NLP-UNED at eRisk 2021: self-harm early risk detection with TF-IDF and linguistic features

Elena Campillo-Ageitos[1], Hermenegildo Fabregat[1], Lourdes Araujo[1,2] and Juan Martinez-Romo[1,2]

[1]NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED), Juan del Rosal 16, Madrid 28040, Spain
[2]IMIENS: Instituto Mixto de Investigación, Escuela Nacional de Sanidad, Monforte de Lemos 5, Madrid 28019, Spain

Abstract

Mental health problems such as depression and anxiety are conditions that can have very serious consequences. Self-harm is a lesser-known symptom mostly associated with young people that has been linked to depression. Research suggests that the way people write can reflect mental well-being and mental health risks, and social media provides a source of user-generated text to study. Early detection is crucial for mental health problems. In this context, the shared task eRisk was proposed. This paper describes the participation of the group NLP-UNED on the 2021 T2 subtask. Participants were asked to create systems to detect early signs of self-harm on users from Reddit. We propose a feature-driven classifier with features based on text data, TF-IDF terms, first-person pronoun use, sentiment analysis and self-harm terminology. The official task results show that a relatively simple model can achieve fast and competent results.

Keywords

early risk detection, self-harm detection, natural language processing, TF-IDF, sentiment analysis, CEUR-WS

1. Introduction

Mental health problems, such as depression, are conditions that affect more people every day. These conditions may go undetected for many years, causing the people who suffer them to not receive adequate medical assistance. Untreated mental health issues can lead to serious consequences, such as addictions or even suicide. Self-harm, also known as Non-Suicidal Self-Injury (NSSI from now on) is a lesser known type of mental health problem that affects primarily young people [1]. Self-harm refers to the act of causing bodily harm to oneself with no suicidal intent, such as cutting, burning, hair pulling, and it has been linked to underlying mental health problems such as depression and anxiety [2]. It is a maladaptive form of coping [3] that causes pain and distress to the person, and could lead to unintentional suicide. Given the severity of the symptoms and the risks, it is important to dedicate efforts to better detect mental health problems in the society so they can better receive the help they need.

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
ecampillo@lsi.uned.es (E. Campillo-Ageitos); gildo.fabregat@lsi.uned.es (H. Fabregat); lurdes@lsi.uned.es (L. Araujo); juaner@lsi.uned.es (J. Martinez-Romo)
0000-0003-0255-0834 (E. Campillo-Ageitos); 0000-0001-9820-2150 (H. Fabregat); 0000-0002-7657-4794 (L. Araujo); 0000-0002-6905-7051 (J. Martinez-Romo)

It has been proven that people who suffer from mental health problems show differences in the way they communicate with other people, and how they write [4, 5]. Natural Language Processing (NLP) can be used to analyze these writings and detect underlying mental health problems. Social media use has been on the rise in the past decades, and the sheer volume of information available in these platforms can be used for these purposes. Recent research has applied NLP techniques to develop systems that automatically detect users with potential mental health issues.

Early detection is key in the treatment of mental health problems, since a fast intervention improves the probabilities of a good prognosis. The longer a mental health problem goes undetected, the more likely serious consequences are to derive from it. Most of the efforts done in the literature focus on detection, but not on early detection. Early detection would allow a faster diagnostic, which would help mental health specialist to do a faster intervention.

In the light of this problem, the shared task eRisk was created. This task focuses on early detection of several mental health problems, such as depression, anorexia, self-harm and gambling on temporal data extracted from Reddit. The 2021 eRisk task [6] proposed three different subtasks: 1) Task 1: early detection of signs of pathological gambling; 2) Task 2: early detection of signs of self-harm; and 3) Task 3: measuring the severity of the signs of depression. Our team participated in Task 2: early detection of signs of self-harm. The dataset for this subtask is a collection of chronological written posts made by different users on Reddit. Each user is tagged as positive or negative, where positive users show signs of self-harm, and negative users do not. The objective of this task was to evaluate the writings sequentially and predict as fast as possible whether each user showed signs of self-harm.

The task was divided in two stages: (i) Training stage: during this phase, a training set was given to prepare and tune each participating system. The training data was composed of 2020's Task 1 (T1) training and testing data, and each user was labelled as either positive (self-harm) or negative (no self-harm). The data was divided into a train set and a validation set. (ii) Test stage: participants connected to a server to obtain the testing data and send the predictions. For each request to the server, one message for each user was obtained, and a prediction for each user had to be sent before being able to make a new request for new writings. Thus, participants had to create a system that interacted with the server and made predictions for every user, one writing at a time. After the test stage, each proposed system was evaluated based on precision, recall, F-measure, and new metrics developed for the sake of this competition that penalize late decisions: Early Detection Error (ERDE) and latency-weighted F-measure. More information on these metrics can be found at [7].

This paper presents our participation in the self-harm subtask T2. The paper is organized as follows: Section 2 shows a review of the related literature; Section 3 describes the task dataset; Section 4 details our proposed model for the task; Section 5 explains the experiment setup; Section 6 summarizes our results for the task; finally, Section 7 presents our conclusions and ideas for future research.

2. Related work

Social media has been previously studied in relation to health [8, 9]. Mental health, and depression in general, is a common focus on works attempting to detect individuals who suffer from that illness [10, 11, 12, 13]. Some work focuses on early prediction of mental illness symptoms [4, 14], but there are very few of them [15].

Studies performed on self-harm are also scarce. Most work has been done on studying the personalities and behavioral patterns of people who self-harm [16, 17], showing common patterns about high negative affectivity, and how it's a maladaptive coping strategy. Some effort has been done on studying self-harm behavior in social media in particular [18, 19, 20, 21], but they focus on studying posting patterns, behaviours, consequences, etc. Their findings show how people who self-harm have different posting patterns than mentally healthy users.

Some researchers focused on identifying self-harm content on social media [22, 23]. They show a mixture of NLP methods, both supervised and unsupervised, and using traditional and deep learning methods. Wang et al. [23] uses a mixture of CNN-generated features and those obtained from analyzing posting patterns: language has different structures, and more negative sentiment, they are more likely to have more interactions with other users but less online friends and posting hours are different, and self-harm content is usually done late at night.

Research done on predicting future self-harm behavior or finding at-risk individuals is rare. While some efforts have been done using methods such as using apps and data from wearable devices [24, 25], there is little research done on predicting this behavior on social media. The eRisk shared task first introduced the early risk detection on 2019 as a subtask, but no training data was given to develop the solutions. Most participants focused on developing their own training data instead of opting for unsupervised methods. The task was reintroduced on 2020 and training data was provided. A succesful system with promising results was team iLab [26]. The researchers proposed a classification system based on BERT and neural networks. In general, participants with deep learning models obtained the best results, but they were also slower and lacking explainability.

3. Dataset description

Our system was trained with the eRisk 2021 shared Task 2 dataset. This section presents this dataset.

The eRisk 2021 dataset was given to the participants by the organizers of the task. It is an early risk detection dataset first presented by Losada et al. at [27]. The 2021 dataset is described at [6]. The data was extracted from Reddit, and it presents a collection of users and messages. Each document corresponds to a different user, and in it, there is an arbitrary amount of messages or posts (submissions done to Reddit).

The dataset is a collection of two groups of Reddit users: those who have explicitly said in the platform that they engage in self-harm (positive users), and a control group (negative users). Those messages that were used to identify positive users were removed from the collection by the organizers. The data came in two groups: train set and test set. We decided to use this division for our train and validation division.

Table 1
Breakdown of eRisk 2021 dataset's number of positive and negative users in train and test set.

eRisk 2021 data	Train	Test	*Total*
Positive users	41	104	*145*
Negative users	299	319	*618*
All users	340	423	*763*

The texts written in these documents are not formal texts; they are written by people from all ages and areas on the social media site known as Reddit. Grammar and spelling are not always going to be correct. Emojis are going to be used, emoticons are going to be used; links, weird expressions, mayus, many vowels, too little vowels, etc. These carry meaning on themselves (for example, writing a text in all caps usually conveys that the person is shouting, excited, etc.). This can be a challenge for machine learning systems.

Table 1 shows a summary of the eRisk 2021 T2 dataset. There are 763 users in total, divided into 145 positive users and 618 negative users. This is a percentage of 19% positive users, which makes the data highly imbalanced. Unbalanced datasets can be a problem while training and evaluating a machine learning model. The data was divided into a training and testing group, with sizes 340 and 423 respectively.

4. Proposed model

We propose a machine learning approach that uses a combination of text-based and TF-IDF-based (term frequency–inverse document frequency) features to predict whether a message belongs to a positive or negative user. These features are fed to a SVM classifier. If the classifier flags a message as positive, the user is classified as positive.

There has been, to the best of our knowledge, no previous team that implemented a combination of text-based and TF-IDF-based features with a SVM classifier to detect early risk of self-harm. Team UDE [28] implemented in 2019 content-based features with a SVM classifier; team BioInfo@UAVR [29] applied Bag of Words (BoW) and TF-IDF based features to a Linear SVM classifier in one of their runs from eRisk 2020 Task 1.

The most challenging part of the eRisk task is the temporal complexity of the problem. The features are calculated taking this into account, and the model is also trained with that in mind.

The model can be divided in three distinct stages: 1) Data pre-processing; 2) feature calculation; and 2) message classification, where the supervised part of the model takes place and each user is categorized as positive (1) or negative (0).

4.1. Data pre-processing

Each user has an arbitrary number of messages, and they are ordered sequentially. During training, we are given all the messages at once. However, during the test stage of the task, we are given one message from each user at a time. First, messages were cleaned, tokenized, and stems and part-of-speech tags were calculated.

To obtain clean text, we separated contractions into their own words and removed hyperlinks, all punctuation characters, and decimal codes (words that started with # followed by numbers and ended with ;). Tokens, stems and part-of-speech tags were obtained from clean text by first removing stop words, then tokenizing, calculating part-of-speech tags and stemming. Uncleaned text was also preserved in order to calculate some text features (exclamation and question marks).

4.1.1. Sliding window

In our system, each new message is not observed in a vacuum. Their context, that is, their surrounding messages are also taken into account. Since future messages are (in the case of testing) unknown, only the previous messages can be used.

For this, we implemented a sliding window. For every new received message, our system combined its text with the previous w messages, where w is a configurable parameter, to form a window. The features were calculated on that window. Depending on the size of this parameter, a longer or shorter user history would be taken into account in each step e.g., a size of 1 only uses the current message, while a size of "all" would use the whole user history. This sliding window was applied both during the training and testing stage.

4.1.2. User subsets

One of our hypothesis while developing our model this year was that the first messages in the sequence from each user carry more information about the risk (or lack of risk) than the last ones. We tested this by training and validating the model with only the first m messages of each user, where m was a configurable size. After testing different values for m (10, 100, 200, 500 and all messages), we found that using only the first 100 messages gave us the best results.

Because of this, we used a subset of the data available, both during the training and test stage. During training and validation, the first 100 messages from each user were selected, and the rest were discarded. During testing, only the first 100 messages were processed to make a new decision.

4.2. Features

We implemented and combined two types of features: 1) text-based, and 2) TF-IDF-based.

4.2.1. Text features

We used different kinds of text-based features. All were normalized by the text length, and then discretized. The number of bins the features were discretized to was a parameter configurable to a d size.

- Grammar-based features: Table 2 shows the list of features.
- Special features: They were tailored to the self-harm dataset. The following section describes them in detail.

Table 2
Text features generated for the eRisk classification

Features		Description
char_count		Length of the title and content of the text, combined.
word_count		Number of words on the title and content, combined.
word density		Density of the words related to the length of the text in chars.
punctuation_count		Number of punctuation characters, such as ".", ",", etc.
upper_case_count		Number of characters in upper case.
questions_count		Number of question marks used.
exclamations_count		Number of exclamation marks used.
smilies		Number of smilies (":)") used.
sad_faces		Number of sad faces (":(") used.
pos_tags		Part-of-speech tags.
	noun_count	Number of nouns used.
	pron_count	Number of pronouns used.
	verb_count	Number of verbs used.
	adj_count	Number of adjectives used.
	adv_count	Number of adverbs used.

To choose the special features, previous work was done in analyzing the eRisk 2019 dataset. We explored the differences between positive and negative users regarding use of self-harm words, first-person pronouns, and sentiment score. It was observed that, in general, positive users did have significant differences from negative users. Our participation from the previous year [30] shows details of this analysis. From this, the following features were developed:

Pronouns There is evidence suggesting that people who use more first-person pronouns on average are more depressed than people who use the third person [31, 32]. There is also evidence linking depression to non-suicidal self-harm [2], so tracking this information would prove beneficial for our task. Furthermore, two sentences on the topic of self-harm are different depending on the pronouns that are used: "I cut myself today" versus "She is thinking about cutting herself". In the first case, the speaker shows clear signs of doing self-harm. In the second case, however, the speaker is seeking advice about a person they know, but they show no evidence about themselves. We can track this difference by counting first-person pronouns.

Sentiment analysis As mentioned previously, it has been observed that people who do self-harm show more negative emotions [2]. Tracking sentiment to detect the users' moods makes sense in this context. We focused only on positive or negative sentiment. This feature shows the sentiment of the window as a numeric score normalized by the length of the texts. A negative score demonstrates a negative sentiment, while a positive score demonstrates a positive emotion.

NSSI words Some self-harmers talk about their experiences online. There is a sub-reddit (a Reddit community) dedicated to self-harm, where users talk about their disorder and support each other. We suppose that some of the users in our dataset use the platform to talk about

self-harm. If this is the case, it proves useful to track the usage of the most common words related to self-harm. We used a word dictionary of terms related to non-suicidal self injury (NSSI words from now onwards) from [2]. This dictionary is divided in several categories: 1) Methods of NSSI; 2) NSSI terms; 3) Instruments used; 4) Reasons for NSSI. In this feature, we tracked the number of words from each category in a text, normalized by the length of that text. Each NSSI category became an independent feature.

Text-based and special features were combined with TF-IDF-based features using Scipy Hparse, by being appended to the end of the TF-IDF features.

4.2.2. TF-IDF features

A TF-IDF featurizer was trained on the positive users of the train data. This featurizer was then used to obtain TF-IDF features for each window. We used 5000 maximum features for this featurizer. We obtain with this the TF-IDF-based features for each message window. This is then passed on to the classifier. We experimented with single word and n-gram based features, but word-based features worked best.

4.3. Message classification module

The features calculated from the window messages are fed to the SVM classifier. This classifier predicts whether these features belong to a message generated by a positive or negative user. For every new message we receive, we have to classify each user as "positive" or "negative". A positive decision is final, but a negative one may be revised later. Besides, the task rewards quick decisions, so the earlier we make a positive decision, the better.

One message classified as positive should not be enough to classify the user as positive, so we follow a decision policy of consecutive alerts. Every time our classifier marks a message as positive, we consider this as an alert, but the user is still classified as negative. If a number a of consecutive alerts is reached, the system classifies the user as positive. The parameter a is configurable. During model development we tested different values for a, but 1 and 2 consecutive alerts gave us the best results.

4.4. Training the model

The system was developed with the data available during the training stage of the 2021 task. This dataset was comprised of the 2020 task train and test data. The 2020 train data was used to train the model, and the 2020 test data was used for validation. As mentioned before, a subset of only the first 100 messages from every user was used.

Class weights The training data is highly unbalanced, as can be seen in section 3. To train the SVM classifier correctly we implemented weights to our data. It is critical to detect positive users as quickly as possible, so the first messages from a positive user are arguably more important than the last message. It is not so important to detect negative users at any point in time, so all messages carry the same weight. With this hypothesis in mind, we devised a training strategy that would give the same weight (1) to all negative user messages, but would give a descending

weight to positive user messages, from 2 to 1, through a fixed range. Equation 1 shows how the weight for a message from user u with golden truth g_{truth} in sequence i was calculated.

$$w_{m_{u,i}} = \begin{cases} 2 - i \cdot (\frac{1}{max}) & \text{if } g_{truth,u} = 1 \\ & \text{and } 0 \leq i < max \\ 1 & \text{otherwise} \end{cases} \quad (1)$$

Where:

$w_{m_{u,i}}$: is the weight of the message m in position i for user u.
max: is the sequence number of the last message that will have a weight greater than 1 for positive users.
$g_{truth,u}$: is the golden truth of the user u.

During the validation phase, several values were tested for the *max* value. Setting this value as a fixed number and not as the last message of each user guaranteed that every positive user message with the same sequence number had the same weight. For the final training of the model, this value was set as 100.

4.5. System applied to the 2021 task

This section explains the particulars of the model during the test phase.

4.5.1. Scores

Participating teams were required to send, for each iteration, scores that represented the estimated risk of each user. We prioritized simplicity, so our system only gave two kinds of scores: 1 if the user was classified as positive, and 0 if the user was classified as negative.

4.5.2. Model parameters

Table 3 shows a list of the configurable parameters that were available for our model. Some of them were set as a fixed value for all runs, while the rest were configured with different values depending on the run.

Sliding window New messages arriving from the test server were converted into windows following the same strategy and window size as during training. Different window sizes were tested during model development and, after careful evaluation, a size of 10 was selected for the final model. The strategy we followed for the first 10 messages was to form the window with only the messages we had received.

User subsets As previously mentioned, only the first 100 messages were observed and sent to the classifier. For the following messages, the system always responded with the last decision (100th) for each user.

Table 3
Model configurable parameters. The table shows the values tested for each parameter during the train stage, and the final value used to train the final model, and during the test stage of the task.

Parameter	Description	Tested values	Final value
Sliding window	The size w of the sliding window in the train and test stages.	1, 3, 5, 10, 20	10
User subsets	The number of messages from each user the system was trained on.	10, 20, 100, 500, all	100
Feature combinations	The features calculated during the train and test stages.		run-dependent
Discretizer bins	The size d of the feature discretizer bins.	10, 20, 50, 100, 200, 500	run-dependent
Consecutive alerts	The number of consecutive alerts a it takes to classify a user as positive.	1, 2, 3, 5, 10	1
Class weights max	The last message for positive users to have a weight greater than 1.	10, 100, 200, all	100
Observed messages	The number of messages to observe during validation and the test stage.	10, 20, 100, 500, all	100

5. Experimental Setup

This section presents the experiments conducted for the official eRisk 2021 task using the model proposed in Section 4.

5.1. Model implementation

The SVM classifier was implemented using a combination of NLTK [1] and Scikit-learn [2]. More specifically, Scikit-learn's SVC implementation of C-Support Vector Classification model was used. Parameters apart from weights were used as default.

NLTK was used for data cleanup and text pre-processing (tokenizing and stemming). Sentiment analysis was also performed with NLTK's Sentiment Intensity Analyzer.

5.2. Submitted runs

Our team participated with five different runs. We selected a different subsect of text features and discretizer sizes for each of these runs (TF-IDF features were used for all runs). Table 4 shows the subset of text features that were calculated for each run. The size of the discretized bins for each run are shown in Table 5.

Each of the runs used a different subset of text and special features (all used TF-IDF features). In order to check the influence of the discretizer, *run 0* and *run 2* both used all text features, but with different discretizer bin sizes. We were not sure if sentiment analysis was useful for this task, so *run 1* and *run 4* used only special features, with the difference that *run 4* used sentiment

[1] https://www.nltk.org/
[2] https://scikit-learn.org/

Table 4
Feature combinations used to train the models and evaluate the results in the different runs.

Features used in each run		#run				
		0	1	2	3	4
Text features						
char_count		x		x	x	
word_count		x		x	x	
word_density		x		x	x	
punctuation_count		x		x	x	
upper_case_count		x		x	x	
questions_count		x		x	x	
exclamations_count		x		x	x	
smilies		x		x	x	
sad_faces		x		x	x	
pos_tag	*noun_count*	x		x		
	pron_count	x		x		
	verb_count	x		x		
	adj_count	x		x		
	adv_count	x		x		
Special features						
Sentiment		x		x	x	x
First-person pronouns		x	x	x	x	x
NSSI words	*methods*	x	x	x	x	x
	nssi_terms	x	x	x	x	x
	instruments	x	x	x	x	x
	reasons	x	x	x	x	x

Table 5
Bin sizes for the features discretizer for each run.

Run id	Discretizer bin sizes
0	50
1	100
2	100
3	50
4	50

analysis, while *run 1* did not. *Run 0* and *run 2*, compared to *run 1* and *run 4*, also allowed us to compare performance with and without non-special text features. Finally, we wanted to test whether part-of-speech tags were useful for the model, so *run 3* used all features except part-of-speech tags.

Table 6
Number of runs, number of user writings processed, and lapse of time taken for the whole process. We show a comparison of our team (NLP-UNED) with other teams with fast or good results. Our team obtained the fastest time.

team	#runs	#user writings processed	lapse of time (from 1st to last response)
NLP-UNED	5	472	07:08:37
Birmingham	5	11	2 days 08:01:32
NuFAST	3	6	17:07:57
UPV-Symanto	5	538	11:56:33
UNSL	5	1999	3 days 17:36:10

6. Results and Discussion

This section shows the official results for the task. The overview for the official results of all teams can be found at [6].

During the evaluation stage of the task, teams had to iteratively request data from a server and send their predictions, one message per user at a time. The model described in Section 4 was developed, trained, and applied to a program that automatically connected to this server and performed the model calculations. The program was launched and run without errors. Instead of allowing it to run until there were no messages left, we decided to stop it early at around 450 iterations (exactly 472 iterations). This was done because we wanted to prioritize speed, and our system only made new decisions for the first 100 messages.

Tables 6, 7 and 8 show the official results for our team received by the task organizers. Results from other teams were added for comparison purposes.

Table 6 shows the time span and number of messages processed. We include information from the fastest and slowest teams, plus the ones that achieved the best results in the official metrics. Our team did not process all messages like *UNSL*, but processed 472 and obtained the best results in terms of speed (7 hours and 8 minutes). However, the server experimented some issues during the evaluation stage, and some teams might have been affected by this.

Table 7 shows the official evaluation metrics for the binary decision task. We added the team runs that achieved the best results for each metric.

Participating teams were also required to send scores estimating the risk of each user. Table 8 shows the official result for our team, compared to the teams with the best results. Standard IR metrics were calculated by the organizers after processing 1 message, 100 messages, 500 messages and 1000 messages. Our team only processed 472 messages, so the 500 and 1000 messages metrics are not given.

This system is a great improvement over our participation from the previous year [30]. All five runs achieved moderately good results, obtaining a latency-weighted F-measure above 0.5. *Run 4* placed our team in the third place in terms of this measure, achieving a 0.564. The precision and recall measures show that a good proportion of users were classified as positive and negative, unlike last year, when most users were classified as negative.

In the case of the text features, quality appears to be more important than quantity. *Run 4*, which used only the special features (plus TF-IDF) achieved the best results in term of latency-

Table 7
eRisk 2021 T2 decision-based evaluation. Our teams' results (NLP-UNED) are compared to the best results for each metric. Our best results for each metric and the overall best results for the rest of the teams are bolded.

team name	run id	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{tp}$	$speed$	$latency\text{-}weighted\ F1$
NLP-UNED	0	.442	.75	.556	**.08**	.042	**6**	**.981**	.545
NLP-UNED	1	.442	.796	.568	.091	.041	11	.961	.546
NLP-UNED	2	.422	.73	.535	.088	.047	7	.977	.522
NLP-UNED	3	.419	.77	.543	.093	.047	10	.965	.524
NLP-UNED	4	**.453**	**.816**	**.582**	.088	**.04**	9	.969	**.564**
Birmingham	2	**.757**	.349	.477	.085	.07	4	.988	.472
UPV-Symanto	1	.276	.638	.385	**.059**	.056	1	1.0	.385
UNSL	0	.336	.914	.491	.125	**.034**	11	.961	.472
UNSL	4	.532	.763	.627	.064	.038	3	.992	**.622**

Table 8
Ranking-based evaluation. Our team's results (NLP-UNED) are compared to the best team's results. Our best results and best results overall are bolded.

		1 writing			100 writings		
team	run	*P@10*	*NDCG@10*	*NDCG@100*	*P@10*	*NDCG@10*	*NDCG@100*
NLP-UNED	0	0.8	**0.82**	**0.47**	0.8	0.74	0.47
NLP-UNED	1	0.7	0.68	0.39	0.8	0.86	0.55
NLP-UNED	2	**0.9**	0.81	0.39	0.6	0.44	0.44
NLP-UNED	3	0.6	0.6	0.37	0.6	0.58	0.47
NLP-UNED	4	0.5	0.47	0.32	**0.9**	**0.94**	**0.55**
UNSL	0	**1**	**1**	0.7	0.7	0.74	**0.82**
UNSL	4	**1**	**1**	0.63	**0.9**	0.81	0.76

Table 8
Ranking-based evaluation. Our team's results (NLP-UNED) are compared to the best team's results. Our best results and best results overall are bolded. *Continuation*

		500 writing			1000 writings		
team	run	*P@10*	*NDCG@10*	*NDCG@100*	*P@10*	*NDCG@10*	*NDCG@100*
NLP-UNED	0	0.0	0.0	0.0	0.0	0.0	0.0
NLP-UNED	1	0.0	0.0	0.0	0.0	0.0	0.0
NLP-UNED	2	0.0	0.0	0.0	0.0	0.0	0.0
NLP-UNED	3	0.0	0.0	0.0	0.0	0.0	0.0
NLP-UNED	4	0.0	0.0	0.0	0.0	0.0	0.0
UNSL	0	**0.8**	**0.81**	**0.8**	0.8	0.81	**0.8**
UNSL	4	**0.9**	**0.81**	0.71	**0.8**	0.73	0.69

weighted F-measure out of the five runs. *Run 1*, which also only used the special features, was the second best run. This run used discretized bins of size 100, while *Run 4* used size 50, which could mean that fewer bins leads to better results. Nonetheless, runs that used all features still obtained good results. The fastest run was *run 0*.

Our team achieved the third position out of all the participant teams in terms of latency-weighted F-measure. It was also the fastest system, taking only 7 hours and 8 minutes to process 472 messages. (Again, we should mention that the test server experienced some problems during the test phase, which could have slowed down some of the participant teams).

The strength of our system lies in its simplicity. While most systems that participate in this task use complex models with neural networks that can take hours if not days to train and evaluate new data, our model can analyze a big amount of information in a small amount of time. This has several reasons. First, the model itself does not require a great amount of computational power. Second, the features are calculated with a sliding window, so past messages are not processed more than x times, x being the size of the window. The model can scale well and stay running indefinitely without running out of memory.

Using text features makes our proposal explainable. Explainability is very important in the field of medicine and mental health: a good alert system will inform why a subject is at risk, so a medical practitioner can make the final decision.

Classification problems in the medical field are also more delicate in terms of recall than precision. False negatives are much more severe than false positives. False negatives mean that at-risk patients are going undetected, while false positives only add extra resources to monitor them. Our system obtains higher scores in recall than in precision, as can be seen in Table 7. *Run 4* obtained the best score, 0.816.

Overall, these evaluations show that even a simple, feature-driven approach can be applied to what looks like a very complex problem and obtain competent results.

7. Conclusions and Future Work

In this paper we presented the NLP-UNED participation on the eRisk 2021 T2 task. We developed a classification system based on TF-IDF, text-based and especially tailored features: first-person pronouns, sentiment analysis and NSSI terms. The official task results show that our system managed to obtain fast, competent results. More work is needed to obtain state-of-art results.

We would like to keep exploring terms and tailored features that can better detect at-risk subjects. We are interested in applying these features to a deep learning system. Finally, there are evidences that suggest that self-harm subjects have different posting patterns than non self-harmers, so we are interested in exploring the temporal differences in the dataset and creating more features.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the DOTT-HEALTH Project (MCI/AEI/FEDER, UE) under Grant PID2019-106942RB-C32, as well as project EXTRAE II (IMIENS 2019) and the research network AEI RED2018-102312-T (IA-Biomed).

References

[1] S. Gluck, Self-Injurers and Their Common Personality Traits | HealthyPlace, accessed 2021-05-27. URL: https://www.healthyplace.com/abuse/self-injury/self-injurers-and-their-common-personality-traits.

[2] M. M. Greaves, A Corpus Linguistic Analysis of Public Reddit and Tumblr Blog Posts on Non-Suicidal Self-Injury, Ph.D. thesis, Oregon State University, 2018. URL: https://ir.library.oregonstate.edu/concern/graduate_thesis_or_dissertations/mp48sk29z.

[3] S. Lewis, D. Santor, Self-harm reasons, goal achievement, and prediction of future self-harm intent, The Journal of nervous and mental disease 198 (2010) 362–9. URL: http://journals.lww.com/00005053-201005000-00008.

[4] M. De Choudhury, S. Counts, E. Horvitz, Social Media as a Measurement Tool of Depression in Populations, in: Proceedings of the 5th Annual ACM Web Science Conference, WebSci '13, Association for Computing Machinery, New York, NY, USA, 2013, pp. 47–56.

[5] J. W. Pennebaker, M. R. Mehl, K. G. Niederhoffer, Psychological Aspects of Natural Language Use: Our Words, Our Selves, Annual Review of Psychology 54 (2003) 547–577.

[6] J. Parapar, M.-R. P., D. E. Losada, F. Crestani, Overview of eRisk 2021: Early Risk Prediction on the Internet, Proceedings of the Twelfth International Conference of the Cross-Language Evaluation Forum for European Language (2021).

[7] D. E. Losada, F. Crestani, J. Parapar, Early detection of risks on the internet: an exploratory campaign, in: 41st European Conference on Information Retrieval, Springer, 2019, pp. 259–266. URL: http://dx.doi.org/10.1007/978-3-030-15719-7_35.

[8] M. J. Paul, M. Dredze, You Are What You Tweet: Analyzing Twitter for Public Health., International AAAI Conference on Weblogs and Social Media (ICWSM) (2011).

[9] M. Park, D. W. McDonald, M. Cha, Perception differences between the depressed and non-depressed users in Twitter, Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013 (2013) 476–485.

[10] M. Conway, D. O'Connor, Social media, big data, and mental health: Current advances and ethical implications., Current Opinion in Psychology 9 (2016) 77 – 82. URL: http://www.sciencedirect.com/science/article/pii/S2352250X16000063.

[11] C. Karmen, R. C. Hsiung, T. Wetter, Screening internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods, Computer Methods and Programs in Biomedicine 120 (2015) 27–36. URL: http://dx.doi.org/10.1016/j.cmpb.2015.03.008.

[12] B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, H. Christensen, Detecting suicidality on twitter, Internet Interventions 2 (2015) 183–188. URL: http://dx.doi.org/10.1016/j.invent.2015.03.005.

[13] W. Yang, L. Mu, GIS analysis of depression among Twitter users, Applied Geography 60 (2015) 217–223. URL: http://dx.doi.org/10.1016/j.apgeog.2014.10.016.

[14] M. Nadeem, Identifying depression on twitter, 2016. URL: http://arxiv.org/abs/1607.07384. arXiv:1607.07384.

[15] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, J. C. Eichstaedt, Detecting depression and mental illness on social media: an integrative review, Current Opinion in Behavioral Sciences 18 (2017) 43–49. URL: http://dx.doi.org/10.1016/j.cobeha.2017.07.005.

[16] I. Baetens, L. Claes, J. Muehlenkamp, H. Grietens, P. Onghena, Non-Suicidal and Suicidal Self-Injurious Behavior among Flemish Adolescents: A Web-Survey, Archives of Suicide Research 15 (2011) 56–67. URL: https://doi.org/10.1080/13811118.2011.540467.

[17] K. A. Nicolai, M. D. Wielgus, A. Mezulis, Identifying Risk for Self-Harm: Rumination and Negative Affectivity in the Prospective Prediction of Nonsuicidal Self-Injury, Suicide and Life-Threatening Behavior 46 (2016) 223–233.

[18] P. A. Cavazos-Rehg, M. J. Krauss, S. J. Sowles, S. Connolly, C. Rosas, M. Bharadwaj, R. Grucza, L. J. Bierut, An analysis of depression, self-harm, and suicidal ideation content on Tumblr, Crisis 38 (2017) 44–52. URL: https://psycnet.apa.org/record/2016-36501-001.

[19] C. Emma Hilton, Unveiling self-harm behaviour: what can social media site twitter tell us about self-harm? a qualitative exploration, Journal of Clinical Nursing 26 (2017) 1690–1704. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jocn.13575. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jocn.13575.

[20] M. A. Moreno, A. Ton, E. Selkie, Y. Evans, Secret Society 123: Understanding the Language of Self-Harm on Instagram, Journal of Adolescent Health 58 (2016) 78–84.

[21] J. W. Patchin, S. Hinduja, Digital Self-Harm Among Adolescents, Journal of Adolescent Health 61 (2017) 761–766.

[22] A. Yates, A. Cohan, N. Goharian, Depression and self-harm risk assessment in online forums, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2017, pp. 2968–2978. URL: https://www.aclweb.org/anthology/D17-1322.

[23] Y. Wang, J. Tang, J. Li, B. Li, Y. Wan, C. Mellina, N. O'Hare, Y. Chang, Understanding and discovering deliberate self-harm content in social media, 26th International World Wide Web Conference, WWW 2017 (2017) 93–102.

[24] N. Lederer, T. Grechenig, R. Baranyi, uncut: bridging the gap from paper diary cards towards mobile electronic monitoring solutions in borderline and self-injury, in: 3nd IEEE International Conference on Serious Games and Applications for Health, SeGAH 2014, Rio de Janeiro, Brazil, May 14-16, 2014, IEEE Computer Society, 2014, pp. 1–8. URL: https://doi.org/10.1109/SeGAH.2014.7067092. doi:10.1109/SeGAH.2014.7067092.

[25] L. Malott, P. Bharti, N. Hilbert, G. Gopalakrishna, S. Chellappan, Detecting self-harming activities with wearable devices, in: 2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), 2015, pp. 597–602.

[26] R. Martínez-Castaño, A. Htait, L. Azzopardi, Y. Moshfeghi, Early risk detection of self-harm and depression severity using bert-based transformers, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/paper_50.pdf.

[27] D. E. Losada, F. Crestani, A Test Collection for Research on Depression and Language Use CLEF 2016, Évora (Portugal), Experimental IR Meets Multilinguality, Multimodality, and Interaction (2016) 28–29. URL: https://tec.citius.usc.es/ir/pdf/evora.pdf.

[28] R. Masood, F. Ramiandrisoa, A. Aker, UDE at Erisk 2019: Early risk prediction on the internet, CEUR Workshop Proceedings 2380 (2019) 9–12.

[29] A. Trifan, P. Salgado, J. L. Oliveira, BioInfo@UAVR at eRisk 2020: on the use of psycholin-

guistics features and machine learning for the classification and quantification of mental diseases (2020) 22–25. URL: http://early.irlab.org/.

[30] E. Campillo-Ageitos, J. Martinez-Romo, L. Araujo, NLP-UNED at eRisk 2020: Self-harm Early Risk Detection with Sentiment Analysis and Linguistic Features, Working Notes of {CLEF} 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020 2696 (2020) 22–25. URL: http://ceur-ws.org/Vol-2696/paper{_}41.pdf.

[31] J. W. Pennebaker, The Secret Life of Pronouns What Our Words Say About Us, Bloomsbury Press, 2011.

[32] T. Edwards, N. S. Holtzman, A meta-analysis of correlations between depression and first person singular pronoun use, Journal of Research in Personality 68 (2017) 63–68. URL: http://dx.doi.org/10.1016/j.jrp.2017.02.005.