NLP-UNED at eRisk 2020: self-harm early risk detection with sentiment analysis and linguistic features

Elena Campillo Ageitos[1][0000−0003−0255−0834], Juan Martinez-Romo[1,2][0000−0002−6905−7051], and Lourdes Araujo[1,2][0000−0002−7657−4794]

[1]NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia (UNED)
[2]Instituto Mixto de Investigación - Escuela Nacional de Sanidad (IMIENS)
ecampillo@lsi.uned.es, juaner@lsi.uned.es, lurdes@lsi.uned.es

Abstract. Mental health problems such as depression are conditions that, going undetected, can have serious consequences. A less-known mental health problem that has been linked to depression is self-harm. There is evidence suggesting that people's writings can reflect these problems, and research has been done to detect these individuals through their content on social media. Early detection is crucial for mental health problems, and for this purpose a shared task named eRisk was proposed. This paper describes NLP-UNED's participation on the 2020 T1 subtask. Participants were asked to create systems that detected early self-harm signs on Reddit users. Our team shows a data analysis of the 2019 T2 subtask and proposes a simple feature-driven classifier with features based on first-person pronoun use, sentiment analysis and self-harm terminology.

Keywords: Early Risk Detection · Self-Harm detection · Sentiment Analysis · Natural Language Processing

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 [7]. Self-harms refer to the act of causing bodily harm to oneself with no suicidal intent, such as cutting, burning, hair pulling, and they have been linked

to underlying mental health problems such as depression and anxiety [8]. Is is a maladaptive form of coping [12] that causes pain and distress to the self-harmer, and could lead to unintentional suicide. It is important to dedicate efforts to better detect mental health problems in the society so they can better receive the help they need.

It has been proven that people who suffer from mental health problems show differences in the way they communicate with other people, and the way they write [4] [24]. Natural Language Processing can be used to analyze these people's 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 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, and self-harm on temporal data extracted from Reddit. The 2020 eRisk task [14] proposed two different subtasks: Task 1 focused on early detection of signs of self-harm, while Task 2 focused on measuring the severity of the signs of depression. Our team participated in Task 1: detecting 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 give a prediction of whether a user showed signs of self-harm or not as fast as possible.

The task was divided in two stages: (i) training stage: during this phase, a training set was given to prepare and tune each team's systems. The training data was composed of 2019's task 2 (T2) training and testing data, and each user was labelled as either positive (self-harm) or negative (no self-harm). (ii) test stage: participants connected to a server to obtain the testing data and send the predictions. For each request to the server, an array of users with one writing each 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. The objective of the task was to detect positive users as early as possible. After the test stage, each team's participation was evaluated based on precision, recall, F1, and new metrics developed for the sake of this competition that penalize late decisions: Early Detection Error (ERDE) and latency-weighted F1. More information on these metrics can be found at [13].

This paper presents our participation in the self-harm subtask. We present an exploratory analysis of the 2019 T2 dataset. The rest of the paper is organized as follows: Section 2 shows a review of the related literature; section 3 details our proposed model for the task; section 4 presents an exploratory analysis of the 2019 T2 dataset we performed previous to developing the system; section 6 summarizes our official results for the task, plus some corrections; finally, section 7 presents our conclusions and ideas for future research.

2 Related Work

Social media has been previously studied in relation to health [22] [20]. Mental health, and depression in general, is a common focus on works attempting to detect individuals who suffer from that illness [3] [10] [19] [26]. Some work focuses on early prediction of mental illness symptoms [4] [17], but there are very few of them [9].

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 [1] [18], 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 [2] [6] [16] [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 [27] [25]. They show a mixture of NLP methods, both supervised and unsupervised, and using traditional and deep learning methods. Wang et al. [25] uses a mixture of CNN-generated features and features obtained from their findings on 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 [11] [15], 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.

3 Proposed model

We propose a machine learning approach that uses text features to predict whether a message belongs to a positive or negative user. These features are fed to a SVM classifier. A decision module takes the classified messages and decides whether an user is positive or negative.

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

The model can be divided in three distinct stages: 1) Pre-processing and feature calculation; 2) Message classification, where the supervised part of the model takes place; and 3) User decision, where each user is categorized as positive (1) or negative (0).

3.1 Features window module

The Window One of the biggest challenges of the dataset for this task is that the golden truth is given for users, but each user has an arbitrary number of posts. It is naïve to assume that any and all messages will give us the same amount of relevant information about whether an user self-harms or not. For once, the user status is known because the user has self-reported (in the case of positive users) in a post. While it is unlikely a person will falsely self-report self-harm, there is no information about when they started doing it, and when or if they stopped. Besides, some users that are classified in the golden truth as negative might do self-harm but have never reported it.

Furthermore, this is a fundamentally temporal task. Each message is not created in isolation: there is a context to them. We are limited in the context information we have about each message, but we do know the date of each post, and therefore the order in which they were created.

Finally, not all messages are equal in "information quality". These messages are posted in a social network, where writing conventions are loose. Some of them may be very short while others are very long in comparison, some of them might only be a media link, some of them might be a copied text not written by the user and so on.

To take all those challenges into consideration and create a hopefully better 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 unknown, only the previous messages can be used.

For this, we implemented a sliding window. For every new received message, our system calculated the features of the text combined from the current message and the previous w messages, where w is a configurable parameter. Depending on the size of this parameter, a longer or shorter user history would be taken into account in each step: a size of 1 only uses the current message, while a size of "*all*" would use the whole user history.

Features For each window of messages, a set of text features was calculated. These features were a mixture of textual and grammatical features (text length, number of words, etc.) and "special" features. Table 1 shows the list of features.

For these special features, previous work was done in analyzing the 2019 dataset to check if we could find differences between the positive and negative

Table 1. Text features generated for the classification system

Feature	Description
Title length (combined)	Length of the title and comment combined
Number of words in title	Number of distinct words in the title
Title length	Length of the title
Number of words in text	Number of words in the comment
Text length	Length of the comment
Punctuation	Number of punctuation marks (',', '.')
Questions	Number of questions marks
Exclamations	Number of exclamation marks
Happy faces	Number of happy emoticons (':)', ':))', ':D', etc.)
Sad faces	Number of sad emoticons (':(', ':((', 'D:', etc.)
Special features	
Sentiment analysis	Emotional score of the title and comment combined
Pronouns	Number of first-person pronouns (I, me, mine, my, myself)
NSSI words	Number of words from the NSSI corpus [8]

users. It was observed that, in general, positive users did have significant differences from negative users, although the difference between single messages was big. Section 4 shows details of this analysis.

First-person 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 [5] [8] [23]. There is also evidence linking depression and non-suicidal self-harm [8], so tracking this information would prove beneficial for our task. Besides, two sentences talking about self-harm are different depending on who the person is talking about: "I cut myself today" VS "She is thinking about cutting herself". In the first case, the user shows clear signs of doing self-harm. In the second case, however, the user 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 is supposed that people who do self-harm show more negative emotions [8]. Tracking sentiment to keep track of 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: Finally, some people who self-harm will surely talk about it. There is a sub-reddit dedicated to self-harm, where users talk about their disorder and support each other. We can suppose that at least some of the users in our dataset will use this sub-reddit, or they will talk about their problem somewhere else. It proves useful to track the usage of the most common words related to self-harm. This feature is linked to the first-person pronouns one. By tracking not only self-harm words, but also who is the subject of those sentences, we know if the user is more likely to be doing self-harm, or they are talking about somebody else. A list of words related to self-harm (NSSI words from now on-

wards) was obtained from [8]. This feature shows the number of words from this list that appear in the window, normalized by the length of the texts.

3.2 Message classification module

The features calculated from the window messages are fed to a previously trained SVM classifier. This classifier predicts whether these features belong to a message generated by a positive or negative user.

3.3 User decision module

In the final step, the outputs from the previous module are fed to the decision module.

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.

Following the same reasoning as with the features window module, however, one positive decision should not be enough to classify one user as positive. We must implement a decision policy.

The decision policy was created as such: for every new message, after receiving the output (positive or negative) of the window, the previous n outputs for that user would be observed, where n is a configurable parameter. If they were all positive, this user would be classified as positive in this iteration. If not, they would be classified as negative.

4 Data Analysis

Before starting the development of the model, we did an exploratory analysis of the 2019 dataset used in the eRisk task the previous year. The results of our findings are presented in this section.

During the model development, the data was divided in train and test data, and the analysis was only performed on the train data. However, we recalculated the analysis with all the 2019 data after the 2020 task was over for the purposes of these working notes.

The categories of the analysis follow the same division as the features explained in section 3.

Table 2 shows how many positive and negative users there exist in the dataset. As was stated before, the data is highly skewed towards negative users. All analytical results have to be taken with this information into account, since there are five times more negative than positive users.

Table 3 shows how the amount of positive and negative users affects to the number of posts that can be found in the dataset. Since there are more negative users, it is no surprising that there are more diversity in the posts from this kind of users. There is a difference of 989 posts between the minimum and maximum

Table 2. Amount of positive and negative users.

Users	Total
Positive	41
Negative	299
Total	340

for positive users, while the difference is 1982 for negative users. Information about the total and average length of posts is also given in Table 3. Although the total length is greater for negative users, the mean shows that posts made by positive users are longer on average, with the median value also being higher. The longest post belongs to a negative user, however. In addition, Table 3 shows total and average number of words used per post. In this table we can see that positive users use, on average, more words per post than negative users.

Table 3. Number of posts, post length and number of words per post for positive and negative users.

Users	Total	Mean	Deviation	Min	Max	Median
Number of posts						
Positive	6927	168.951	260.282	8	997	50
Negative	163506	546.843	544.145	10	1992	340
Post length						
Positive	1290174	186.253	334.899	1	5880	81
Negative	23605461	144.371	394.172	1	37555	59
Number of words per post						
Positive	1543682	111.425	252.631	0	5880	36
Negative	27929951	85.410	289.123	0	37555	26

Following the data analysis of this section, we decided to explore the use of first, second and third-person pronouns and how they differed between positive and negative users. Table 4 shows our findings. These values are normalized by post length. It can be seen that, on average, positive users use more pronouns per post, and the greater difference can be seen in first-person pronouns.

The same analysis was performed for the use of NSSI words. Table 5 shows the statistics in the use of NSSI words. These values are also normalized by post length. A notable difference is observed once again between positive and negative users, with positive users using more NSSI words on average. Figures 1 and 2 show the frequency distribution of the NSSI words for positive and negative users, respectively. Table 6 shows the same statistics with NSSI words divided in their different categories.

Finally, Table 7 shows the differences found when applying the sentiment analysis between positive and negative users. The values are normalized by post length, and a greater value equals a more positive sentiment. Unfortunately, there are no observable differences between them.

Table 4. Use of first, second and third-person pronouns per post normalized by post length.

Users	Mean	Deviation	Median
First-person			
Positive	1.036E-02	1.490E-02	4.329E-03
Negative	6.486E-03	2.299E-02	0
Second-person			
Positive	4.211E-03	1.074E-02	0
Negative	3.169E-03	9.594E-03	0
Third-person			
Positive	3.208E-03	7.679E-03	0
Negative	2.295E-03	6.908E-03	0

Table 5. Use of NSSI words according to positive and negative users.

Users	Total	Mean	Deviation	Median	Min	Max
Positive	2223	1.423E-03	5.774E-03	0	0	.143
Negative	22450	9.979E-04	6.367E-03	0	0	.333

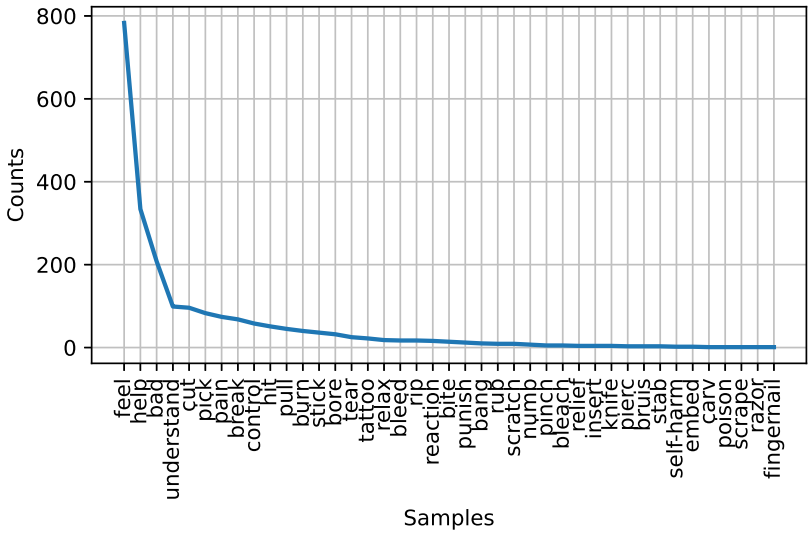

Fig. 1. Positive users. Frequency distribution of the NSSI words.

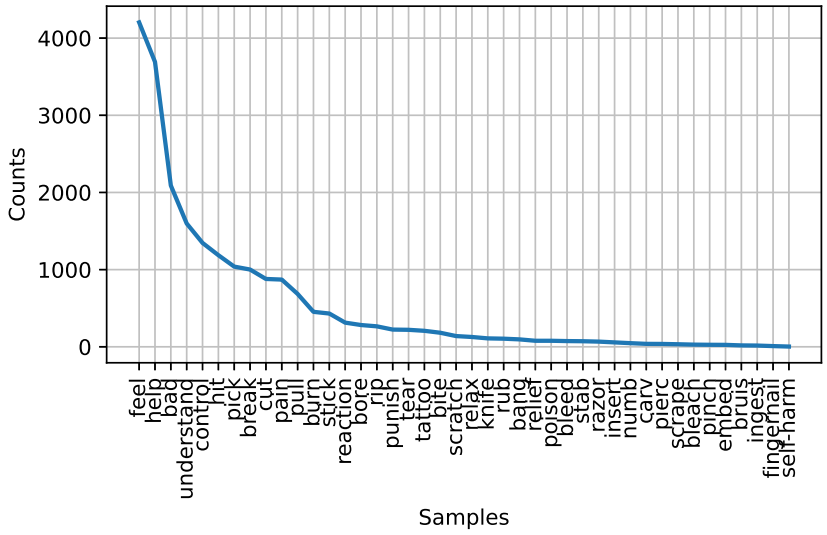

Fig. 2. Negative users. Frequency distribution of the NSSI words.

Table 6. Use of NSSI words divided by categories per post normalized by post length.

Users	Mean	Deviation
Methods of NSSI		
Positive	3.981E-04	2.880E-03
Negative	3.991E-04	4.564E-03
Cutting-specific terms		
Positive	6.347E-05	1.185E-03
Negative	4.093E-05	1.122E-03
NSSI Terms		
Positive	1.822E-07	1.101E-05
Negative	3.615E-08	1.029E-05
Instruments used		
Positive	2.275E-05	1.186E-03
Negative	1.267E-05	6.094E-04
Reasons for NSSI		
Positive	1.002E-03	4.888E-03
Negative	5.861E-04	4.391E-03

Table 7. Sentiment analysis score per post normalized by post length.

Users	Total	Mean	Deviation	Min	Max	Median
Positive	920.074	2.693E-03	1.525E-02	-.148	.229	0
Negative	18269.027	2.338E-03	1.531E-02	-.345	.293	0

5 Experimental Setup

This section presents the experiments conducted for the official eRisk 2020 task using the model proposed in section 3.

5.1 Model implementation

The SVM classification model was implemented using a combination of NLTK [1] and Scikit-learn [2]. More specifically, Scikit-learn's LinearSVC implementation of C-Support Vector Classification model was used. The amount of positive and negative users available for training was highly unbalanced in favour of the negative users, so the "class weight balanced" was used during training. Other parameters 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.

Training and testing The SVM classifier was trained with data from the eRisk 2019 task. During the model evaluation, this data was divided in training and testing data, and for the current task evaluation, a new classifier was trained with the whole 2019 data collection.

5.2 Submitted runs

Our team participated with five different runs. We were interested in observing the differences in performance by combining three factors: 1) The window size during training, 2) The window size during testing and 3) the decision window size during testing (the amount of consecutive positive messages before declaring an user as positive). Every run used a different combination of these factors. Table 8 shows the configuration of each run.

Table 8. Configuration for the runs. **All* denotes that the size of the window is the total of posts for each user.

Run id	Training window	Testing message window	Decision window
0	1	10	5
1	1	10	3
2	1	20	3
3	All*	10	3
4	All*	20	3

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

6 Results and Discussion

This section shows the official results for the task, plus some additional tests performed independently by our team. The overview for the official results of all teams can be found at [14].

During the evaluation stage of the task, teams were to iteratively request data from a server and send their predictions, one message per user at a time. After implementing our model, a program was implemented that automatically connected to this server and performed the model calculations. This program was launched on May 30th, and let run for 24 hours.

Some problems were encountered during the evaluation stage of the task. The program halted for 12 hours and had to be relaunched, which caused the number of processed messages to be lesser than expected. Furthermore, a bug in the code caused an issue with the differentiation of the five distinct runs. After the official results were given and our implementation error was fixed, we rerun the predictions again in order to show more realistic results in these working notes.

Tables 9, 10 and 12 show the official results for our team received by the task organizers. Results from other teams were added for comparison purposes.

Table 9 shows the time span and number of messages processed. We include information from the fastest and slowest teams, plus the one that achieved the best results in the official metrics. Our team, which took 1 day to process 554 messages, is amongst the faster teams, especially considering 12 hours were lost.

Table 9. User writings processed and time lapsed.

team	runs	user writings processed	lapse of time
NLP-UNED	5	554	1 day
SSN_NLP	5	222	3 hs
hildesheim	5	522	72 days + 22 hs
iLab	5	954	20 hs

Table 10 shows the official evaluation metrics for the binary decision task, plus our own calculations for the results of our fixed system. The results of the runs that achieved the best results for each metric are also added, and it is interesting to note that all belong to the same team. Table 11 also shows additional information about the number of users that were classified as positive or negative by our fixed system.

Participating teams were also required to send, for each iteration, scores that represented the estimated risk of each user. Table 12 shows the official result for our team, and the best results. Standard IR metrics were calculated after processing 1 message, 100 messages, 500 messages and 1000 messages. Our team only processed 554 messages, so the 1000 messages metrics are not given.

The testing window appears to have little effect on the result metrics. This could be due to the difference between the window sizes being too small (10

Table 10. Official eRisk 2020 T1 results compared with the recalculated results and the best of each metric for comparison. The best results for each metric and our best results are bolded. All team's results are available at [14]

team name	run id	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{tp}$	$speed$	$latency-weighted F1$
T1 official results									
NLP-UNED	0	.237	.913	.376	.423	.199	11	.961	.362
NLP-UNED	1	**.246**	**1**	**.395**	**.210**	**.185**	1	1	**.395**
NLP-UNED	2	**.246**	**1**	**.395**	**.210**	**.185**	1	1	**.395**
NLP-UNED	3	**.246**	**1**	**.395**	**.210**	**.185**	1	1	**.395**
NLP-UNED	4	**.246**	**1**	**.395**	**.210**	**.185**	1	1	**.395**
T1 fixed results									
NLP-UNED	0	.234	.875	.369	.332	.204	5	.984	.363
NLP-UNED	1	.237	.942	.379	.255	.197	3	.992	.376
NLP-UNED	2	.238	.942	.380	.255	.197	3	.992	.377
NLP-UNED	3	**.246**	**1**	**.395**	.213	**.185**	3	.992	**.392**
NLP-UNED	4	**.246**	**1**	**.395**	.213	**.185**	3	.992	**.392**
T1 best results									
iLab	0	.833	.577	.682	.252	.111	10	.965	**.658**
iLab	1	**.913**	.404	.560	.248	.149	10	.965	.540
iLab	2	.544	.654	.594	**.134**	.118	2	.996	.592
iLab	3	.564	.885	.689	.287	**.071**	45	.830	.572
iLab	4	.828	.692	**.754**	.255	.255	100	.632	.476

Table 11. Fixed results information about number of positives, negatives, and confusion matrix.

run id	# positives	# negatives	# true positives	# true negatives	# false positives	# false negatives
0	389	34	91	21	298	13
1	413	10	98	4	315	6
2	412	11	98	5	314	6
3	423	0	104	0	319	0
4	423	0	104	0	319	0

Table 12. Ranking official results next to the best results for comparison.

		1 writing			100 writings			500 writings		
team	run	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100
NLP-UNED	0	.7	.69	.49	.6	.73	.26	.6	.73	.24
NLP-UNED	1	.6	.62	.27	.2	.27	.18	.2	.27	.16
NLP-UNED	2	.6	.62	.27	.2	.27	.18	.2	.27	.16
NLP-UNED	3	.6	.62	.27	.2	.27	.18	.2	.27	.16
NLP-UNED	4	.6	.62	.27	.2	.27	.18	.2	.27	.16
iLab	3	.9	.94	.66	1	1	.83	1	1	.84

and 20). The decision window size affected the latency, which can be seen more clearly in the fixed results: The run with window size 5 had a latency of 5, while the runs with window size 3 had a latency of 3. The biggest difference was found for the training window size. Runs 3 and 4, trained with window size "All", obtained better results for the evaluation metrics, but they also classified every user as positive. Runs 0, 1 and 2, which were trained with window size 1, classified more than 10 users as negative.

While our system was a simple approach, it achieved modest results. Latency-weighted F1 is an interpretable metric that estimates the "goodness" of the solution, and our team scored, on average, more than half that the winning team achieved. This shows that even a simple, feature-driven approach can tackle what looks like a very complex problem with promising results.

Furthermore, in this kind of problem, recall is a more important metric than precision. This is because, ideally, this system would be used as a tool to raise alarm about users, but an expert would review each case in a separate basis. For this reason, it is very important to detect each and every one of the true positive cases. Table 10 shows that, while our precision score is low, our recall score is very high. Table 11 shows that this is partially because some runs categorize all users as positive, but we believe some tuning in the decision window size would somewhat fix this problem.

Speed and latency are important metrics in early risk detection, and our system achieved high scores for both of them. It is also important to note that our decisions are fast and not very heavy on computing resources. Past messages are not iterated more than x, being x the size of the window, so the model can continue forever with no extra cost.

7 Conclusions and Future Work

In this paper we present the NLP-UNED participation on the eRisk 2020 T1 task. We perform a data analysis of the 2019 T2 self-harm data and use our findings to construct features for a system to perform early predictions of signs of self-harm on users extracted from Reddit data. Our analysis shows that subjects who self-harm, on average, write longer posts, use more first-person pronouns, and mention more words related to NSSI. The official eRisk results show that our system, while simple, manages to achieve modest but fast results, but more work is needed to obtain state-of-art results.

We are interested in finding if fine-tuning the window sizes using in our system could significantly improve results. Implementing the same window policy during the training phase as the testing phase could yield better results as well. Finally, there are evidences 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 projects PROSA-MED (TIN2016-77820-C3-2-R), DOTT-HEALTH (PID2019-106942RB-C32), and EXTRAE II (IMIENS 2019).

References

1. Baetens, I., Claes, L., Muehlenkamp, J., Grietens, H., Onghena, P.: Non-Suicidal and Suicidal Self-Injurious Behavior among Flemish Adolescents: A Web-Survey. Archives of Suicide Research **15**(1), 56–67 (2011), https://doi.org/10.1080/13811118.2011.540467
2. Cavazos-Rehg, P.A., Krauss, M.J., Sowles, S.J., Connolly, S., Rosas, C., Bharadwaj, M., Grucza, R., Bierut, L.J.: An analysis of depression, self-harm, and suicidal ideation content on Tumblr. Crisis **38**(1), 44–52 (2017), https://psycnet.apa.org/record/2016-36501-001
3. Conway, M., O'Connor, D.: Social media, big data, and mental health: Current advances and ethical implications. Current Opinion in Psychology **9**, 77 – 82 (2016), http://www.sciencedirect.com/science/article/pii/S2352250X16000063
4. De Choudhury, M., Counts, S., Horvitz, E.: Social Media as a Measurement Tool of Depression in Populations. In: Proceedings of the 5th Annual ACM Web Science Conference. pp. 47–56. WebSci '13, Association for Computing Machinery, New York, NY, USA (2013)
5. Edwards, T., Holtzman, N.S.: A meta-analysis of correlations between depression and first person singular pronoun use. Journal of Research in Personality **68**, 63–68 (2017), http://dx.doi.org/10.1016/j.jrp.2017.02.005
6. Emma Hilton, C.: Unveiling self-harm behaviour: what can social media site twitter tell us about self-harm? a qualitative exploration. Journal of Clinical Nursing **26**(11-12), 1690–1704 (2017), https://onlinelibrary.wiley.com/doi/abs/10.1111/jocn.13575
7. Gluck, S.: Self-injurers and their common personality traits, https://www.healthyplace.com/abuse/self-injury/self-injurers-and-their-common-personality-traits
8. Greaves, M.M.: A Corpus Linguistic Analysis of Public Reddit and Tumblr Blog Posts on Non-Suicidal Self-Injury. Ph.D. thesis, Oregon State University (2018), https://ir.library.oregonstate.edu/concern/graduate_thesis_or_dissertations/mp48sk29z
9. Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H., Eichstaedt, J.C.: Detecting depression and mental illness on social media: an integrative review. Current Opinion in Behavioral Sciences **18**, 43–49 (2017), http://dx.doi.org/10.1016/j.cobeha.2017.07.005
10. Karmen, C., Hsiung, R.C., Wetter, T.: Screening internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods. Computer Methods and Programs in Biomedicine **120**(1), 27–36 (2015), http://dx.doi.org/10.1016/j.cmpb.2015.03.008
11. Lederer, N., Grechenig, T., Baranyi, R.: UnCUT: Bridging the gap from paper diary cards towards mobile electronic monitoring solutions in borderline and self-injury. In: SeGAH 2014 - IEEE 3rd International Conference on Serious Games and Applications for Health, Books of Proceedings. Institute of Electrical and Electronics Engineers Inc. (2014)

12. Lewis, S., Santor, D.: Self-harm reasons, goal achievement, and prediction of future self-harm intent. The Journal of nervous and mental disease **198**, 362–9 (05 2010), http://journals.lww.com/00005053-201005000-00008
13. Losada, D.E., Crestani, F., Parapar, J.: Early detection of risks on the internet: an exploratory campaign. In: 41st European Conference on Information Retrieval. pp. 259–266. Springer (2019), http://dx.doi.org/10.1007/978-3-030-15719-7_35
14. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2020: Early Risk Prediction on the Internet. In: Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020). Springer International Publishing (2020)
15. Malott, L., Bharti, P., Hilbert, N., Gopalakrishna, G., Chellappan, S.: Detecting self-harming activities with wearable devices. In: 2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops). pp. 597–602 (2015)
16. Moreno, M.A., Ton, A., Selkie, E., Evans, Y.: Secret Society 123: Understanding the Language of Self-Harm on Instagram. Journal of Adolescent Health **58**(1), 78–84 (2016)
17. Nadeem, M.: Identifying depression on twitter (2016), http://arxiv.org/abs/1607.07384
18. Nicolai, K.A., Wielgus, M.D., Mezulis, A.: Identifying Risk for Self-Harm: Rumination and Negative Affectivity in the Prospective Prediction of Nonsuicidal Self-Injury. Suicide and Life-Threatening Behavior **46**(2), 223–233 (2016)
19. O'Dea, B., Wan, S., Batterham, P.J., Calear, A.L., Paris, C., Christensen, H.: Detecting suicidality on twitter. Internet Interventions **2**(2), 183–188 (2015), http://dx.doi.org/10.1016/j.invent.2015.03.005
20. Park, M., McDonald, D.W., Cha, M.: Perception differences between the depressed and non-depressed users in Twitter. Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013 pp. 476–485 (2013)
21. Patchin, J.W., Hinduja, S.: Digital Self-Harm Among Adolescents. Journal of Adolescent Health **61**(6), 761–766 (2017)
22. Paul, M.J., Dredze, M.: You Are What You Tweet: Analyzing Twitter for Public Health. International AAAI Conference on Weblogs and Social Media (ICWSM) (2011)
23. Pennebaker, J.W.: The Secret Life of Pronouns What Our Words Say About Us. Bloomsbury Press (2011)
24. Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G.: Psychological Aspects of Natural Language Use: Our Words, Our Selves. Annual Review of Psychology **54**(1), 547–577 (2003)
25. Wang, Y., Tang, J., Li, J., Li, B., Wan, Y., Mellina, C., O'Hare, N., Chang, Y.: Understanding and discovering deliberate self-harm content in social media. 26th International World Wide Web Conference, WWW 2017 pp. 93–102 (2017)
26. Yang, W., Mu, L.: GIS analysis of depression among Twitter users. Applied Geography **60**, 217–223 (2015), http://dx.doi.org/10.1016/j.apgeog.2014.10.016
27. Yates, A., Cohan, A., Goharian, N.: Depression and self-harm risk assessment in online forums. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2968–2978. Association for Computational Linguistics (2017), https://www.aclweb.org/anthology/D17-1322