OPI@LT-EDI-ACL2022: Detecting Signs of Depression from Social Media Text using RoBERTa Pre-trained Language Models

Rafał Poświata & Michał Perełkiewicz
National Information Processing Institute, 00-608 Warsaw, Poland
{rposwiata, mperelkiewicz}@opi.org.pl

Abstract

This paper presents our winning solution for the Shared Task on Detecting Signs of Depression from Social Media Text at LT-EDI-ACL2022. The task was to create a system that, given social media posts in English, should detect the level of depression as 'not depressed', 'moderately depressed' or 'severely depressed'. We based our solution on transformer-based language models. We fine-tuned selected models: BERT, RoBERTa, XLNet, of which the best results were obtained for $\text{RoBERTa}_{\text{large}}$. Then, using the prepared corpus, we trained our own language model called DepRoBERTa (RoBERTa for Depression Detection). Fine-tuning of this model improved the results. The third solution was to use the ensemble averaging, which turned out to be the best solution. It achieved a macro-averaged F1-score of 0.583. The source code of prepared solution is available at https://github.com/rafalposwiata/depression-detection-lt-edi-2022.

1 Introduction

Depression (major depressive disorder) is a common and serious medical illness that, according to **World Health Organization** (WHO), already affects about **322 million** people worldwide (WHO, 2017). The main symptoms of depression include: feeling sad or having a depressed mood, loss of interest or pleasure, feeling worthless or guilty, insomnia or hypersomnia, thoughts of death and suicidal ideation or suicide attempts (American Psychiatric Association, 2013). When diagnosed and treated quickly, it can greatly improve quality of live and in some cases even save it. Such rapid detection of depression signs is possible, for example, based on the social media posts of the individual (De Choudhury et al., 2013). Following this assumption, Sampath et al. (2022) organized at **LT-EDI-ACL2022** the **Shared Task on Detecting Signs of Depression from Social Media Text**. The task was to create a system that, given social media posts in English, should classify the level of depression as **'not depressed'**, **'moderately depressed'** or **'severely depressed'**.

In this paper we present our solution for this competition. The paper is organized as follows. Section 2 describes related work with particular emphasis on issues of depression detection in social media. Section 3 presents the dataset and its modification. The process of developing our solution is explained in Section 4. The next section shows performed experiments, the results, along with the error analysis. Finally, Section 6 concludes this paper.

2 Related Work

De Choudhury et al. (2013) authored one of the first papers on detecting depression based on social media posts. In their work, they collected a group of **Twitter**[1] users diagnosed with depression whose one-year posts were used to create a statistical classifier to estimate the risk of depression. Tsugawa et al. (2015) prepared the dataset in a similar way but for Japanese users, and then trained a Support Vector Machines (SVM) classifier to estimate the presence of active depression. Wolohan et al. (2018) created a dataset based on **Reddit**[2] posts in which users were assigned to one group: depressed or control. Then, among other things, they analyzed their posts using the Linguistic Inquiry and Wordcount Tool (LIWC) (Pennebaker et al., 2015). Pirina and Çöltekin (2018) also used Reddit as a data source and with other datasets they verified how training data can affect the quality of a SVM-based model to identify depression. Tadesse et al. (2019) use different types of approaches to text encoding (the LIWC dictionary, Latent Dirichlet Allocation (LDA) topics or N-grams) to explore the users' linguistic usage in the depressive posts. Arora and Arora (2019) analyze tweets for depression and anxiety by using Multinomial Naive Bayes

[1] https://twitter.com
[2] https://www.reddit.com

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 276 - 282

PID	Text	Label
train_pid_6035	Happy New Years Everyone : We made it another year	not depression
train_pid_35	My life gets worse every year : That's what it feels like anyway....	moderate
train_pid_8066	Words can't describe how bad I feel right now : I just want to fall asleep forever.	severe

Table 1: Samples from the dataset.

and Support Vector Regression (SVR) Algorithm as a classifier. Lin et al. (2020) create **SenseMood** system to detect depression from tweets based on visual and textual features using Convolutional Neural Network (CNN) and BERT language model. Zogan et al. (2021) propose novel summarization boosted deep framework for depression detection called **DepressionNet**. Other works worth mentioning include Aswathy et al. (2019); Haque et al. (2021); William and Suhartono (2021).

For text-based classification, the last few years have been primarily a time of deep learning and large pre-trained transformer-based language models (Min et al., 2021). This kind of solutions achieve state-of-the-art results for numerous classification tasks (Devlin et al., 2019; Liu et al., 2019; Chan et al., 2020; Dadas et al., 2020).

3 Dataset

The dataset used in the competition consists of English posts from Reddit, where each was annotated with one of the labels: **not depression**, **moderate** and **severe** (Kayalvizhi and Thenmozhi, 2022). The first label indicates a case where no signs of depression were identified. The other two labels show that symptoms in the post indicate moderate or severe depression respectively. Example texts with labels from the dataset are presented in Table 1. The dataset was divided into three parts: train, dev, and test. Labels for the test part were not provided by the organizers, as this one was the part on which the solutions were verified. To verify the quality of the collections used to prepare the solution (train, dev), we first verified their diversity by removing duplicate records containing the same posts. As a result of this step, we noticed that the train set consists of a large number of the same examples, and the unique ones are only **2,720** (out of **8,891** total). In the case of the dev set, the difference was much smaller, i.e., **4,481** unique against **4,496** all. It is good practice to make the train set larger than the dev or test set. This is especially important when using machine learning or deep learning methods where the quality of the model directly depends on the number and variety of samples during training. Therefore, we decided to use part of the dev set for training, leaving **1,000** examples for verification (we kept the class distribution close to the original one). As a result, the train set we used in our experiments counted **6,006** unique examples (the final number is due to the fact that there were overlaps between the original train and dev sets). The whole process of preparing the dataset, including class distribution, is shown in Figure 1. What is worth noting is that the dataset is unbalanced, and the severe class is underrepresented.

4 Our solution

We organized the work on our solution into three steps, which will be presented in the following subsections.

4.1 Fine-tuned Transformer-Based Language Models

First, we fine-tune several commonly used English pre-trained language models. We use the standard fine-tuning procedure like Devlin et al. (2019), which involves training pre-trained language model with classification head on top (a linear layer on top of the pooled output). The following models were utilized: **BERT** (Devlin et al., 2019), **RoBERTa** (Liu et al., 2019) and **XLNet** (Yang et al., 2019). Both in base and large version. All models were downloaded from the Hugging Face hub[3]. The best result on the dev set was achieved by **RoBERTa$_{\textbf{large}}$**, which will be further described in Section 5.3.

4.2 Pre-trained and Fine-tuned Domain Specific Transformer-Based Language Model

The models used in the previous step were pre-trained on general domain corpora (e.g. English Wikipedia or BooksCorpus). It can be assumed that most of the texts from these corpora did not manifest symptoms of depression. Inspired by Lee et al. (2019), we decided to pre-train our own

[3]https://huggingface.co/models

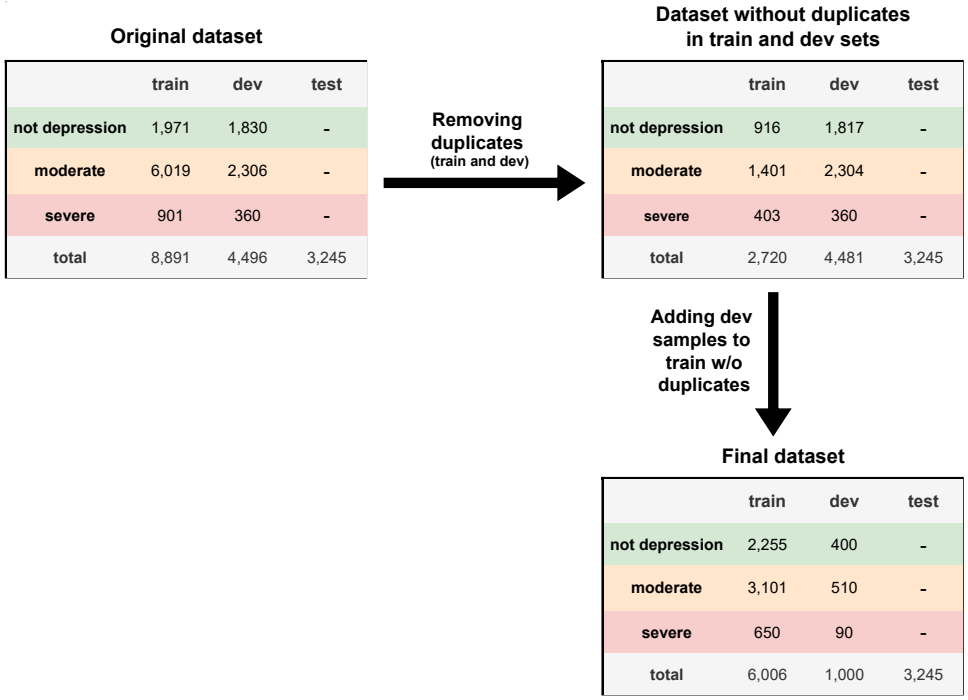

Figure 1: The process of preparing the dataset including the distribution of classes at each step. The dashes (-) are due to the lack of labels for the test set.

$$\mathbf{y_{ensemble}} = \mathbf{arg\ max}\left(\frac{\mathbf{softmax}(\mathbf{y'_{RoBERTa_{large}}}) + \mathbf{softmax}(\mathbf{y'_{DepRoBERTa}})}{\mathbf{2}} \right) \quad (1)$$

language model on texts mainly expressing depression. We built a corpus based on the **Reddit Mental Health Dataset** (Low et al., 2020) and a dataset of **20,000** posts from **r/depression** and **r/SuicideWatch** subreddits[4]. We filtered the data appropriately, leaving mainly those related to **depression (31,2%)**, **anxiety (20,5%)** and **suicide (18.1%)**, which resulted in a corpora consisting of **396,968** posts. We used a **further pre-training** technique where the model weights were initialized with the RoBERTa$_{\text{large}}$ model weights, since it was the fine-tuning of this particular model that gave the best results in the first step. We called the resulting model **DepRoBERTa** (RoBERTa for Depression Detection). For more information on the corpus statistics and the pre-training process, we refer you to the appendices. Then, as with the models in Section 4.1, we performed DepRoBERTa fine-tuning on the train set.

4.3 Ensemble

In the last step, we combined the best models obtained in the previous steps using **ensemble averaging** (Naftaly et al., 1999). This method involves averaging the predictions from a group of models, and its implementation in our case is presented in Equation 1. Where $\mathbf{y'_{RoBERTa_{large}}}$ and $\mathbf{y'_{DepRoBERTa}}$ are vectors of raw (non-normalized) predictions generated by fine-tuned RoBERTa$_{\text{large}}$ and DepRoBERTa, respectively.

Parameter	Value
Optimizer	AdamW
Learning rate	5e-6
Batch size	16
Dropout	0.1
Weight decay (L2)	0.1
Epochs	10
Validation after no. steps	100
Max sequence length	300

Table 2: Hyper-parameters used when fine-tuning models.

[4]https://www.kaggle.com/xavrig/reddit-dataset-rdepression-and-rsuicidewatch

Model	Accuracy	Precision	Recall	F1-score
BERT$_{base}$	0.627	0.586	0.574	0.579
BERT$_{large}$	0.606	0.568	0.566	0.566
RoBERTa$_{base}$	0.622	0.567	0.573	0.570
RoBERTa$_{large}$	**0.664**	0.629	0.591	**0.605**
XLNet$_{base}$	0.654	**0.632**	0.576	0.590
XLNet$_{large}$	0.639	0.611	**0.597**	0.602
DepRoBERTa	**0.661**	**0.628**	**0.607**	**0.616**
Ensemble	**0.695**	**0.663**	**0.621**	**0.637**

Table 3: Results of each model on the dev set. Bolded and underlined values indicate the best and second-best scores for models from each of the three steps for a given measure.

Model	Accuracy	Precision	Recall	F1-score
RoBERTa$_{large}$	0.614	0.583	0.564	0.552
DepRoBERTa	0.626	0.575	0.588	0.571
Ensemble	**0.658**	**0.586**	**0.591**	**0.583**

Table 4: Results of submitted models on the test set (official competition results made available by the competition organisers). Bolded and underlined values indicate the best and second-best scores for the measure, respectively.

5 Experiments and Results

5.1 Experimental Setup

We utilized Simple Transformers library (Rajapakse, 2019) to perform experiments, including models fine-tuning and pre-training the DepRoBERTa model. Used hyper-parameters are presented in Table 2. The fine-tuning procedure for each model was repeated 5 times using the train and dev sets described in Section 3. All experiments were run on a single GPU Tesla V100.

5.2 Metrics

The metrics used during the experiments are accuracy, macro-averaged precision, macro-averaged recall and macro-averaged F1-score across all the classes. The macro-averaged F1-score was the main measure when evaluating solutions.

5.3 Results

Table 3 shows the results on the dev set. Among the fine-tuned transformer-based language models, **RoBERTa$_{large}$** model was the best in terms of accuracy (**0.664**) and F1-score (**0.605**). In the other two measures, **XLNet** models were better, respectively **XLNet$_{base}$** for precision (**0.632**) and **XLNet$_{large}$** for recall (**0.597**). **RoBERTa$_{large}$** was second in these cases. We improved the F1-score by **0.011** using the **DepRoBERTa** fine-tuned model. This was mainly due to the high score for the recall measure (**0.607**), as the results for the other measures were worse than **RoBERTa$_{large}$**. **Ensemble** proved to be the best approach by achieving the highest scores on each measure, having an F1-score of **0.637** (an improvement of **0.021** over **DepRoBERTa**). Due to these results, we have chosen as our official competition solutions: **RoBERTa$_{large}$**, **DepRoBERTa** and **Ensemble**. The results they achieved on the test set are presented in Table 4. As expected, **Ensemble** proved to be the best by achieving an F1-score of **0.583**. This score gave our team the **1st** place among the **31** participating teams.

5.4 Errors Analysis

To be able to evaluate the errors and strengths of our models, we created the confusion matrices shown in Figure 2. Each model specializes in one class, i.e. it achieves the best results for a different class. **RoBERTa$_{large}$** performs best for the **not depression** class, **DepRoBERTa** for the **severe** class, and **Ensemble** for the **moderate** class. The most common mistake is to assign a **severe** class to a post originally tagged as **moderate**. A mistake that also often occurs is confusion between **not depression** and **modereate** classes. The analysis was carried out on the dev set as the competition organisers did not provide labels for the test set.

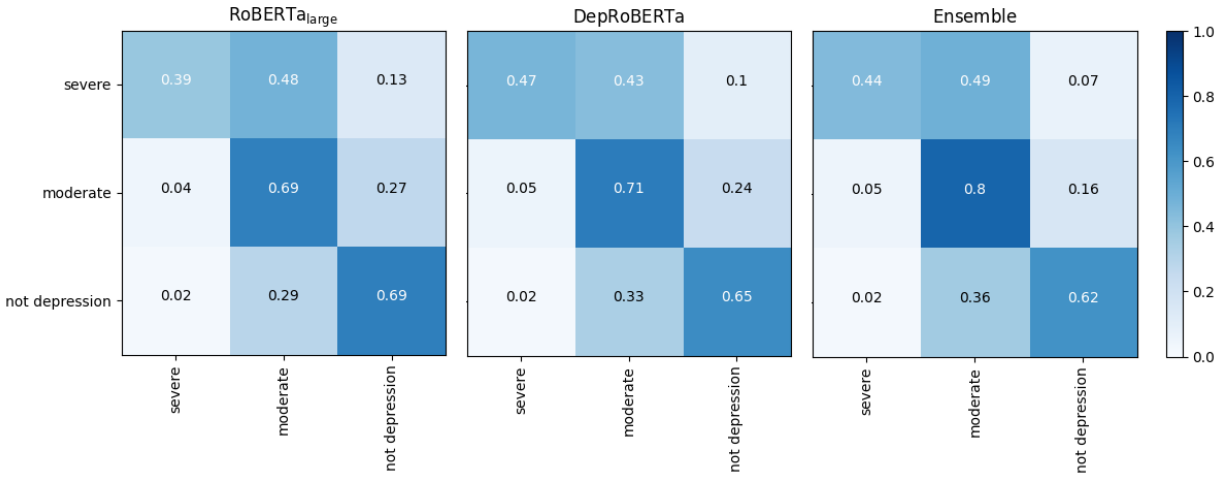

Figure 2: Normalized confusion matrices for RoBERTa$_{large}$, DepRoBERTa and their ensemble on the dev set.

6 Conclusion

In this paper, we presented a solution to the Shared Task on Detecting Signs of Depression from Social Media Text at LT-EDI-ACL2022. The use of ensemble averaging previously fine-tuned language models proved to be the best. As part of this work, in addition to the models designed for this competition, we also prepared a new pre-train language model, DepRoBERTa. In the future it can be used for other depression detection tasks. We plan to pre-train it further on a larger corpus of texts expressing depression, as an extension of this work.

The code of our solution and prepared models are available online at https://github.com/rafalposwiata/depression-detection-lt-edi-2022.

References

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*. American Psychiatric Association, Arlington, VA.

Priyanka Arora and Parul Arora. 2019. Mining twitter data for depression detection. In *2019 International Conference on Signal Processing and Communication (ICSC)*, pages 186–189.

K S Aswathy, P C Rafeeque, and Reena Murali. 2019. Deep learning approach for the detection of depression in twitter. In *Proceedings of the International Conference on Systems, Energy Environment (ICSEE)*.

A. T. BECK, C. H. WARD, M. MENDELSON, J. MOCK, and J. ERBAUGH. 1961. An Inventory for Measuring Depression. *Archives of General Psychiatry*, 4(6):561–571.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Sławomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2020. Pre-training polish transformer-based language models at scale. In *Artificial Intelligence and Soft Computing*, pages 301–314. Springer International Publishing.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):128–137.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. 2021. Deep learning for suicide and depression identification with unsupervised label correction. In *Artificial Neural Networks and Machine Learning – ICANN*

2021, pages 436–447, Cham. Springer International Publishing.

S Kayalvizhi and D Thenmozhi. 2022. Data set creation and empirical analysis for detecting signs of depression from social media postings. *arXiv preprint arXiv:2202.03047*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics (Oxford, England)*, 36.

Chenhao Lin, Pengwei Hu, Hui Su, Shaochun Li, Jing Mei, Jie Zhou, and Henry Leung. 2020. *SenseMood: Depression Detection on Social Media*, page 407–411. Association for Computing Machinery, New York, NY, USA.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

David E. Losada, Fabio Crestani, and Javier Parapar. 2020. Overview of erisk 2020: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings*, page 272–287, Berlin, Heidelberg. Springer-Verlag.

David E. Losada, Fabio A. Crestani, and Javier Parapar. 2017. Clef 2017 erisk overview: Early risk prediction on the internet: Experimental foundations. In *CLEF*.

David E. Losada, Fabio A. Crestani, and Javier Parapar. 2018. Overview of erisk 2018: Early risk prediction on the internet (extended lab overview). In *CLEF*.

David E. Losada, Fabio A. Crestani, and Javier Parapar. 2019. Overview of erisk at clef 2019: Early risk prediction on the internet (extended overview). In *CLEF*.

Daniel M Low, Laurie Rumker, John Torous, Guillermo Cecchi, Satrajit S Ghosh, and Tanya Talkar. 2020. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of medical Internet research*, 22(10):e22635.

Bonan Min, Hayley H. Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *ArXiv*, abs/2111.01243.

Ury Naftaly, Nathan Intrator, and David Horn. 1999. Optimal ensemble averaging of neural networks. *Network: Computation in Neural Systems*, 8.

Javier Parapar, Patricia MartÃn, David E. Losada, and Fabio Crestani. 2021. Overview of eRisk 2021: Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction Proceedings of the Twelfth International Conference of the CLEF Association (CLEF 2021)*. Springer International Publishing.

James W. Pennebaker, Ryan Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. University of Texas at Austin.

Inna Pirina and Çağrı Çöltekin. 2018. Identifying depression on Reddit: The effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 9–12, Brussels, Belgium. Association for Computational Linguistics.

Daniel Preoţiuc-Pietro, Maarten Sap, H. Andrew Schwartz, and Lyle Ungar. 2015. Mental illness detection at the world well-being project for the CLPsych 2015 shared task. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 40–45, Denver, Colorado. Association for Computational Linguistics.

T. C. Rajapakse. 2019. Simple transformers. `https://github.com/ThilinaRajapakse/simpletransformers`.

Philip Resnik, William Armstrong, Leonardo Claudino, and Thang Nguyen. 2015. The University of Maryland CLPsych 2015 shared task system. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 54–60, Denver, Colorado. Association for Computational Linguistics.

Kayalvizhi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, and Jerin Mahibha C. 2022. Findings of the shared task on Detecting Signs of Depression from Social Media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Michael M. Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing depression from twitter activity. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*.

WHO. 2017. Depression and other common mental disorders: global health estimates. World Health Organization.

David William and Derwin Suhartono. 2021. Text-based depression detection on social media posts: A systematic literature review. *Procedia Computer Science*, 179:582–589. 5th International Conference on Computer Science and Computational Intelligence 2020.

JT Wolohan, Misato Hiraga, Atreyee Mukherjee, Zeeshan Ali Sayyed, and Matthew Millard. 2018. Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 11–21, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. 2021. Depressionnet: A novel summarization boosted deep framework for depression detection on social media. *ArXiv*, abs/2105.10878.

Appendix

A Previous competitions

The Shared Task on Detecting Signs of Depression from Social Media Text at LT-EDI-ACL2022 was not the first competition to address the topic of depression detection. To the best of our knowledge, the first was the **CLPsych 2015 Shared Task: Depression and PTSD on Twitter** (Coppersmith et al., 2015). The shared task consisted of three tasks, two of which related to depression: identifying depressed users from a control group and distinguishing depressed users from those with PTSD (post-traumatic stress disorder). The SVM classifier and its variants have proven to be the best and most popular solution (Resnik et al., 2015; Preoţiuc-Pietro et al., 2015). This was followed by a series of **eRisk** competitions as part of the **CLEF** conference (Losada et al., 2017, 2018, 2019, 2020; Parapar et al., 2021). In the first two editions (2017-2018), the problem was defined as an early risk detection task. So, in addition to identifying depression, the system should be able to do so by having the shortest possible list of posts or chunks of a user's posting history. In subsequent editions (2019-2021), participants were asked to create systems that would determine a user's severity of depression based on their posts by predicting their responses to a standard depression questionnaire derived from the Beck's Depression Inventory (BDI) (BECK et al., 1961). In the case of eRisk contests, the datasets created were based on Reddit posts.

B Reddit Depression Corpora

subreddit	**# posts**	**%**
depression	123,824	31.2
suicidewatch	71,816	18.1
anxiety	53,797	13.6
bpd	21,836	5.5
lonely	21,399	5.4
socialanxiety	19,648	4.9
fitness	10,000	2.5
jokes	10,000	2.5
legaladvice	10,000	2.5
parenting	10,000	2.5
personalfinance	10,000	2.5
relationships	10,000	2.5
healthanxiety	7,847	2.0
ptsd	7,551	1.9
bipolarreddit	5,186	1.3
teaching	4,064	1.0

Table 5: Statistics of the corpus formed to pre-train DepRoBERTa.

C DepRoBERTa

Parameter	**Value**
Optimizer	AdamW
Learning rate	4e-5
Batch size	50
Dropout	0.1
Epochs	10
Training samples	389,028
Validation samples	7,940
Validation after no. steps	5,000

Table 6: Configuration used when pre-training DepRoBERTa.