UMUTeam@LT-EDI-ACL2022: Detecting Signs of Depression from text

José Antonio García-Díaz and **Rafael Valencia-García***
Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain
{joseantonio.garcia8,valencia}@um.es

Abstract

Depression is a mental condition related to sadness and the lack of interest in common daily tasks. In this working-notes, we describe the proposal of the UMUTeam in the LT-EDI shared task (ACL 2022) concerning the identification of signs of depression in social network posts. This task is somehow related to other relevant Natural Language Processing tasks such as Emotion Analysis. In this shared task, the organisers challenged the participants to distinguish between moderate and severe signs of depression (or no signs of depression at all) in a set of social posts written in English. Our proposal is based on the combination of linguistic features and several sentence embeddings using a knowledge integration strategy. Our proposal achieved the 6th position, with a macro f1-score of 53.82 in the official leader board.

1 Introduction

The automatic analysis of depression is a medium that allows people to support their mental health (Evans-Lacko et al., 2018). The shared-task DepSign LT-EDI (ACL-2022) (Sampath et al., 2022) aims to measure the ability of neural networks and Natural Language Processing (NLP) tools to detect signs of depression from social media posts written in English. It is worth noting that this is not the first shared task concerning the identification of depression. In (Losada et al., 2017), the organisers of eRisk 2017 develop a pilot project which main purpose is the identification of early risk detection of depression.

In this shared task, the organisers proposed a multi-classification challenge that consists of identifying whether a moderate or severe sign of depression is observed in a short text or, on the contrary, no sign of depression is observed. For this, the performance of all participants is ranked using the macro averaged precision, recall and f1-score. The details of the dataset compilation can be found at (Kayalvizhi and Thenmozhi, 2022). The dataset is distributed into three folds: training, validation, and testing. We decided to use this distribution and not to merge train and validation to make a custom training-validation split. Table 1 depicts the label distribution per split. We can observe that the dataset is imbalanced, with many instances that reflect moderate signs of depression.

	Train	Validation	Test
Not depressed	1971	1830	-
Moderate	6019	2306	-
Severe	901	360	-
Total	8891	4496	3245

Table 1: Label distribution

Our research group has experience in Emotion Analysis. Specifically, we participated in the EmoEvalEs shared task (Plaza-del Arco et al., 2021), organised in the IberLEF 2021 workshop. This shared-task is about a multi-classification task of identification of emotions in Spanish (based on Ekman's basic emotions). Our participation is detailed at (García-Díaz et al., 2021b). Besides, we released the Spanish MisoCorpus 2021 and evaluated with different feature sets and neural network models (García-Díaz et al., 2022). In the same line, we evaluated in (García-Díaz et al., 2022) how to combine different feature sets and state-of-the-art neural network architectures for improving automatic hate-speech detectors. Specifically, we tested two strategies for combining the features: knowledge integration and ensemble learning. In this work we evaluate these strategies as well. Besides, as part of the doctoral thesis of one of the members of the team, we evaluate a subset of language-independent linguistic features in order to observe if they contribute to improve the performance of state-of-the-art embeddings.

*Corresponding author

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 145 - 148

2 Methodology

Our pipeline can be summarised as follows. First, documents are pre-processed by removing punctuation symbols, spaces, emojis, and punctuation. Second, four feature sets are extracted from the documents: linguistic features (LF), sentence embeddings from FastText (SE), BERT (BF), and RoBERTa (RF). Third, several neural networks with different combinations of the feature sets are trained using hyperparameter tuning. Forth, two additional ensembles are created to combine the features. Finally, we use the best neural network to get the final submission with the official test.

Next, we describe the feature extraction stage. The linguistic features (LF) are extracted with the UMUTextStats tool (García-Díaz and Valencia-García, 2022). The linguistic features are related to stylometry (for instance, word and sentence length, or Type-Token ratio), Part-of-Speech, emojis and generic social network jargon. The main advantage of linguistic features versus state-of-the-art embeddings is that linguistic features are easy to interpret at the same time they achieve promising results, specially in Author Analysis tasks (García-Díaz et al., 2021a). The sentence embeddings from FastText (SE) are extracted with the FastText tool (Mikolov et al., 2018). These sentence embeddings are not contextual. That is, the same word has the same representation, regardless of its context. Finally, the sentence embeddings from BERT (BF) and RoBERTa (RF) are extracted from distilled models (Sanh et al., 2019). We use the distilled versions because they require less computational resources. To obtain the sentence embeddings from BERT or RoBERTa, a hyperparameter selection stage of 10 models is conducted to obtain a good configuration of the models. Next, the sentence embeddings from BERT and RoBERTa are obtained from the `[CLS]` token (using the approach described at (Reimers and Gurevych, 2019)). During the hyperparameter selection stage, we use Tree of Parzen Estimators (TPE) (Bergstra et al., 2013) for determining the best parameters (weight decay, batch size, warm-up speed, number of epochs, and learning rate).

The next step is the training of several neural networks. We train a neural network for each feature set (LF, SE, BF, RF), and a neural network that combines all feature sets (LF + SE + BF + RF). All these neural networks are trained with hyperparameter selection. For this, we rely on Ray Tune (Liaw et al., 2018). For each training, we evaluate different number of hidden layers, neurons, batch size, learning rate or regularisation mechanisms. We distinguish between (1) shallow neural networks, that are simple neural networks composed of one or two hidden layers with the same number of neurons in each layer; and (2) deep neural networks, that have 3, 4, 5, 6, 7 or 8 hidden layers. Besides, the layers of deep neural networks are evaluated with different number of neurons disposed in several shapes (brick, triangle, diamond, rhombus, and funnel). For the rest of the parameters, we evaluate large batch sizes due to class imbalance, a dropout mechanism for regularisation (in different ratios), and small and large learning rates.

The results for the hyperparameter optimisation stage are shown in Table 2. We can observe that the best neural network that combines all features consisted in a shallow neural network composed of 2 wide hidden layers, with 128 neurons each. The batch size is large (512), the learning rate is large (0.01) and there is no activation function (is linear). Besides, this network uses a small dropout ratio of .1.

3 Results and discussion

We report the results achieved with the validation split. Table 3 depicts the macro average precision, recall, and f1-score of each feature set separately and combined with ensemble learning and two ensemble learning strategies: one based on the mode of the predictions and another based on averaging the predictions.

From the results achieved with the feature sets separately, BF is the one that achieves better results (77.27% of f1-score). This result is similar to RF (76.91% of f1-score) and outperforms largely SE and LF. With the knowledge integration strategy, the results outperform the ones achieved separately, with a f1-score of 77.90. Besides, when the results are combined with ensembles, the results are larger with the average of the probabilities (mean) achieving a macro f1-score of 78.69.

We decided to use for the final submission the predictions obtained with the knowledge integration strategy. This decision is taken because in past competitions we have achieved better results with this strategy with the official test (that is, we suspect this strategy generalises better than ensemble learning). Accordingly, we show the classification report of the validation split in Table 4 and its con-

	shape	\# of layers	first_neuron	dropout	lr	activation
LF	brick	1	48	0.1	0.001	relu
SE	brick	2	128	False	0.010	relu
BF	brick	1	48	0.1	0.010	relu
RF	brick	1	128	0.3	0.001	relu
K.I.	brick	2	128	0.1	0.010	linear

Table 2: Results for the best hyperparameters for each feature set separately or combined using knowledge integration. We include the shape of the neural network, the number of layers, the number of neurons in the first hidden layer, the dropout ratio, the learning rate, and the activation function

Feature set	**P**	**R**	**F1**
LF	61.42	61.44	60.44
SE	70.02	69.89	69.92
BF	78.80	75.97	77.27
RF	76.98	76.86	76.91
K. I.	79.88	76.30	77.90
Ensemble (Mode)	**80.52**	71.70	75.12
Ensemble (Mean)	80.47	**77.18**	**78.69**

Table 3: Macro average precision (P), recall (R), and f1-score (F1) of each feature set (LF, SE, BF and RF), the knowledge integration strategy (K.I) and the two ensemble learning strategies (mode and mean) with the validation split

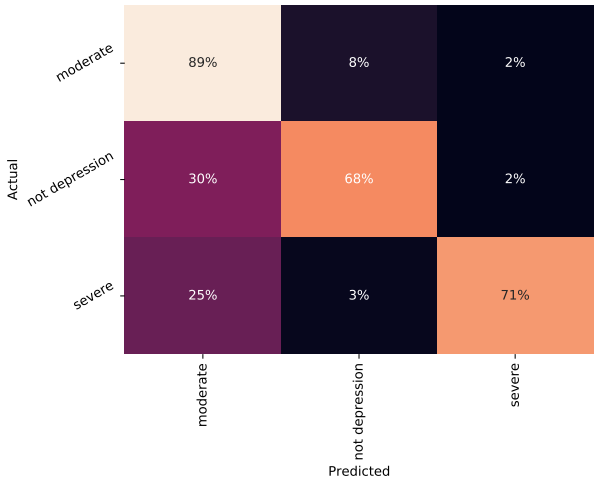

Figure 1: Confusion matrix of knowledge integration strategy with the validation split

fusion matrix in Figure 1. We can observe that the precision and recall of all labels are competitive, achieving a macro f1-score of 79.90% and a weighted f1-score of 81.41%. Moderate sign of depression (the majority label) is the one that achieves better precision and recall. Concerning the confusion matrix, we can observe that most wrong classifications occur between not depression and moderate depression and between severe and moderate depression. This means that our system does not mismatch severe failures, such as classifying severe signs of depression as not depression.

	P	**R**	**F1**
moderate	83.61	89.49	86.45
not depression	77.78	68.11	72.63
severe	78.26	71.29	74.61
macro avg	79.88	76.30	77.90
weighted avg	81.45	81.70	81.41

Table 4: Classification report of the knowledge integration strategy with the validation split, showing the precision (P), recall (R) and f1-score (F1) of each label and the macro and weighted scores

Next, Table 5 shows the official results in the leader board. We achieved 6th position in the task in a total of 31 teams. We achieve 53.82 of macro f1-score (4.48% below the best result).

Team	**R**	**P**	**F1**
OPI (1)	59.12	58.60	58.30
NYCU_TWD (2)	57.32	53.94	55.23
ARGUABLY (3)	57.20	53.03	54.67
BERT4EVER (4)	58.06	52.18	54.26
KADO (5)	57.04	52.63	54.22
UMUTeam (6)	55.75	52.48	**53.82**

Table 5: Official results, including the team name and the rank, the recall (R), precision (P), and the macro f1-score (f1)

4 Conclusions and promising research lines

Here we have described the participation of UMUTeam in the LT-EDI-ACL2022 shared task, concerning the identification of moderate and severe signs of depression in short texts. We achieved 6th position from a total of 31 participants with a system that combines linguistic features and three

forms of sentence embeddings using knowledge integration. We are proud of our participation as it has allowed us to evaluate a subset of language-independent linguistic features. Accordingly, we will continue to adapt our methods to English. Specifically, we will include linguistic features from figurative language, as the ones described at (del Pilar Salas-Zárate et al., 2020).

Acknowledgements

This work is part of the research project LaTe4PSP (PID2019-107652RB-I00) funded by MCIN/ AEI/10.13039/501100011033. This work is also part of the research project PDC2021-121112-I00 funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. In addition, José Antonio García-Díaz is supported by Banco Santander and the University of Murcia through the Doctorado Industrial programme.

References

James Bergstra, Daniel Yamins, and David Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123. PMLR.

María del Pilar Salas-Zárate, Giner Alor-Hernández, José Luis Sánchez-Cervantes, Mario Andrés Paredes-Valverde, Jorge Luis García-Alcaraz, and Rafael Valencia-García. 2020. Review of english literature on figurative language applied to social networks. *Knowledge and Information Systems*, 62(6):2105–2137.

Sara Evans-Lacko, Sergio Aguilar-Gaxiola, Ali Al-Hamzawi, Jordi Alonso, Corina Benjet, Ronny Bruffaerts, WT Chiu, Silvia Florescu, Giovanni de Girolamo, Oye Gureje, et al. 2018. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the who world mental health (wmh) surveys. *Psychological medicine*, 48(9):1560–1571.

José Antonio García-Díaz, Ricardo Colomo-Palacios, and Rafael Valencia-García. 2021a. Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020. *Future Generation Computer Systems*.

José Antonio García-Díaz, Ricardo Colomo-Palacios, and Rafael Valencia-Garcia. 2021b. Umuteam at emoevales 2021: Emosjon analysis for spanish based on explainable linguistic features and transformers.

José Antonio García-Díaz, Salud María Jiménez-Zafra, Miguel Angel García-Cumbreras, and Rafael Valencia-García. 2022. Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers. *Complex & Intelligent Systems*, pages 1–22.

José Antonio García-Díaz and Rafael Valencia-García. 2022. Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers. *Complex & Intelligent Systems*, pages 1–14.

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*.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.

David E Losada, Fabio Crestani, and Javier Parapar. 2017. erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 346–360. Springer.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Flor Miriam Plaza-del Arco, Salud M Jiménez Zafra, Arturo Montejo Ráez, M Dolores Molina González, Luis Alfonso Ureña López, and María Teresa Martín Valdivia. 2021. Overview of the emoevales task on emotion detection for spanish at iberlef 2021.

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.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.