An End-to-End Set Transformer for User-Level Classification of Depression and Gambling Disorder

Ana-Maria Bucur[1,2], Adrian Cosma[3], Liviu P. Dinu[4,5] and Paolo Rosso[2]

[1]*Interdisciplinary School of Doctoral Studies, University of Bucharest, Romania*
[2]*PRHLT Research Center, Universitat Politècnica de València, Spain*
[3]*Politehnica University of Bucharest, Romania*
[4]*Faculty of Mathematics and Computer Science, University of Bucharest, Romania*
[5]*Human Language Technologies Research Center, University of Bucharest, Romania*

Abstract

This work proposes a transformer architecture for user-level classification of gambling addiction and depression that is trainable end-to-end. As opposed to other methods that operate at the post level, we process a set of social media posts from a particular individual, to make use of the interactions between posts and eliminate label noise at the post level. We exploit the fact that, by not injecting positional encodings, multi-head attention is permutation invariant and we process randomly sampled sets of texts from a user after being encoded with a modern pretrained sentence encoder (RoBERTa / MiniLM). Moreover, our architecture is interpretable with modern feature attribution methods and allows for automatic dataset creation by identifying discriminating posts in a user's text-set. We perform ablation studies on hyper-parameters and evaluate our method for the eRisk 2022 Lab on early detection of signs of pathological gambling and early risk detection of depression. The method proposed by our team BLUE obtained the best $ERDE_5$ score of 0.015, and the second-best $ERDE_{50}$ score of 0.009 for pathological gambling detection. For the early detection of depression, we obtained the second-best $ERDE_{50}$ of 0.027.

Keywords

set transformer, sentence encoder, gambling disorder detection, depression detection, social media

1. Introduction

How much can one know about someone from their social media interactions? Billions of people[1] use social media sites like Facebook, Instagram, Twitter, and Reddit every day. While some sites like Facebook and Instagram encourage users to use their real names, websites such as Reddit are often praised for enabling users to hide between a pseudonym, offering the illusion of privacy. Under the guise of anonymity, users tend to post more personal information related to their lives and their everyday struggles instead of striving to maintain an image and a persona when their identities are open [1]. Many aspects of a user's personal life can be uncovered in their posting history. Of course, not one single post can be all-encompassing, but rather the information is scattered across many unrelated comments and posts. For instance, on

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
ana-maria.bucur@drd.unibuc.ro (A. Bucur); cosma.i.adrian@gmail.com (A. Cosma); ldinu@fmi.unibuc.ro (L. P. Dinu); prosso@dsic.upv.es (P. Rosso)
0000-0003-2433-8877 (A. Bucur); 0000-0003-0307-2520 (A. Cosma); 0000-0002-7559-6756 (L. P. Dinu); 0000-0002-8922-1242 (P. Rosso)

[1]https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

the *r/relationship_advice*[2] subreddit a user might reveal their gender and age when discussing intimate relationship struggles, while on *r/depression*[3] a user might provide clues for their internal conflicts and experiences.

In the task of mental health disorders detection from social media text, many approaches operate on the post-level [2, 3, 4], considering that, for instance, if a user is depressed, then all their posts might contain some information regarding this issue. However, we posit that this method of post-level classification is unsuitable - many posts are unrelated and uninformative to the particular task. Their interaction, however, might contain clues to the mental well-being of a user.

As such, we propose an architecture that performs user-level classification by processing a set of posts from a user. We exploit the fact that the multi-head attention operation in transformers is permutation invariant and inputs multiple texts from a single user into the network, modeling their interaction and classifying the user. This approach has several advantages: *(i)* it is trainable end-to-end, mitigating the need for hand-crafted construction of global user features *(ii)* it is robust to label noise, as some posts might be uninformative, the network learns to ignore them in the decision and *(iii)* it is interpretable, using feature attribution methods [5] we can extract the most important posts for the decision.

The Early Risk Prediction on the Internet (eRisk)[4] Lab started in 2017 with one pilot task and, since then, tacked the early risk detection of several mental illnesses: depression, self-harm, eating disorders, and pathological gambling. This work showcases team BLUE's proposed approach for Tasks 1 and 2 of eRisk 2022 Lab [6], of gambling and depression detection, respectively.

The paper makes the following contributions:

1. We propose a set-based transformer architecture for user-level classification, which makes a decision by processing multiple texts of a particular user.
2. We show that our architecture is robust to label noise and is interpretable with modern feature attribution methods, allowing it to be used as a dataset filtering tool.
3. We obtained promising results on the eRisk 2022 tasks on early risk detection of pathological gambling (best $ERDE_5$[5] score of 0.015 and the second-best $ERDE_{50}$ score of 0.009) and depression detection (second-best $ERDE_{50}$ of 0.027).

2. Related Work

Pathological Gambling For the detection of gambling disorder, the eRisk Lab is the first to use social media data for the assessment of gambling risk. Usually, the automated methods use data from behavioral markers [7, 8] or personality biomarkers [9]. In the first iteration of the task for gambling addiction detection, the best-performing systems were developed by Maupomé et al. [10] and Loyola et al. [11]. Maupomé et al. [10] used a user-level approach based on the similarity distance between the vector of topic probabilities of the users' texts to be assessed for pathological gambling risk and testimonials or items from a self-evaluation questionnaire for

[2] https://www.reddit.com/r/relationship_advice/
[3] https://www.reddit.com/r/depression/
[4] https://erisk.irlab.org/
[5] Early Risk Detection Error, introduced in Section 5.1

compulsive gamblers. By using this method, the authors obtain the best $ERDE_5$ of 0.048. Loyola et al. [11] attain the best $ERDE_{50}$ (0.020) and latency-weighted F1 (0.693) through a post-level rule-based early alert policy on bag-of-words text representation classified with SVM.

Depression Depression detection from social media data is an interdisciplinary topic, and efforts have been made by researchers from both NLP and Psychology to detect different markers of depression found in the online discourse of individuals. Some depression cues found in language are: greater use of the first-person singular pronouns *"I"* [12], lesser use of first-person plural *"we"* [13], increased use of negative or absolutist terms (e.g., "never", "forever") [14], greater use of verbs at past tense [15].

For the task of early detection of depression, the best systems from the first iteration of the task (eRisk 2017) used as input linguistic meta information extracted from the texts such as LIWC [16], readability and hand-crafted features [17] obtaining the best $ERDE_5$ (12.70%) or a combination of linguistic information and temporal variation of terms from users' posts [18] achieving the best $ERDE_{50}$ (9.68%). The best-performing systems from eRisk 2018 were the ones from Funez et al. [19] and Trotzek et al. [20]. Funez et al. [19] propose a user-level approach using an SVM classifier on semantic representations that take into account the temporal variation of terms between the users' posts and achieve an $ERDE_5$ of 8.78%. On the other hand, the best $ERDE_{50}$ (6.44%) is attained by Trotzek et al. [20] using a chunk-level[6] approach using an ensemble of logistic regression classifiers on bag-of-words features. The dataset from the depression detection task from the eRisk Lab was an important resource later used in different research articles tackling the detection problem using approaches such as a neural network architecture on topic modeling features [21], SVM or deep learning architectures using fine-grained emotions features [22] or deep learning methods using content, writing style and emotion features [23].

3. Method

The transformer encoder, as proposed by Vaswani et al. [24], essentially consists of multiple sequential layers of multi-head attention. Scaled dot-product attention of a query Q relative to a set of values V and a set of keys K is computed using the following equation (d_k is the dimensionality of the query and keys):

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

As such, multi-head attention consists of multiple applications of the attention mechanism to the same input. The multi-head attention is defined as:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \text{head}_2 \dots \text{head}_h)W^O \\ \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2}$$

In this formulation, multi-head attention is *permutation invariant*, and the current way to inject temporal information into the input sequence is by employing positional encodings [25]. This is useful when processing sequential data such as texts. However, by omitting positional

[6] in 2018 the test data was released in chunks of posts, not one post at a time as it is the case in this year's tasks

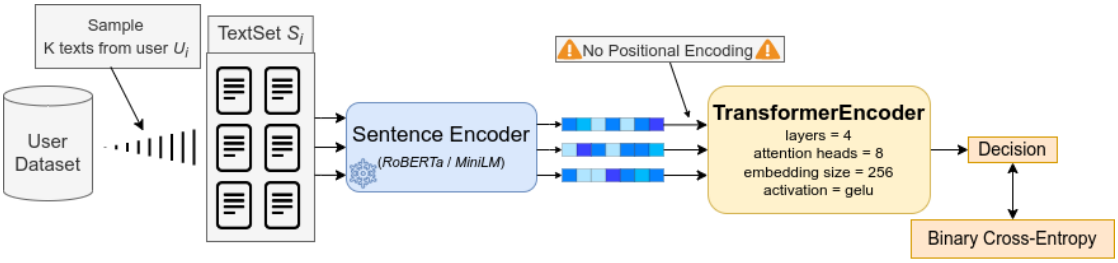

Figure 1: Proposed model architecture. We perform user-level classification by operating on a sample of K texts from a user. Texts are encoded with a pretrained sentence encoder and processed by a permutation-invariant transformer network. Binary cross-entropy loss is applied at the user level for a text-set.

encodings, the transformer essentially acts as a *set encoder*. Lee et al. [26] introduced the Set Transformer, in which they prove that multi-head attention is permutation invariant and that the Set Transformer is a universal approximator of permutation invariant functions. We make use of this fact to perform user-level classification by processing *sets of texts* (in the form of social media posts) from a particular user. The intuition behind processing a set of texts from a user is that no single social media post is sufficiently informative for a classifier decision, but rather their interaction and the user behavior as a whole. Moreover, through mean pooling, the inevitable noise (in terms of unrelated posts) is dampened, which aids classification in weakly-supervised scenarios, such as ours, in which a user is labeled rather than all of their posts.

We consider a user i to contain multiple social media posts U_i. A set of K texts t are randomly sampled from U_i, which defines our text-set $S_i = \{t^j \sim U_i, j \in (1 \dots K)\}$. We sample K posts from the user's history, instead of processing all of them due to memory limitations - some individuals have thousands of posts while others have only in the order of tens. Moreover, stochasticity is introduced in the training procedure, which prevents overfitting. As such, for training, an input batch of size n is defined by the concatenation of n such text-sets: $B = \{S_{b_1}, S_{b_2}, \dots S_{b_n}\}$. We do not consider the relative order of the texts for a particular user, and text-sets are fed into the transformer encoder without using positional encoding. Since some users have a total number of texts smaller than K, creating a batch of text-sets is impossible without padding and masking. However, to alleviate this problem, we train with an effective batch size of 1 and chose to employ gradient accumulation to simulate a larger batch size.

Figure 1 showcases our proposed model architecture for user-level classification. Each text in a text-set is embedded into a fixed-size vector using available pretrained sentence encoder models (i.e., RoBERTa / MiniLM). The text embeddings are fed into the transformer encoder network, and after processing, we perform mean pooling and output the decision. We compute binary cross-entropy at the user-level, for a text-set. The pretrained sentence encoder is frozen and not updated during training.

Baytas et al. [27] proposed to use a T-LSTM to process social media posts sequentially as a time-series. The authors modify the LSTM architecture to include a relative time component. However, in our case it is unclear how to incorporate such a mechanism into the transformer

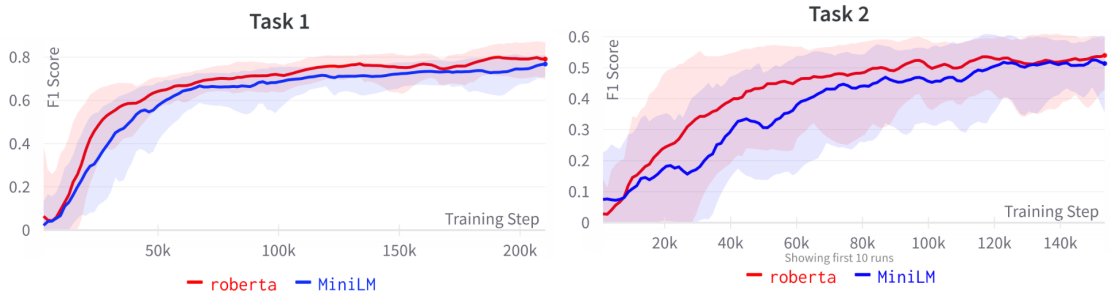

Figure 2: Performance of our model across training steps, in terms of F_1 score, for different sentence encoders (RoBERTa / MiniLM). We show the mean and standard deviation of F_1 score across multiple values of K. For both tasks, RoBERTa yields consistent superior performance compared to MiniLM. Best viewed in color.

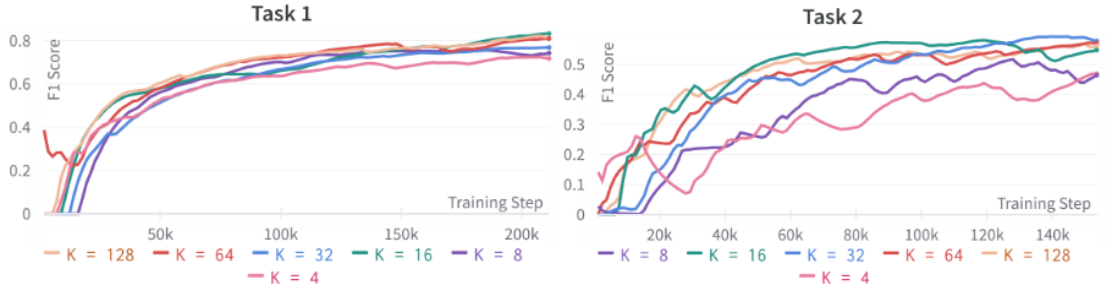

Figure 3: Performance of our model across training steps, in terms of validation F_1 score, for RoBERTa sentence embeddings and varying the K, the number of texts per user. For Tasks 1 and 2, the best performance is attained with $K = 16$ and $K = 32$, respectively. Best viewed in color.

architecture, aside from using a relative positional encoding [28], which ignores long-ranged dependencies between posts. As such, we chose to ignore the temporal order of the posts and process them directly as a set. The main reason for considering the posts as a set is that in a user's post history, many posts are uninformative to the modeling task, and by processing a set of texts, label noise is reduced naturally as a direct consequence of the attention mechanism, which assigns more importance to informative posts. However, training with a sufficiently large dataset might achieve the same effect, but previous attempts at post-level classification have proven ineffective [4].

In order to assess the impact of the sentence representations, we chose two different sentence encoders: RoBERTa [29] and MiniLM [30]. We chose RoBERTa since it is one of the best performing English language models in downstream tasks [29], and MiniLM, a multi-lingual model, since some users have social media posts in languages other than English. Figure 2 showcases the performance gap between the two sentence encoders, averaged across multiple values of K. RoBERTa yields a consistently superior performance across training steps. Similarly, to assess the impact of the text-set size K, we performed an ablation study, as shown in Figure 3. We kept the sentence encoder fixed to RoBERTa, and vary the number of texts per user $K \in \{4, 8, 16, 32, 64, 128\}$. The best performance was achieved with $K = 16$ and $K = 32$ for Tasks 1 and 2, respectively.

In our final submission, we chose RoBERTa as a sentence encoder and sampled $K = 16$ texts

per user for Task 1 and $K = 32$ for Task 2. We used the standard formulation of the transformer network [24], with 4 encoder layers, 8 attention heads each and a dimensionality of 256. Both networks were trained for 120 epochs, with AdamW optimizer [31], with a cyclical learning rate [32] ranging from 0.00001 to 0.0001 across 6 epochs and a batch size of 128. To account for class imbalance, we computed balanced class weights with respect to each dataset and adjusted the loss function accordingly. Finally, we opted for a very high threshold when predicting the final decision.

Our proposed architecture can be easily interpretable using modern explainability methods for feature attribution [33, 34, 5], such as Integrated Gradients [5]. It automatically identifies social media posts containing signs of mental health disorders and filters out uninformative posts.

4. Interpretability

Since our model operates on sets of social media texts from a particular user, we can employ model explainability methods to assess the importance of a piece of text to the model decision. Through this, automatic filtering and selection of the most indicative posts of a user can be made for use in dataset creation. This idea is similar to Ríssola et al. [3], which employed a series of heuristics to recognize posts portraying depression symptoms for use in constructing a post-level training set from existing depression datasets annotated at the user level. As such, we use Integrated Gradients [5] to compute attribution scores for a text-set. The integrated gradients method has been used in NLP to explore the contribution of individual words and phrases to a decision made by a classifier. Since we are not operating on words, but rather on whole texts, this method computes the most important text to the classifier decision.

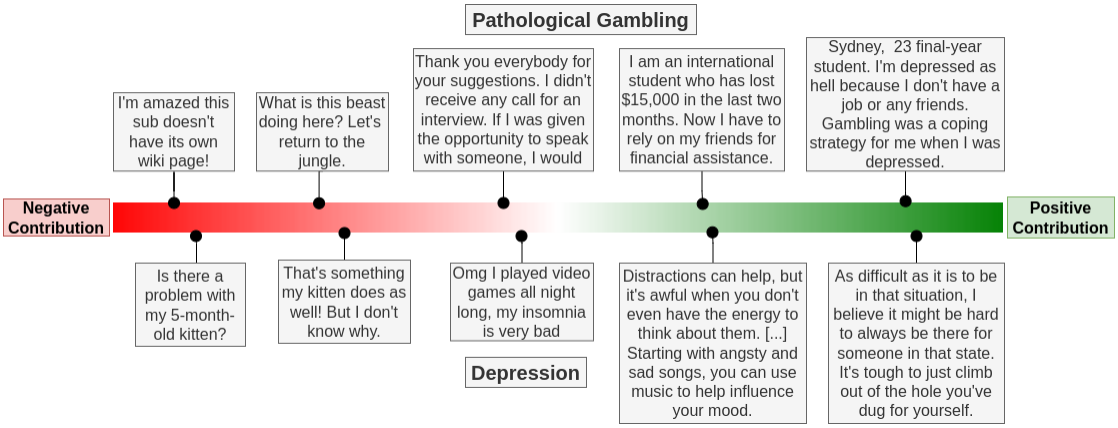

Figure 4: Texts from a particular user, relatively ranked in terms of attribution scores (contribution to a positive decision by the model) computed with the Integrated Gradients method. For each task, all texts belong to a single text-set of a user. The model is able to identify posts with a clear discriminating information for each task. Best viewed in color. Examples have been paraphrased for anonymity.

Figure 4 showcases selected samples ordered by their attribution score from the validation set of each task. All samples belong to the same user for each task, and the attribution scores are

computed within the respective text-set. Posts with a high positive contribution to the decision contain more explicit descriptions of symptoms, while posts with more negative contributions are mainly unrelated to the particular mental illness. We use the integrated gradients method in one of our runs to select the most important posts in the user history. However, we emphasize that the best application of this approach is for automatic dataset creation in scenarios of weak supervision, which we aim to explore in future work.

5. Results

5.1. Evaluation

There are two kinds of evaluation used for measuring the performance of the systems, decision-based and raking-based. The **decision-based evaluation** is used for quantifying the capacity of a system to perform the binary classification and predicting if a user is from the positive class (i.e., pathological gambling or depression) or the negative one. It is comprised of standard measures for classification (Precision, Recall, F1) and measures for this specific task of early detection that consider the delay and the speed of the decision. The early risk detection error (ERDE) [35] measures the correct predictions considering a late decision penalty (for predictions taken after the 5 or 50 first submissions of a user). To overcome the limitations of this metric [36], the *latency-weighted F1* score [37] was also proposed to measure the performance of early risk detection. *Latency* measures the delay in detecting true positives based on the median number of submissions seen by the system before taking a decision. The *speed* of a system that correctly predicts true positives from the first submission is equal to 1, while a slow system which decides after processing hundreds of texts. The latency-weighted F1 combines the F1-score with the delay in decision-taking for true positives. A perfect system should achieve a latency-weighted F1 of 1. Besides the binary classification decisions, the participating teams were asked to also submit a score for estimating the risk of users for the **ranking-based evaluation**. These scores are used to rank users' risk for pathological gambling or depression. Standard IR metrics (P@10, NDCG@10, and NDCG@100) are used to measure the models' ranking-based performance after processing 1, 100, 500, or 1000 submissions.

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

The first task proposes the detection of gambling addiction from social media data. This being the second edition of this task, the organizers provided the last year's test data for training the systems. The dataset was collected from Reddit, following the methodology described by Losada and Crestani [35] and contains a chronological sequence of posts from each user. The training dataset was comprised of 164 pathological gamblers, with a total of 54,674 submissions, and 2,184 control users with 1,073,883 submissions. The test dataset contains 81 users with gambling addiction, summing 14,627 posts, and 1,998 control users with a total of 1,014,122 posts. For the testing phase, the submissions of users were released sequentially, the systems proposed by the participating teams received one submission at a time from all the users. We submitted three runs for the early detection of pathological gambling: **Run 0** is comprised of the text-set transformer model using the most recent $K = 16$ posts for prediction; the system

for **Run 1** is the same text-set transformer model using as input the set of $K = 16$ texts that are most important in a user's history, selected with Integrated Gradients; **Run 2** is a baseline run, using the proposed model architecture for predicting at post-level, on one sample at a time.

Table 1
Decision-based evaluation on Task 1: Early Detection of Signs of Pathological Gambling. We show the performance of our systems compared to the best-performing run from each team.

Team	Run ID	P	R	F1	$ERDE_5$	$ERDE_{50}$	$Latency_{TP}$	Speed	Latency-Weighted F1
BLUE	0	0.260	0.975	0.410	**0.015**	0.009	**1.0**	**1.000**	0.410
BLUE	1	0.123	0.988	0.219	0.021	0.015	**1.0**	**1.000**	0.219
BLUE	2	0.052	**1.000**	0.099	0.037	0.028	**1.0**	**1.000**	0.099
UNED-NLP	4	0.809	0.938	**0.869**	0.020	**0.008**	3.0	0.992	**0.862**
SINAI	1	0.575	0.802	0.670	**0.015**	0.009	**1.0**	**1.000**	0.670
BioInfo_UAVR	4	0.192	0.988	0.321	0.033	0.011	5.0	0.984	0.316
RELAI	2	0.052	0.963	0.099	0.036	0.029	**1.0**	**1.000**	0.099
BioNLP-UniBuc	4	0.046	**1.000**	0.089	0.032	0.031	**1.0**	**1.000**	0.089
UNSL	1	0.461	0.938	0.618	0.041	**0.008**	11.0	0.961	0.594
NLPGroup-IISERB	4	**1.000**	0.074	0.138	0.038	0.037	41.5	0.843	0.116
stezmo3	4	0.160	0.901	0.271	0.043	0.011	7.0	0.977	0.265

Table 2
Ranking-based evaluation on Task 1: Early Detection of Signs of Pathological Gambling.

		1 writing			100 writings			500 writings			1000 writings		
Team	Run ID	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100
BLUE	0	**1.00**	**1.00**	**0.76**	**1.00**	**1.00**	0.81	**1.00**	**1.00**	0.89	**1.00**	**1.00**	0.89
BLUE	1	**1.00**	**1.00**	**0.76**	**1.00**	**1.00**	0.89	**1.00**	**1.00**	0.91	**1.00**	**1.00**	0.91
BLUE	2	**1.00**	**1.00**	0.69	**1.00**	**1.00**	0.40	0.00	0.00	0.02	0.00	0.00	0.01
UNED-NLP	4	**1.00**	**1.00**	0.56	**1.00**	**1.00**	0.88	**1.00**	**1.00**	**0.95**	**1.00**	**1.00**	**0.95**
UNSL	0	**1.00**	**1.00**	0.68	**1.00**	**1.00**	**0.90**	**1.00**	**1.00**	0.93	**1.00**	**1.00**	**0.95**

Table 1 showcases the performance of the systems measured using the decision-based measures. Regarding ERDE, our first run (Run 0), using the transformer architecture on the most recent texts from each user, manages to achieve the best $ERDE_5$ score of 0.015, and the second-best $ERDE_{50}$ score of 0.009, demonstrating that the system could detect early the true positive cases. The perfect scores for *latency*$_{TP}$ and *speed* show that our models were successful at detecting the true positive cases after the first writing. As expected, the baseline run using a post-level approach (Run 2) has the lowest performance. Regarding Run 2, we expected it to achieve the best performance from our submitted runs, as this approach is more aggressive in taking decisions by using for classification the most informative posts from users' history. Furthermore, our best run from this year's task surpasses all the runs from our participation in the first iteration of the task in 2021 [4], showing that a user-level approach considering a set of texts from each individual is more suitable than a post-level approach. In Table 2 we show the results of the ranking-based evaluation, in which each team had to submit the rankings of users' risk for pathological gambling. Our team has excellent results for NDCG and P@10 in all the situations (after 1, 100, 1000, 5000 writings).

5.3. Task 2: Early Detection of Depression

This year marks the third iteration of the early detection of depression task, continuing the 2017 T1 and 2018 T2 tasks. The organizers provided the data from the previous two editions for training the models. Users from the depression class were labeled by their mention of diagnosis on their Reddit posts (e.g., "I was diagnosed with depression"). In contrast, users from the control class are users who do not have any mention of diagnosis in their posts [35]. The training dataset comprises 214 users diagnosed with depression with 270,666 submissions and 1493 control users with a total of 2,959,080 submissions. The test set contains 98 users with depression with 35,332 posts, and 1,302 users in the control group with a total of 687,228 posts. The texts for making the predictions for the testing phase were released sequentially, and the systems from the participating teams had to decide on firing a decision for a specific user or waiting for more data. We submitted three runs for the early detection of depression: **Run 0** is the text-set transformer model using the most recent $K = 32$ posts for prediction; for **Run 1** we employ the same text-set transformer model using as input the set of $K = 32$ texts that are most important in a user's history, selected with Integrated Gradients; **Run 2** is a baseline run, using the proposed model architecture for predicting at post-level, on one sample at a time.

Table 3
Decision-based evaluation on Task 2: Early Detection of Depression. We show the performance of our systems compared to the best-performing run from each team.

Team	Run ID	P	R	F1	$ERDE_5$	$ERDE_{50}$	$Latency_{TP}$	Speed	Latency-Weighted F1
BLUE	0	0.395	0.898	0.548	0.047	0.027	5.0	0.984	0.540
BLUE	1	0.213	0.939	0.347	0.054	0.033	4.5	0.986	0.342
BLUE	2	0.106	**1.000**	0.192	0.074	0.048	4.0	0.988	0.190
CYUT	0	0.165	0.918	0.280	0.053	0.032	3.0	0.992	0.277
LauSAn	4	0.201	0.724	0.315	**0.039**	0.033	**1.0**	**1.000**	0.315
BioInfo_UAVR	4	0.378	0.857	0.525	0.069	0.031	16.0	0.942	0.494
TUA1	4	0.159	0.959	0.272	0.052	0.036	3.0	0.992	0.270
NLPGroup-IISERB	0	0.682	0.745	**0.712**	0.055	0.032	9.0	0.969	**0.690**
RELAI	0	0.085	0.847	0.155	0.114	0.092	51.0	0.807	0.125
UNED-MED	1	0.139	0.980	0.244	0.079	0.046	13.0	0.953	0.233
Sunday-Rocker2	1	0.355	0.786	0.489	0.068	0.041	27.0	0.899	0.439
SCIR2	3	0.316	0.847	0.460	0.079	**0.026**	44.0	0.834	0.383
UNSL	2	0.400	0.755	0.523	0.045	0.026	3.0	0.992	0.519
E8-IJS	0	**0.684**	0.133	0.222	0.061	0.061	1.0	1.000	0.222
NITK-NLP2	3	0.149	0.724	0.248	0.049	0.039	2.0	0.996	0.247

In Table 3 we present the performance of the systems using the decision-based metrics. Our best performing run is the transformer architecture using the most recent texts from users (Run 0), followed by the system that considers only the most informative submissions from each user for the model's decisions (Run 1). The post-level system (Run 2) has the worst performance. Our three submitted runs achieve high Recall at the expense of lower Precision scores. The precision of our models can be improved by incorporating a mechanism for weighting user posts according to the prevalence of signs of depression [38]. As such, a text-set containing few posts with signs of depression will not induce a positive prediction. Regarding the early detection evaluation, our team has the second-best score on the $ERDE_{50}$ metric (0.027), while our $ERDE_5$ score is close to the best one. Compared to the best metrics from the 2018 edition

of this task, when the best $ERDE_5$ and $ERDE_{50}$ were 0.087 and 0.064, respectively, current systems surpass these scores due to more data being available for training the models and the advancements in the field of machine learning in the last few years. Regarding the standard metrics for classification, a slight improvement was made in terms of F1 score, from 0.64 in 2018 to 0.71 in 2022. The ranking-based evaluation performance from Table 4 shows that for 1 and 1000 writings, our systems attain some of the best scores for P@10 and NDCG.

Table 4
Ranking-based evaluation on Task 2: Early Detection of Depression.

		1 writing			100 writings			500 writings			1000 writings		
Team	Run ID	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100
BLUE	0	**0.80**	**0.88**	**0.54**	0.60	0.56	0.59	0.80	0.81	0.66	**0.80**	0.80	0.68
BLUE	1	**0.80**	**0.88**	**0.54**	0.70	0.64	**0.67**	0.80	0.84	**0.74**	**0.80**	**0.86**	**0.72**
BLUE	2	**0.80**	0.75	0.46	0.40	0.40	0.30	0.30	0.35	0.20	0.30	0.38	0.16
NLPGroup-IISERB	0	0.00	0.00	0.02	**0.90**	0.92	0.30	**0.90**	**0.92**	0.33	0.00	0.00	0.00
Sunday-Rocker2	1	0.70	0.81	0.39	**0.90**	**0.93**	0.66	**0.90**	0.88	0.65	0.00	0.00	0.00

6. Conclusion

In this work, we proposed a transformer architecture that performs user-level classification of gambling addiction and depression detection. For each individual, the transformer processes a set of texts encoded by a pretrained sentence encoder to model the interactions between posts and mitigate noise in the dataset. Our network is interpretable and allows for automatic dataset creation by filtering uninformative posts in a user's history. Our method is a promising approach, especially for social media text processing, where a user has many texts: some informative and some unrelated to the particular modeling task. However, their interaction is indicative of the mental state of the user. We attained the best $ERDE_5$ score of 0.015, and the second-best $ERDE_{50}$ score of 0.009 for pathological gambling detection. For the early detection of depression, we obtained the second-best $ERDE_{50}$ (0.027).

For future work, we aim to extend our method and construct a mechanism for encoding the relative order of a user's posts with a modified version of relative positional embeddings [39]. While we chose an approach that ignores temporal ordering and processes posts as a set, preserving order is a natural way to increase the expressive power in modeling a user's entire social media interactions, similar to architectures such as the time-aware LSTM [27].

Acknowledgments

The work of Ana-Maria Bucur was in the framework of the research project NPRP13S-0206-200281. The work of Paolo Rosso was in the framework of the research project PROMETEO/2019/121 (DeepPattern) by the Generalitat Valenciana. The authors thank the EU-FEDER Comunitat Valenciana 2014–2020 grant IDIFEDER/2018/025.

References

[1] M. De Choudhury, S. De, Mental health discourse on reddit: Self-disclosure, social support, and anonymity, in: Eighth international AAAI conference on weblogs and social media, 2014.

[2] M. M. Tadesse, H. Lin, B. Xu, L. Yang, Detection of suicide ideation in social media forums using deep learning, Algorithms 13 (2019) 7.

[3] E. A. Ríssola, S. A. Bahrainian, F. Crestani, A dataset for research on depression in social media, in: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, 2020, pp. 338–342.

[4] A.-M. Bucur, A. Cosma, L. P. Dinu, Early risk detection of pathological gambling, self-harm and depression using bert, in: CLEF (Working Notes), 2021.

[5] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.

[6] J. Parapar, P. M. Rodilla, D. E. Losada, F. A. Crestani, Overview of erisk 2022: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association, CLEF 2022, Springer International Publishing, 2022.

[7] K. S. Philander, Identifying high-risk online gamblers: A comparison of data mining procedures, International Gambling Studies 14 (2014) 53–63.

[8] X. Deng, T. Lesch, L. Clark, Applying data science to behavioral analysis of online gambling, Current Addiction Reports 6 (2019) 159–164.

[9] A. Cerasa, D. Lofaro, P. Cavedini, I. Martino, A. Bruni, A. Sarica, D. Mauro, G. Merante, I. Rossomanno, M. Rizzuto, et al., Personality biomarkers of pathological gambling: A machine learning study, Journal of neuroscience methods 294 (2018) 7–14.

[10] D. Maupomé, M. D. Armstrong, F. Rancourt, T. Soulas, M.-J. Meurs, Early detection of signs of pathological gambling, self-harm and depression through topic extraction and neural networks, in: CLEF (Working Notes), 2021.

[11] J. M. Loyola, S. Burdisso, H. Thompson, L. Cagnina, M. Errecalde, Unsl at erisk 2021: A comparison of three early alert policies for early risk detection, in: CLEF (Working Notes), 2021.

[12] S. Rude, E.-M. Gortner, J. Pennebaker, Language use of depressed and depression-vulnerable college students, Cognition & Emotion 18 (2004) 1121–1133.

[13] A.-M. Bucur, I. R. Podină, L. P. Dinu, A psychologically informed part-of-speech analysis of depression in social media, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), 2021, pp. 199–207.

[14] S. Fekete, The internet-a new source of data on suicide, depression and anxiety: a preliminary study, Archives of Suicide Research 6 (2002) 351–361.

[15] D. Smirnova, P. Cumming, E. Sloeva, N. Kuvshinova, D. Romanov, G. Nosachev, Language patterns discriminate mild depression from normal sadness and euthymic state, Frontiers in psychiatry 9 (2018) 105.

[16] J. W. Pennebaker, M. E. Francis, R. J. Booth, Linguistic inquiry and word count: Liwc 2001, Mahway: Lawrence Erlbaum Associates 71 (2001) 2001.

[17] 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.
[18] M. L. Errecalde, M. P. Villegas, D. G. Funez, M. J. G. Ucelay, L. C. Cagnina, Temporal variation of terms as concept space for early risk prediction, in: CLEF (Working Notes), 2017.
[19] D. G. Funez, M. J. G. Ucelay, M. P. Villegas, S. Burdisso, L. C. Cagnina, M. Montes-y Gómez, M. Errecalde, Unsl's participation at erisk 2018 lab., in: CLEF (Working Notes), 2018.
[20] 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: CLEF (Working Notes), 2018.
[21] A. Bucur, L. P. Dinu, Detecting early onset of depression from social media text using learned confidence scores, in: Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021, volume 2769 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.
[22] M. E. Aragon, A. P. Lopez-Monroy, L.-C. G. Gonzalez-Gurrola, M. Montes, Detecting mental disorders in social media through emotional patterns-the case of anorexia and depression, IEEE Transactions on Affective Computing (2021).
[23] 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 124 (2021) 480–494.
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[25] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y. N. Dauphin, Convolutional sequence to sequence learning, in: International Conference on Machine Learning, PMLR, 2017, pp. 1243–1252.
[26] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, Y. W. Teh, Set transformer: A framework for attention-based permutation-invariant neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 3744–3753.
[27] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, J. Zhou, Patient subtyping via time-aware lstm networks, in: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017, pp. 65–74.
[28] S. M. Kazemi, R. Goel, S. Eghbali, J. Ramanan, J. Sahota, S. Thakur, S. Wu, C. Smyth, P. Poupart, M. Brubaker, Time2vec: Learning a vector representation of time, arXiv preprint arXiv:1907.05321 (2019).
[29] 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.
[30] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, Advances in Neural Information Processing Systems 33 (2020) 5776–5788.
[31] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: ICLR (Poster), 2015. URL: http://arxiv.org/abs/1412.6980.
[32] L. N. Smith, Cyclical learning rates for training neural networks, in: 2017 IEEE winter conference on applications of computer vision (WACV), IEEE, 2017, pp. 464–472.
[33] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances

in neural information processing systems 30 (2017).
[34] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
[35] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2016, pp. 28–39.
[36] D. E. Losada, F. A. Crestani, J. Parapar, Overview of erisk at clef 2019: Early risk prediction on the internet (extended overview), in: CLEF (Working Notes), 2019.
[37] F. Sadeque, D. Xu, S. Bethard, Measuring the latency of depression detection in social media, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 2018, pp. 495–503.
[38] E. Ríssola, S. A. Bahrainian, F. Crestani, A dataset for research on depression in social media, in: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, 2020, pp. 338–342. doi:10.1145/3340631.3394879.
[39] A. Qu, J. Niu, S. Mo, Explore better relative position embeddings from encoding perspective for transformer models, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 2989–2997.