

UPV-Symanto at eRisk 2021: Mental Health Author Profiling for Early Risk Prediction on the Internet

Angelo Basile¹, Mara Chinea-Rios², Ana-Sabina Uban^{3,4}, Thomas Müller²,
Luise Rössler², Seren Yenikent¹, Berta Chulví³, Paolo Rosso³ and
Marc Franco-Salvador²

¹Symanto Research, Nuremberg, Germany

²Symanto Research, Valencia, Spain

³PRHLT Research Center, Universitat Politècnica de València

⁴Human Language Technologies Research Center, University of Bucharest

Abstract

This paper presents the contributions of the UPV-Symanto team, a collaboration between Symanto Research and the PRHLT Center, in the eRisk 2021 shared tasks on gambling addiction, self-harm detection and prediction of depression levels. We have used a variety of models and techniques, including Transformers, hierarchical attention networks with multiple linguistic features, a dedicated early alert decision mechanism, and temporal modelling of emotions. We trained the models using additional training data that we collected and annotated thanks to expert psychologists. Our emotions-over-time model obtained the best results for the depression severity task in terms of ACR (and second best according to ADODL). For the self-harm detection task, our Transformer-based model obtained the best absolute result in terms of ERDE₅ and we ranked equal first in terms of speed and latency.

Keywords

risk detection, depression, self-harm, pathological gambling, social media, hierarchical networks, transformer

1. Introduction

The availability of user-generated texts on social media such as Reddit and Twitter, makes it possible to organize an early reaction to risks and threats as these are mentioned in conversations between users. It has been shown that social media language data can be used for detecting natural risks such as floods and earthquakes [1], predicting public health issues such as influenza [2], and analyzing riots and protest events [3]. In this work, we focus on predicting individual risk of mental disorder, within the context of our participation to the eRisk 2021 CLEF evaluation campaign [4]. We participate in all the three shared tasks proposed by the organizers: Early Detection of Signs of Pathological Gambling (Task 1), Early Detection of Signs of Self-Harm


CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

✉ angelo.basile@symanto.com (A. Basile); ana.uban+acad@gmail.com (A. Uban); luise.roessler@symanto.com (L. Rössler); seren.yenikent@symanto.com (S. Yenikent); proso@dsic.upv.es (P. Rosso); marc.franco@symanto.com (M. Franco-Salvador)

🆔 0000-0002-3312-9359 (A. Basile); 0000-0002-2313-9633 (M. Chinea-Rios); 0000-0003-2197-3947 (A. Uban); 0000-0002-8360-4189 (T. Müller); 0000-0003-4834-5326 (S. Yenikent); 0000-0001-7946-6601 (M. Franco-Salvador)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

(Task 2) and Measuring the Severity of Signs of Depression (Task 3). All three tasks are framed as author profiling tasks, i.e., some personal characteristics of an author have to be inferred from their writings.

Task 1 For Task 1, we built a system for classifying Reddit users as potential pathological gamblers based on their writings. This task was organized as an "only-test" task, with no training data released by the organizers. The test data is collected from Reddit following the procedure described in [5]. All the texts contained in the dataset are in English. Before the system submission deadline, no other information about the test data is known. Since the task focuses on early detection of signs of pathological gambling, the evaluation metrics take into account the number of posts processed before providing a positive prediction for each user.

Task 2 For this task, participants were asked to develop a system for predicting early signs of self-harm. The task was framed as a binary classification task (self-harm, no self-harm). As for Task 1, the data consists of Reddit comments in English, collected following [5]. For this task, the organizers provided a training dataset with posts from 763 labelled users, of which 145 belonging to the positive (i.e., SELF-HARM) class. This task is evaluated in the same way as Task 1.

Task 3 In contrast to both Task 1 and Task 2, Task 3 is not focused on early prediction, but on estimating the severity of the users' depression. As for Task 1 and Task 2, the data source is Reddit and all the texts are in English; for each Reddit user in the dataset (90 in total), the organizers collected their Reddit post history; furthermore, the organizers provided the answers to a depression questionnaire as filled by each user included in the dataset. The goal of Task 3 consists in estimating users' response to the questionnaire given their history of Reddit comments.

We approached the three tasks using a combination of neural models and manually engineered features, developed by domain experts. We collected additional data from Reddit and hired expert psychologists to annotate it. We obtained the best results in Task 2 and Task 3 according to several key metrics.

2. Data

We train our models in a supervised fashion using all the data released by the organizers. In addition to that, we augment the released training splits by collecting additional data from Reddit.¹ For Task 1, we build a training and development set from scratch, since no data was released by the organizers for this task. We follow the strategy described in [5] and run a series of queries looking for occurrences of the following strings in all of Reddit:

- *I was diagnosed with depression*
- *[I am]|[I'm] a problem gambler*
- *[I am]|[I'm] addicted to gambling*

We then collect all the comments and submissions from all the Reddit users who posted a text matching the queries. Furthermore, we collect all the comments and submissions posted to

¹We use the PushShift API [6].

a manually compiled list of subreddits.² We collected in total approximately 16 million texts.

Data Annotation Since we ran our initial queries against all of Reddit and since we collected indiscriminately all the posts and comments posted to specific subreddits, we expected the data to be noisy and to contain a lot of false positives.³ Considering the large size of the collected dataset, sampling a random portion of the data for annotation would have probably lead to a highly imbalanced label distribution. For these reasons, we adopted a "search as labelling" approach: all the data was dumped into a PostgreSQL database [7] and indexed using trigrams, allowing annotators to use their expertise for finding instances of the positive and negative class using free text queries. Two psychologists were hired to annotate the collected data. Annotators used the criteria provided in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) for each task assigned for each disorder; gambling disorder, non-suicidal self-injury, and major depressive disorder [8]. To annotate instances for the control groups, we used two different sets of labels. We used one (`definitely-no-gambler` and `definitely-no-self-harm`) for instances whose authors can safely be classified as a part of the control group, e.g., requests for help for someone other than the post author or news articles using a vocabulary that partially overlaps with the one used in the target group (e.g., articles on financial investment risk): when such an instance is found, we automatically labeled all the text from its author with the same label. We believe these instances to be the most challenging for the models. Another set of labels (`maybe-no-gambler` and `maybe-no-self-harm`) is used to annotate those instances that do not belong to the positive class, but nothing can be safely inferred about the author: in this case, we don't label all the texts from the same author as belonging to the negative class. Table 1 shows the label distribution in the annotated dataset. The most common feedback from the annotators involved the comorbidity issue. Accordingly, in all three tasks, respective disorders were observed to be co-occurring with signals from other types of disorders. For instance, in the depression task, it was observed that signals for anxiety disorders, post-traumatic stress disorder, and self-harm would commonly co-occur. While alcohol and drug addiction were the major comorbid conditions in the gambling task, self-harm was accompanied by depressive and anxiety signals. In some cases, the signals from the co-occurring conditions might have been stronger in text and led to misclassifications. Although this may be considered as a hindering effect for the training of the models, it in fact showcases the real-time conditions. Comorbidity of disorders is a common issue in mental health conditions [9]. Thus, this observation in our dataset is quite expected, and suggests that the models developed for such tasks should reflect this scenario.

3. Task 1: Early Detection of Signs of Pathological Gambling

We approach Task 1 in two steps. First, we train a text classifier on each text of each user independently, propagating the users' labels (e.g., *pathological gambler* or *not pathological gambler*) to their writings, assuming that the information contained in a single post can potentially be

²Each subreddit is a forum dedicated to a particular topic. The complete list of crawled subreddits can be found in Appendix A.

³For example, we noticed that many submissions on depression-related subreddits are from people asking for help for their loved ones instead of being ill themselves.

Table 1

Overview of the label distribution in the in-house annotated dataset. For task 2, these data was merged with the official dataset released by the organizers.

| Task | Label | # posts | # users |
|--------|--------------|---------|---------|
| Task 1 | gambler | 3143 | 722 |
| | no gambler | 1655 | 178 |
| Task 2 | self-harm | 104 | 28 |
| | no self-harm | 209 | 18 |

enough for classifying its author. Second, we build an alert-emitting system which computes the probability of a user being ill based on the averaged probabilities assigned to each processed post. To develop the models we use our manually annotated corpus.

3.1. Models

3.1.1. Transformer Model with Alert-Emitting System

For modelling task 1, we train a Transformer-based text classifier using a pretrained, small English uncased Bert model [10, 11].⁴ We report the hyper-parameter settings in Appendix B.1.

To build the alert-emitting system, we followed the work of the participants that obtained the best results in the 2020 edition of this shared task [12]. The system emits a risk alert if the average probability of the positive class is higher than a certain value θ , having considered a number of user posts between a minimum and maximum ψ and δ , respectively. We find the best values for θ , ψ , and δ according to 5 key metrics using a black-box optimization approach based on a Gaussian process.⁵ We tuned an early alert decision maker for each of these metrics: F1-score, latency-weighted F1-score, ERDE₅, ERDE₅₀, and an equally-weighted combination of all of them. Table 2 highlights the results. The models differ only with respect to θ , ψ , and δ .

Table 2

Development results for Task 1, with the automatically optimized hyper-parameters for the alert-emitting system: minimum number of posts (ψ), maximum number of posts (δ), minimum threshold for emitting a prediction (θ). The bold values in each column denote the optimized metric.

| model | ψ | δ | θ | P | R | F1 | latency _{F1} | ERDE ₅ | ERDE ₅₀ | norm avg |
|---------------|--------|----------|----------|--------|--------|---------------|-----------------------|-------------------|--------------------|---------------|
| UPV-Symanto 0 | 1 | 5 | 0.20 | 88.32% | 96.56% | 92.25% | 92.61% | 3.86% | 3.51% | 94.38% |
| UPV-Symanto 1 | 1 | 50 | 0.21 | 88.49% | 96.31% | 92.24% | 92.59% | 3.96% | 3.61% | 94.32% |
| UPV-Symanto 2 | 1 | 89 | 0.38 | 43.78% | 79.28% | 56.41% | 56.63% | 13.59% | 11.03% | 72.10% |
| UPV-Symanto 3 | 1 | 5 | 0.25 | 37.26% | 88.29% | 52.41% | 52.61% | 11.72% | 11.47% | 70.46% |
| UPV-Symanto 4 | 1 | 50 | 0.35 | 42.01% | 82.88% | 55.76% | 55.98% | 13.61% | 10.76% | 71.84% |

⁴Specifically, we use the *bert_en_uncased_L-2_H-128_A-2* model available from the <https://tfhub.dev> model repository.

⁵We use the implementation available in *scikit-optimize*.

3.2. Evaluation and Results

The official evaluation setup of Task 1 is composed of two set of metrics: one for a *decision-based evaluation* and one for a *ranking-based evaluation*. The decision-based evaluation provides an estimation of models’ performance at classifying at-risk users, while ranking-based evaluation is used to asses the goodness of a model at sorting users by their level of risk. For the decision-based evaluation, the standard classification metrics are used, i.e., precision (P), recall (R) and f-measure (F1), together with a set of metrics which take into account the time required to emit an alert. We measure classification performance considering the number of posts required to emit a correct prediction for each user (using $ERDE_5$ and $ERDE_{50}$, requiring 5 and 50 posts respectively) and considering the number of writings required by a system for finding true positive instances (using $latency_{TP}$ and latency-weighted F1). Table 3 shows the official results as computed by the organizers on the test set. A detailed description of the metrics can be found in [5] and [13]. The ranking-based evaluation is conducted using the $P@N$ and $NDCG@N$ metrics.⁶

Table 3

Decision-based evaluation for Task 1. For comparison, we include the runs that obtained the best results in each metric, as reported in [4].

| run | P | R | F1 | $ERDE_5$ | $ERDE_{50}$ | $latency_{TP}$ | speed | $latency_{F1}$ |
|----------------|-------------|-------------|-------------|-------------|-------------|----------------|----------|----------------|
| UPV-Symanto 0 | .422 | .415 | .077 | .088 | .087 | 1 | 1 | .077 |
| UPV-Symanto 1 | .420 | .457 | .074 | .097 | .091 | 1 | 1 | .074 |
| UPV-Symanto 2 | .030 | .238 | .053 | .093 | .091 | 1 | 1 | .053 |
| UPV-Symanto 3 | .035 | .409 | .064 | .098 | .097 | 1 | 1 | .064 |
| UPV-Symanto 4 | .028 | .256 | .051 | .098 | .095 | 1 | 1 | .051 |
| UNSL 2 (Best) | .586 | .939 | .721 | .073 | .020 | 11 | .961 | .693 |
| RELAI 0 (Best) | .138 | .988 | .243 | .048 | .036 | 1 | 1 | .243 |

4. Task 2: Early Detection of Signs of Self-Harm

We experimented with two types of models for approaching Task 2. The first model mirrors our work for Task 1, using a Transformer model to classify each post individually and then predicting a label for a user based on the probabilities assigned to their writings. A second model consists of a hierarchical LSTM-based architecture with attention (HAN) using a set of hand-crafted features. For both types of models, at inference time we use the alert-emitting system described in Section 3.1.1. Runs UPV-SYMANTO 0, 2 and 3 are based on HAN, while runs UPV-SYMANTO 1 and 4 are based on a Transformer.

⁶We don’t report here on the official ranking-based evaluation for Task 1, since due to a bug our model always predicts the negative class and thus all the metrics are equal to 0.

4.1. Models

4.1.1. Transformer Model

The Transformer-based architectures that we use for modelling Task 2 are the same that we used for Task 1, with the same hyper-parameters described in Appendix B.1.

4.1.2. Hierarchical Attention Network with Composite Features

In Task 2, we used a Hierarchical Attention Network (HAN)[14] with multiple linguistic features. Here we describe the features used, the experimental setup and the network architecture.

Content features. We include a general representation of text content by transforming each text into word sequences. Preprocessing of texts includes lowercasing and tokenizing, removing punctuation and numbers; function words are not excluded. Most frequent 20,000 words were selected to form the vocabulary, and words not in the vocabulary were represented as a special "unknown" token. When passed as input to the neural networks, words within a sequence were encoded as embeddings of dimension 300. In order to initialize the weights of the embedding layers, we started from pre-trained GloVe embeddings [15]⁷.

Style features. We aim at representing the stylistic level of texts through including function word and pronoun features. Function words have traditionally been used as stylistic markers, whereas increased use of pronouns, especially first person pronouns, has been shown to correlate with mental disorder risk [16]. We include two separate stylistic features: firstly, we extract from each text a numerical vector representing function words frequencies as bag-of-words. We complement these with features extracted from the LIWC lexicon [17], including pronoun usage and other syntactical features, as described below.

LIWC features. The LIWC [17]⁸ is a lexicon mapping words in the English vocabulary to lexico-syntactic features of different kinds. It has been widely used in computational studies for analysing how suffering from mental disorders manifests in authors' writings. LIWC categories have the capacity to capture different levels of language: including style (through syntactic categories), emotions (through affect categories) and topics (through content-oriented categories such as words referring to cognitive or analytical processes, or words referring to topics such as money, health or religion). We use LIWC 2015 and include in our analysis all 64 categories in the lexicon, and represent them as numerical vectors by computing for each category the ratio of words in a text that are related to the category, according to the lexicon.

Emotions and sentiment features. We dedicate a few features to represent emotional content in our texts, since the emotional state of a user is known to be highly correlated with their mental health. Several of the LIWC categories aim to capture sentiment polarity and emotion content (*negative emotion, positive emotion, affect, sadness, anxiety*). We additionally include a second lexicon, the NRC emotion lexicon [18], which is dedicated exclusively to emotion representation, containing 8 different emotion categories, as well as the 2 sentiment categories: *anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise, trust*.⁹ We represent

⁷<http://nlp.stanford.edu/data/glove.840B.300d.zip>

⁸<http://www.liwc.net/>

⁹Time limitations at the inference stage prevented us from using the more sophisticated sentiment and emotion model available through Symanto's Text Analysis API (<https://www.symanto.com/api/>), which we intend to explore

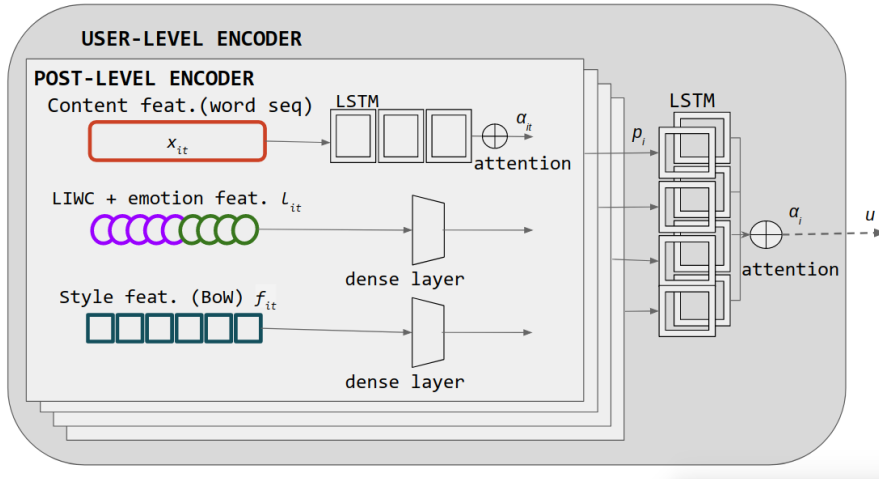


Figure 1: Hierarchical attention network architecture.

NRC features similarly to LIWC features, by computing for each category the proportion of words in the text which are associated with that category.

Experimental setup. We trained our models on the eRisk 2020 [19] training data for the self-harm task. For one of the models (UPV-SYMANTO 3), we used transfer learning by first pretraining all the model’s parameters on eRisk data annotated for anorexia [20], and subsequently training them on the self-harm data.

During the training phase, we do not consider social media posts individually as datapoints, since they are too short to be sufficiently predictive. Instead, we generate our datapoints by grouping sequences of c chronologically consecutive posts into larger chunks, to obtain more consistent samples of text as our datapoints. Features are computed at chunk-level. We use different values for c : UPV-SYMANTO 0 uses chunks of 80 posts, while UPV-SYMANTO 2 uses 10 posts per chunk, and UPV-SYMANTO 3 uses 50 posts per chunk.

Hierarchical Attention Network. Hierarchical attention networks were introduced in [14] where they were used for review classification, by representing a text as a hierarchical structure where a document is comprised of sentences and a sentence is comprised of words. We propose that social media data in our setup is well suited to such a hierarchical representation; in our case the hierarchy consists of user post histories, which are composed of social media posts, which are in turn composed of word sequences. Especially since the evolution of the mental state of a user is in itself a relevant indicator for the development of a disorder, as shown in [21], user-level representations are expected to be natural and useful for modelling this problem.

In the hierarchical setup, posts within a chunk (datapoint) are stacked to form a hierarchical structure: word sequences (truncated at 256 words), as well as the rest of vectorial numerical and bag-of-words features, are stacked to form bi-dimensional vectors. Bag-of-words and numerical features also follow a hierarchical structure, with a set of features extracted for each post in the group, and stacked together into bi-dimensional vectors. The hierarchical network is composed

in future work.

of two components: a *post-level encoder*, which produces a representation of a post, and a *user-level encoder*, which generates a representation of a user’s post history. Each of the posts in the input datapoint is encoded with the post-level encoder, and then they are stacked to form a bi-dimensional representation, which is then concatenated with the other features, and passed to the user-level encoder. We choose to model the user-level encoder as an LSTM layer with attention, with 32 units. The output of the user encoder is connected to the output layer which generates the final prediction. A depiction of the hierarchical architecture is shown in Figure 1.

We use batch normalization and L2 regularization. Binary cross-entropy is used as a loss function. More details on the network’s configuration are found in the Appendix.

Table 4

Development results for Task 2, with the automatically optimized hyper-parameters for the alert-emitting system: minimum number of posts (ψ), maximum number of posts (δ), minimum threshold for emitting a prediction (θ). The bold values in each column denote the optimized metric. **H**: Hierarchical Attention Network; **T**: Transformer model.

| | model | ψ | δ | θ | P | R | F1 | latency _{F1} | ERDE ₅ | ERDE ₅₀ | norm avg |
|----------|---------------|--------|----------|----------|--------|--------|--------------|-----------------------|-------------------|--------------------|----------------|
| H | UPV-Symanto 0 | 5 | 46 | 0.40 | 66.30% | 60.60% | 63.3% | 62.60% | 17.90% | 11.60% | 74.10% |
| T | UPV-Symanto 1 | 1 | 5 | 0.05 | 53.70% | 62.50% | 57.80% | 58.0% | 12.80% | 12.50% | 72.63% |
| H | UPV-Symanto 2 | 1 | 50 | 0.05 | 61.70% | 63.50% | 62.60% | 62.60% | 14.0% | 11.40% | 74.93% |
| H | UPV-Symanto 3 | 4 | 100 | 3 | 53.90% | 73.10% | 62.00% | 61.60% | 17.70% | 10.5% | 73.86% |
| T | UPV-Symanto 4 | 3 | 100 | 0.10 | 44.60% | 75.00% | 55.90% | 55.70% | 16.00% | 12.30% | 70.848% |

4.2. Evaluation and Results

The evaluation of task 2 mirrors exactly the setup of task 1. Table 4 reports the development results and Table 5 highlights the official results on the test set as released by the organizers.

Table 5

Official test results for the decision-based evaluation for Task 2.

| | run | P | R | F1 | ERDE ₅ | ERDE ₅₀ | latency _{TP} | speed | latency _{F1} |
|----------|---------------------|-------------|------------|-------------|-------------------|--------------------|-----------------------|------------|-----------------------|
| H | UPV-Symanto 0 | .307 | .678 | .422 | .097 | .051 | 5 | 1.0 | .416 |
| T | UPV-Symanto 1 | .276 | .638 | .385 | .059 | .056 | 1 | .996 | .385 |
| H | UPV-Symanto 2 | .313 | .645 | .422 | .072 | .053 | 2 | .984 | .420 |
| H | UPV-Symanto 3 | .301 | .770 | .433 | .089 | .044 | 5 | .992 | .426 |
| T | UPV-Symanto 4 | .198 | .711 | .310 | .082 | .063 | 3 | .961 | .307 |
| | UNSL 4 (Best) | .532 | .763 | .627 | .064 | .038 | 3 | .992 | .622 |
| | Birmingham 2 (Best) | .757 | .349 | .477 | .085 | .07 | 4 | .988 | .472 |
| | CeDRI 2 (Best) | .105 | 1.0 | .19 | .096 | .094 | 1 | 1.0 | .19 |

5. Task 3: Measuring the Severity of Signs of Depression

Task 3 consists of filling a questionnaire with 21 questions related to the user’s mental state from the user’s Reddit post history.

Table 6

Ranking-based results for Task 2 computed using 1 and 10 posts per users.

| # writings | | run | P@10 | NDCG@10 | NDCG@100 |
|------------|----------|---------------|------------|-------------|-------------|
| 1 | H | UPV-Symanto 0 | 0.8 | 0.83 | 0.53 |
| | T | UPV-Symanto 1 | 0.8 | 0.88 | 0.5 |
| | H | UPV-Symanto 2 | 0.8 | 0.82 | 0.55 |
| | H | UPV-Symanto 3 | 0.6 | 0.7 | 0.51 |
| | T | UPV-Symanto 4 | 0.9 | 0.93 | 0.53 |
| 100 | H | UPV-Symanto 0 | 0.9 | 0.94 | 0.67 |
| | T | UPV-Symanto 1 | 0.8 | 0.69 | 0.64 |
| | H | UPV-Symanto 2 | 0.8 | 0.83 | 0.59 |
| | H | UPV-Symanto 3 | 0.9 | 0.94 | 0.69 |
| | T | UPV-Symanto 4 | 0.9 | 0.81 | 0.65 |

5.1. Models

5.1.1. Emotions over Time Model

As one of our models, we chose an approach based on the evolution of emotions and certain psycho-linguistic features over time. Unlike other models used in this task, this approach models users not by extracting static features from their writings, but instead as time series describing their communication style related to emotions and self-expression over time.

Features. We use, to this effect, some of the features introduced in Section 4.1, namely: the 10 emotion categories in the NRC lexicon, and in addition 3 categories of the LIWC lexicon related to self-reference: *I* (usage of first person singular pronoun), *we* (usage of first person plural pronoun) and *ppron* (overall usage of personal pronouns). We compute scores for each of these categories for each post in the dataset, in a similar way to the feature extraction step for Task 2: for a given text and feature (lexicon category), we compute the number of words in the text corresponding to that category, normalized by the text length.

In order to obtain time series for each of the considered features, we compute the scores for a given user aggregated at the day level (computed over all texts posted in one day by a given user). In this way, we allow a fair comparison between users who have different habits in terms of frequency of posting, but who might nevertheless exhibit similar patterns in terms of emotion evolution over time. We also apply a rolling average of 100 days over the obtained scores, so as to reduce noise.

User Similarity over Time. In order to obtain predictions for a given user, we use the computed time series to define a similarity metric between users, and then predict answers to the questionnaire by imitating the answers of similar users in the training set. The similarity metric between users includes two factors: the static scores for the extracted features for the two users, as well as the correlations over time for the two users.

1. Static scores. We compute the average score for each of the considered features for a given user across all their writings (as a score between 0-1). The distance between two users will be computed as the arithmetic difference between the scores for two users (If d

is the distance between two users, the similarity between two users is then $1 - d$).

2. Correlations over time. These are computed between the time series of feature scores corresponding to two given users. Since the time series for any two users are not guaranteed to have the same length, we attempt to "align" the two time series by finding the maximum correlation between them. We use a sliding window of the length of the smaller of the time series, and compute correlation scores for all possible alignments between the two time series, then take the maximum correlation score as the similarity between the two time series.

The final similarity score between two users is computed as the sum between the static and temporal component, both factors contributing with equal weight.

Predictions for a given user are computed separately for each question in the questionnaire, as a weighted mean to the answers of the most similar 15 users in the training data, weighing the answers (as integers) with the corresponding similarity scores, and rounding the result to the nearest integer (in order to obtain a valid answer to the question¹⁰). In this way, we approach the prediction of answers as a regression problem, by considering a continuous range of possible answers, and are able to obtain good approximations for the overall level of depression (obtaining high ADODL and ACR scores), even when the exact answer is not correctly predicted.

5.1.2. Classification with Reddit BERT model

This model is trained in a two step approach. We first crawl posts from subreddits related to mental health issues such as depression, self harm and anxiety. We group the data into 13 categories related to mental health and an additional category consisting of random posts. We then train a balanced classifiers to discriminate between these 14 classes. We train a classifier based on *distilroberta-base* [22] and another one based on *roberta-base* [23] we call these models SUBREDDIT14 and SUBREDDIT14-ROBERTA-BASE, respectively.

In a second step we extract the [CLS] embedding of the pre-trained model for every post in the training and test datasets as well as the probability of the *depression* class. While the first is used as the main representation for classification the second one gives us a notion of relevance.

For every user in the training and tests sets we then average the embeddings of their posts to obtain the final user representation. Given this representation we train a classifier for each of the 21 questions. Since the dataset contains a small number of users we find it helpful to create multiple examples per users by sampling 80% of the user's posts. This can be understood as a form of random dropout. Additionally we find it beneficial to restrict to the posts where the pre-trained model predicts a probability of > 0.07 ¹¹ for the *depression* class. The motivation is that many posts are unrelated to the user's mental state and that this filtering removes the noise introduced by these irrelevant posts.

¹⁰For questions where one answer had two variations (*a* and *b*), we ignored the variation and only considered the integer value.

¹¹Recall that we trained a balanced classifier on 14 classes so that the average probability assigned to a class is ≈ 0.07

Table 7

Development results for Task 3 (trained on the 2019 data and evaluated on the 2020 data).

| model | AHR | ACR | ADODL | DCHR | MEAN |
|---|--------------|--------------|--------------|--------------|-------|
| RANDOM ¹² | 28.81 | 63.38 | 80.15 | 27.29 | 49.90 |
| Emotion over time model | 27.14 | 74.35 | 83.19 | 33.53 | 54.55 |
| SVM (SUBREDDIT14) | 38.23 | 69.30 | 81.18 | 24.29 | 53.25 |
| SVM (SUBREDDIT14-ROBERTA-BASE) | 39.25 | 70.25 | 83.04 | 35.71 | 57.06 |
| SVM (SUBREDDIT14, most recent 30 posts) | 35.78 | 67.96 | 82.43 | 35.71 | 55.47 |
| Random-Forest (SUBREDDIT14, most recent 30 posts) | 35.99 | 68.78 | 83.52 | 35.71 | 56.00 |
| UPV 2020 System 1 [24] | 34.56 | 67.44 | 80.63 | 35.71 | 54.59 |
| UPV 2020 System 2 (Best) [24] | 36.94 | 69.02 | 81.72 | 31.53 | 54.80 |
| BioInfo@UAVR (Best) [25] | 38.30 | 69.21 | 76.01 | 30.00 | 53.38 |
| iLab run2 (Best) [12] | 37.07 | 69.41 | 81.70 | 27.14 | 53.83 |
| relai_lda_user (Best) [26] | 36.39 | 68.32 | 83.15 | 34.29 | 55.54 |

Table 8

Official Evaluation results for Task 3

| model | AHR | ACR | ADODL | DCHR | MEAN |
|---|--------------|--------------|--------------|-------|-------|
| Emotion over time model | 34.17 | 73.17 | 82.42 | 32.50 | 55.57 |
| SVM (SUBREDDIT14-ROBERTA-BASE) | 32.20 | 66.05 | 77.28 | 26.25 | 50.45 |
| SVM (SUBREDDIT14) | 34.58 | 67.32 | 75.62 | 26.25 | 50.94 |
| SVM (SUBREDDIT14, most recent 30 posts) | 33.15 | 66.05 | 75.42 | 23.75 | 49.59 |
| Random-Forest (SUBREDDIT14, most recent 30 posts) | 33.09 | 66.39 | 76.87 | 23.75 | 50.03 |
| RELAI etm (BEST) | 38.78 | 72.56 | 80.27 | 35.71 | 56.83 |
| CYUT run2 (BEST) | 32.62 | 69.46 | 83.59 | 41.25 | 56.73 |

5.2. Evaluation and Results

Following the setup used in the shared task we use the following metrics: Average Hit Rate (AHR), Average Closeness Rate (ACR), Average Difference between Overall Depression Levels (ADODL), Depression Category Hit Rate (DCHR) and the average of the former four metrics (MEAN).

6. Discussion

In Task 3, we have considered that among the features used for mental disorder detection in previous literature, LIWC categories, emotions and personal pronouns usage have consistently been shown to be relevant for this task [27, 28, 29, 30, 31, 32, 33]. In previous work [34] we have noticed that the expression of emotions in relation with the use of personal pronouns reveals a specific pattern in users diagnosed with a mental disorders. For example, the use of “I” and personal pronouns in general present differences in the correlation with all positive

¹²Random answer drawn from the train distribution of each questions. Metrics are averaged over 10 runs.

emotions between depressed and not depressed people: the more depressed people use “I” and personal pronouns, the more they express positive emotions like joy, anticipation and trust, and the opposite happens with not depressed people. On this basis, in task 3 we considered that similarity in the expression of emotions and in the use of personal pronouns could be a predictor of responses in the Beck Depression Inventory. The good results obtained on the ACR and ADODL reinforce the idea that the expression of emotions, both in a static way and in its evolution over time, are a strong sign of development of depression.

6.1. Negative Results

We experimented with modelling all the three tasks simultaneously with a single Multi-Task Learning (MTL) model [35]. We aggregated all the datasets and trained a *roberta-base* [23] model, using a masked version of the cross-entropy loss, which does not penalize the model when training on instances with missing labels. These experiments did not provide good results.

For Task 1, the promising results obtained during development degraded terribly on the official test set due to a bug introduced right before the submission.

The choice of post chunking for training our hierarchical attention network for self-harm detection (Task 2) was motivated by preliminary experiments showing that classifying individual posts (using a network with a comparable architecture) does not achieve reasonable performance on the development data. We also experimented with different number of posts per chunk (including training and prediction), ranging between 10 and 90 posts per chunk and have seen that, overall, prediction performance (in terms of F1-score) increases proportionally with the size of chunks, while the models using larger chunks lose some performance in terms of latency-based metrics (latency-weighted F1 and ERDE scores). We tuned the early-alert decision mechanism applied to the trained models’ predictions (in terms of the different metrics evaluated on the development data) for choosing the configurations of our official runs.

In Task 3, an interesting question is why we obtain good scores in ACR and ADODL (that informs about scores predictions item by item and about overall depression level), but we fail to predict the depression category. What is observed is that we lose a lot of precision when we move from an ordinal scale, such as ACR and ADODL, to a categorical scale such as DCHR. We consider for future work the possibility to examine what patterns in ACR and ADODL are present in users who have been well classified and what patterns are present in those who have been misclassified in terms of categories for depression. It could be possible to apply a correction factor to certain extreme scores on certain items if we observe that these scores and these items play a major role in the errors of user’s classification to a certain level or category. Another interesting observation is that the classification-based models perform well on the development setup but not on test. A possible explanation is that there is a discrepancy of users with minimal and moderate depression: 24% on test, 47% on dev and 40% on train (Appendix C). However, it is unclear why this affects the classification-based models more than other approaches.

6.2. Motivation and Intended Usage

Newly emerging AI-supported services are in a promising position in the use of early detection and prevention of mental health conditions. The necessity to implement such services into everyday life is becoming more relevant, especially considering recent incidents like the CoViD-19 pandemic where many people suffered from psychological problems [36]. Digital mental health solutions (e.g. therapy programs, chat bots, smart device applications) is one of the largest use case areas in this sense. Early detection is a vital aspect of therapeutic interventions which makes the process more effective and prevents an aggravation of symptoms [37]. AI-supported systems could be used as a preliminary analysis tool in clinical settings to enable an early and preventive way of determining the type of the condition, severity of symptoms, and recommendations for a successful therapy concept [38]. Furthermore, digital solutions could solve the issue of accessibility, and stigmatization while providing individuals a healthy and unbiased way of self-help [39]. Besides the application in a clinical environment, AI-supported tools would be of use for general well-being practices. A holistic understanding of mental health requires not only the detection of problems but also the positive build-up of human psychology. Linguistic social media data is able to reflect different components of well-being, thus provide important insights into the everyday representations of mental health topics [40]. Early detection models developed by considering such insights could help to raise awareness on self-reflection and foster preventive lifestyle interventions. Academic and organizational institutions, which possess large application and impact areas on individuals, could use such tools to predict well-being of students and employees, and spot and support at-risk individuals [41, 42].

7. Related Work

Apart from the overviews of the previous editions of the eRisk shared task [20, 19], the closest literature to our work can probably be found in the review of CLPsych 2015 shared task on predicting depression and PTSD from Twitter data [43].

There are many studies in both computational linguistics and psychology which approach the problem of analyzing the language of people suffering from a mental disorder, especially depression. Many of these studies perform simple quantitative analyses or use traditional machine learning models (such as logistic regression). Recently, more studies have started employing deep learning for mental disorder detection, generally using word sequences as features [44, 45, 46]. A few recent studies also use pre-trained transformers for detection of mental health disorders [47, 48].

Hierarchical attention networks have successfully been used for mental health disorder detection in the past, including previous editions of the eRisk shared task: Mohammadi et al. [49] use HANs for anorexia detection (obtaining best results at the eRisk 2019 shared task [20]). Recently, Rao et al. [50] use hierarchical networks for depression detection, and Amini and Kosseim [51] use them for anorexia detection. All previously mentioned studies use HANs with standard word embedding features. Hierarchical attention networks using multiple linguistic features have previously been used for self-harm detection in eRisk 2020 [24], as well as for studying the detection of other mental health disorders as well as the model's explainability,

including for depression, anorexia and post-traumatic stress disorders [52].

Most computational studies model mental disorder symptoms as static phenomena, whereas the evolution of mental disorder markers, as well as their prevalence in texts posted by a user, is an important indicator of mental disorder risk. We mention one previous study [21] in which the authors attempt to classify time series representing the mood of social media users in order to predict occurrence of anorexia, with promising results. One recent study attempts a more in-depth analysis of emotions and other psycho-linguistic features over time [34].

Emotions have been previously shown to be relevant for modelling mental disorders, but not many go beyond simple quantitative analyses. We mention an approach focused on a fine-grained analysis of emotions, published three studies on depression, anorexia and self-harm detection [53, 54, 55]. Starting from Plutchik's eight basic emotions [56], the authors use word embedding spaces to automatically identify sub-emotions, which they use as features for their classifiers, trained to detect depression [53], anorexia [54] and self-harm [55] respectively.

The "search as labelling" approach that we used to annotate our internal Reddit corpus is described in [57].

8. Conclusion

In this paper we presented the contributions of the UPV-Symanto team in the eRisk 2021 shared tasks: gambling addiction and self-harm detection and the prediction of depression levels, based on social media text data. We have used a variety of models and techniques, including Transformers, hierarchical attention networks with multiple linguistic features, a dedicated early alert decision mechanism, and temporal modelling of emotions. We ranked first in terms of ACR and second in terms of ADODL for Task 3, exceeding the previous state-of-the-art for this eRisk shared task [19, 4], as well as best results for Task 2 in terms of $ERDE_5$ score. We conclude that our methods are promising, encouraging the use of emotion and linguistic features, temporal modelling, and dedicated early detection mechanisms.

Acknowledgements

The authors from Universitat Politècnica de València thank the EU-FEDER Comunitat Valenciana 2014-2020 grant IDIFEDER/2018/025. The work of Paolo Rosso was in the framework of the research project PROMETEO/2019/121 (DeepPattern) by the Generalitat Valenciana. We would like to thank the two anonymous reviewers who helped us improve this paper.

Appendix

A. List of crawled subreddits

For augmenting the training data, we collected all the posts and comments from the following list of subreddits:

- r/ADHD/
- r/Anxiety/
- r/aspergers/
- r/bipolar/
- r/BipolarReddit/
- r/BPD/
- r/CPTSD/
- r/depression/
- r/GamblingAddiction
- r/mentalhealth/
- r/OCD/
- r/problemgambling/
- r/schizophrenia/
- r/selfharm/
- r/SuicideWatch/

B. Hyper-Parameters

B.1. Transformer-based Model

- batch size = 8
- optimizer = Adam
- dropout = 0.1
- learning rate = $5e-5$
- early stopping patience = 7
- epochs = 15
- maximum sequence length = 512

B.2. Hierarchical Attention Network

- LSTM units (post encoder) = 128
- dense BoW units = 20
- dense lexicon units = 20
- LSTM units (user encoder) = 32
- dropout = 0.3

- $l_2 = 0.00001$
- optimizer = Adam
- learning rate = $1e-4$
- early stopping patience = 5
- epochs = 25
- maximum sequence length = 256
- posts per chunk = 80

C. Task 3 - Risk Category Distribution

Table 9

Risk category distribution for Task 3

| name | minimal | mild | moderate | severe |
|--------------|---------|------|----------|--------|
| train (2019) | 0.20 | 0.20 | 0.20 | 0.40 |
| dev (2020) | 0.14 | 0.33 | 0.26 | 0.27 |
| test (2021) | 0.08 | 0.16 | 0.34 | 0.43 |

References

- [1] K. Kireyev, L. Palen, K. Anderson, Applications of topics models to analysis of disaster-related twitter data, in: NIPS workshop on applications for topic models: text and beyond, volume 1, Canada: Whistler, 2009.
- [2] E. Aramaki, S. Maskawa, M. Morita, Twitter catches the flu: Detecting influenza epidemics using Twitter, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Edinburgh, Scotland, UK., 2011, pp. 1568–1576. URL: <https://www.aclweb.org/anthology/D11-1145>.
- [3] J. Sech, A. DeLucia, A. L. Buczak, M. Dredze, Civil unrest on Twitter (CUT): A dataset of tweets to support research on civil unrest, in: Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020), Association for Computational Linguistics, Online, 2020, pp. 215–221. URL: <https://www.aclweb.org/anthology/2020.wnut-1.28>. doi:10.18653/v1/2020.wnut-1.28.
- [4] J. Parapar, M.-R. Patricia, D. E. Losada, F. Crestani, Overview of erisk 2021: Early risk prediction on the internet, in: Proceedings of the Twelfth International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2021.
- [5] D. Losada, F. Crestani, A test collection for research on depression and language use, in: Proc. of Experimental IR Meets Multilinguality, Multimodality, and Interaction, 7th International Conference of the CLEF Association, CLEF 2016, Evora, Portugal, 2016, pp. 28–39. URL: https://citius.usc.es/sites/default/files/publicacions_postprints/clef.pdf.
- [6] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, J. Blackburn, The pushshift reddit dataset, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 14, 2020, pp. 830–839.

- [7] M. Stonebraker, L. A. Rowe, The design of postgres, *ACM Sigmod Record* 15 (1986) 340–355.
- [8] A. P. Association, et al., *Diagnostic and statistical manual of mental disorders (DSM-5®)*, American Psychiatric Pub, 2013.
- [9] O. Plana-Ripoll, C. B. Pedersen, Y. Holtz, M. E. Benros, S. Dalsgaard, P. De Jonge, C. C. Fan, L. Degenhardt, A. Ganna, A. N. Greve, et al., Exploring comorbidity within mental disorders among a danish national population, *JAMA psychiatry* 76 (2019) 259–270.
- [10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: <https://www.aclweb.org/anthology/N19-1423>. doi:10.18653/v1/N19-1423.
- [11] I. Turc, M.-W. Chang, K. Lee, K. Toutanova, Well-read students learn better: On the importance of pre-training compact models, *arXiv preprint arXiv:1908.08962* (2019).
- [12] R. Martínez-Castaño, A. Htait, L. Azzopardi, Y. Moshfeghi, Early risk detection of self-harm and depression severity using bert-based transformers: ilab at clef erisk 2020, *Early Risk Prediction on the Internet (2020)*. URL: http://ceur-ws.org/Vol-2696/paper_50.pdf.
- [13] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk: early risk prediction on the internet, in: *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2018, pp. 343–361.
- [14] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, 2016*, pp. 1480–1489.
- [15] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
- [16] M. Trotzek, S. Koitka, C. M. Friedrich, Linguistic metadata augmented classifiers at the clef 2017 task for early detection of depression., in: *CLEF (Working Notes)*, 2017.
- [17] J. W. Pennebaker, M. E. Francis, R. J. Booth, *Linguistic inquiry and word count: Liwc 2001*, Mahway: Lawrence Erlbaum Associates 71 (2001) 2001.
- [18] S. M. Mohammad, P. D. Turney, *Nrc emotion lexicon*, National Research Council, Canada 2 (2013).
- [19] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk at clef 2020: Early risk prediction on the internet (extended overview) (2020). URL: http://ceur-ws.org/Vol-2696/paper_253.pdf.
- [20] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk 2019 early risk prediction on the internet, in: *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2019, pp. 340–357. URL: http://www.dei.unipd.it/~ferro/CLEF-WN-Drafts/CLEF2019/paper_248.pdf.
- [21] W. Ragheb, J. Azé, S. Bringay, M. Servajean, Attentive multi-stage learning for early risk detection of signs of anorexia and self-harm on social media., in: *CLEF (Working Notes)*, 2019.
- [22] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller,

- faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).
- [23] 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, CoRR abs/1907.11692 (2019). URL: <http://arxiv.org/abs/1907.11692>. arXiv:1907.11692.
 - [24] A.-S. Uban, P. Rosso, Deep learning architectures and strategies for early detection of self-harm and depression level prediction, in: CEUR Workshop Proceedings, volume 2696, Sun SITE Central Europe, 2020, pp. 1–12.
 - [25] L. Oliveira, Bioinfo@ uavr at erisk 2020: on the use of psycholinguistics features and machine learning for the classification and quantification of mental diseases (2020).
 - [26] D. Maupomé, M. D. Armstrong, R. Belbahar, J. Alezot, R. Balassiano, M. Queudot, S. Mosser, M.-J. Meurs, Early mental health risk assessment through writing styles, topics and neural models (2020).
 - [27] M. De Choudhury, S. Counts, E. J. Horvitz, A. Hoff, Characterizing and predicting postpartum depression from shared facebook data, in: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, 2014, pp. 626–638.
 - [28] M. De Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media, in: Seventh international AAAI conference on weblogs and social media, 2013.
 - [29] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, J. C. Eichstaedt, Detecting depression and mental illness on social media: an integrative review, Current Opinion in Behavioral Sciences 18 (2017) 43–49.
 - [30] M. Trotzek, S. Koitka, C. M. Friedrich, Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia., in: L. Cappellato, N. Ferro, J. Nie and L. Soulier (eds.) CLEF 2018 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org, volume 2125, 2018.
 - [31] M. Conway, D. O’Connor, Social media, big data, and mental health: current advances and ethical implications, Current opinion in psychology 9 (2016) 77–82.
 - [32] P. Resnik, A. Garron, R. Resnik, Using topic modeling to improve prediction of neuroticism and depression in college students, in: Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 1348–1353.
 - [33] J. C. Eichstaedt, R. J. Smith, R. M. Merchant, L. H. Ungar, P. Crutchley, D. Preoțiu-Pietro, D. A. Asch, H. A. Schwartz, Facebook language predicts depression in medical records, Proceedings of the National Academy of Sciences 115 (2018) 11203–11208.
 - [34] A. S. Uban, B. Chulvi, P. Rosso, An emotion and cognitive based analysis of mental health disorders from social media data, Future Generation Computer Systems (In press) (2021).
 - [35] R. Caruana, Multitask learning, Machine learning 28 (1997) 41–75.
 - [36] B. Pfefferbaum, C. S. North, Mental health and the covid-19 pandemic, New England Journal of Medicine 383 (2020) 510–512.
 - [37] S. Graham, C. Depp, E. E. Lee, C. Nebeker, X. Tu, H.-C. Kim, D. V. Jeste, Artificial intelligence for mental health and mental illnesses: an overview, Current psychiatry reports 21 (2019) 1–18.
 - [38] M. Ewbank, R. Cummins, V. Tablan, A. Catarino, S. Buchholz, A. Blackwell, Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts, Psychotherapy Research (2020) 1–13.

- [39] C. A. Lovejoy, Technology and mental health: the role of artificial intelligence, *European Psychiatry* 55 (2019) 1–3.
- [40] H. A. Schwartz, M. Sap, M. L. Kern, J. C. Eichstaedt, A. Kapelner, M. Agrawal, E. Blanco, L. Dziurzynski, G. Park, D. Stillwell, et al., Predicting individual well-being through the language of social media, in: *Biocomputing 2016: Proceedings of the Pacific Symposium*, World Scientific, 2016, pp. 516–527.
- [41] E. Pogrebtsova, G. F. Tondello, H. Premasukh, L. E. Nacke, Using technology to boost employee wellbeing? how gamification can help or hinder results., in: *PGW@ CHI PLAY*, 2017.
- [42] S. Volkova, K. Han, C. Corley, Using social media to measure student wellbeing: a large-scale study of emotional response in academic discourse, in: *International Conference on Social Informatics*, Springer, 2016, pp. 510–526.
- [43] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, M. Mitchell, 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*, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 31–39. URL: <https://www.aclweb.org/anthology/W15-1204>. doi:10.3115/v1/W15-1204.
- [44] F. Sadeque, D. Xu, S. Bethard, Uarizona at the clef erisk 2017 pilot task: linear and recurrent models for early depression detection, in: *CEUR workshop proceedings*, volume 1866, NIH Public Access, 2017.
- [45] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, W. Zhu, Depression detection via harvesting social media: A multimodal dictionary learning solution., in: *IJCAI*, 2017, pp. 3838–3844.
- [46] Y.-T. Wang, H.-H. Huang, H.-H. Chen, A neural network approach to early risk detection of depression and anorexia on social media text., in: L. Cappellato, N. Ferro, J. Nie and L. Soulier (eds.) *CLEF 2018 Labs and Workshops*, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org, volume 2125, 2018.
- [47] M. Matero, A. Idnani, Y. Son, S. Giorgi, H. Vu, M. Zamani, P. Limbachiya, S. C. Guntuku, H. A. Schwartz, Suicide risk assessment with multi-level dual-context language and bert, in: *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 2019, pp. 39–44.
- [48] A. Zirikly, P. Resnik, O. Uzuner, K. Hollingshead, Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts, in: *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, 2019, pp. 24–33.
- [49] E. Mohammadi, H. Amini, L. Kosseim, Quick and (maybe not so) easy detection of anorexia in social media posts., in: *CLEF (Working Notes)*, 2019. URL: http://ceur-ws.org/Vol-2380/paper_74.pdf.
- [50] G. Rao, Y. Zhang, L. Zhang, Q. Cong, Z. Feng, Mgl-cnn: A hierarchical posts representations model for identifying depressed individuals in online forums, *IEEE Access* 8 (2020) 32395–32403.
- [51] H. Amini, L. Kosseim, Towards explainability in using deep learning for the detection of anorexia in social media, in: *International Conference on Applications of Natural Language to Information Systems*, Springer, 2020, pp. 225–235.
- [52] A. S. Uban, B. Chulvi, P. Rosso, On the explainability of automatic predictions of mental

- disorders from social media data, in: International Conference on Applications of Natural Language to Information Systems (In press), Springer, 2021.
- [53] M. E. Aragón, A. P. López-Monroy, L. C. González-Gurrola, M. Montes, Detecting depression in social media using fine-grained emotions, 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), 2019, pp. 1481–1486.
- [54] M. E. Aragón, A. P. López-Monroy, M. Montes-y Gómez, Inaoe-cimat at erisk 2019: Detecting signs of anorexia using fine-grained emotions., in: L. Cappellato, N. Ferro, D. Losada and H. Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org, volume 2380, 2019.
- [55] M. E. Aragón, A. P. López-Monroy, M. Montes-y Gómez, Inaoe-cimat at erisk 2020: Detecting signs of self-harm using sub-emotions and words 2696 (2020).
- [56] R. Plutchik, The emotions, University Press of America, 1991.
- [57] J. Attenberg, F. Provost, Why label when you can search? alternatives to active learning for applying human resources to build classification models under extreme class imbalance, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, Association for Computing Machinery, New York, NY, USA, 2010, p. 423–432. URL: <https://doi.org/10.1145/1835804.1835859>. doi:10.1145/1835804.1835859.