From Posts to Predictions: A User-Aware Framework for Faithful and Transparent Detection of Mental Health Risks on Social Media

Hessam Amini and Leila Kosseim
Computational Linguistics at Concordia (CLaC) Laboratory
Department of Computer Science and Software Engineering
Concordia University, Montreal, Canada
hessam.amini@mail.concordia.ca; leila.kosseim@concordia.ca

Abstract

We propose a user-aware attention-based framework for early detection of mental health risks from social media posts. Our model combines *DisorBERT*, a mental health–adapted transformer encoder, with a user-level attention mechanism that produces transparent post-level explanations. To assess whether these explanations are faithful, i.e., aligned with the model's true decision process, we apply adversarial training and quantify attention faithfulness using the *AtteFa* metric. Experiments on four eRisk tasks (depression, anorexia, self-harm, and pathological gambling) show that our model achieves competitive latency-weighted F1 scores while relying on a sparse subset of posts per user. We also evaluate attention robustness and conduct ablations, confirming the model's reliance on high-weighted posts. Our work extends prior explainability studies by integrating faithfulness assessment in a real-world high-stakes application. We argue that systems combining predictive accuracy with faithful and transparent explanations offer a promising path toward safe and trustworthy AI for mental health support.

1 Introduction

Mental health disorders are major public health concerns, and early identification of conditions such as depression, anorexia, self-harm, and pathological gambling is critical. As social media becomes a widespread platform for self-expression, online posts have emerged as valuable data sources for detecting mental health risks. Prior shared tasks, notably the *eRisk* challenges, have facilitated research on early risk prediction from Reddit posts, encouraging models that can issue timely and accurate alerts for at-risk individuals (Losada et al., 2017, 2018, 2019, 2020; Parapar et al., 2021, 2022, 2023, 2024).

However, deploying such systems in practice requires not just predictive accuracy but also *trustworthy explanations*. Black-box decisions can be problematic in sensitive domains like mental health, where users and clinicians need to understand the reasoning behind predictions. Attention mechanisms have been widely used in neural NLP models as a means of offering transparency, producing weights over input segments that are often interpreted as explanations.

Yet, recent work has called this interpretability into question. Jain and Wallace (2019) showed that attention weights are often not faithful—that is, they may not align with the model's actual decision logic. Wiegreffe and Pinter (2019) argued that attention can still be useful if properly validated. The distinction between *plausibility* (i.e., what seems intuitive to humans) and *faithfulness* (what the model actually uses) has since become central in explainability research (Jacovi and Goldberg, 2020; Moradi et al., 2021). In high-stakes settings like mental health, faithfulness is essential.

In this work, we propose a user-aware attention-based framework for early detection of mental health risks on social media. Our model combines DisorBERT (Aragon et al., 2023), a BERT model adapted to mental health and social media domains, with a user-level attention mechanism that identifies the most informative posts as explanations. To evaluate the faithfulness of these explanations, we further employ adversarial training to evaluate attention faithfulness using *AtteFa* (Amini and Kosseim, 2022), a metric that quantifies how accurately the attention weights reflect the true importance of input features in the model's actual decision-making process.

We validate our framework on four eRisk tasks (*depression*, *anorexia*, *self-harm*, and *pathological gambling*), achieving competitive latency-weighted F1 scores. Our analysis confirms that the model typically relies on a sparse subset of posts to make predictions, underscoring both the rarity of strong mental health signals and the value of faithful ex-

Proceedings of Recent Advances in Natural Language Processing,pages 75–84
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-009

planations in this domain.

Our main contributions are as follows:

- We propose a novel user-aware attention-based architecture for early detection of mental health risks from social media, integrating DisorBERT with user-level attention for post-level explainability.
- We quantitatively evaluate the faithfulness of user-level attention weights through adversarial training and the use of the *AtteFa* metric.
- We conduct extensive experiments on four eRisk shared tasks (*depression*, *anorexia*, *self-harm*, *pathological gambling*), achieving competitive latency-weighted F1 scores and demonstrating that only a small subset of posts are typically required for accurate classification.

2 Related Work

2.1 Mental Health Detection from Social Media

Early work demonstrated the feasibility of using linguistic and behavioral patterns from social media to detect mental health conditions (De Choudhury et al., 2013; Coppersmith et al., 2014, 2015). The CLEF eRisk shared tasks (Losada et al., 2017, 2018, 2019, 2020; Parapar et al., 2021, 2022, 2023, 2024) formalized early detection tasks across several disorders, emphasizing timely decision-making over Reddit data.

More recent approaches leverage transformer-encoder-based models, fine-tuned on Reddit or Twitter datasets. Murarka et al. (2021) used RoBERTa (Zhuang et al., 2021) to detect multiple mental disorders with strong performance. Aragon et al. (2023) introduced DisorBERT, a variant of BERT (Devlin et al., 2019) adapted first to social media language and then to mental health content, showing improved accuracy across the eRisk datasets.

User-level modeling has also gained traction, often using hierarchical architectures to aggregate post representations. Hierarchical attention networks (HANs; Yang et al., 2016) allow models to weight each post, highlighting the most indicative content. Zogan et al. (2022) proposed a multi-aspect HAN that augmented attention-based signals with user metadata, improving both interpretability and performance for depression detection.

While these models provide some degree of transparency through their built-in attention mechanism, few studies have systematically assessed whether explanations reflect the model's decision process. This gap motivates a deeper examination of attention *faithfulness* in the detection of mental health risks using social media.

2.2 Explainability and Attention Faithfulness

Attention mechanisms are often viewed as providing interpretable insights into neural predictions. However, Jain and Wallace (2019) showed that attention distributions can be manipulated without affecting model outputs, questioning their explanatory power. Wiegreffe and Pinter (2019) argued that attention can serve as explanation if validated using diagnostic tests, including adversarial attention training.

Other studies proposed ways to evaluate *faithfulness* of the attention mechanism. Serrano and Smith (2019) and Chen et al. (2020) used input ablation to measure the impact of attended tokens. Chrysostomou and Aletras (2021) introduced Task-Specific Scaling to bias attention weights toward task-relevant tokens. More recently, Amini and Kosseim (2022) proposed ATTEFA, a metric that quantifies how well attention weights reflect a model's prediction under adversarial perturbations.

Our work builds on these efforts by enforcing faithfulness through adversarial training and evaluating explanation integrity via ATTEFA. We extend attention-based interpretability to the user level in a real-world task, showing that faithful explanations can enhance trust in systems for mental health support.

3 Methodology

In this section, we describe our user-aware attention-based framework for early detection of mental health risks from social media. The model operates at two levels: a post-level encoder generates contextualized embeddings for each user post, and a user-level attention mechanism aggregates these embeddings to produce a prediction, while simultaneously providing post-level interpretability. We employ DisorBERT as the base encoder, leveraging its domain adaptation to social media and mental health language. To ensure that attention-based explanations are faithful to the model's decision process, we apply adversarial training and quantify explanation faithfulness using the AtteFa

metric. We also outline our training setup, adversarial objective, datasets, and evaluation protocol.

3.1 Model

Figure 1 shows the overall architecture of the model, together with the inputs and outputs for each layer. The model classifies each user as *at risk* or *without risk* based on its social media posts. In the following subsections, we will review each layer separately.

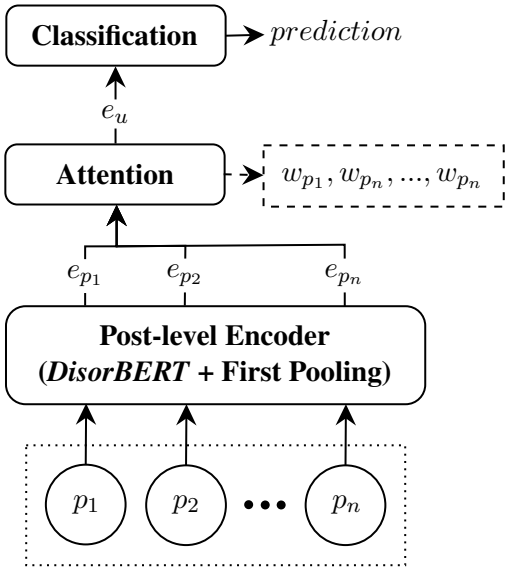

Figure 1: Model Architecture. $p_1, p_2, \ldots p_n$ represent the posts by user u, e_{p_i} and w_{p_i} denote the embedding and the attention weight for the post p_i, and e_u is the final embedding of the user u that is output by the attention.

3.1.1 Post-level Encoder

The post-level encoder layer is responsible for computing the 1D embedding (e_{p_i}) for each individual post by the user (p_i).

As the encoding model, we used *DisorBERT* (Aragon et al., 2023), which is a BERT-based model that is domain-adapted to both the mental health domain and social media language. We applied first pooling to extract the representation of the `[CLS]` token as the embedding for each post (e_{p_i}).

3.1.2 User-level Attention

Given a user u with posts $p_1, p_2, \ldots, p_n$, we first compute the logit l_{p_i} for each post p_i using Equation 1:

$$l_{p_i} = \text{F2}(\tanh(\text{F1}(e_{p_i}))) \quad (1)$$

where e_{p_i} is the embedding for the post p_i, computed using the post-level encoder (see Section 3.1.1). F1 is a fully connected layer, with $|e_{p_i}|$ input and $|e_{p_i}|/2$ output nodes. F2 is also a fully connected layer, with $|e_{p_i}|/2$ input nodes and 1 output node[1].

Having the logits for all posts, the attention weight w_{p_i} for each post p_i is computed through softmax, as shown in Equation 2:

$$w_{p_i} = \frac{\exp(l_{p_i})}{\sum_{j=1}^{n} \exp(l_{p_j})} \quad (2)$$

Lastly, the embedding e_u for the user u is computed through a weighted average over $e_{p_1}, e_{p_2}, \ldots, e_{p_n}$, with $w_{p_1}, w_{p_2}, \ldots, w_{p_n}$ as weights:

$$e_u = \sum_{i=1}^{n} w_{p_i} e_{p_i} \quad (3)$$

The attention layer returns both e_u (fed to the classification head), along with the vector $[w_{p_1}, w_{p_2}, \ldots, w_{p_n}]$, which we use to determine the informativeness of each post.

3.1.3 Classification Head

The classification head is composed of a dropout (with $p = 0.1$), followed by a feed-forward layer with one output node, followed by a sigmoid activation. Having e_u as the input, this component is used to predict the likelihood of the user u being *at risk* for the corresponding mental health problem.

3.2 Training Process

In this section, we describe the training procedures for the base and adversarial models. The base model is trained to detect mental health risks from user posts. In contrast, the adversarial model is specifically optimized to replicate the base model's predictions while assigning different attention weights to the user's posts. This approach enables a rigorous evaluation of whether the attention mechanism faithfully reflects the model's true decision process.

3.2.1 Base Model

Training is done using PyTorch (Paszke et al., 2019) and HuggingFace Transformers (Wolf et al., 2020). We initialize the *Post-level Encoder* from the pretrained checkpoint of *DisorBERT*. The rest of the parameters in the model are randomly initialized.

[1] In our experiments, $|e_{p_i}|$ was equal to 768 (i.e. the hidden size of *DisorBERT*).

Training is done using binary cross-entropy as the loss function. While the loss is back-propagated on a per-sample basis, we use gradient accumulation to leverage an effective batch size of 4. In order not to overflow the GPU memory, we enable gradient checkpointing (Chen et al., 2016), and feed the posts in minibatch sizes of 32. The text for each post is created through joining the title and the body of the corresponding post with a newline character (\n)[2].

AdamW (Loshchilov and Hutter, 2019) is used as the optimizer. We set the learning rate to $1e{-}5$, and use the default values in Pytorch for the remaining optimizer hyperparameters.

Training is continued until the area under precision-recall curve (AUC-PR) over the test dataset does not improve in 10 consecutive epochs, or a maximum of 50 epochs is reached. After the completion of training, the trained model checkpoint is picked at the epoch where the maximum AUC-PR over the test dataset was reached. We then set the prediction threshold so that it maximizes the F1 score on the test data.

3.2.2 Adversarial Model

We follow a similar approach to Amini and Kosseim (2022) to train the adversarial model. Using the trained base model, we compute the prediction and the user-level attention weights for each user in the dataset.

For each sample, the prediction loss is initially calculated as the absolute difference between the outputs of the base and adversarial models:

$$\mathcal{L}_y^i = |y_a^i - y_b^i| \tag{4}$$

where y_a^i and y_b^i denote the outputs of the adversarial and the base models, respectively, on sample i.

We then compute the Jensen-Shannon Divergence (JSD) between the user-level attention weights between the adversarial and the base models. With α_a^i and α_b^i representing the attention weights of the adversarial and the base models, respectively, for sample i, and $\bar{\alpha_i} = \frac{\alpha_a^i + \alpha_b^i}{2}$, the JSD between α_a^i and α_b^i is computed using Equation 5:

$$\text{JSD}(\alpha_a^i, \alpha_b^i) = \tfrac{1}{2}\text{KLD}(\alpha_a^i || \bar{\alpha_i}) + \tfrac{1}{2}\text{KLD}(\alpha_b^i || \bar{\alpha_i}) \tag{5}$$

To compute the Kullback–Leibler Divergence (KLD), we use $\epsilon = 1e - 10$ to avoid a mathematical error when computing the $\log(0)$. With this ϵ, the maximum possible value for JSD is ~ 0.6931, which we will later refer to as JSD_{max}.

We then compute the second portion of the loss, using the following equation:

$$\mathcal{L}_\alpha^i = \frac{\text{JSD}(\alpha_a^i, \alpha_b^i)}{JSD_{max}} \tag{6}$$

The division by JSD_{max} ensures the same theoretical range (i.e., between 0 and 1) for both $\mathcal{L}_i^y$ and $\mathcal{L}_\alpha^i$. With $\mathcal{L}_y^i$ and $\mathcal{L}_\alpha^i$ at hand, the adversarial loss for sample i is computed using Equation 7:

$$\mathcal{L}^i = \mathcal{L}_y^i + \mathcal{L}_\alpha^i \tag{7}$$

During training, the final loss to be back-propagated for each batch is computed by averaging the individual values of adversarial loss ($\mathcal{L}^i$) for each sample within the batch.

We use the average adversarial loss over the test data as the early stopping criterion. Aside from the computation of loss and the early stopping criterion, the remaining setup for the adversarial training phase is similar to the training of the base model.

Amini and Kosseim (2022) observed that maximizing the JSD between the user-level attention weights of the adversarial and the base model is a simpler task compared to minimizing the distance between the outputs of the two models. We argue that this would increase the chances of convergence to a local minima, where the $\mathcal{L}_y^i$ component of the loss is not fully optimized. To avoid such a problem, we have performed the adversarial training from both random initialization and the trained base checkpoints, with the earlier case using a similar setup as Amini and Kosseim (2022). In the latter scenario, $\mathcal{L}_y^i$ starts from a value close to zero[3], which would theoretically guide the adversarial model to retain close predictions to the trained base model.

3.3 Datasets and Tasks

We evaluate our setup using the following 4 datasets from the eRisk shared task series (Losada et al., 2017, 2018, 2019, 2020; Parapar et al., 2021, 2022, 2023). The datasets contain a collection of

[2]To accelerate training, we limited each post to its first 200 tokens.

[3]Due to the use of regularization techniques during training, the value of $\mathcal{L}_y^i$ is not exactly zero.

reddit posts and comments[4], grouped by users and sorted in chronological order. Each user is labelled as *at risk* or *without risk* for the corresponding mental health issue, using the methodology proposed in Losada and Crestani (2016). Table 1 shows the sources for each dataset, and Table 2 shows the distribution of the number of positive (i.e., *at risk*) and negative (i.e., *without risk*) samples (users) in the training and test datasets for each task.

Dataset	Subset	Source(s)
Anorexia	Train	eRisk 2018 - T2 (Losada et al., 2018)
	Test	eRisk 2019 - T1 (Losada et al., 2019)
Depression	Train	eRisk 2017 - T1 (Losada et al., 2017) eRisk 2018 - T1 (Losada et al., 2018)
	Test	eRisk 2022 - T2 (Parapar et al., 2022)
Gambling	Train	eRisk 2021 - T1 (Parapar et al., 2021) eRisk 2022 - T1 (Parapar et al., 2022)
	Test	eRisk 2023 - T2 (Parapar et al., 2023)
Self-harm	Train	eRisk 2019 - T2 (Losada et al., 2019) eRisk 2020 - T1 (Losada et al., 2020)
	Test	eRisk 2021 - T2 (Parapar et al., 2021)

Table 1: Sources of the datasets.

Dataset	Train		Test	
	pos	neg	pos	neg
Anorexia	61	411	73	742
Depression	214	1,493	98	1,302
Gambling	245	4,182	103	2,071
Self-harm	145	618	152	1,296

Table 2: Distribution of the number of users in each dataset.

3.4 Evaluation Method

We evaluate the precision, recall, F1, accuracy, and the PR-AUC for the trained base models. Furthermore, in order to determine how well our trained model performs compared to the top-performing models in each shared task, we compute the *latency-weighted F1 score* of our models, using decision-based evaluation protocol proposed by the eRisk shared task organizers (Losada et al., 2019, 2020; Parapar et al., 2021, 2022):

1. For each user in the test data, the model processes their posts in chronological order and issues a binary prediction on whether they are at risk or not. Once a user is detected as at risk by the model, the prediction cannot be reverted back after observing more posts.

2. At each time step, the F1 score is computed based on the correctness of the prediction.

3. For each true positive (i.e., a correct prediction for an at-risk user), we record the number of posts k_u that the system has to process in order to make this prediction.

4. A latency penalty to each true positive is computed using Equation 8:

$$\text{penalty}(k_u) = -1 + \frac{2}{1 + e^{-p(k_u - 1)}} \quad (8)$$

where $p = 0.0078$, calibrated so the penalty equals 0.5 at the median number of posts in the dataset.

5. The overall detection speed is computed using Equation 9:

$$\text{speed} = 1 - \text{median}\{\text{penalty}(k_u)\} \quad (9)$$

which ranges from 1 (instance detection) to 0 (very late detection).

6. The *latency-weighted F1* is computed using Equation 10:

$$lw_{F1} = \text{F1} \cdot \text{speed} \quad (10)$$

Lastly, we compute *AtteFa* (Amini and Kosseim, 2022) on the test datasets to assess the faithfulness of the user-level attention weights, using Equation 11:

$$\text{AtteFa} = \frac{1}{N} \sum_{i=1}^{N} \min \left(\frac{|y_a^i - y_b^i| \cdot JSD_{max}}{\text{JSD}(\alpha_a^i, \alpha_b^i)}, 1 \right) \quad (11)$$

where y_b^i and y_a^i correspond to the output predictions of the base and the adversarial models, respectively, for user i; α_b^i and α_a^i represent the user-level attention weights; N is the total number of users in the test dataset; and $JSD_{max} \approx 0.6931$, as stated in Section 3.2.2.

4 Results and Discussion

In this section, we present the results of our proposed framework on four mental health detection tasks from the eRisk benchmark: anorexia, depression, pathological gambling, and self-harm. We first evaluate the predictive performance of our base model using standard and latency-weighted metrics, comparing its competitiveness against top-performing systems from the shared tasks. We then analyze the faithfulness of the user-level attention weights via the AtteFa metric.

[4]For simplicity, the term "post" is used throughout this paper to encompass both posts and comments.

4.1 Base Model Evaluation

Table 3 shows the precision, recall, F1, accuracy, and PR-AUC of the trained based models on each task. The results show that the proposed approach is capable of detecting mental health issues with a high accuracy and a relatively high F1 score.

Dataset	P	R	F1	A	PR-AUC
Anorexia	0.896	0.822	0.857	0.975	0.921
Depression	0.734	0.816	0.773	0.966	0.786
Gambling	0.971	0.981	0.976	0.998	0.998
Self-harm	0.860	0.684	0.762	0.955	0.807

Table 3: Base Results on test datasets.

To assess the competitiveness of our approach compared to the best-performing models in each eRisk shared task, Table 4 provides the ranking of our trained model according to *latency-weighted F1*, compared to the participating models in the corresponding shared task. The results show that our approach is able to provide better or competitive results to the best models, showing promise in our approach.

4.2 Faithfulness of the User-level Attention

Table 5 shows the AtteFa scores for each model, along with the average $\mathcal{L}_y$ and $\mathcal{L}_\alpha$ over the test datasets.

Dataset	Initialization	AtteFa	$\mathcal{L}_y$	$\mathcal{L}_\alpha$
Anorexia	Random	0.527	0.476	0.916
	Trained	0.524	0.481	0.931
Depression	Random	0.497	0.464	0.943
	Trained	0.487	0.455	0.944
Gambling	Random	0.512	0.493	0.968
	Trained	0.731	0.581	0.777
Self-harm	Random	0.538	0.495	0.931
	Trained	0.539	0.490	0.920

Table 5: Adversarial Results.

The AtteFa scores reported in Table 5 average around 0.5. In light of the findings by Amini and Kosseim (2022), we hypothesize that, on average, only 10% or fewer of a user's posts are informative for detecting a mental health problem.[5] This aligns with prior studies (Song et al., 2018; Gui et al., 2019; Amini and Kosseim, 2020), which indicate that signs of mental health issues are sparse in social media data. The results also reflect a high degree of faithfulness: if the model attends to a different subset of posts, its predictions deviate considerably from the optimal. This is further supported by the observed $\mathcal{L}_y$ values (approximately 0.5 on average), suggesting that the adversarial models behave similarly to random predictors when attention weights are perturbed.

Another notable observation in Table 5 is the high values of $\mathcal{L}_\alpha$ for both types of weight initialization. This indicates that the initialization of the weights does not play a significant role in the convergence, and in almost all cases, maximizing the divergence on the attention weights is a significantly easier objective compared to minimizing the distance between the predictions. The only exception where $\mathcal{L}_\alpha$ does not approach its upper bound can be observed for the Gambling task with the trained base checkpoint as the initialization. We hypothesize that this is due to missing the optimal model checkpoint during adversarial training, as the Gambling dataset is larger than the others and we only saved and evaluated checkpoints at the end of each epoch. We believe that a more frequent checkpointing and evaluation would help bridge the gap on the final value of $\mathcal{L}_\alpha$ between this dataset and the others.

5 A Closer Look at the Attention Weights

Amini and Kosseim (2020) demonstrated that attention weights typically correlate with the strength of signals for mental health issues. In their work, they sorted posts in a descending order according to their corresponding attention weights, and obtained the model's prediction after passing only the top N posts to the model. They observed that increasing the number of posts results in an increasing trend in the precision and a decreasing trend in the recall, indicating that the model usually tends to predict a user as at-risk, if the model only observed the very few highest-weighted posts for that user.

In our study, we perform a variation of their experiment: Instead of starting with only the highest-weighted posts and gradually including the lower-weighted ones, we start by feeding all of the posts to the model and gradually remove the highest-weighted posts one at a time and monitor the prediction flips. Our hypotheses were the following:

1. For users predicted as *without risk*, the model should consistently emit a no-risk prediction. This is because the model did not predict those

[5] Amini and Kosseim (2022) reported an AtteFa score of 0.5 on a synthetic dataset in which only 10 out of 100 tokens per example carried sentiment weights.

Rank	**Anorexia** (55)		**Depression** (63)		**Gambling** (50)		**Self-harm** (56)	
	Team (Run)	lw_{F1}	Team (Run)	lw_{F1}	Team (Run)	lw_{F1}	Team (Run)	lw_{F1}
1	**OURS**	**0.766**	NLPGroup-IISERB (0)	0.690	ELiRF-UPV (0)	0.927	UNSL (4)	0.622
2	CLaC (4)	0.690	**OURS**	**0.558**	NLP-UNED-2 (3)	0.877	**OURS**	**0.612**
3	CLaC (1)	0.690	BLUE (0)	0.540	NLP-UNED-2 (1)	0.876	UNSL (3)	0.583
4	CLaC (3)	0.680	UNSL (2)	0.519	Xabi_EHU (0)	0.875	BLUE (2)	0.578
5	CLaC (2)	0.680	NLPGroup-IISERB (3)	0.511	NLP-UNED-2 (2)	0.875	NLP-UNED (4)	0.564
6	INAOE-CIMAT (3)	0.630	LauSAn (3)	0.498	**OURS**	**0.848**	Birmingham (0)	0.551
7	lirmm (0)	0.630	NLPGroup-IISERB (1)	0.496	Xabi_EHU (3)	0.844	NLP-UNED (1)	0.546
8	INAOE-CIMAT (0)	0.620	BioInfo_UAVR (4)	0.494	Xabi_EHU (1)	0.839	NLP-UNED (0)	0.545
9	lirmm (1)	0.620	Sunday-Rocker2 (1)	0.439	NLP-UNED-2 (0)	0.833	BLUE (3)	0.534
10	INAOE-CIMAT (4)	0.610	UNSL (1)	0.426	Xabi_EHU (4)	0.823	NLP-UNED (3)	0.524

Table 4: Model rankings for each task based on the *latency-weighted F1* score. The number of participating models in each task is indicated in parentheses in the header row. For details on the corresponding eRisk shared tasks, please refer to the test dataset sources listed in Table 1.

users as *at risk*, even when it observed their highest-weighted posts.

2. For users predicted as *at risk*, there should be a cut-off point in the number of posts, at which the model starts considering them as *without risk*. Removing more high-weighted posts should not change the model's prediction.

3. As we expect that, on average, less than 10% of posts contain signs of mental health issues (see Section 4.2), the decision flip should often happen after the removal of <10% of the posts.

In our experiment, the first hypothesis is shown to be true for all instances in all 4 datasets. The second hypothesis also turns out to be correct, with the exception of 2 instances in the *Depression* dataset, where the model's prediction flips again from *without risk* to *at risk* after removing one extra post after the cut-off. But even in those 2 cases, the model starts predicting *without risk* again after another removal, which remains constant throughout the rest of the process. We believe that this happens due to the model not being fully optimized on the test dataset, resulting in a small number of exceptions where the attention weights do not fully correlate with the degree of the sign of mental health issues. But the fact that this occurs in fewer than 1% of cases highlights the consistency with which the user-level attention weights align with the degree of mental health signal in the posts.

For the third hypothesis, for each test user that was predicted as *at risk*, we calculate the cut-off point at which the decision flip happens. In Table 6, we report the average, median, and the standard deviation of the median of the proportion of posts (in percentage) that need to be removed in order to observe a decision flip (from *at risk* to *without risk*). Also in Figure 2, we provide the violin to visually show the distribution of the cut-off points in proportion to the total number of posts by each user.

Dataset	**% Posts**		
	Avg	**Med**	**Std**
Anorexia	6.27	3.18	7.71
Depression	4.92	1.04	7.99
Gambling	16.03	11.27	16.77
Self-harm	4.61	1.60	6.86

Table 6: Average, median, and standard deviation of the percentage of the highest weighted posts per user that need to be removed until a decision flip happens from 1 (*at risk*) to 0 (*without risk*).

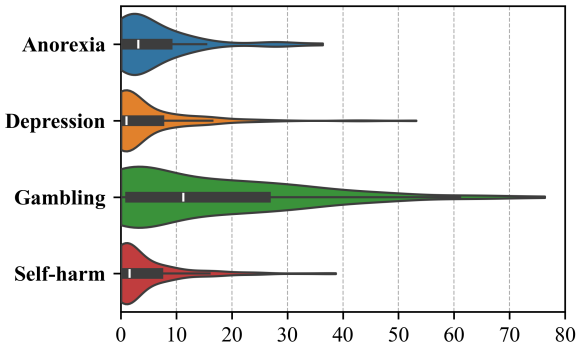

Figure 2: Distribution of the percentage of the highest weighted posts per user that need to be removed until a decision flip happens from 1 (*at risk*) to 0 (*without risk*).

As Table 6 shows, for the *Anorexia*, *Depression*, and *Self-harm* datasets, the average cut-off is under 10%. The *median + standard deviation* is $\approx$ 10% for the *Anorexia* dataset, while being lower than 10% for *Depression* and *Self-harm*. These

are also observable in the violin plots, where the distribution is mostly centered below 10%. This observation is aligned with the findings of Amini and Kosseim (2022), where they observed that an AtteFa of ≈ 0.5 occurs when only 10% or fewer samples within a data point are informative to the task.

This behavior, however, is not exhibited in the *Gambling* dataset, where the average and median cut-off is above 10%. Further analysis is needed to understand why this dataset exhibits this pattern. While this may occur due to the model not being fully optimized on this task (due to the infrequent checkpointing and evaluation during training, as stated in Section 5), it would be useful to perform a deeper study to understand if an inherent characteristic of this dataset has also played a role.

6 Limitations

As stated in Section 1, the focus of our study was to assess the explainability of the user-level attention mechanism, as opposed to achieving the state of the art. Therefore, we aimed to develop a model with comparable performance to the best performing models in the eRisk shared tasks through minimal hyperparameter tuning and leveraging the test dataset for early stopping and final checkpoint selection. In order to understand the full extent of performance and generalization capability for our approach, the following work remains:

For a fair comparison, the test dataset should remain entirely unseen during training. This also allows for a more accurate assessment of the model's generalization ability to unseen data. Furthermore, our system should be benchmarked on a wider set of tasks/datasets and against a larger number of systems in the literature. Lastly, we need to further adjust the model or training hyperparameters, and experiment with different encoder models, in order to achieve the most optimal results.

In terms of *explainability assessment*, our current study focuses only on faithfulness and transparency (i.e., what posts were deemed important by the model). In order to fully claim that our system is explainable, we should also assess the notions of plausibility and sufficiency, which are the other two pillars in explainability (Wiegreffe and Pinter, 2019). Such assessments require human assessment, potentially by experts in the field.

Lastly, in order to assess the feasibility of using such a system in a real-life scenario, studies should be done in terms of computational cost and inference latency. In addition, further optimizations are necessary in order to improve the system on those fronts. An additional study would also be necessary to evaluate the practical risks of deploying such a system, in terms of privacy and ethical considerations, and mitigate or reduce such risks as much as possible.

7 Conclusion

In this paper, we introduced a user-aware, attention-based framework for detecting mental health risks from social media, designed to generate faithful and transparent predictions. Our architecture combines DisorBERT for post-level encoding with a user-level attention mechanism that provides both predictive performance and a potential means for interpretability. To assess the faithfulness of the attention weights, we employed adversarial training and computed *AtteFa*, a metric designed to quantify how faithfully attention reflects decision-relevant input. Across four tasks in the eRisk shared task series, our system achieved competitive latency-weighted F1 scores and demonstrated high attention faithfulness, particularly in domains where signals are rare and sparsely distributed across user posts.

Our experimental analysis further confirmed that a small fraction of posts – often under 10% – tend to drive the model's predictions, supporting the hypothesis that mental health signals in social media are both rare and concentrated. These findings underscore the need for transparent and interpretable systems, especially in sensitive domains like mental health, where trust and explainability are essential for practical adoption.

While addressing the limitations presented in Section 6 is a prominent future direction, another line of work would be to further move from a system that can detect individual mental health problems to a more general system to assess mental health issues in social media. One promising starting point is to train the model in a multi-task learning fashion (e.g., Kendall et al., 2018) on a combination of datasets related to the detection of mental health issues, allowing the model to leverage shared knowledge between different tasks.

Acknowledgment

This work was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Hessam Amini and Leila Kosseim. 2020. Towards explainability in using deep learning for the detection of anorexia in social media. In *Proceedings of the 2020 International Conference on Applications of Natural Language to Information Systems (NLDB 2020)*, pages 225–235.

Hessam Amini and Leila Kosseim. 2022. How (un)faithful is attention? In *Proceedings of the 2022 Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackBoxNLP 2022)*, pages 119–130, Abu Dhabi, UAE.

Mario Aragon, Adrian Pastor Lopez Monroy, Luis Gonzalez, David E. Losada, and Manuel Montes. 2023. DisorBERT: A double domain adaptation model for detecting signs of mental disorders in social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, pages 15305–15318, Toronto, Canada.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020. Generating hierarchical explanations on text classification via feature interaction detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 5578–5593, Online.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *CoRR*, abs/1604.06174.

George Chrysostomou and Nikolaos Aletras. 2021. Improving the faithfulness of attention-based explanations with task-specific information for text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021) and the 11th International Joint Conference on Natural Language Processing (IJCNLP 2021) – Volume 1: Long Papers*, pages 477–488, Online.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2014): From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA.

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

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the 7th International AAAI Conference on Web and Social Media (ICWSM 2013)*, pages 128–137, Cambridge, Massachusetts, USA.

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

Tao Gui, Qi Zhang, Liang Zhu, Xu Zhou, Minlong Peng, and Xuanjing Huang. 2019. Depression detection on social media with reinforcement learning. In *Proceedings of Chinese Computational Linguistics: 18th China National Conference (CCL 2019)*, page 613–624, Kunming, China.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 4198–4205, Online.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pages 3543–3556, Minneapolis, Minnesota, USA.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of 2018 Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, Salt Lake City, Utah, USA.

David E. Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *Proceedings of CLEF 2016: Conference and Labs of the Evaluation Forum*, Évora, Portugal.

David E Losada, Fabio Crestani, and Javier Parapar. 2017. eRisk 2017: CLEF lab on early risk prediction on the internet: Experimental foundations. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*, Dublin, Ireland.

David E Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of eRisk 2018: Early risk prediction on the internet (extended lab overview). In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, Avignon, France.

David E. Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of eRisk 2019: Early Risk Prediction on the Internet (extended overview). In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*, Lugano, Switzerland.

David E. Losada, Fabio Crestani, and Javier Parapar. 2020. Overview of eRisk at CLEF 2020: Early Risk Prediction on the Internet (extended overview). In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, Thessaloniki, Greece.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*, New Orleans, LA, USA.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. 2021. Measuring and improving faithfulness of attention in neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021): Main Volume*, pages 2791–2802, Online.

Ankit Murarka, Balaji Radhakrishnan, and Sushma Ravichandran. 2021. Classification of mental illnesses on social media using RoBERTa. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis (LOUHI 2021)*, pages 59–68, online.

Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2021. Overview of eRisk at CLEF 2021: Early Risk Prediction on the Internet (extended overview). In *Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum*, Bucharest, Romania.

Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2022. Overview of eRisk at CLEF 2022: Early Risk Prediction on the Internet (extended overview). In *Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, Bologna, Italy.

Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2023. Overview of eRisk at CLEF 2023: Early Risk Prediction on the Internet (extended overview). In *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*, Thessaloniki, Greece.

Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2024. Overview of eRisk at CLEF 2024: Early Risk Prediction on the Internet (extended overview). In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*, Grenoble, France.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the 2019 Conference on Neural Information Processing Systems (NeurIPS 2019)*, volume 32, Vancouver, Canada.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 2931–2951, Florence, Italy.

Hoyun Song, Jinseon You, Jin-Woo Chung, and Jong C. Park. 2018. Feature attention network: Interpretable depression detection from social media. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation (PACLIC 32)*, pages 613–622, Hong Kong.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pages 11–20, Hong Kong, China.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): System Demonstrations*, pages 38–45, Online.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. 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 (NAACL-HLT 2016)*, pages 1480–1489, San Diego, California, USA.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China.

Hamad Zogan, Imran Razzak, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. 2022. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web*, 25(1):281–304.