Attentive Multi-stage Learning for Early Risk Detection of Signs of Anorexia and Self-harm on Social Media

Waleed Ragheb[1,2], Jérôme Azé[1,2], Sandra Bringay[1,3], and Maximilien Servajean[1,3]

[1] LIRMM UMR 5506, CNRS, University of Montpellier, Montpellier, France
[2] IUT de Béziers, University of Montpellier, Béziers, France
[3] AMIS, Paul Valéry University - Montpellier 3 , Montpellier, France
{first.last}@lirmm.fr

Abstract. Three tasks are proposed at CLEF eRisk-2019 for predicting mental disorder using users posts on Reddit. Two tasks (T1 and T2) focus on early risk detection of signs of anorexia and self-harm respectively. The other one (T3) focus on estimation of the severity level of depression from a thread of user submissions. In this paper, we present the participation of LIRMM (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier) in both tasks on early detection (T1 and T2). The proposed model addresses this problem by modeling the temporal mood variation detected from user posts through multi-stage learning phases. The proposed architectures use only textual information without any hand-crafted features or dictionaries. The basic architecture uses two learning phases through exploration of state-of-the-art deep language models. The proposed models perform comparably to other contributions.

Keywords: Classification, LSTM, Attention, Temporal Variation, Bayesian Variational Inference, Anorexia, Self-harm

1 Introduction

Anorexia is consider one of the most common eating disorder. It is characterized by low weight, worry of gaining weight, and a powerful need to be skinny, leading to food restriction. Many who suffer from eating disorder see themselves as overweight although they could be thin [8]. Individuals with eating disorders have also been shown to have lower employment rates, in addition to an overall loss of earnings. Eating disorder sufferers who are experiencing an overall loss in earnings associated with their illness are also magnified by the excess of healthcare costs. According to the National Eating Disorder Association (NEDA), up

to 70 million people worldwide suffer from eating disorders [1]. Eating disorder symptoms are beginning earlier in both males and females. As estimated, 1.1 to 4.2 percent of women suffer from anorexia at some point in their lifetime [6]. Young people between the ages of 15 and 24 with anorexia have 10 times the risk of dying compared to their same-aged peers.

Self-harm is a very common problem, and many people are struggling to deal with it [9]. Several illnesses are associated with self-harm, including borderline personality disorder, depression, eating disorders, anxiety or emotional distress [3]. Self-harm occurs most often during the teenage and young adult begin around age 14 and carry on into their 20s, though it can also happen later in life [9]. There is also an increased risk of suicide in individuals who self-harm and it is found in 40% to 60% of suicides [5].

Social media is becoming increasingly used not only by adults but also at different age stages. Mental disordered patients also turn to online social media and web forums for information on specific conditions and emotional support. Even though social media can be used as a very helpful tool in changing a person's life, it may cause such conflicts that can have a negative impact. This puts responsibilities for content and community management for monitoring and moderation. With the increasing number of users and their contents, these operations turn out to be extremely difficult. Many social media try to deal with this problem by reactive moderation. In reactive moderation, users report any inappropriate, negative or risky user generated contents. However it may reduce the workload or the cost of moderating, it is not enough especially for handling mental disordered user's threads or posts.

Previous researches on social media have established the relationship between an individual's psychological state and his\her linguistic and conversational patterns [19, 18]. This motivate the task organizers to initiate the pilot task for detecting depression from user posts on Reddit[1] in eRisk-2017 [11]. In eRisk-2018 the extension of the study was planned to include detection of anorexia. In eRisk-2019, a continuation of anorexia tasks in addition to two other tasks are proposed. One task is for early detection of signs of self-harm (T2). In this task no training dataset is provided. Also, another new task for detection of severity level of depression (T3) is presented. Tasks organizers proposed new evaluation measures than what were used before.

In this paper, we present the participation of LIRMM (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier) in both tasks for early detection of anorexia and self-harm in eRisk-2019. The originality of our approach is to perform the detection through two main learning phases. In the first learning phase. we proposed Deep Mood Evaluation Module (DMEM) that uses attention based deep learning models to construct a time series representing temporal mood variation through users posts or writings. The second phase is either to use machine learning or Bayesian inference model to obtain

[1] Reddit is an open-source platform where community members (red-ditors) can submit content (posts, comments, or direct links), vote submissions, and the content entries are organized by areas of interests (subreddits).

the proper decision. The main idea is to give a decision once the models detect clear signs of mental disorder from current and previous mood extracted from the content.

The rest of the paper is organized as follows. In Section 2, the related work is introduced. Then in Section 3, a brief tasks (T1 and T2) description of early risk detection and used datasets are presented. Section 4 presents the proposed models. The experimental setup and all model variants used are introduced in Section 5. In Section 6, the evaluation results and discussions are presented. We conclude the study and experiments in Section 7.

2 Related Work

Recent psychological studies showed the correlation between person's mental status and mood variation over time [11]. It is also evident that some mental disordered may have chronic week-to-week mood instability. It is a common presenting symptom for people with a wide variety of mental disorders, with as many as 8 of 10 patients reporting some degree of mood instability during assessment. These studies suggest that clinicians should screen for temporal mood variation across most common mental health disorders.

Concerning text representation, traditional Natural Language Processing (NLP) modules start with feature extraction from text such as the count or frequency of specific words, predefined patterns, Part-of-Speech tagging, etc. These hand-crafted features should be selected carefully and sometimes with an expert view. However these features are interesting [22], sometimes they loose the sense of generalization. Another recent trend is the use of word and documents vectorization methods. These strategies that convert either words, sentences or even overall documents into vectors take into account all the text not just parts of it. There are many ways to transform a text to high-dimensional space such as term frequency and inverse document frequency (TF-IDF), Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), etc [14]. This direction was revolutionized by Mikolov et al. [16, 17] who proposed the Continuous Bag Of Words (CBOW) and skip-gram models known as Word2vec. It is a probabilistic based model that makes use of a two layered neural network architecture to compute the conditional probability of a word given its context. Based on this work Le et al. [10] propose Paragraph Vector model. The algorithm which is also known as Doc2vec learns fixed-length feature representations from variable-length pieces of texts, such as sentences, paragraphs, and documents. Both word vectors and documents vectors are trained using stochastic gradient descent and back-propagation shallow neural network language models. The development of Universal Language Model Fine Tuning (ULMFiT) is considered like moving from shallow to deep contextual pre-training word representation [7]. This idea has been proved to achieve Computer Vision (CV)-like transfer learning for many NLP task. ULMFiT make use of the state-of-the art language model AWD-LSTM (Average stochastic gradient descent - Weighted Dropout LSTM)

proposed by Merity et al. in 2017 [15]. The same 3-layer LSTM recurrent architecture with the same hyperparameters and no additions other than tuned dropout hyperparameters are used. The classifier layers above the base LM encoder is simply a pooling layer (maximum and average pool) followed by three fully-connected linear layers. The overall models signicantly outperforms the state-of-the-art on six text classication tasks including three tasks for sentiment analysis. In this paper, we will use these techniques for text representations.

Attention mechanism is considered as one of the recent trends in NLP models [2]. It can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. This can be seen as take a collection of vectors, whether it could be a sequence of vectors representing a sequence of words, or an unordered collections of vectors representing a collection of attributes and summarize them into a single vector. This summarization is done by scoring each input sequence with a probability-like scores obtained from the attention. This helps the model to pay close attention to the sequence items with higher attention scores. In this paper, we will evaluate the effect of attention mechanisms on the model.

In this paper, we will use deep attention based modification of ULMFiT classifier to construct a time series representing temporal mood variation. We the used classical machine learning and statistical models to get the final decisions.

3 Tasks Description

In CLEF eRisk 2019, three tasks are presented [13]. The first task (T1) is for early detection of signs of anorexia. It is a continuation of the same task in eRisk-2018. The second one (T2) is a new task in 2019 for early detection of signs of self-harm. No training data is provided for this task. Another task was proposed (T3) for measuring the severity of the signs of depression. In this section we will describe the first two tasks (T1 and T2) that we have participated on.

Both tasks are considered as a binary classification problem. The datasets are a dated textual data of user posts and comments -posts without titles- on Reddit. The training and testing datasets are provided in stream of user writings (posts and comments). The stream is ordered chronologically. A brief statistics and summary for these datasets are provided in Table 1. Task organizers set up a server that iteratively gives user writings to the participating teams. The goal is not only to perform classification but also to do it as early as possible using minimum amount of writings for each user. A decision must be sent after processing each user writing to continue receiving more. This decision could be positive risk case or postponed for future writings. A detailed description of the tasks and used evaluation metrics can be found in the corresponding task description paper [13].

Table 1: Summary of early risk detection tasks (T1 and T2) Datasets

	T1		T2
	Training	Testing	Testing
No. of Users (At-risk/Controlled)	472 (61/411)	815 (73/742)	340(41/299)
No. of writings	253,341	570,510	170,698
Avg. No. of writings/User	536.74	700.01	502.05
Avg. writings Size (words)	35.38	34.83	33.15
Vocabulary Size	117,090	210,763	105,448

4 Proposed Models

The temporal aspects of the eRisk tasks inspired us to model the temporal mood variation through user's text content. The average number of days ranging from the first submission to the last submission is approximately 600 days. So, determining the way in which user's posts and comments vary from positive to negative and vice versa through time is worth inspecting. In the proposed models, the main idea is to process user writings for each user and determine the probability of how positive or negative it is. A detailed description of our model can be found in the working notes paper of eRisk 2018 [20]. The proposed architecture of our models comes in three main steps.

Step 1 - Text Vectorization Module: It is considered as language modeling step. The input of this step is the textual training datasets and the output is text vectorization model.

Step 2 - Mood Evaluation Module: This step is considered as the first supervised learning phase. Assign to each writing a probability like score representing how positive (risky) the submission is. The output of this step is a time series representing the mood variability over time. These time series will be the training set of the second learning phase.

Step 3 - Temporal Modeling Module: Another learning phase is to build machine learning models to learn some patterns from these time series to come up with the final classification model.

We tried to encapsulate text vectorization and mood evaluation modules and proposed Deep Mood Evaluation Module (DMEM). This module is based on ULMFiT architecture [7] and the idea of transfer learning for language modeling in addition to using attention layers for classifications. In addition, we tried Bayesian Variational Inference (BVI) [21] for the second learning phase.

4.1 Deep Mood Evaluation Module (DMEM)

We propose a modification of the basic architecture of the ULMFiT by adding attention to the model. The proposed architecture will help the model to focus on the important parts of the text that influence the network decision. Figure 1 shows the proposed model and the separation between encoder layers (text vectorization module) and classifier layers (mood evaluation module).

The input sequence is passed to the embedding layer then the three Bi-LSTM layers to form the output of the encoder. The encoder output has the form of $X_i = \{x_1^i, x_2^i, x_3^i, \ldots, x_N^i\}$ where N is the sequence length. The attention layer takes the encoded input sequence and computes the attention scores S^i. The attention layer can be viewed as a linear layer without bias.

$$\begin{aligned} \alpha^i &= \{W^i.X^i\} \\ S^i &= log[\frac{exp(\alpha^i)}{\sum_{j=1}^{N} exp(\alpha_j^i)}] \end{aligned} \tag{1}$$

Where W^i is the weight of the attention layer of the i^{th} sequence. The attention scores S^i is used to compute the scored sequence $O^i = \{o_1^i, o_2^i, o_3^i, \ldots, o_N^i\}$ which has the same length as the input sequence.

$$O^i = S^i \odot X^i \tag{2}$$

Since the input sequence to the attention layer (encoder output) resulted from Bi-LSTM layers, the last element in the scored output S_N^i can be used for representing the whole sequence. The whole sequence is represented by the weighted sum of all output sequences $\bar{O}^i$.

$$\bar{O}^i = \sum_{<N>} S^i \odot X^i \tag{3}$$

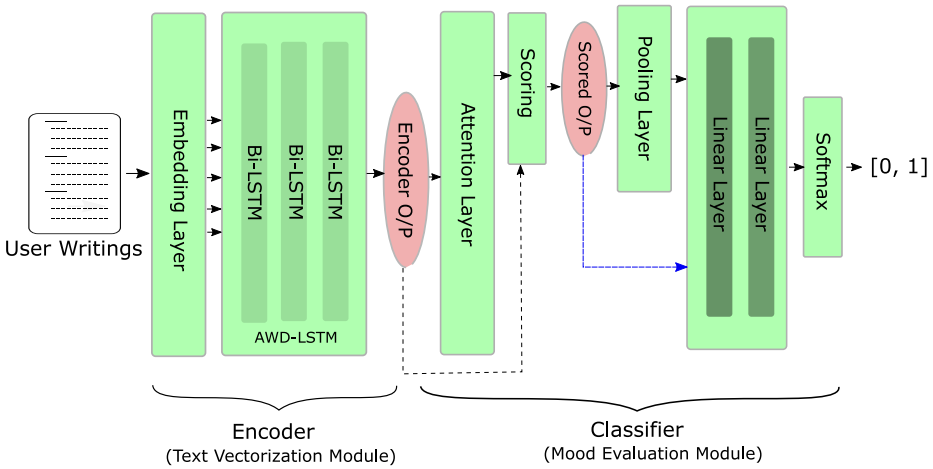

Fig. 1: Deep Mood Evaluation Module (DMEM)

For classification layers, a simple concatenation between the maximum and average pooling in addition to the scored output is inputted to a group of two

different sizes fully connected linear layers. The output of the last linear layer is passed to the Softmax to form the network decision.

Training the over whole models comes into three main steps proposed in [7].

1. The LM is initialized by training the encoder on a general-domain corpus (Wikitext-103 dataset [23]). This helps to capture general features of the language. Preserve low-level representations and adapt high-level ones
2. The pre-trained LM is fine-tuned using the training datasets for both tasks.
3. The classifier and the encoder is fine-tuned on the target task using different strategies for each layer group.

The training of the architecture is done using slanted triangular learning rates (STLR), discriminative fine-tuning (Discr) and layers gradual unfreezing proposed for ULMFiT with the same hyperparameter settings [7]. We train the model on the forward language models for both the general-domain and task specific datasets. Training the attention layer uses the same learning rates and cycles used in the classification layers group.

4.2 Bayesian Variational Inference (BVI)

We can represent the problem of classifying users from the already classified (observed) writings as a variant of independent Bayesian classifier combination [21]. Figure 2 shows the graphical model for the proposed BVI where the observed random variable W_i^k represents if the i^{th} writing for the k^{th} user if it is classified as positive or negative such that:

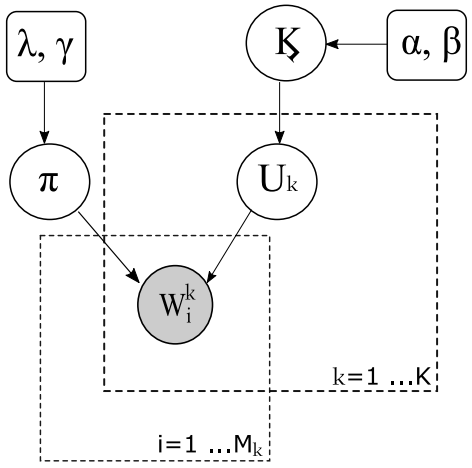

Fig. 2: Graphical Model for BVI: The shaded node represents observed values, circular nodes are variables with a distribution and rectangular nodes are instantiated variables

$$\begin{aligned} W_i^k &\sim \text{Bernoulli}(\pi_{u_k}) \\ \pi_i &\sim \text{Beta}(\lambda, \gamma) \end{aligned} \tag{4}$$

The hidden variable u_k represents if the user will be classified as at-risk (anorexia, self-harm) or not. So we can say:

$$\begin{aligned} u_k &\sim \text{Bernoulli}(\kappa) \\ \kappa &\sim \text{Beta}(\alpha, \beta) \end{aligned} \tag{5}$$

The variables λ, γ, α and β are the hyper-parameters reflecting our *a priori* belief about the proportion of positive and negative users.

We are interested in the *posterior* distribution of the random variable U_k, that defines if the user is positive or negative, which is unfortunately intractable. We use a variational inference approach to compute an approximation such as in [21]. The approximation is obtained by solving the following equation for all variables Z_i conditionned on the observed data X:

$$\log\, q_i(Z_i|X) = \mathbb{E}_{j \neq i}[\log\, p(Z, X)] + \text{const.} \tag{6}$$

So, we start from a number of positive and negative user writings (N^d) where $d \in \{+, -\}$ for positives and negatives respectively. More specifically:

$$N^+ = \sum_{k,i} 1[W_i^k = 1], \qquad N^- = \sum_{k,i} 1[W_i^k = 0] \tag{7}$$

Then, the expected number of positive and negative writings for positive users can be represented by N_1^+ and N_1^- respectively. The same for negative users is N_0^+ and N_0^-. These values are computed as:

$$N_r^d = \sum_{k,i} \mathbb{E}[1[u_k = d]].[1[w_i^k = r]], \quad d \in \{+, -\}, r \in \{0, 1\} \tag{8}$$

We can estimate the expectation of the log of the probability to observe positive writings independently of the user category as $\mathbb{E}[\ln(\kappa)]$ and for negative writings as $\mathbb{E}[\ln(1 - \kappa)]$ such that:

$$\begin{aligned} \mathbb{E}[\ln(\kappa)] &= \psi(\alpha + N^+) - \psi(\alpha + \beta + N^+ + N^-) \\ \mathbb{E}[1 - \ln(\kappa)] &= \psi(\beta + N^-) - \psi(\alpha + \beta + N^+ + N^-) \end{aligned} \tag{9}$$

Where ψ is the digamma function defined as the logarithmic derivative of the gamma function. In addition, we can estimate the expectation of the log probability for positive users to write positive writings as $\mathbb{E}[\ln(\pi_1)]$ and for negative users as $\mathbb{E}[\ln(\pi_0)]$ where:

$$\begin{aligned}\mathbb{E}[\ln(\pi_i)] &= \psi(\lambda + N_i^+) - \psi(\lambda + \gamma + N_i^+ + N_i^-) \\ \mathbb{E}[1 - \ln(\pi_i)] &= \psi(\gamma + N_i^-) - \psi(\lambda + \gamma + N_i^+ + N_i^-)\end{aligned} \tag{10}$$

So, the expectation of a user to be positive or negative can be obtained as:

$$\begin{aligned}\ln(\rho_j^k) = \sum_i^{M_k} W_i^k \, \mathbb{E}[\ln(\pi_j)] + (1 - W_i^k) \, \mathbb{E}[\ln(1 - \pi_j)] \\ + (\alpha - 1) \, \mathbb{E}[\ln(\kappa)] + (\beta - 1) \, \mathbb{E}[\ln(1 - \kappa)] \\ \mathbb{E}[1[U_k = j]] = \frac{\rho_i^k}{\sum_j \rho_i^k}\end{aligned} \tag{11}$$

Where $\mathbb{E}[1[U_k = j]]$ is a normalized value for the two types of users (at-risk or controlled). We can evaluate an optimal value for it iteratively by first initializing all factors, then updating each in turn using the expectations with respect to the current values of the other factors [21].

5 Experimental Setup

For each task, each team could participate with different five runs. We create different variants of our proposed architecture. In this section, we will present all these variants, training procedures and model hyperparameters.

5.1 Proposed Model Variants

All the proposed model variants for both tasks are based on two supervised learning phases (step 2 and step 3 in temporal mood variation model). For self-harm detection task (T2), as there is no training data, we train our models on the depression and anorexia datasets of eRisk-2018 [12]. We assumed that if a person with a clear signs of depression and/or anorexia could think about harm himself. We used the DMEM module as the first learning phase an all the variants and tried different machine learning and statistical methods as the second learning phase. Table 2 shows the used model for the second learning phase in all the runs for both tasks. MLP stands for Multi-Layer Perceptrons and RF is for Random Forest. All models that do not employ another learning phase are marked by dashes. In these runs, we used simple counting thresholds for successive positive classified writings.

5.2 Model Training and hyperparameters

We processed the training and testing streams of user writings by moving window concatenation of size (N). In other words, to give a decision about the

Table 2: Summary of the proposed model variants

Run ID	Model Name	2^{nd} Learning Phase	
		T1	T2
0	LIRMMA	MLP	—
1	LIRMMB	RF	—
2	LIRMMC	—	MLP
3	LIRMMD	—	RF
4	LIRMME	BVI	BVI

current writing at time (t), we process all user writing starting from $(t-N+1)$. This gives more information about the context of a writing and reduce the effect of noisy and irrelevant ones. Experiments show that $(N = 5)$ to be a reasonable choice for the window size.

For DMEM, we use the same set of hyperparameter of AWD-LSTM proposed by [15] replacing the LSTM with Bi-LSTM and keep the same embedding size of 400 and 1150 hidden activations. We used weighted dropout of 0.2 and 0.25 as the input embedding dropout and the learning rate is 0.004. We fine-tuned the LM by eithr anorexia or depression training datasets provided. We train the LM for 14 epochs using batch size of 128 and limit the number of vocabulary to all token that appear more than twice. For classifier, we used masked self-attention layers and concatenation of maximum and average pooling. For the linear block, we used hidden linear layer of size 100 and apply dropout of 0.4. We used Adam optimizer [4] with $\beta_1 = 0.8$ and $\beta_2 = 0.99$. The base learning rate is 0.01. We used the same batch size used in training LMs. For training the classifier, we create each batch using weight random sampling to handle the problem of imbalance in the datasets. We train the classifier on training set for 30 epochs and select the best model on validation set to get the final model. For T2 training, we combine the training datasets for depression and anorexia of eRisk-2018.

In the second learning phase, the used architecture of the MLP had two hidden layers with ten neurons each. Concerning the RF classifier, ten estimators were used. These models are used to classify time series of (N) points. For MLP, RF and BVI models in T1, positive users were reported for those with classification probability higher than 0.8. This value increases to 0.9 in T2. We set both thresholds to 0.6 in the last rounds. For some model variants (LIRMMC and LIRMMD in T1 and LIRMMA and LIRMMB in T2), we apply counting of successive positive writings and give a decision after either 5 or 10 following writings respectively.

6 Results & Discussions

In eRisk-2019 two different types are used for model evaluation. The first one is decision-based evaluations; where the classical classification measures - precision (P), Recall (R) and (F1) - are computed for positive (at-risk) user. In addition to these and due to the drawbacks of *ERDE* measure, a new latency weighted F1 measure is introduced [13]. The other complementary evaluation is ranking-based evaluation. Beside the fired decision, scores are computed and used to build a ranking of users in decreasing estimation of risk. We participated only for decision-based evaluation. Tables 3 and 4 show the evaluation results of all our proposed variants for both tasks. It is clear that using MLP for the second learning phase is the best choice for both tasks. However, the usage of high threshold in T2 make the models predict most of the positive user in late writings. Also, applying BVI gets more comparable results than the runs with simple counting of positive writings. But it needs more precise choice of threshold for early detection in both tasks.

Table 3: Results of proposed runs for anorexia task (T1)

	P	R	F1	latency-weighted F1
LIRMMA	0.74	0.63	**0.68**	0.**63**
LIRMMB	**0.77**	0.60	0.68	0.62
LIRMMC	0.66	0.70	0.68	0.60
LIRMMD	0.74	0.42	0.54	0.48
LIRMME	0.57	**0.75**	0.65	—

Table 4: Results of proposed runs for self-harm task (T2)

	P	R	F1	latency-weighted F1
LIRMMA	**0.57**	0.29	0.39	**0.35**
LIRMMB	0.53	0.22	0.31	0.29
LIRMMC	0.48	**0.49**	**0.48**	—
LIRMMD	0.47	0.44	0.46	—
LIRMME	0.52	0.41	0.46	—

Tables 5 and 6 show some statistics of other participants runs compared to our proposed models. The ranks of the best run for each evaluation metric are also included. The statistics of the anorexia task are for 54 runs of 13 teams. The self-harm task statistics on results are for 33 runs of 8 teams. However the proposed architecture does not include any hand-crafted features, it seems to be comparable with other contributions for both tasks. Also, combining anorexia and past eRisk depression training datasets for detecting signs of self-harm is very competitive.

Table 5: Statistics on 54 participating runs results and our ranks for T1

	P	R	F1	latency-weighted F1
Max	0.77	0.99	0.71	0.69
Min	0.11	0.15	0.20	0.19
Average	0.45	0.63	0.48	0.46
Standard Deviation	0.17	0.24	0.17	0.15
Rank	1	14	5	5

Table 6: Statistics on 33 participating runs results and our ranks for T2

	P	R	F1	latency-weighted F1
Max	0.71	1.00	0.52	0.52
Min	0.12	0.22	0.22	0.17
Average	0.29	0.73	0.32	0.29
Standard Deviation	0.18	0.29	0.11	0.10
Rank	3	17	3	4

7 Conclusions

In this paper we present the participation of LIRMM in the CLEF eRisk-2019 T1 and T2 tasks. Both tasks are for early detection of signs of anorexia and self-harm from users posts on Reddit respectively. We proposed five runs for each task and the results are interesting and comparable to other contributions. The proposed framework architecture used the text without any handcrafted features. It performs the classification through two phases of supervised learning using state-of-the-art deep language modeling neural network. The first learning phase builds a time series representing the mood variation using attention-based modification of the ULMFiT model. The second learning phase is another classification model that learns patterns from these time series to detect early signs of such mental disorders. In this phase, We tried set of machine learning (MLP and RF) and statistical (BVI) models.

Combining anorexia and previous eRisk depression datasets to detect early signs of self-harm (T2) is interesting and shows the correlation of such mental disorders. However, the proposed models need tuning of second learning phase classification thresholds for earlier risk detection.

Acknowledgments

We would like to acknowledge La Région Occitanie and l'Agglomération Béziers Méditerranée which finance the thesis of Waleed Ragheb as well as INSERM and CNRS for their financial support of CONTROV project.

References

1. The national eating disorders association (NEDA).: Envisioning a world without eating disorders. In: The newsletter of the National Eating Disorders Association. Issue 22 (2009)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: International Conference on Learning Representations (ICLR). vol. abs/1409.0473 (Sep 2014)
3. Doyle, L., Treacy, M.M.P., Sheridan, A.J.: Self-harm in young people: Prevalence, associated factors, and help-seeking in school-going adolescents. International journal of mental health nursing **24 6**, 485–94 (2015)
4. Dozat, T., Manning, C.D.: Deep biaffine attention for neural dependency parsing. vol. abs/1611.01734 (2017)
5. Hawton, K., Zahl, D., Weatherall, R.: Suicide following deliberate self-harm: long-term follow-up of patients who presented to a general hospital. British Journal of Psychiatry **182**(6), 537542 (2003)
6. Hoek, H.: Review of the worldwide epidemiology of eating disorders. In: Current Opinion in Psychiatry. vol. 29 (2016)
7. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 328–339 (2018)
8. Joyce, D., L. Sulkowski, M.: The diagnostic and statistical manual of mental disorders: Fifth edition (dsm-5) model of impairment. In: Assessing Impairment: From Theory to Practice. pp. 167–189 (2016)
9. Klonsky, E.D.: The functions of deliberate self-injury: A review of the evidence. Clinical psychology review **27**, 226–39 (04 2007). https://doi.org/10.1016/j.cpr.2006.08.002
10. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: ICML. JMLR Workshop and Conference Proceedings, vol. 32, pp. 1188–1196. JMLR.org (2014)
11. Losada, D.E., Crestani, F., Parapar, J.: erisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations. In: 8th International Conference of the CLEF Association. pp. 346–360. Springer Verlag (2017)
12. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk – Early Risk Prediction on the Internet. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). Avignon, France (2018)
13. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2019: Early Risk Prediction on the Internet. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Association, CLEF 2019. Springer International Publishing, Lugano, Switzerland (2019)
14. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. pp. 142–150. HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
15. Merity, S., Keskar, N.S., Socher, R.: Regularizing and optimizing LSTM language models. In: International Conference on Learning Representations (2018)
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26. pp. 3111–3119. Curran Associates, Inc. (2013)

17. Mikolov, T., Yih, S.W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013). Association for Computational Linguistics (2013)
18. Moulahi, B., Azé, J., Bringay, S.: Dare to care: A context-aware framework to track suicidal ideation on social media. In: Bouguettaya A. et al. (eds) Web Information Systems Engineering - WISE 2017.,Lecture Notes in Computer Science,. Springer, Cham. vol. 10570 (2017)
19. Paul, M.J., Dredze, M.: You are what you tweet: Analyzing twitter for public health. In: Adamic, L.A., Baeza-Yates, R.A., Counts, S. (eds.) ICWSM. The AAAI Press (2011)
20. Ragheb, W., Moulahi, B., Azé, J., Bringay, S., Servajean, M.: Temporal mood variation: at the CLEF erisk-2018 tasks for early risk detection on the internet. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018. (2018)
21. Simpson, E., Roberts, S., Psorakis, I., Smith, A.: Dynamic bayesian combination of multiple imperfect classifiers. In: Decision Making and Imperfection. pp. 1–35. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
22. Trotzek, M., Koitka, S., Friedrich, C.: Linguistic metadata augmented classifiers at the clef 2017 task for early detection of depression. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum. vol. CEUR-WS 1866 (2017)
23. Wang, H., Keskar, N.S., Xiong, C., Socher, R.: Identifying generalization properties in neural networks. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=BJxOHs0cKm