# CIMAT-NLP@LT-EDI: Finegrain Depression Detection by Multiple Binary Problems Approach

**María de J. García Santiago, Fernando Sánchez Vega, A. Pastor López-Monroy**
*Mathematics Research Center, Jalisco S/N Valenciana, 36023 Guanajuato, GTO México*
{maria.garcia, fernando.sanchez, pastor.lopez}@cimat.mx

## Abstract

This work described the work of the team CIMAT-NLP on the Shared task of Detecting Signs of Depression from Social Media Text at LT-EDI 2023 Sampath et al. (2023), which consists of depression classification on three levels: ”not depression”, ”moderate” depression and ”severe” depression on text from social media. In this work, we proposed two approaches: (1) a transformer model which can handle big text without truncation of its length, and (2) an ensemble of six binary Bag of Words. Our team placed fourth in the competition and found that models trained with our approaches could place second.

## 1 Introduction

Approximately 280 million persons suffer depression around the world, and suicide is the fourth cause of death according to World Health Organization (2022).

Other studies have highlighted the impact of social media on adolescents, introducing a phenomenon known as ”Facebook depression” O'Keeffe et al. (2011). This term refers to the symptoms of depression that young people may experience when they spend significant time on social media platforms.

Additionally, a study on college students in Afghanistan revealed a correlation between social media addiction and depression Haand and Shuwang (2020). The findings suggested that individuals experience more severe symptoms of depression as their social media usage increases.

Given the increasing number of people affected by depression, developing systems to detect individuals with this mental illness is crucial. One notable effort in this direction is the Shared Task on Detecting Signs of Depression from Social Media Texts at LT-EDI Sampath et al. (2023).

Our team, CIMAT-NLP, proposed two approaches for this task. Firstly, we divided significant texts into sub-packages and utilized the RoBERTa transformer Liu et al. (2019). Secondly, we employed an ensemble of six binary Bags of Words (BOW) models with different characteristics.

The remaining sections of this paper are organized as follows: Section 2 discusses related works on detecting depression on social media. Section 3 provides an overview of the competition and data distribution. In Section 4, we describe the methods we developed for the task. Section 5 presents the results obtained by our models. Finally, in Section 6, we draw conclusions based on our work.

## 2 Related work

Detecting depression presents a challenging task due to the intricate nature of this mental disorder. The complexity of this mental illness makes screening for depression a demanding task. In this field, various workshops are dedicated to this cause, such as the Early Internet Risk Prediction workshop (CLEF eRisk) Parapar et al. (2022). This workshop focuses on developing methods for automatic systems for online risk prevention. In the context of eRisk, proposals have predominantly focused on the use of Bag of Words (BOW)-based machine learning models together with SVM classifiers or deep neural networks, due to the proven effectiveness of both approaches. Notable examples include the top three best ranks in the 2018 eRisk competition (Losada et al. (2018), Trotzek et al. (2018),Funez et al. (2018)), demonstrating the effectiveness of BOW in representing text for tasks related to detecting mental illness. Our approach follows suit, as we choose to implement BOW with several specialized classifiers, targeting different levels of depression.

*LT-EDI 2023 – Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 124–130, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-084-7_018

It is evident that transformers have surpassed the state of the art in various NLP tasks. However, a drawback of transformers is their inability to process large inputs, which is a common scenario in author profiling tasks, due to the high computational resources required. In Martínez-Castaño et al. (2021), this issue was addressed by segmenting the text into subchunks per category during training and averaging the prediction probabilities of these subchunks for an overall prediction. Another limitation of transformers lies in the variability of their predictions, which stems from variations in their initialization seeds during training. To mitigate this variability and leverage the benefits of these results, multiple transformer ensemble techniques have been proposed. Poświata and Perełkiewicz (2022), Janatdoust et al. (2022), and Wang et al. (2022) have introduced such techniques, emphasizing that sets of classifiers can offer improved predictions compared to a single one. Inspired by the ensemble's design philosophy, we have extended it to methods based on Bag of Words (BOW) and SVM classifier approaches

## 3 Dataset

The DepSign-LT-EDI@RANLP-2023 dataset consists of texts collected from various social media networks. All the texts are in English and have been classified into three labels: ”not depression,” ”moderate,” and ”severe.”

The competition was divided into two phases. In the first phase, the organizing committee provided participants with the training and development data to work. In the second phase, participants were given the test data to make predictions and submit their results.

| | Training | Dev | Test |
|---|---|---|---|
| Total users | 7006 | 3233 | 499 |
| ”not depression” label | 2667 | 844 | 135 |
| ”moderate” label | 3584 | 2161 | 275 |
| ”severe” label | 745 | 228 | 89 |

Table 1: Initially, the training dataset had 7201 instances where 195 were duplicated. In the case of the dev dataset, only 12 instances were duplicated.

## 4 Method

In this section, we described the approaches used for the task. In the first approach, we proposed a transformer model. Because the text is in English, we used BERT Devlin et al. (2018a), RoBERTa Liu et al. (2019), MentalBERT and MentalRoBERTa (Ji et al., 2021) for our experiments. The second approach is an ensemble of six Bags of Words, each specializing in diverse detection with different characteristics.

### 4.1 Transformer based approach

Most text in the training and dev dataset does not pass for 124 tokens. Therefore, this length was set as the maximum for tokenizing the instances.

During the training phase, if a text exceeds 124 tokens, it is truncated to 124 tokens, and the remaining text is divided into subtexts of 124 tokens each. These subtexts are then added as new instances to the training dataset. As a result, the final number of instances in the training dataset amounts to $13,238$.

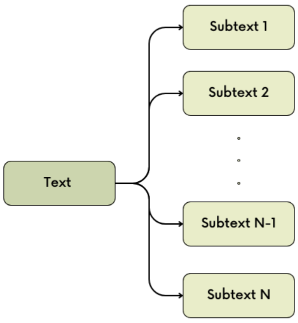

Figure 1: For example, if a text has a length of 368 tokens, the total number of subtexts is three, where the first and second have the same length and the last one has 120 tokens.

In the case of dev and test data, the datasets were not modified, however in the inference part, the process is,

- If the text to classify has a length less than 124, tokens are passed on to the model and predict its class.
- In other cases, the text is divided into sub-text with the length set before. Each sub-text class is predicted, and the final prediction is made with a voting scheme. In Fig.2, the process is illustrated.

#### 4.1.1 Voting scheme

A count of the number of subtext predicted for each subtext is made in the process.

Let be Counter Control ($CC$), the number of subtexts predicted as ”not depression”, Counter

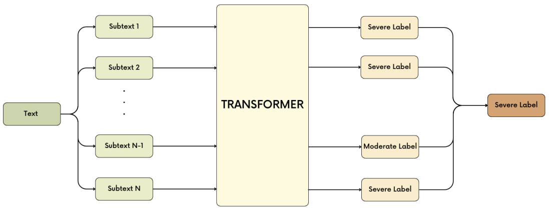

Figure 2: For each sequence of tokens representing the subtexts, the unique tokens of [CLS] and [SEP] are added.

Moderate ($CM$) for ”moderate”, and Counter Severe ($CS$) for ”severe”.

- If $CC > CM$ and $CC > CS$ then the final prediction is labeled as ”not depression”.
- If $CM > CS$ and $CM > CC$ then the final prediction is labeled as ”moderate”.
- If $CS > CM$ and $CS > CC$ then the final prediction is labeled as ”severe”.
- If $CC > CS$ and $CC = CM$ then the final prediction is labeled as ”moderate”.
- If $CS > CC$ and $CM = CS$ then the final prediction is labeled as ”severe”.
- If $CC > CM$ and $CC = CS$ then the final prediction is labeled as ”severe”.
- If $CC = CM = CS$ then the final prediction is labeled as ”moderate”.

Most of the subchunks from a text are from one specific label, then is correct to give that classification to the whole text. The problem arises when there is an equal quantity of subchunks from two o more classes; this happens in one of three cases: an equal number of not-depressing chunks as severe-depression classified chunks, an equal number of not-depressing chunks as moderate depression and finally if we have an equal number of severe depression chunks as moderate depression chunks. The majority of these cases are marked as severe because we prefer to make false positives instead of false negatives, as we think that if the text presented various parts of the severe-label text is because the user that wrote the original text could have symptoms of depression.

### 4.2 Multiple binary BOW approach

Instead of making a multiclass BOW, we decided to make an ensemble of binary BOWs, each of which was trained in different datasets. We decided to use binary BOWs because BOW has outstanding performance in binary classification tasks, as in the work of Ortega-Mendoza et al. (2022).

The training datasets were made from the original training data provided by the committee organizer; the strategy was the following:

- Two labels are merged into one label, and the third is left untouched. Using this strategy, we made three datasets: ”moderate-severe” vs ”not depression”, ”moderate-not depression” vs ”severe”, and ”severe-not depression” vs ”moderate”.
- The second strategy only uses two labels and discards the third one from the training data. Using this strategy, we made three datasets: ”moderate” vs ”not depression”, ”moderate” vs ”severe”, and ”severe” vs ”not depression”.

In total, we created six different data sets for training the BOW on the six binary decisions. The same strategy was followed for the dev data for the corresponding case. For each dataset, we construct a specific BOW using the $\chi$-square function to select the best attributes (in Section 5, we talk about the weight and number of $n$-grams used for the construction). Each BOW is passed on to its classifier and gets the prediction from all the text in the dev dataset. The fusion of the BOW is made using an ensemble; the process is now on the text level, as the decision is made for each of them. Let be the $text_j$, for this text are six predictions: $Prediction\_k_j$ with $1 \leq k \leq 6$.

Depending if the $Prediction\_k_j$ is positive or not, we add a specific weight to the three counter variables: $CS$, $CM$ and $CC$ variables for ”severe”, ”moderate”, and ”not depression” respectively. The next step is to pass these variables into the voting scheme[1] for the final prediction.

## 5 Results

In this section, we present the results in the dev dataset for each approach to choosing the model and hyperparameters for the submissions in the competitions. In the second part of this section, we present the competition results for the two models.

[1] The voting scheme contains the same rules described in Subsubsection 4.1.1, except that when $CC = CM$ and $CC > CS$, the final label is ”moderate”.

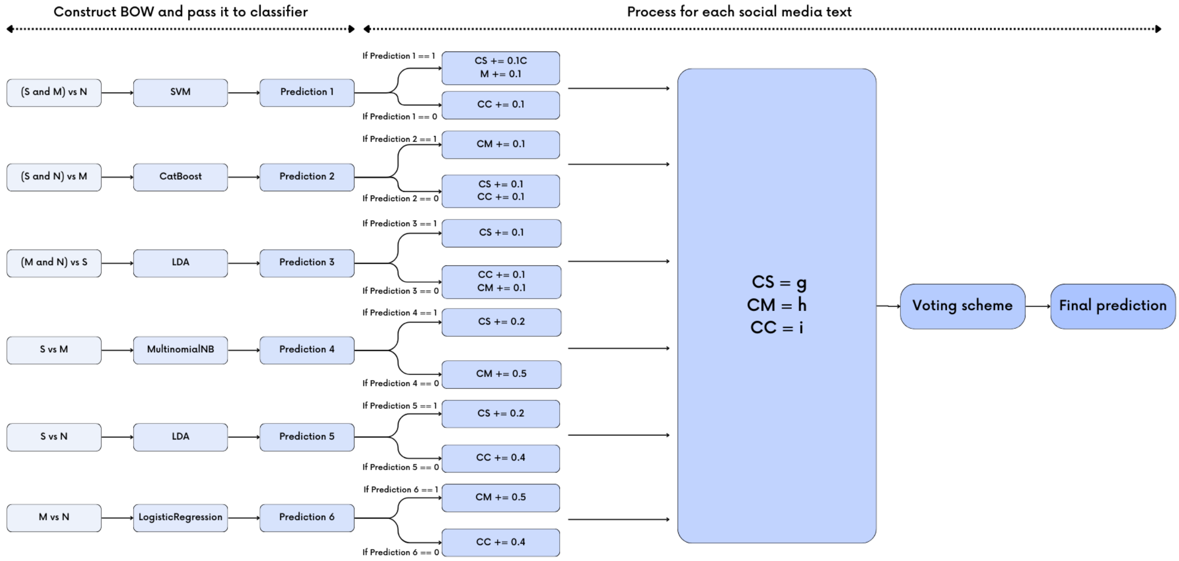

Figure 3: In this ensemble, we refer to ”severe” as $S$, ”moderate” as $M$, and ”not depression” as $N$. The counter variables $CS$, $CM$ and $CC$ are for ”severe”, ”moderate”, and ”not depression”, respectively. The counter variables, voting scheme and final prediction are per user.

### 5.1 Method validations and hyperparameter selection

Because the target is multi-class the F1-score macro metric was used to select the best models for the competition. The dev data was used as test data for this part.

### 5.1.1 Transformer based approach

For the transformer approach, we made different experiments using pre-trained base models of BERT Devlin et al. (2018b), RoBERTa Liu et al. (2019), MentalBERT and MentalRoBERTa Ji et al. (2021), using different learning rates and batch train for experimentation with fixed seed 42.

In Table 2, the best five models per transformer model are described; most of these models are ensemble models of three transformers with the same lr and train batch size. This ensemble models use majority voting for the final prediction. We decided to use RoBERTa-32, which is a single base pre-trained RoBERTa trained with learning rate $1e^{-5}$ and train batch size 32 because it was the model with the best F1-macro score.

### 5.1.2 Multiple binary BOW approach

Considering that six different BOWs conform to the ensemble, the best hyperparameters and classificators were used; in each BOW, the best attributes were selected by $\chi^2$ function [2].

The hyperparameter and classification algorithm for each one of the six binary subproblems is:

- BOW 1 (S and M vs N): This BOW was created with unigrams and bigrams, 200 attributes, tf-idf weighting and SVM classifier.
- BOW 2 (S and N vs M): This BOW was created with unigrams and bigrams, 700 attributes, tf-idf weighting and CatBoostClassifier classifier.
- BOW 3 (M and N vs S): This BOW was created with unigrams, bigrams, tri-grams, 100 attributes, tf weighting and LinearDiscriminantAnalysis classifier.
- BOW 4 (S vs M): This BOW was created with unigrams and bigrams, 500 attributes, binary weighting and MultinomialNB classifier.

[2] All the BOWs were implemented using CountVectorizer for binary weighting and TfidfVectorizer for the other BOWs; the two functions are implemented in the sklearn library. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html, https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

| Model | lr | Train Batch Size | F1-score macro |
|---|---|---|---|
| BERT ensemble | $1e^{-5}$ | 16 | 0.5312 |
| BERT ensemble | $1e^{-5}$ | 32 | 0.5221 |
| BERT ensemble | $1e^{-5}$ | 64 | 0.5145 |
| BERT | $1e^{-5}$ | 16 | 0.5098 |
| RoBERTa | $1e^{-5}$ | 32 | **0.5475** |
| RoBERTa ensemble | $1e^{-5}$ | 32 | 0.5438 |
| RoBERTa ensemble | $1e^{-5}$ | 64 | 0.5383 |
| RoBERTa | $1e^{-5}$ | 64 | 0.5199 |
| MentalBERT ensemble | $1e^{-5}$ | 16 | 0.5358 |
| MentalBERT ensemble | $1e^{-5}$ | 128 | 0.5270 |
| MentalBERT ensemble | $1e^{-5}$ | 32 | 0.5257 |
| MentalBERT | $1e^{-5}$ | 16 | 0.5105 |
| MentalRoBERTa | $1e^{-5}$ | 64 | 0.5451 |
| MentalRoBERTa | $1e^{-5}$ | 16 | 0.5412 |
| MentalRoBERTa | $1e^{-5}$ | 128 | 0.5355 |
| MentalRoBERTa | $1e^{-5}$ | 16 | 0.5241 |

Table 2: For each experiment, three models were made with the same characteristics and then used for the ensemble models.

- BOW 5 (S vs N): This BOW was created with unigrams, bigrams, tri-grams, 100 attributes, tf weighting and LinearDiscriminantAnalysis classifier.
- BOW 6 (M vs N): This BOW was created with unigrams and bigrams, 1700 attributes, tf weighting and a LogisticRegression classifier.

In the second stage of this ensemble, we used grid search to set the best weights for the ensemble. This grid search is done into six parameters affected by the predictions of each BOW. The final values used for the submissions are described in Fig. 3. The performance of this ensemble in the dev dataset was 54.73 of F1-score macro.

### 5.2 Results in the competition

In this subsection, we present the performance obtained in the competition; the best places are shown as performance references and other strategies were added to the comparison. In Table 3, we add the BOW-multiclass, these BOWs were constructed using the sklearn implemention, with the difference that this BOW has a multiclass target because they are trained with the dataset with all the labels.

- BOW-mutliclass 1: This BOW was created with unigrams, 100 attributes, tf-idf weighting and SVM classifier.
- BOW-multiclass 2: This BOW was created with unigrams, bigrams, tri-grams, 100 attributes, tf-idf weighting and SVM classifier.
- BOW-multiclass 3: This BOW was created with unigrams, bigrams, 100 attributes, tf weighting and SVM classifier.
- BOW-multiclass 4: This BOW was created with unigrams, bigrams, tri-grams, 200 attributes, tf without stopwords weighting and SVM classifier.

| | F1-score macro |
|---|---|
| 1st place | **0.474** |
| 2nd place | 0.446 |
| 3rd place | 0.441 |
| 4th place | 0.439 |
| RoBERTa-32 (4th place) | <u>0.439</u> |
| BOW ensemble | 0.432 |
| BOW-multiclass 1 | 0.460 |
| BOW-multiclass 2 | 0.451 |
| BOW-multiclass 3 | 0.450 |
| BOW-multiclass 4 | 0.437 |
| RoBERTa ensemble 32 | 0.443 |

Table 3: Our best model is underlined. The BOW-multiclass were not proposed to submission because their performance in the dev dataset did not surpass the proposed models. BOW-multiclass 1 would be placed second at the competition.

Our model RoBERTa-32 was placed fourth on the competition. The ranking provided by the organizers take the best run from each team, so our second model could be placed on top fifteen of the models. The RoBERTa ensemble 32 refers to the ensemble of RoBERTa models with lr $1e^{-5}$ and batch size 32, this model surpass our best model with little difference, as we see only one of them have performance similar with less computational resources.

In the case of BOW-multiclass, we did not include it in our proposal submissions as in previous experiments, and they did not obtain better performances than transformers and an ensemble of BOW.

## 6 Ethical issues

The automatic detection of mental illness, such as depression, using user-generated data raises several critical ethical considerations. One key aspect is the need to prioritize and ensure the anonymity of the users whose text is recorded for training and development purposes.

Crowd-sourcing is commonly employed in the context of labelling the data, where multiple annotators assess and assign labels to the instances. However, this process introduces a level of subjectivity, and there is no guarantee of perfect accuracy or consistency in the labelling. Annotators may have different interpretations or judgments, leading to potential discrepancies in the assigned labels. Consequently, the reliability and consistency of the labelled data may be influenced by the subjective opinions of the annotators.

The data used for automatic detection should ideally be collected with explicit user consent, where individuals knowingly and willingly provide their data for such purposes. However, in some cases, the data might have been obtained without users' explicit permission or awareness. This raises concerns about privacy violations and the potential discomfort or distress users may feel upon discovering their data is being used without their knowledge.

It is essential to prioritize data anonymization and protection of user identities throughout the entire data collection and storage process. Furthermore, efforts should be made to obtain explicit user consent when collecting data, ensuring individuals are fully aware of how their data will be used and can deny the use if they wish.

## 7 Conclusion

As we see in the previous subsection, multiclass classification is a difficult task for the complexity of depression detection. Our proposed models obtained performances similar to other teams in better places in the competition. The BOW ensemble obtained an F1-macro score close to our best-proposed model using less computational resources than a transformer model, as this model does not need GPU or a large amount of storage to be used.

The BOW-multiclass surpassed too the transformer model and the ensemble, even though the dev dataset did not surpass the proposed models; this could be because the test dataset is smaller than the dev dataset, and Machine Learning models as BOW tend to function better with fewer data.

In future work, we plan to explore better strategies for the values in the weights for the ensemble models of BOW and the rules made for the final predictions.

## Acknowledgments

This research was funded by *Consejo Nacional de Humanidades Ciencia y Tecnología* (CONAHCyT) master's degree grant #1141296. The authors thank to CONACyT, CIMAT and *Instituto Nacional de Astrofísica, Óptica y Electrónica* (INAOE) for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies (*Plataforma de aprendizaje profundo para tecnologías del lenguaje*) and CIMAT Bajio Supercomputing Laboratory (#300832). Sanchez-Vega acknowledges CONACyT for its support through the Program "Investigadoras e Investigadores por México" by the project "Desarrollo de Inteligencia Artificial aplicada a la prevención de violencia y salud mental." (ID.11989, No. 1311).

## References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Darío Gustavo Funez, Maria José Garciarena Ucelay, Maria Paula Villegas, Sergio Gastón Burdisso, Leticia Cecilia Cagnina, Manuel Montes y Gómez, and Marcelo Luis Errecalde. 2018. Unsl's participation at erisk 2018 lab. In *Conference and Labs of the Evaluation Forum*.

Rahmatullah Haand and Zhao Shuwang. 2020. The relationship between social media addiction and depression: a quantitative study among university students in khost, afghanistan. *International Journal of Adolescence and Youth*, 25(1):780–786.

Morteza Janatdoust, Fatemeh Ehsani-Besheli, and Hossein Zeinali. 2022. KADO@LT-EDI-ACL2022: BERT-based ensembles for detecting signs of depression from social media text. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 265–269, Dublin, Ireland. Association for Computational Linguistics.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *CoRR*, abs/2110.15621.

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. *CoRR*, abs/1907.11692.

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

Rodrigo Martínez-Castaño, Amal Htait, Leif Azzopardi, and Yashar Moshfeghi. 2021. Bert-based transformers for early detection of mental health illnesses. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings*, page 189–200, Berlin, Heidelberg. Springer-Verlag.

Gwenn Schurgin O'Keeffe, Kathleen Clarke-Pearson, Council on Communications, and Media. 2011. The Impact of Social Media on Children, Adolescents, and Families. *Pediatrics*, 127(4):800–804.

Rosa María Ortega-Mendoza, Delia Irazú Hernández-Farías, Manuel Montes y Gómez, and Luis Villaseñor-Pineda. 2022. Revealing traces of depression through personal statements analysis in social media. *Artificial Intelligence in Medicine*, 123:102202.

Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2022. Overview ofnbsp;erisk 2022: Early risk prediction onnbsp;thenbsp;internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings*, page 233–256, Berlin, Heidelberg. Springer-Verlag.

Rafał Poświata and Michał Perełkiewicz. 2022. OPI@LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pretrained language models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 276–282, Dublin, Ireland. Association for Computational Linguistics.

Kayalvizhi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Jerin Mahibha C, Kogilavani Shanmugavadivel, and Pratik Anil Rahood. 2023. Overview of the second shared task on detecting signs of depression from social media text. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Marcel Trotzek, Sven Koitka, and C. Friedrich. 2018. Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia. In *Conference and Labs of the Evaluation Forum*.

Wei-Yao Wang, Yu-Chien Tang, Wei-Wei Du, and Wen-Chih Peng. 2022. NYCU_TWD@LT-EDI-ACL2022: Ensemble models with VADER and contrastive learning for detecting signs of depression from social media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 136–139, Dublin, Ireland. Association for Computational Linguistics.

World Health Organization. 2022. Depression. `https://www.who.int/news-room/fact-sheets/detail/depression`. 18 june of 2023.