Semantic Similarity Models for Depression Severity Estimation

Anonymous ACL submission

Abstract

Depressive disorders constitute a severe public health issue worldwide. However, public health systems have limited capacity for case detection and diagnosis. In this regard, the widespread use of social media has opened up a way to access public information on a large scale. Computational methods can serve as support tools for rapid screening by exploiting this user-generated social media content. This paper presents an efficient semantic pipeline to study depression severity in individuals based on their social media writings. We select test user sentences for producing semantic rankings over an index of representative training sentences corresponding to depressive symptoms and severity levels. Then, we use the sentences from those results as evidence for predicting symptoms severity. For that, we explore different aggregation methods to answer one of four Beck Depression Inventory (BDI-II) options per symptom. We evaluate our methods on two Reddit-based benchmarks, achieving 30% improvement over state of the art in terms of measuring depression level[1].

1 Introduction

Around two-thirds of all cases of depression remain undiagnosed according to conservative estimates (Epstein et al., 2010). To help with this problem, governments and agencies have launched programs to raise awareness of mental health in their citizens (Arango et al., 2018). In this context, detecting and receiving appropriate treatment in the early stages of these diseases is essential to reduce their impact and case escalation (Picardi et al., 2016). However, the insufficient resources of the public health systems severely limit their capacity for case detection and diagnosis.

As an alternative to public health systems, social platforms are a promising channel to assess risks in an unobtrusive manner (De Choudhury et al., 2013), where people tend to consider these platforms as comfortable media to express their feelings and concerns (Chancellor and De Choudhury, 2020). Exploiting this type of user-generated content, NLP techniques have shown promising results in terms of identifying depressive patterns and linguistic markers (Ríssola et al., 2021). Due to the growing popularity of the mental health detection models, the community also produced diverse datasets (Yates et al., 2017; Cohan et al., 2018), with the Early Risk Prediction on the Internet (eRisk) (Crestani et al., 2022), and the Computational Linguistics and Clinical Psychology (CLPsych) (Zirikly et al., 2022) being the two most popular benchmarks in the field. They produce task definitions, datasets and evaluation methodologies to encourage research in this domain.

Depression identification from social media posts faces challenges considering their integration into clinical settings (Walsh et al., 2020). Previous studies formulated this task as a binary classification problem (i.e., depressed vs control users) (Ríssola et al., 2021). Despite achieving remarkable results under this setting, ignoring different levels of depression limits the capacity to prioritize users with higher risks (Naseem et al., 2022). Moreover, most existing approaches focused on the use of engineered features, which may be more difficult to interpret than other clinical markers[2], such as the integration of recognized depressive symptoms (Mowery et al., 2017). Similarly, the black-box nature of deep learning models also limits the ability to understand their decisions, especially by domain experts like clinicians.

In this paper, we perform a fine-grained analysis of depression severity using semantic features to detect the presence of symptom markers. Our methods adhere to accepted clinical protocols by automatically completing the BDI-II (Dozois et al., 1998), a questionnaire used to measure depression.

[1] Implementation at *(restricted for anonymity)*.

[2] An observable sign indicative of a depressive tendency.

The BDI-II includes 21 recognized symptoms, such as sadness, fatigue or sleep issues. Each symptom has four alternative responses scaled in severity from 0 to 3. Using a sentence-based pipeline, we build 21 different symptom-classifiers for estimating the user responses to the symptoms. For this purpose, we employ eRisk collections related to depression levels (Losada et al., 2019, 2020; Parapar et al., 2021). In our pipeline, we explore selection algorithms to filter relevant sentences to each BDI-II symptom from training users. Once filtered, we index these training sentences with the user responses as labels (0 - 3) as examples of how people with different severity speak about the symptom. Then, to predict test users responses, we select their relevant sentences, which serve as queries to produce a semantic ranking over the indexed training sentences. Finally, we construct two aggregation methods based on the ranking results to estimate the symptoms severity.

The main contributions of this work are: 1) We present a semantic retrieval pipeline to perform a fine-grained classification of the severity of depressive symptoms. Following the symptoms covered by the BDI-II, our methods also consider different depression severity levels, distinguishing between lower and higher risks. 2) We propose a data selection process using a range of unsupervised and semi-supervised selection strategies to filter relevant sentences for the symptoms. 3) Experiments using different variants of our pipeline achieved remarkable results in two eRisk collections, outperforming state of the art considering depression severity in both datasets.

2 Related Work

Extensive research investigated the use of engineered features to identify linguistic markers and patterns related to mental disorders from different social platforms (Gaur et al., 2021; Coppersmith et al., 2014; Yates et al., 2017). For instance, the LIWC writing analysis tool (Pennebaker et al., 2003), equipped with psychological categories, revealed remarkable differences in writing style between depression and control groups (De Choudhury et al., 2013). Other studies used depression and emotional lexicons to determine depression markers (Cacheda et al., 2019), whereas Trotzek et al. (2018) examined additional distinctive features, leveraging profile metadata (e.g. posting hours or posts length) and social activity to examine the mental state of individuals.

The recent advances in contextualized embeddings significantly impacted many NLP-related tasks, including depression detection in social media. These deep learning models have consistently outperformed engineered features on diverse datasets (Jiang et al., 2020; Nguyen et al., 2022). However, they lack the interpretability clinicians require to rely on the results from automated screening methods (Amini and Kosseim, 2020). To enhance interpretability, the works proposed to the *eRisk depression estimation shared task* based their efforts on predicting BDI-II symptoms responses (Uban and Rosso, 2020; Spartalis et al., 2021; Basile et al., 2021). This study is directly related to eRisk, as we work with the eRisk collections and follow the same evaluation methodology. In contrast to these approaches, our methods highlight the user posts that lead to every symptom decision, which may be helpful for further inspection of model predictions.

Besides the works presented to eRisk, two recent studies explored the use of depressive symptoms to screen social media posts. Zhang et al. (2022a) aggregated symptoms from different questionnaires into a BERT-base model to calculate symptom risk at post level. Nguyen et al. (2022) experimented with various methods using symptom markers to detect depression, demonstrating their potential to improve the generalization and interpretability of their approaches. In this case, authors considered the symptoms from the PH9Q questionnaire (Kroenke et al., 2001) to define manual pattern-based strategies and train symptom-classifiers at post level. Both approaches formulated their methods with a binary classification setting, while our approach considers different severity levels. We also differ in that we pre-compute dense representations of training posts, rather than relying on pre-trained language models, which may be slow for many practical cases (Reimers and Gurevych, 2019). Our approach only needs a few post encodings and cosine similarity calculations, improving the efficiency of our solutions.

3 Method

Problem definition: We aim to estimate the depression severity level for users based on the Writing History (WH) of their social media posts. We define depression severity levels following the clinical classification schema of the BDI-II score (Lasa

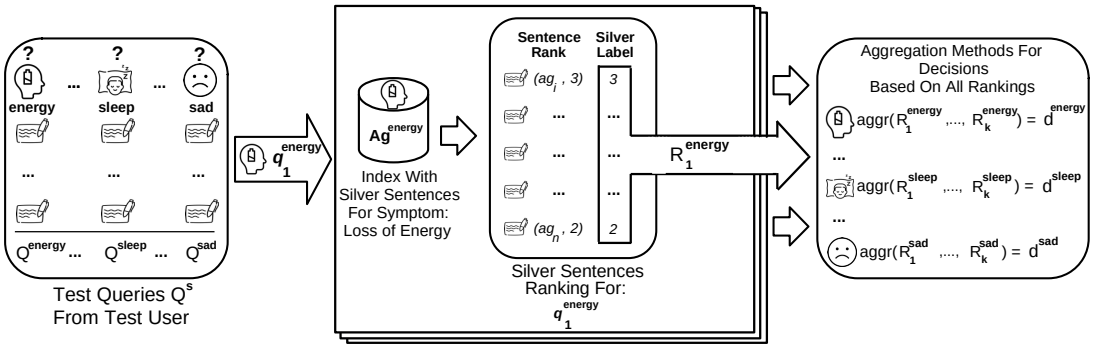

Figure 1: Retrieval pipeline to predict symptom options for a test user. R_1^{energy} is the list with the top ranked silver sentences for the query q_1^{energy}. Each silver sentence from the rank has a silver label associated (0-3). d^{energy} represents the option decision for that symptom based on the ranking retrieved for all the test queries, Q^{energy}.

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

Table 1: Depressive levels related to the BDI-II score.

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

Table 2: BDI-II options for the symptom *Loss of energy*.

et al., 2000). The score is the sum of the option responses to the 21 symptoms covered by this questionnaire, and it is associated with four depression levels. Table 1 shows these levels.

Instead of relying on a unique classifier to calculate that score, we build 21 different symptom-classifiers (i.e., one for each symptom). For this purpose, we categorize the symptoms into one of its four response options. Therefore, we formulate each classifier as a multi-class classification problem. Table 2 provides an example of the option descriptions for the symptom *Loss of energy*. To estimate the depression level for a user, we aggregate the predicted responses of all the symptom-classifiers.

Our approach relies on two critical components: 1) a semantic retrieval pipeline (§3.1) and 2) silver sentences selection (§3.2). The symptom-classifiers follow a semantic retrieval pipeline to predict every symptom decision. This pipeline searches for semantic similarities over an index of silver sentences for a specific symptom s, denoted as Ag^s. These silver sentences are considered relevant to s, and each one has a label corresponding to the symptom options[3], o. Formally, Ag^s is the set containing the pairs of the silver sentences ag_i and their corresponding label o_i for the symptom s, where $Ag^s = \{(ag_i, o_i)\}$, and $o_i \in \{0, 1, 2, 3\}$.

To the best of our knowledge, there are no datasets in the literature where sentences are relevant to the symptom and labelled by their severity. For this reason, we propose a selection process to create the silver sentences, Ag^s, where we use as training data the eRisk collections (§3.2). In our experiments, we explore the performance of the semantic pipeline with our generated silver sentences. However, we could apply this pipeline to any similar datasets. In the following subsections, we explain both components in detail.

3.1 Semantic Retrieval Pipeline

Using the writing history from a test user as input, our semantic retrieval pipeline classifies its label severity for a specific symptom s. From the publications of the test user, we first select the sentences that are relevant to s, which will serve as queries. We denote these relevant sentences as the symptom test queries Q^s, where $Q^s = \{q_1^s, ..., q_k^s\}$, since we select a top k of them. In the next subsection, we explain our sentence selection algorithms (§3.2). The top k of queries are the input to our semantic pipeline. Figure 1 illustrates this process, exemplified for one test query, q_1^{energy}, corresponding to the symptom *Loss of energy*:

1) The first step consists in calculating a semantic ranking for each test query for the symptom s, defined as q_i^s. To calculate that ranking, we en-

[3]Throughout the rest of the paper, we will refer to these severity options as the labels of the symptoms.

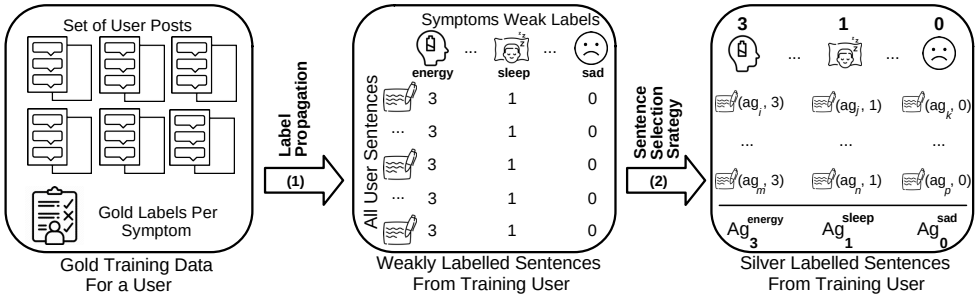

Figure 2: From the responses of the eRisk training users to the symptom options (0-3), the silver selection process creates one different set of silver sentences relevant to each symptom s and option o, denoted as Ag^s_o.

code q^s_i and all the silver sentences in Ag^s for the symptom as embeddings. Then, we use k-Nearest Neighbours (kNN) to compute the semantic similarity of each silver sentence w.r.t the test query q^s_i. The semantic similarity sm for a silver sentence ag_j, belonging to Ag^s, and a test query q^s_i, is the cosine similarity between their embeddings (ϕ):

$$sm(ag_j, q^s_i) = cos\big(\phi(ag_j), \phi(q^s_i)\big) \quad (1)$$

Computing sm, we produce a ranking of silver sentences, R^s_i, for each test query q^s_i. The silver sentences in the ranking have an associated silver label. For example, the position j of the ranking contains the pair: $R^s_i[j] = \{(ag_j, o_j)\}$, with $o_j \in \{0, 1, 2, 3\}$. To select the cut-off of the rankings R^s_i, we experimented with a varying number of similarity thresholds. To calculate the embeddings, we use a pre-trained model based on RoBERTa[4] using sentence-transformers (SBERT).

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

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

$$f_{av}(o) = \sum_{i \in R^s_i} \sum_{j=1}^{n} \begin{cases} 1 & \{(ag_j, o_j) | o_j = o\} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Accumulative Recall: For each ranking R^s_i, compute the recall for each label o. That is, the fraction of silver sentences in the ranking out of all the available silver sentences from that label, denoted as Ag^s_o, where $Ag^s_o = \{(ag_i, o_i) | o_i = o\}$. Then, we accumulate the recall over the rankings R^s_i. The final decision is $d^s = \underset{o}{\text{argmax}} f_{ar}(o)$ with:

$$f_{ar}(o) = \sum_{i \in R^s_i} \frac{\sum_{j=1}^{n} \begin{cases} 1 & \{(ag_j, o_j) | o_j = o\} \\ 0 & \text{otherwise} \end{cases}}{|Ag^s_o|} \quad (3)$$

3.2 Silver Sentences Selection

We design a process to select relevant sentences for each symptom s, and the severity labels o (previously denoted as Ag^s_o), defined as silver sentences. For this purpose, we use the eRisk collections as training data. These collections contain users from the Reddit platform, and have two main elements: i) the user responses to the BDI-II symptoms and ii) their posts from Reddit. We use the option responses from the users as the severity labels for each symptom $(0-3)$. Therefore, the training labels are initially available at user level. Details of eRisk datasets are provided in Section 4.1.

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

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

non-informative w.r.t the assigned weak label. We implement two different strategies:

Option descriptions as queries: This strategy works in an unsupervised manner, since we consider the option descriptions from the symptoms as queries to select the silver sentences. Table 2 shows an example of the descriptions for the symptom *Loss of energy*. We use each option description as one different query. Based on the sentences retrieved from these queries, we select a top of sentences from the eRisk training users who answered the same option used as the query. Following this approach, we perform lexical and semantic retrieval variants. For lexical search, we use BM25 (Robertson et al., 1995) to retrieve relevant sentences for each training user. In the semantic variant, we calculate the similarity based on a semantic threshold, as described in the semantic ranking (§3.1), using the same RoBERTa model for the semantic search.

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

We used the eRisk2019 training users to obtain the golden sentences. Following the approach by Karisani and Agichtein (2018), three experts in the field conducted the annotation process. The number of golden sentences was low, averaging 35 per symptom. The data augmentation process consisted of, for every golden sentence belonging to a specific option, following the semantic ranking (§3.1) over the rest of the weakly-labelled sentences from that same option. The final set of relevant sentences combines the golden and the silver sentences that surpass the similarity threshold. Table 3 shows an example of a golden sentence along with the top 3 augmented silver sentences. The golden sentence corresponds to the option 3 for the symptom *Pessimism in the future*, and the augmented silver sentences correspond to other training users who reported the same option[5].

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

Table 3: Examples of augmented silver sentences with highest semantic similarity to the golden sentence.

4 Experimental Settings

We evaluate the performance of our methods in the eRisk2020 and 2021 collections. In eRisk2020, we use 2019 as training data. In eRisk2021, we use the 2019 and 2020 collections as training. The competing methods used the same collection splits, while some of them also considered external datasets (§4.2). In our experiments, we study the two components of our approach: i) the performance of the semantic retrieval pipeline (§3.1) and ii) the effectiveness of the sentence selection strategies (§3.2). For this reason, our methods consist of combinations of these components. We consider three hyperparameters: 1) The value k of the number of test queries, $Q^s = \{q_1^s, ..., q_k^s\}$. 2) The semantic threshold to select the cut-off of the rankings, R_i^s. 3) The number of silver sentences to generate the silver dataset, Ag^s. The specific hyperparameters and the tuning process are described in Appendix B.

4.1 Datasets

The collections selected for experiments correspond to the data delivered for the *eRisk depression severity estimation task* in 2019, 2020 and 2021 editions (Losada et al., 2019, 2020; Parapar et al., 2021). We rely on these collections as they are adopted as the benchmark for this task and contain real answers from Reddit users to the BDI-II. Table 5 summarizes the main statistics of these datasets.

4.2 Evaluation

Evaluation Metrics. We use the official metrics proposed in the eRisk benchmark (Losada et al., 2019) to keep a fair comparison against the competing methods. These metrics assess the quality

[5]Information about the dataset construction and the annotation process can be found in Appendix A.

Collection	Model		External Dataset	Questionnaire Metrics			Symptom Metrics	
				DCHR (↑)	ADODL (↑)	RMSE (↓)	AHR (↑)	ACR (↑)
eRisk 2020	BioInfo (Trifan et al., 2020) (a)		✓	30.00	76.01	18.78	38.30	69.21
	ILab (Castaño et al., 2020) (b)		✓	27.14	81.70	14.89	37.07	69.41
	Relai (Maupomé et al., 2020) (c)		✓	34.29	83.15	14.37	36.39	68.32
	UPV (Uban and Rosso, 2020) (d)		✗	35.71	80.63	15.40	34.56	67.44
	Sense2vec (Pérez et al., 2022) (e)		✗	37.14	82.61	12.40	**38.97**	**70.10**
	Aggregation	**Silver Sentences Selection**						
	Accum Voting	BM25	✗	$38.57^{a,b,c,d,e}$	85.19	12.37	35.24	68.37
	Accum Recall	BM25	✗	$40.00^{a,b,c,d,e}$	84.65	12.13	35.71	67.60
	Accum Voting	SBERT	✗	$42.86^{a,c,d}$	83.08	14.25	34.83	65.90
	Accum Recall	SBERT	✗	$42.86^{a,b,c,e}$	84.51	12.37	33.33	66.05
	Accum Voting	Aug Dataset	✗	$47.14^{a,b,c,d,e}$	**85.33**	**11.87**	35.24	67.41
	Accum Recall	Aug Dataset	✗	$\mathbf{50.00}^{a,b,c,d}$	85.24	12.09	35.44	67.23
eRisk 2021	DUTH (Spartalis et al., 2021) (a)		✗	15.00	73.97	19.60	**35.36**	67.18
	Symanto (Basile et al., 2021) (b)		✓	32.50	82.42	14.46	34.17	**73.17**
	CYUT (Wu and Qiu, 2021) (c)		✗	41.25	**83.59**	**12.78**	32.62	69.46
	Aggregation	**Silver Sentences Selection**						
	Accum Voting	BM25	✗	$45.00^{a,b,c}$	82.16	14.11	30.97	64.54
	Accum Recall	BM25	✗	$42.50^{a,b,c}$	80.62	15.13	28.03	62.92
	Accum Voting	SBERT	✗	$45.00^{a,b,c}$	81.92	14.15	29.67	64.27
	Accum Recall	SBERT	✗	$41.25^{a,b,c}$	81.86	14.2	27.47	62.89
	Accum Voting	Aug Dataset	✗	$46.25^{a,b,c}$	81.72	14.83	27.95	62.40
	Accum Recall	Aug Dataset	✗	$\mathbf{51.25}^{a,b,c}$	81.65	14.96	27.66	61.72

Table 4: Results on eRisk collections. The numbers of the official metrics are in percentage. Best values are bolded. Methods using external datasets for training the model are marked. Statistical significant differences in the severity level category assignment according to the Stuart-Maxwell marginal homogeneity test w.r.t to the baselines are super-scripted (p-values < 0.05). For the remaining metrics, we found no statistically significant differences.

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

Table 5: Statistics of the eRisk collections.

of a questionnaire estimated by a system compared to the real one reported by the user. They include two evaluations: 1) At questionnaire level, the Depression Category Hit Rate (**DCHR**) computes the percentage of depressive levels correctly estimated, and the Difference Between Overall Depression Levels (**ADODL**) computes the overall estimations of the BDI-II score. 2) On the other hand, The Average Hit Rate (**AHR**) and Average Closeness Rate (**ACR**) assess the results at symptom level. Apart from the eRisk evaluation, we include one additional error metric: the Root-Mean-Square Error (**RMSE**) (Chai and Draxler, 2014) to compare the models predictions of the BDI-II score. Thus, the lower the value reported by RMSE, the lower the difference between predictions and real scores are.

Competing Methods. We consider the best prior works for each metric for the eRisk2020/2021 collections. We refer the reader to the corresponding shared task surveys for a detailed analysis (Losada et al., 2020; Parapar et al., 2021). In eRisk2020, BioInfo (Trifan et al., 2020) and Relai (Maupomé et al., 2020) methods obtained their own datasets to perform standard ML classifiers using engineered features as linguistic markers. Other deep learning approaches, such as ILab (Castaño et al., 2020) and UPV (Uban and Rosso, 2020), focused their efforts on the use of large language models (LLMs) explicitly trained for depression severity estimation. Finally, a recent work by Pérez et al. (2022) (Sense2vec) designed different word embedding models for each of the symptoms and achieved state-of-the-art results in this dataset. In eRisk2021, Symanto (Basile et al., 2021) team trained a neural model with additional data annotated by psychologists and combined it with a set of engineered features, whereas Wu and Qiu (2021) (CYUT) experimented with different RoBERTa classifiers. Similar to our work, Spartalis et al. (2021) (DUTH) used semantic features with sentence transformers to extract one dense representation per user, which is then fed as input, experimenting with various classifiers. Although insightful, eRisk approaches

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

(a) Number of golden and the augmented silver sentences for each severity level and their F1 using *Accum Recall-Aug Dataset*.

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

(b) F1 results in a binary classification scenario using the *Accum Recall-Aug Dataset* variant and the best prior model.

Table 6: (a) Data augmentation effects and (b) F1 results considering a binary classification setting.

cannot evidence the sentences that lead to symptom decisions.

5 Results and Discussion

Table 4 compares the results of all the variants of our approach against the competing methods. These variants are the combination of our two aggregation methods (*Accum Voting* and *Accum Recall*) and the sentences selection strategies (*BM25*, *SBERT* and the augmented dataset, *Aug Dataset*). The comparison is based on the use of questionnaire and symptom level metrics (§4.2).

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

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

Finally, with respect to the sentence selection strategies, we can observe that using the options descriptions as queries (*BM25* and *SBERT*) performs worse than the augmented dataset (*Aug Dataset*). This emphasizes the importance of a precise candidate selection. Moreover, despite the distribution of depression levels varies in both collections (see Table 5), our methods show robustness as we keep achieving good performance in DCHR.

5.1 Effect of Data Augmentation Strategy

To better understand the performance of the data augmentation, we report the number of augmented silver sentences along with the F1 metric for each depression level. Table 6a shows the F1 results of our best variant using the augmented dataset, *Accum Recall-Aug Dataset*, in eRisk2020 and 2021. Looking at the statistics, we see more presence in golden sentences of high-risk levels (moderate and severe). In addition, the number of silver sentences augmented for each of them is also higher. For example, using eRisk2019 as the training set, an average of three silver sentences were augmented from each golden one in the minimal level ($\frac{310}{98} \approx 3$). In contrast, the average of silver sentences augmented from the severe category is 7 ($\frac{2414}{354} \approx 7$). This suggests that users with higher depressive levels tend to manifest more explicit thoughts related to the symptoms. As a result, our augmentation method finds pieces of evidence in these levels easier.

If we observe the F1 results in Table 6a, we

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

Table 7: Example of the top query sentences from a test user for two symptoms along with the golden option response of that user and the predicted option of our method.

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

Table 8: Examples of retrieved silver sentences with their assigned label from two test queries from a user.

also see considerable variability among depressive levels. In both collections, we achieve better results for higher risk categories. This seems to be related to the number of golden sentences. Therefore, if we obtain more samples belonging to the lower risk levels, there may be an improvement in these categories. Finally, we examine our results with a binary classification setting. For this purpose, we categorize the four depression levels into only two: 1) *low risk* (minimal + mild levels) and 2) *high risk* (moderate + severe levels). Table 6b shows the results for the *Accum Recall-Aug Dataset* variant along with the best prior work under this setting. Our results suggest the effectiveness of our method, which distinguishes with fair accuracy between higher and lower risks.

5.2 Interpretability - Case Study

The lack of reliable clinical markers is one of the barriers to the practical use of mental health prediction models (Walsh et al., 2020; Amini and Kosseim, 2020). By considering a more refined grain in the symptom presence, we provide valuable information that may be strong clinical markers. Table 7 showcases how our approach offers interpretability of the symptom decisions, showing three query sentences from an anonymized test user. The symptoms in the Table are *Sleep problems* and *Loss of pleasure*, and the user declared the option 1 and 3 for them, respectively. We can see that these test queries are robust indicators of symptom concerns. Following this approach, clinicians may inspect sentences as a first step towards further diagnosis or monitoring methods during treatment.

In addition, Table 8 displays some of the silver sentences retrieved for the same test queries selected from the anonymized user. The silver sentences are related to the content of the query, and clinicians may evaluate the justifications for every symptom decision by reviewing their labels. Moreover, in our method, false positive/negative predictions can still be helpful for future inspection. For example, for the symptom *Sleep problems*, the test user reported the option 1, but our method retrieved more silver sentences with the option 2. While the prediction may be incorrect (golden label $(1) \neq$ predicted label (2)), the risk may still be present.

6 Conclusions

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

7 Ethical Statement

The collections used in this work are publicly available following the data usage policies. They were collected in a manner that falls under the exempt status outlined in Title 45 CFR §46.104. Exempt research includes research involving the collection or study of existing data, documents, records, or specimens if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified. We adhered to the corresponding policies and took measures to ensure that personal information could not be identified from the data. The data is available by filling a user agreement according to the eRisk shared task policies[6]. In this context, all users have an anonymous state. We paraphrased the reproduced writings to preserve their privacy.

In terms of impact in real-world settings, there is still work to be done to produce effective depression screening tools. The development of such technologies should be approached with caution to ensure that their use is ethical and respects patient privacy and autonomy. Our work aims to supplement the efforts of health professionals rather than replace them. We acknowledge the validation gap between mental health detection models and their clinical applicability. Our goal is to develop automated technologies that can complement current online screening approaches. To ensure safe implementation and as future work, we collaborate with clinicians to validate and obtain a more in-depth analysis of the limitations of these systems.

We took several measures to ensure the objectivity and reliability of our annotations, including providing the same guidelines to all annotators and using a majority vote system to resolve any disagreements. While one of the annotators is also an author of this study, we want to emphasize that they were not given any preferential treatment or guidelines that differed from the other annotators. Moreover, the high agreement percentage among the three annotators (as reported in our study) further supports the reliability and objectivity of our process. Overall, our annotation process was conducted objectively and reliably and the potential for bias was minimized to the greatest extent possible.

Despite being experts in the field, we recognize that annotating depressive symptoms may have an impact on annotators. We provided them with the necessary breaks and did not subject them to any time constraints. Annotators did not report any negative effects after their work. In addition, they were not biased in scoring a higher or lower number of positive sentences.

We recognize that the application of NLP models in real-world scenarios requires careful consideration and analysis due to potential risks and limitations. These models should not be immediately deployed as decision support systems without further studies, including participant recruitment, trials, and regulatory approval. To address potential risks, we consulted with clinical experts to validate the possible dual-use risks before conducting this work and involved them in designing annotation guidelines and all stages of the study. One of the main risks associated with such systems is their performance, as they may produce false positive/negative cases. However, it is important to note that diagnostic discrepancies and their risks associated are common in the clinical setting (Regier et al., 2013).

Therefore, our system is intended to be used in conjunction with health professionals to obtain a more accurate diagnosis. The final decision must always be supported by the validation of a health professional. Our study highlights the potential of NLP-based approaches in assisting clinicians with diagnosis, but further research and testing are needed before it can be considered for clinical deployment.

8 Limitations

We recognize that the performance of our solutions is far from ideal to be integrated directly in clinical settings. Moreover, it lacks external validity (Ernala et al., 2019), as they were never tested in real clinical scenarios. The dataset used in this study (corresponding to the eRisk collections) has a limited size (170 users in total) and diversity, since it only covers one social platform, Reddit. This is partly due to the protection necessary for securing sensitive data related to mental health (Harrigian et al., 2020). We chose the BDI-II questionnaire for our study because it is the only questionnaire with an available dataset that contains both (1) user responses on each symptom and (2) their writing history on social media (Reddit, in this case). Other questionnaires in clinical practice are also widely used (CES-D, GDS, HADS, and PHQ-9). However, we do not have a dataset containing the respondents' answers to the symptoms. As future work,

[6] https://erisk.irlab.org/2021/eRisk2021.html

we will work on extending this same pipeline to other related questionnaires.

We are also aware that social media platforms provide an imperfect representation of the population, which is a clear limitation that must be accounted for when using these approaches for public health screening. For this reason, our methods are likely to be modified when other data sources (i.e., other social platforms) are considered. Moreover, when processing data from different clinical contexts (e.g., clinical records), the models may generalize inadequately (Harrigian et al., 2020). Another limitation of our work is the low performance of the symptom evaluation compared with the questionnaire level (related to depression severity levels). As we previously commented, we did not focus our efforts on tuning individual symptom-classifiers but rather to use them as a proxy to estimate depression levels, since we do not have enough training data for most of the symptoms. We also believe that certain errors in these symptom estimations may be due to a lack of awareness of the individual or stigmas associated with the different symptoms.

Despite the gap between mental health prediction models and actual clinical practice, many recent studies (Zhang et al., 2022b,a; Yates et al., 2017; Pérez et al., 2022) investigated approaches to identifying and detecting depression using reliable clinical markers. We can also see other studies that adhered to clinical questionnaires to investigate other related features such as personality detection (Yang et al., 2021). These studies seek to propose solutions that can be a proxy between health professionals and NLP methods. Our study aims to contribute to this area of research and advance the development of reliable solutions for health professionals.

References

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

Celso Arango, Covadonga M Díaz-Caneja, Patrick D McGorry, Judith Rapoport, Iris E Sommer, Jacob A Vorstman, David McDaid, Oscar Marín, Elena Serrano-Drozdowskyj, Robert Freedman, et al. 2018. Preventive strategies for mental health. *The Lancet Psychiatry*, 5(7):591–604.

Angelo Basile, Mara Chinea-Rios, Ana Sabina Uban, Thomas Müller, Luise Rössler, Seren Yenikent, Berta Chulvi, Paolo Rosso, and Marc Franco-Salvador. 2021. UPV-Symanto at eRisk 2021: Mental Health Author Profiling for Early Risk Prediction on the Internet. In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 908–927. CEUR-WS.org.

Fidel Cacheda, Diego Fernandez, Francisco J Novoa, Victor Carneiro, et al. 2019. Early detection of depression: social network analysis and random forest techniques. *Journal of medical Internet research*, 21(6):e12554.

Rodrigo Castaño, Amal Htait, Leif Azzopardi, and Yashar Moshfeghi. 2020. Early risk detection of self-harm and depression severity using BERT-based transformers: iLab at CLEF eRisk 2020. *Early Risk Prediction on the Internet*.

Tianfeng Chai and Roland R Draxler. 2014. Root mean square error (rmse) or mean absolute error (mae). *Geoscientific Model Development Discussions*, 7(1):1525–1534.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):1–11.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomedical Informatics Insights*, 10:117822261879286.

Fabio Crestani, David E. Losada, and Javier Parapar, editors. 2022. *Early Detection of Mental Health Disorders by Social Media Monitoring*. Springer International Publishing.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social Media as a Measurement Tool of Depression in Populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, page 47–56, New York, NY, USA. Association for Computing Machinery.

David JA Dozois, Keith S Dobson, and Jamie L Ahnberg. 1998. A psychometric evaluation of the Beck Depression Inventory–II. *Psychological assessment*, 10(2):83.

Ronald M Epstein, Paul R Duberstein, Mitchell D Feldman, Aaron B Rochlen, Robert A Bell, Richard L Kravitz, Camille Cipri, Jennifer D Becker, Patricia M Bamonti, and Debora A Paterniti. 2010. "I didn't know what was wrong": How people with undiagnosed depression recognize, name and explain their distress. *Journal of general internal medicine*, 25(9):954–961.

Sindhu Kiranmai Ernala, Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. CHI '19, page 1–16, New York, NY, USA. Association for Computing Machinery.

Manas Gaur, Vamsi Aribandi, Amanuel Alambo, Ugur Kursuncu, Krishnaprasad Thirunarayan, Jonathan Beich, Jyotishman Pathak, and Amit Sheth. 2021. Characterization of time-variant and time-invariant assessment of suicidality on reddit using c-ssrs. *PloS one*, 16(5):e0250448.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3774–3788, Online. Association for Computational Linguistics.

Zhengping Jiang, Sarah Ita Levitan, Jonathan Zomick, and Julia Hirschberg. 2020. Detection of mental health from Reddit via deep contextualized representations. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 147–156, Online. Association for Computational Linguistics.

Payam Karisani and Eugene Agichtein. 2018. Did You Really Just Have a Heart Attack? Towards Robust Detection of Personal Health Mentions in Social Media. WWW '18, page 137–146, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

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

Lourdes Lasa, Jose L Ayuso-Mateos, Jose L Vázquez-Barquero, FJ Dıez-Manrique, and Christopher F Dowrick. 2000. The use of the Beck Depression Inventory to screen for depression in the general population: a preliminary analysis. *Journal of affective disorders*, 57(1-3):261–265.

David E Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of erisk 2019 early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 340–357. Springer.

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

Diego Maupomé, Maxime D Armstrong, Raouf Moncef Belbahar, Josselin Alezot, Rhon Balassiano, Marc Queudot, Sébastien Mosser, and Marie-Jean Meurs. 2020. Early Mental Health Risk Assessment through Writing Styles, Topics and Neural Models. In *CLEF (Working Notes)*.

Danielle Mowery, Hilary Smith, Tyler Cheney, Greg Stoddard, Glen Coppersmith, Craig Bryan, Mike Conway, et al. 2017. Understanding depressive symptoms and psychosocial stressors on twitter: a corpus-based study. *Journal of medical Internet research*, 19(2):e6895.

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

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8446–8459. Association for Computational Linguistics.

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

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

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

Anxo Pérez, Javier Parapar, and Álvaro Barreiro. 2022. Automatic depression score estimation with word embedding models. *Artificial Intelligence in Medicine*, 132:102380.

Darrel A Regier, William E Narrow, Diana E Clarke, Helena C Kraemer, S Janet Kuramoto, Emily A Kuhl, and David J Kupfer. 2013. Dsm-5 field trials in the united states and canada, part ii: test-retest reliability of selected categorical diagnoses. *American journal of psychiatry*, 170(1):59–70.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Paul Richter, Joachim Werner, Andrés Heerlein, Alfred Kraus, and Heinrich Sauer. 1998. On the validity of the beck depression inventory. *Psychopathology*, 31(3):160–168.

Esteban A. Ríssola, David E. Losada, and Fabio Crestani. 2021. A survey of computational methods for online mental state assessment on social media. *ACM Trans. Comput. Healthcare*, 2(2).

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp*, 109:109.

Christoforos Spartalis, George Drosatos, and Avi Arampatzis. 2021. Transfer Learning for Automated Responses to the BDI Questionnaire. In *Working Notes of CLEF 2021 – Conference and Labs of the Evaluation Forum*, volume 2936, pages 1046–1058, Bucharest, Romania.

Alina Trifan, Pedro Salgado, and José Luís Oliveira. 2020. BioInfo@UAVR at eRisk 2020: on the Use of Psycholinguistics Features and Machine Learning for the Classification and Quantification of Mental Diseases. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Marcel Trotzek, Sven Koitka, and Christoph Friedrich. 2018. Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32:588–601.

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

Colin G Walsh, Beenish Chaudhry, Prerna Dua, Kenneth W Goodman, Bonnie Kaplan, Ramakanth Kavuluru, Anthony Solomonides, and Vignesh Subbian. 2020. Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence. *JAMIA Open*, 3(1):9–15.

Shih-Hung Wu and Zhao-Jun Qiu. 2021. A RoBERTa-based model on measuring the severity of the signs of depression. In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 1071–1080. CEUR-WS.org.

Feifan Yang, Tao Yang, Xiaojun Quan, and Qinliang Su. 2021. Learning to answer psychological questionnaire for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 1131–1142. Association for Computational Linguistics.

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

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Q. Zhu. 2022a. Psychiatric scale guided risky post screening for early detection of depression. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5220–5226. ijcai.org.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Q Zhu. 2022b. Symptom identification for interpretable detection of multiple mental disorders. *arXiv preprint arXiv:2205.11308*.

Ayah Zirikly, Dana Atzil-Slonim, Maria Liakata, Steven Bedrick, Bart Desmet, Molly Ireland, Andrew Lee, Sean MacAvaney, Matthew Purver, Rebecca Resnik, and Andrew Yates, editors. 2022. *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics, Seattle, USA.

A Manual Dataset and Annotation Process

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

To annotate the golden sentences, we used as data source the training users from the eRisk2019 collection of depression severity (Losada et al., 2019). However, the large size of the eRisk collection requires an exhaustive filter for reasonable annotation efforts. For this purpose, we leveraged the data selection strategy of using the option descriptions as queries (§3.2). In particular, we applied the semantic retrieval variant (SBERT). Using this strategy, we selected candidate sentences

for annotating each BDI-II symptom. We have considered this strategy following a recent study that has shown great results in identifying diverse expressions of symptoms for candidate retrieval annotation (Zhang et al., 2022b). Previous studies on symptom annotation (Zhang et al., 2022b) demonstrated a high variance in the distribution of each symptom. For some of them, it is much easier to find representative sentences than for others. To keep the number of annotations per symptom stable, we fixed a similarity threshold of 0.6 to filter out sentences. However, this similarity threshold still produced too many candidate sentences for some symptoms. For this reason, we further restricted the annotator's work to the first 750 sentences in the symptoms with too many candidates.

More specifically, 17.15% of the candidate sentences have been labelled positive following the semantic retrieval strategy from the total of 5004 candidates. From the same labelled sentences, using keyword matching with BM25 reduced this percentage to 4%. With a random retrieval strategy, it dropped to 0.01% due to the small number of relevant sentences compared to the size of the entire pool. These findings align with previous research indicating that pattern matching is not effective in retrieving diverse sentences relevant to depressive symptoms (Mowery et al., 2017). Instead, a semantic similarity-based strategy is better suited to retrieve representative sentences without relying on specific keywords covered in the clinical questionnaires.

Following the above candidate annotation schema, we constructed a small dataset for all the BDI-II symptoms. The annotation task was carried out by two psychologists and two PhD students with knowledge in the field. Before the annotation process, we removed all supplementary metadata to avoid bias in the annotators, such as the severity option label $(0-3)$ of the user who wrote the sentence. We followed the same annotation procedure as Karisani and Agichtein (2018) to validate the annotation outcomes. This procedure consisted of two phases: 1) First, an initial annotator answered the following question in a binary setting (Positive/Negative): *Does the sentence refer to the symptom, and the user talks about himself/herself (first person)?*. This first annotator labelled a total of 738 positive sentences from the candidate sentences. We considered all the sentences annotated as positive for each symptom to obtain our final labels corresponding to the option levels $(0-3)$. Subsequently, we label these positive sentences with the severity option reported by the user who wrote them. Therefore, for each option o and symptom s, we obtained a different set of golden labels, G_o^s, where the sentences come from the eRisk users that answered the BDI-II symptoms.

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

Table 9 and 10 show the main statistics of our manual dataset. Visualizing these tables, we can extract several findings. We note that, for all the symptoms, the number of sentences associated with the option 0 is very low. In some symptoms, even none of the sentences corresponded to option 0. This suggests retrieving sentences representing positive feelings towards the symptom is more complicated. We attribute this fact to two main reasons, (i) the descriptions of BDI-II options 0 are not entirely appropriate for the candidate retrieval process (most of them are just negations of a negative feeling), and (ii) users are not as likely to talk about positive as they do with negative feelings. To address this, for the symptoms that lacked sentences with option 0, we manually included between 1 and 3 sentences that provide a positive description of the symptom and labelled them with option 0.

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

B Detailed Experimental Settings

We experimented with different hyperparameters to validate the results from our two main components: the semantic retrieval pipeline (§3.1) and the sentence selection process (§3.2). As we do not have a validation set, we performed leave-one-

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

Table 9: Annotations statistics of the first ten BDI-II symptoms.

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

Table 10: Annotations statistics of the last eleven BDI-II symptoms.

out cross-validation using the training set available to calculate the optimal values of all hyperparameters. The metric maximized was *DCHR*. When evaluating our methods in eRisk2020, the training set was the eRisk2019 dataset. When using as test collection the eRisk2021 dataset, the training set was the eRisk2019 and 2020 collections. Table 11 presents the hyperparameters and the optimal values for each method used in our experiments [7].

Semantic retrieval pipeline (§3.1). In the semantic pipeline, we experimented with two hyperparameters:

1) **The value k of the number of test queries.** We explored with selecting a different number of top k values of the user test queries, $Q^s = \{q_1^s, ..., q_k^s\}$. To select these test queries, we used the data selection strategies of using the option descriptions as queries (§3.2). Using BM25, the k values explored were: $[5, 10, 15, 20, 25, 30, 40]$. We also experimented with the same k values using the semantic variant. However, we did not include those results in the paper as they could not improve the use of BM25.

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

Silver sentences selection (§3.2). Additionally, we also experimented with a filtering hyperparameter for creating more or less restrictive filters when generating the silver dataset, denoted as *selection threshold*. Depending on the selection strategy (BM25, SBERT or Aug Dataset), we used the next sentence selection hyperparameters:

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

With respect to the sentence transformers models (Reimers and Gurevych, 2019), we experimented

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

with different pre-trained models: *msmarco-bert-base-dot-v5*[8], *msmarco-distilbert-cos-v5*[9], *all-roberta-large-v1*[10] *and stsb-roberta-large*[11] via the huggingface transformers library. All these models were fine-tuned on diverse semantic similarity datasets. In pilot experiments, the best results were obtained with the model *all-roberta-large-v1*. Thus, all our reported results correspond to the use of that model.

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

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

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

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

Test Set	Method	Number k of User Test Queries	Semantic Ranking Threshold	Sentence Selection Threshold
eRisk 2020	Accum Voting - BM25	25	0.55	Top k = 100
	Accum Recall - BM25	25	0.60	Top k = 100
	Accum Voting - SBERT	25	0.50	0.45
	Accum Recall - SBERT	25	0.50	0.35
	Accum Voting - Aug Dataset	40	0.55	0.50
	Accum Recall - Aug Dataset	35	0.55	0.50
eRisk 2021	Accum Voting - BM25	25	0.55	Top k = 100
	Accum Recall - BM25	25	0.55	Top k = 100
	Accum Voting - SBERT	25	0.50	0.50
	Accum Recall - SBERT	30	0.55	0.40
	Accum Voting - Aug Dataset	25	0.55	0.50
	Accum Recall - Aug Dataset	25	0.55	0.50

Table 11: Best hyperparameter values for all the variants considered in our methods. These values were obtained by performing leave-one-out cross-validation in the training set by maximizing the DCHR metric.