a phrase which forms part of the symptom's name. In contrast, the latter sentence while expressing a semantic that is relevant to "Sadness" provides greater elaboration and does not contain any explicit mentions of the symptom. Since the ground truths for the test dataset were derived from the top-k sentences pooled from all participating teams [28, 29] and given that most runs likely ranked short-explicit sentences in the top-k, configurations that favored more expressive sentences in the rankings (e.g. Configuration 3) could have been undervalued. This might explain the overall lower performance of Configuration 3 with respect to Configuration 2. Still, the latter sentence in the example above shows that expressive sentences could provide greater detail and practical value to a therapy than the short explicit ones. Therefore, future work could consider evaluation methods that more fairly address these expressive sentences in the rankings.

5. Conclusion

This paper summarizes our team's (*NUS-IDS*) contributions and methods developed for the "eRisk2024 Task 1: Search for Symptoms of Depression" competition. Configuration 5, an ensemble consisting of unique base predictors of model and exemplars, performed best in three out of four evaluation metrics among all 29 submitted runs in both Majority voting and Unanimity evaluations. Additionally, our Configuration 2 and 4 outperformed the best run from the other teams along three metrics in Unanimity evaluation and two metrics in Majority voting evaluation. These results highlight the effectiveness of our configurations in ranking and identifying symptoms of depression from online texts. However, there are still unresolved questions when comparing the performance amongst our configurations, indicating opportunities for future research. Firstly, further investigation is needed to unravel the complexities among the base predictors of an ensemble and their influence over the overall performance.

Secondly, future studies could explore evaluation methods that more fairly address expressive yet less explicit sentences in the rankings for a BDI symptom.

Acknowledgments

References

[1] G. Y. Lim, W. W. Tam, Y. Lu, C. S. Ho, M. W. Zhang, R. C. Ho, Prevalence of depression in the community from 30 countries between 1994 and 2014, Scientific reports 8 (2018) 2861.

[2] R. Barzilay, T. M. Moore, D. M. Greenberg, G. E. DiDomenico, L. A. Brown, L. K. White, R. C. Gur, R. E. Gur, Resilience, covid-19-related stress, anxiety and depression during the pandemic in a large population enriched for healthcare providers, Translational psychiatry 10 (2020) 291.

[3] S. Shorey, E. D. Ng, C. H. Wong, Global prevalence of depression and elevated depressive symptoms among adolescents: A systematic review and meta-analysis, British Journal of Clinical Psychology 61 (2022) 287–305.

[4] V. Patel, S. Saxena, C. Lund, G. Thornicroft, F. Baingana, P. Bolton, D. Chisholm, P. Y. Collins, J. L. Cooper, J. Eaton, et al., The lancet commission on global mental health and sustainable development, The lancet 392 (2018) 1553–1598.

[5] A. Karasz, F. Gany, J. Escobar, C. Flores, L. Prasad, A. Inman, V. Kalasapudi, R. Kosi, M. Murthy, J. Leng, et al., Mental health and stress among south asians, Journal of immigrant and minority health 21 (2019) 7–14.

[6] G. Thornicroft, Stigma and discrimination limit access to mental health care, Epidemiology and Psychiatric Sciences 17 (2008) 14–19.

[7] J. A. Naslund, K. A. Aschbrenner, L. A. Marsch, S. J. Bartels, The future of mental health care: peer-to-peer support and social media, Epidemiology and psychiatric sciences 25 (2016) 113–122.

[8] B. Ridout, A. Campbell, The use of social networking sites in mental health interventions for young people: systematic review, Journal of medical Internet research 20 (2018) e12244.

[9] K. Stawarz, C. Preist, D. Coyle, et al., Use of smartphone apps, social media, and web-based resources to support mental health and well-being: online survey, JMIR mental health 6 (2019) e12546.

[10] J. A. Naslund, A. Bondre, J. Torous, K. A. Aschbrenner, Social media and mental health: benefits, risks, and opportunities for research and practice, Journal of technology in behavioral science 5 (2020) 245–257.

[11] A. Whiting, D. Williams, Why people use social media: a uses and gratifications approach, Qualitative market research: an international journal 16 (2013) 362–369.

[12] K.-Y. Lin, H.-P. Lu, Why people use social networking sites: An empirical study integrating network externalities and motivation theory, Computers in human behavior 27 (2011) 1152–1161.

[13] P. B. Brandtzæg, J. Heim, Why people use social networking sites, in: Online Communities and Social Computing: Third International Conference, OCSC 2009, Held as Part of HCI International 2009, San Diego, CA, USA, July 19-24, 2009. Proceedings 3, Springer, 2009, pp. 143–152.

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

[15] S.-H. Wu, Z.-J. Qiu, Cyut at erisk 2022: Early detection of depression based-on concatenating representation of multiple hidden layers of roberta model., in: CLEF (Working Notes), 2022, pp. 1014–1025.

[16] A.-M. Bucur, A. Cosma, L. P. Dinu, P. Rosso, An end-to-end set transformer for user-level classification of depression and gambling disorder, arXiv preprint arXiv:2207.00753 (2022).

[17] A. Säuberli, S. Cho, L. Stahlhut, Lausan at erisk 2022: Simply and effectively optimizing text classification for early detection., in: CLEF (Working Notes), 2022, pp. 995–1004.

[18] X. Wang, S. Chen, T. Li, W. Li, Y. Zhou, J. Zheng, Q. Chen, J. Yan, B. Tang, et al., Depression risk prediction for chinese microblogs via deep-learning methods: Content analysis, JMIR medical informatics 8 (2020) e17958.

[19] S. Devika, M. Pooja, M. Arpitha, R. Vinayakumar, Bert-based approach for suicide and depression identification, in: Proceedings of Third International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2022, Springer, 2023, pp. 435–444.

[20] R. Martínez-Castaño, A. Htait, L. Azzopardi, Y. Moshfeghi, Bert-based transformers for early detection of mental health illnesses, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12, Springer, 2021, pp. 189–200.

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

[22] M. Ten Have, F. Lamers, K. Wardenaar, A. Beekman, P. de Jonge, S. van Dorsselaer, M. Tuithof, M. Kleinjan, R. de Graaf, The identification of symptom-based subtypes of depression: A nationally representative cohort study, Journal of affective disorders 190 (2016) 395–406.

[23] M. Ratcliffe, Experiences of depression: A study in phenomenology, OUP Oxford, 2014.

[24] B. G. McNair, N. J. Highet, I. B. Hickie, Exploring the perspectives of people whose lives have been affected by depression, Medical Journal of Australia 176 (2002) S69–S69.

[25] T. Nguyen, A. Yates, A. Zirikly, B. Desmet, A. Cohan, Improving the generalizability of depression detection by leveraging clinical questionnaires, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 8446–8459.

[26] Z. Zhang, S. Chen, M. Wu, K. Zhu, Symptom identification for interpretable detection of multiple mental disorders on social media, in: Proceedings of the 2022 conference on empirical methods in natural language processing, 2022, pp. 9970–9985.

[27] T. Liu, D. Jain, S. R. Rapole, B. Curtis, J. C. Eichstaedt, L. H. Ungar, S. C. Guntuku, Detecting symptoms of depression on reddit, in: Proceedings of the 15th ACM Web Science Conference 2023, 2023, pp. 174–183.

[28] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2024: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, 15th International Conference of the CLEF Association, CLEF 2024, Springer International, Grenoble, France, 2024.

[29] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2024: Early risk prediction on the internet (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum CLEF 2024, Grenoble, France, September 9th to 12th, 2024, CLEF 2024, CEUR Workshop Proceedings, 2024.

[30] A. T. Beck, R. A. Steer, G. K. Brown, Bdi-ii, beck depression inventory: manual: Psychological corp, San Antonio, TX 3 (1996) 601–608.

[31] A. Latif, J. Kim, Evaluation and analysis of large language models for clinical text augmentation and generation, IEEE Access (2024).

[32] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, L. Zhao, S. Xu, W. Liu, N. Liu, et al., Auggpt: Leveraging chatgpt for text data augmentation, arXiv preprint arXiv:2302.13007 (2023).

[33] F. Piedboeuf, P. Langlais, Is chatgpt the ultimate data augmentation algorithm?, in: The 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

[34] A.-M. Bucur, Utilizing chatgpt generated data to retrieve depression symptoms from social media (2023).

[35] J. E. Young, J. S. Klosko, M. E. Weishaar, Schema therapy: A practitioner's guide, guilford press, 2006.

[36] A. Cormier, B. Jourda, C. Laros, V. Walburg, S. Callahan, Influence between early maladaptive schemas and depression, L'Encephale 37 (2011) 293–298.

[37] A. Bishop, R. Younan, J. Low, P. D. Pilkington, Early maladaptive schemas and depression in

adulthood: A systematic review and meta-analysis, Clinical Psychology & Psychotherapy 29 (2022) 111–130.

[38] O. Sagi, L. Rokach, Ensemble learning: A survey, Wiley interdisciplinary reviews: data mining and knowledge discovery 8 (2018) e1249.

[39] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, P. N. Suganthan, Ensemble deep learning: A review, Engineering Applications of Artificial Intelligence 115 (2022) 105151.

[40] X. Dong, Z. Yu, W. Cao, Y. Shi, Q. Ma, A survey on ensemble learning, Frontiers of Computer Science 14 (2020) 241–258.

[41] N. Recharla, P. Bolimera, Y. Gupta, A. K. Madasamy, Exploring depression symptoms through similarity methods in social media posts (2023).

[42] J. Martinez-Romo, L. Araujo, X. Larrayoz, M. Oronoz, A. Pérez, Obser-menh at erisk 2023: deep learning-based approaches for symptom detection in depression and early identification of pathological gambling indicators, Working Notes of CLEF (2023) 18–21.

[43] Y. Wang, D. Inkpen, uottawa at erisk 2023: search for symptoms of depression, Working Notes of CLEF (2023) 18–21.

[44] E. Bezerra, L. d. F. Santos, R. F. Nascimento, R. P. Lopes, G. P. Guedes, Nailp at erisk 2023: search for symptoms of depression, in: 4th Working Notes of the Conference and Labs of the Evaluation Forum (CLEF-WN), volume 3497, CEUR-WS, 2023, pp. 639–661.

[45] H. Thompson, L. Cagnina, M. Errecalde, Strategies to harness the transformers' potential: Unsl at erisk 2023 (2023).

[46] F. A. Sakib, A. A. Choudhury, O. Uzuner, Mason-nlp at erisk 2023: Deep learning-based detection of depression symptoms from social media texts (2023).

[47] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084.

[48] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, Advances in neural information processing systems 33 (2020) 18661–18673.

[49] B. H. Ang, S. D. Gollapalli, S.-K. Ng, Unraveling online mental health using early maladaptive schemas: An ai-enabled study of online mental health forum communities, 2024. Under review.

[50] S. D. Gollapalli, B. H. Ang, S.-K. Ng, Identifying early maladaptive schemas from mental health question texts, in: The 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

[51] A. E. Harris, L. Curtin, Parental perceptions, early maladaptive schemas, and depressive symptoms in young adults, Cognitive therapy and research 26 (2002) 405–416.

[52] A. Tariq, C. Reid, S. W. Chan, A meta-analysis of the relationship between early maladaptive schemas and depression in adolescence and young adulthood, Psychological Medicine 51 (2021) 1233–1248.

[53] M. Balsamo, L. Carlucci, M. R. Sergi, K. Klein Murdock, A. Saggino, The mediating role of early maladaptive schemas in the relation between co-rumination and depression in young adults, PloS one 10 (2015) e140177.

[54] G. L. Flett, A. L. Goldstein, P. L. Hewitt, C. Wekerle, Predictors of deliberate self-harm behavior among emerging adolescents: An initial test of a self-punitiveness model, Current psychology 31 (2012) 49–64.

[55] Y. Li, Z. Liu, C. Xiong, Z. Liu, More robust dense retrieval with contrastive dual learning, in: Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval, 2021, pp. 287–296.

[56] H. Zhang, A. Sun, W. Jing, G. Nan, L. Zhen, J. T. Zhou, R. S. M. Goh, Video corpus moment retrieval with contrastive learning, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 685–695.

A. Early Maladaptive Schemas (EMS)

Table 5
Descriptions of EMS and their Examples [35]

EMS	Description	Example
Abandonment/Instability	The perception that others will not be reliable to provide emotional support or connection, leading to a sense of instability and insecurity.	I worry that people I feel close to will leave me or abandon me.
Approval-Seeking/Recognition-Seeking	A constant desire for approval, recognition, or acceptance from others to the extent that one compromises their authentic self.	Unless I get a lot of attention from others, I feel less important.
Defectiveness/Shame	A deep-seated belief that one is flawed, unlovable, or inferior, leading to feelings of shame and inadequacy.	I cannot understand how anyone could love me.
Dependence/Incompetence	A belief in one's inability to handle daily responsibilities or make appropriate decisions independently.	I feel that I need someone I can rely on to give me advice about practical issues.
Emotional Deprivation	The belief that one's emotional needs will not be adequately met by others.	For much of my life, I haven't felt that I am special to someone.
Emotional Inhibition	The suppression or inhibition of emotions, often due to a fear of losing control or being overwhelmed by intense feelings.	People see me as uptight emotionally.
Enmeshment/Undeveloped Self	An overly intense emotional attachment and proximity to one or more important individuals (typically parents), which hinders the process of becoming an independent individual and impedes normal social development.	I often found myself constantly turning to my parents for approval in every decision I made.
Entitlement/Grandiosity	An exaggerated sense of self-importance and a belief that one deserves special treatment.	I get very irritated when people won't do what I ask of them.
Failure to Achieve	The belief that one has failed or will inevitably fail in achieving personal and practitioner goals, leading to a sense of underachievement.	I am humiliated by my failures and inadequacies in the work (or school) sphere.
Insufficient Self-Control/Self-Discipline	An inability to control impulses, leading to difficulty in achieving long-term goals.	I often do things impulsively that I later regret.
Mistrust/Abuse	A belief that others are likely to deceive, take advantage of, or harm the person emotionally or physically.	I have been physically, emotionally, or sexually abused by important people in my life.
Negativity/Pessimism	A pervasive focus on the negative aspects of life, while neglecting the positive aspects.	People close to me consider me a worrier.
Punitiveness	The belief that harsh punishment is deserved for perceived mistakes or failures.	I'm a bad person who deserves to be punished.
Self-Sacrifice	The belief that one must meet the needs of others at the expense of one's own needs and desires.	No matter how much I give, I feel it is never enough.
Social Isolation/Alienation	The belief that one is different from others, leading to a sense of isolation and difficulty connecting with others.	I always feel on the outside of groups.
Subjugation	A tendency to prioritize others' needs and desires over one's own because one feels coerced.	I let other people have their way, because I fear the consequences.
Unrelenting Standards/Hyper-criticalness	The belief that one must meet high standards of performance, usually to avoid criticism.	I can't let myself off the hook easily or make excuses for my mistakes.
Vulnerability to Harm or Illness	The belief that catastrophic events are imminent, leading to excessive worry and Anxiety about potential harm or illness.	I feel that the world is a dangerous place.