# uOttawa at eRisk 2023: Search for Symptoms of Depression

Notebook for the eRisk Lab at CLEF 2023

Yuxi Wang*, Diana Inkpen

*University of Ottawa, 800 King Edward, Ottawa, ON, K1N 6N5, Canada*

### Abstract

This paper introduces the University of Ottawa's participation in Task 1 of the eRisk 2023 shared task at CLEF 2023. As early intervention of depression becomes more and more important, we are striving to build a system to search for depression symptoms. By participating, we could evaluate the effectiveness of our search techniques and identify areas for improvement. Our methods focused on extracting relevant sentences for each symptom in the Beck's Depression Inventory questionnaire and providing a ranking for further investigation. To rank the sentences, we represented them as neural embedding vectors, then we computed their cosine similarity to query embedding vectors. We constructed one query for each of the 21 symptoms of interest, based on the corresponding question and possible answers in the questionnaire.

### Keywords

depression detection, social media analysis, information retrieval, natural language processing

## 1. Introduction

Social media has become an essential part of everyone's daily life, people use it as a platform to express their feelings about almost everything. Since depression has become a prevalent mental health issue, early detection of symptoms could greatly improve the chances of proper treatment. Traditional methods of detection, usually human-led, are expensive to conduct and might be individually biased. Our team, as a participant in Task 1 of the eRisk 2023 shared task at CLEF 2023, is aiming to design a method to analyze social media sentences and then help identify potential symptoms of depression as well as support early intervention.

We considered the task as a search/information retrieval task, where user-written sentences are stored as documents. The 21 questions from the Beck's Depression Inventory (BDI) questionnaire were transformed into 21 queries. The aim of the task is to retrieve the top-1000 relevant sentences for each query, and also compute their rankings (rank 1 being the most relevant).

Several text embedding methods were used for transfer learning, including contextual text embedding methods such as DistilBERT, and distributional word embedding method GloVe.

*CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece*

*Corresponding author.

ywan1225@uOttawa.ca (Y. Wang); diana.inkpen@uottawa.ca (D. Inkpen)
https://www.site.uottawa.ca/~diana/ (D. Inkpen)
0000-0002-6831-014X (Y. Wang); 0000-0002-0202-2444 (D. Inkpen)

**Table 1**
Statistics of the dataset

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

```
1     Q0    s-000-000-0      0001    0.967    uOttawa-Method1
1     Q0    s-111-111-1      0002    0.966    uOttawa-Method1
1     Q0    s-222-222-2      0003    0.961    uOttawa-Method1
1     Q0    s-333-333-3      0004    0.961    uOttawa-Method1
1     Q0    s-444-444-4      0005    0.956    uOttawa-Method1
...
21    Q0    s-1111-1111-1    0996    0.875    uOttawa-Method1
21    Q0    s-2222-2222-2    0997    0.875    uOttawa-Method1
21    Q0    s-3333-3333-3    0998    0.875    uOttawa-Method1
21    Q0    s-4444-4444-4    0999    0.875    uOttawa-Method1
21    Q0    s-5555-5555-5    1000    0.874    uOttawa-Method1
```

**Figure 1:** An example of generated results.

Combined with a semantic distance measure cosine similarity, our system extracts relevant sentences for each query and provides a rank for the top-1000 sentences for each of 21 questions/symptoms in the BDI questionnaire.

## 2. Task Description

### 2.1. General Requirements

We focused on Task 1 of the eRisk 2023: Early Risk Prediction on the Internet [1]. Participants are given files in the TREC format containing the sentences of each user (subject). Each document has a document ID number as well as the text of the document. The aim of the task is to extract the top-1000 relevant sentences for each of 21 symptoms in the BDI questionnaire and provide rankings for the extracted sentences. The statistics of the dataset are shown in Table 1.

### 2.2. Evaluations

For each designed method, the results are saved in separate files, which will then be submitted for evaluation. The format of result files is shown through an example in Figure 1.

The sentences were annotated by the shared task organizers with the help of human annotators. There were three annotators, two computer scientists with background in this research area and one psychologist. Based on the relevance of the sentences to the 21 symptoms in the BDI questionnaire, the annotators were guided to label sentences into 2 categories: relevant or irrelevant.

There are 2 types of evaluations, according to the provided relevance judgements (qrels): majority (using majority voting among the available human judgements) and unanimity. The performance is evaluated with 4 standard information retrieval metrics: Average Precision (AP), R-Precision (R-PREC), Precision at 10, and NDCG at 1000. More details including the resulting number of relevant sentences could be found in the overview paper [1].

# 3. Methods

Information retrieval is usually the task that, given a query from a system user, the system searches and returns a ranked list of documents that are matching or are related to the specified query. Therefore we employed information retrieval techniques.

In a typical information retrieval system, an "index" is used to store for each term a list of documents containing the term. This inverted index is first constructed and then it is used for ranking based on some metrics (similarity formula). A search engine collects the documents before the information retrieval step and needs frequent updates. The index needs to be updated too. In our case, the collection of documents is provided and it is static. Since we do not need to update our collection and the system is not designed for frequent searches, we did not construct the inverted index. At this stage, the search space was restricted to a set of documents at hand. To accelerate the calculation of contextual representations, we selected keywords from the questions in the BDI questionnaire for filtering out unrelated documents.

## 3.1. Document Acquisition

All the documents, which in our case are more than 4 million sentences collected from social media and provided in the CLEF eRisk 2023 shared task, are downloaded and stored on an online storage platform Google Drive. Since we are not crawling documents from external sources, no crawler is needed. Before the pre-processing steps, the system connects to the Google Drive and extracts all the documents from the files in the dataset.

## 3.2. Data Normalization and Text Processing

The sentences are stored in files containing documents with DOCNO (document number) and TEXT (textual content). The sentences are extracted, additional information related to the data format is discarded. Depending on the model being used, different pre-processing steps were applied to the texts: for obtaining word embeddings through GloVe [2], we applied tokenization, lowercasing, stopword and punctuation removal; when getting vector representations using transformer-based models, we filtered out sentences that did not contain symptom-related keywords, used transfer learning and did not apply those pre-processing methods (we allowed the specific tokenization used by each contextual embedding model). The normalization and processing steps are applied on both documents and queries.

**Table 2**
Queries and Keywords for Each Question

| Question | Keywords | Query |
|---|---|---|
| Q1 | sadness, sad, unhappy | Sadness. I feel sad unhappy cannot stand it. |
| Q2 | pessimism, discouraged, hopeless | Pessimism. I feel discouraged about my future is hopeless and will get worse. |
| Q3 | failure, fail | Past Failure. I have failed. |
| Q4 | pleasure, enjoy | Loss of Pleasure. I don't enjoy things. |
| Q5 | guilty | Guilty Feelings. I feel guilty. |
| Q6 | punishment, punish | Punishment Feelings. I am being punished. |
| Q7 | confidence, disappointed | Self-Dislike. I have lost confidence. I am disappointed in myself. |
| Q8 | criticalness, critical, criticize, blame, fault | Self-Criticalness. I criticize myself blame myself for my faults. |
| Q9 | suicidal, suicide, kill | Suicidal Thoughts or Wishes. I kill myself. |
| Q10 | crying, cry | Crying. I cry. |
| Q11 | agitation, agitate, restless | Agitation. I am restless or agitated keep moving. |
| Q12 | interest, interested | Loss of Interest. It's hard to get interested. |
| Q13 | indecisiveness, decision, decide | Indecisiveness. I find it difficult to make decisions. |
| Q14 | worthlessness, worthless, worthwhile, useful | Worthlessness. I feel worthless not useful. |
| Q15 | energy, energetic | Loss of Energy. I don't have enough energy. |
| Q16 | sleep, sleeping | Changes in Sleeping Pattern. I sleep more or less than usual. |
| Q17 | irritability, irritable, angry | Irritability. I am irritable. |
| Q18 | appetite, food, eat | Changes in Appetite. My appetite is greater or less. |
| Q19 | concentration, concentrate | Concentration Difficulty. It's hard to keep my mind. I can't concentrate. |
| Q20 | tiredness, fatigue, tired | Tiredness or Fatigue. I am tired or fatigued. |
| Q21 | sex | Loss of Interest in Sex. I am less interested in sex. |

### 3.3. Searching with Contextual Representations

We used transformer-based models to obtain contextual representations of documents and queries, which is a type of embedding that looks at all the words in a sentence in the same time [3]. We filtered out the documents that did not contain certain keywords, to reduce the size of the dataset to accelerate computation. These keywords were picked from the 21 questions in the BDI questionnaire. All of the documents (4,264,693 sentences) were loaded for processing, 111,982 sentences were kept after filtering, and 4,152,711 sentences were filtered out using keywords. The queries we built for each question used both the text of the questions and the text of the possible answers. Information about the keywords and the queries we used is shown in Table 2.

### 3.3.1. DistilBERT with Cosine Similarity

We used DistilBERT [4], a distilled version of BERT with a smaller model and competitive performance. It is faster to train, and lighter to load. After the vector representations of sentences and queries were collected, the cosine similarity was used for calculating semantic similarity between the query and the document (sentence in our case). The ranking of document relevance was then saved.

### 3.3.2. RoBERTa with Cosine Similarity

RoBERTa is an improved version of BERT, with a more carefully designed pretraining [5]. Similar with the method using DistilBERT, we used the cosine similarity to compute the text similarity and the record ranks for 21 the questions (queries).

### 3.3.3. Universal Sentence Encoder with Cosine Similarity

The Universal Sentence Encoder is a text encoder that directly encodes sentences into vectors. It is specifically designed for transfer learning of various types of NLP tasks. The encoder based on the transformer architecture was trained in the following way: the word representations acquired through the transformer were converted to a fixed-length encoding vector by summing the element-wise representations at each word position, and then the vector was divided by the square root of the length of the sentence to reduce sentence length effects. The inputs to the encoder are lowercased strings that tokenized using Penn Treebank Tokenizer (PTB), and the outputs are 512 dimensional vector representations. Since the model was designed to be of general purpose, multi-task learning was conducted [6]. The model has good performance with minimal training data [7]. We used the model to obtain embeddings of queries and sentences, and calculated cosine distance between them to obtain rankings.

### 3.4. Searching with Distributional Word Representations

We used GloVe to get distributional embeddings of sentences and queries. Unlike transformers, GloVe creates co-occurrence matrices of texts, and then applies matrix factorization on the global matrix to shapes with various dimensionalities. As mentioned before, traditional pre-processing steps such as tokenization, lowercasing, stopword and punctuation removal were conducted on sentences. After data were cleaned, we chose 2 versions of GloVe: density of 50-dimension and 100-dimension. The GloVe embeddings were acquired for both documents and queries and then used for cosine similarity calculations. Ranks based on similarities are saved, as before.

## 4. Results and Discussion

The expected criteria for sentence relevance judgements are introduced through examples. In Figure 2, some examples from the task overview paper [1] are given, to illustrate topic relevance.

Participated systems are evaluated using the majority-based qrels and unanimity-based qrels. In total, 10 teams participated and 37 system runs were submitted for this shared task. The results of the majority voting evaluation for the 5 runs submitted by us are presented in Table 3, Judging for symptom: Loss of Energy

| | |
|---|---|
| I cannot control my energy these days. | Relevant (1) |
| My sister has no energy at all. | Irrelevant (0) |
| The book was about a highly energetic man. | Irrelevant (0) |
| I feel more tired than usual. | Relevant (1) |
| The football team is named Top Energy. | Irrelevant (0) |
| I am totally lonely. | Irrelevant (0) |
| I've just recharged my batteries. | Relevant (1) |
| I am lost. | Irrelevant (0) |

**Figure 2:** Examples of sentence relevance.

**Table 3**
Results for submitted 5 runs (majority voting)

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

and the unanimity evaluation for the runs are shown in Table 4. Our team ranked 3rd among the 10 participating teams, and our best performance was achieved by the method that employed the Universal Sentence Encoder with Cosine Similarity (USESim).

The results show that, overall, the universal text representation USE performed better than the other contextual representation techniques such as DistilBERT, for this task. Also, the contextual representation methods performed better on the metrics Precision at 10 and NDCG at 1000, compared to the distributed representation methods based on GloVe. A much larger search space was applied when using the method with GloVe (GloveSim and Glove100Sim) since all the sentences were checked for similarity (we did not filter out sentences in this method since the computation was fast enough). We think a lower performance could be due to the removal of stopwords when using the GloVe-based methods, for example, pronouns that are referring the participant in the discourse (the agent) were removed, but they could contain relevant information.

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

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

In Table 5 and Table 6, we compare our results with the best results from the shared task, for

**Table 4**
Results for submitted 5 runs (unanimity)

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

**Table 5**
Our results compared to the best results in the shared task (majority voting)

| Metric | Run | Rank (Out of 37) | Our best | Best in shared task |
|---|---|---|---|---|
| AP | USESim | 5 | 0.160 | 0.319 |
| R-PREC | USESim | 6 | 0.248 | 0.375 |
| P@10 | USESim | 7 | 0.600 | 0.861 |
| NDCG@1000 | USESim | 6 | 0.382 | 0.596 |

**Table 6**
Our results compared to the best results in the shared task (unanimity)

| Metric | Run | Rank (Out of 37) | Our best | Best in shared task |
|---|---|---|---|---|
| AP | USESim | 5 | 0.139 | 0.293 |
| R-PREC | USESim | 5 | 0.232 | 0.360 |
| P@10 | USESim | 7 | 0.438 | 0.709 |
| NDCG@1000 | USESim | 6 | 0.380 | 0.615 |

the four metrics.

## 5. Conclusion and Future Work

This paper presented several methods for searching relevant sentences for symptoms in the BDI questionnaire. We showed that contextual representations, especially universal text representations perform better than distributed representations; we showed that when used for acquiring embeddings for documents, Universal Sentence Encoder gives better results and would also simplify the calculation since no need to combine word representations to form a single vector for a document. We also showed that transfer learning with pretrained models could be used for even detailed differentiation within a domain such as depression.

In the future, we consider distilling and enhancing our symptom-related keyword collections, and fine-tuning our filtering steps, so that more informative sentences could be retained. Since we conducted the experiments in a limited time, we expect the results to be better if more configurations are tested and more sentences are consumed by systems. Also for the USE, since multiple versions of models trained with different goals are available, we are considering experimenting with other variants of the model. Further pre-training the model on more specific and related corpus is also considerable.

## Acknowledgments

We thank the Natural Sciences and Engineering Research Council of Canada (NSERC) for supporting our research. We also want to thank the CLEF eRisk 2023, for providing a great opportunity for us to explore the task of early risk detection.

## References

[1] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, eRisk 2023: Depression, Pathological Gambling, and Eating Disorder Challenges, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Lecture Notes in Computer Science, Springer Nature Switzerland, Cham, 2023, pp. 585–592. doi:10.1007/978-3-031-28241-6_67.

[2] J. Pennington, R. Socher, C. Manning, GloVe: Global Vectors for Word Representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. URL: https://aclanthology.org/D14-1162. doi:10.3115/v1/D14-1162.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, I. Polosukhin, Attention is All you Need, in: Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[4] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2020. URL: http://arxiv.org/abs/1910.01108. doi:10.48550/arXiv.1910.01108, arXiv:1910.01108 [cs].

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019. URL: http://arxiv.org/abs/1907.11692. doi:10.48550/arXiv.1907.11692, arXiv:1907.11692 [cs].

[6] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, R. Kurzweil, Universal Sentence Encoder, 2018. URL: http://arxiv.org/abs/1803.11175. doi:10.48550/arXiv.1803.11175, arXiv:1803.11175 [cs].

[7] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, R. Kurzweil, Universal Sentence Encoder for English, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 169–174. URL: https://aclanthology.org/D18-2029. doi:10.18653/v1/D18-2029.