

Early Mental Health Risk Assessment through Writing Styles, Topics and Neural Models

Diego Maupomé, Maxime D. Armstrong, Raouf Belbahar,
Josselin Alezot, Rhon Balassiano, Marc Queudot,
Sébastien Mosser^[0000-0001-9769-216X], and
Marie-Jean Meurs^[0000-0001-8196-2153]

University of Quebec in Montreal UQAM
meurs.marie-jean@uqam.ca

Abstract. This paper describes the participation of the RELAI team in the eRisk 2020 tasks. The 2020 edition of eRisk proposed two tasks: (T1) Early assessment of risk of self-harm and (T2) Signs of depression in social media users. The second task focused on automatically filling a depression questionnaire given user writing history. The RELAI team participated in both tasks, and addressed them using topic modeling algorithms (LDA and Anchor), neural models with three different architectures (Deep Averaging Networks (DANs), Contextualizers, and Recurrent Neural Networks (RNNs)), and an approach based on writing styles. For the second task related to early detection of depression, the system based on LDA performed well according to all the evaluation metrics, and achieved the best performance among participants according to the Average Difference between Overall Depression Levels (ADODL) with a score of 83.15%. Overall, the submitted systems achieved promising results, and suggest that evidence extracted from social media could be useful for early mental health risk assessment.

Keywords: Early Risk Detection · Topic Modeling · Neural Networks · Mental Health Risk Assessment.

1 Introduction

The global goal of the eRisk challenges is the early detection of at-risk people from their textual production on social media, using Natural Language Processing (NLP) techniques. In 2020, two different tasks were put forth: early detection of signs of self-harm (T1), and measuring the severity of the signs of depression (T2) using textual data from related Reddit subreddits¹. These tasks are follow-ups of tasks 2 and 3 from 2019, respectively. T1 consists in sequentially

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2020, 22-25 September 2020, Thessaloniki, Greece.

¹ <https://www.reddit.com>.

Table 1. Summary statistics of the training and test sets for T1 (Self-Harm). Distributions in training and test sets are reported with % in parentheses.

	Training			Test		
	Self-Harm	Control	Total	Self-Harm	Control	Total
Nb. users	41 (12%)	299 (88%)	340	104 (25%)	319 (75%)	423

processing writings from a set of social media users, and detecting signs of self-harm [12], classifying users as at-risk or not. The goal is not only to perform this classification but also to do it as early as possible, i.e., based on as few writings per user as possible. T2 consists in automatically filling the Beck’s Depression Inventory (BDI) [3] for a set of users, based on a history of their postings on social media. In this work, we describe the participation of the RELAI team from University of Quebec in Montreal (UQAM) at the Conference and Labs of the Evaluation Forum (CLEF) 2020 eRisk tasks for early detection of signs of self-harm and depression [12]. The article is organized as follows. Sections 2 and 3 describe the proposed approaches and the research background they rely on, present the applied methodologies, experimental setup and results obtained on T1 and T2 respectively. Each Section concludes with a discussion about the results, and suggests possible future improvements.

2 Early Signs of Self-Harm (T1)

Self-harm is thought to affect about 12% of adolescents [9]. In 2014-15, hospitalizations due to self-inflicted injuries in Canada were thrice as numerous as suicides [18]. While it is co-morbid with other mental health disorders [5], its peculiar characteristics and cyclical nature have caused non-suicidal self-injury to be included as an independent disorder in the DSM-V [22]. Further, only a small fraction of young people will seek professional help either before or after engaging in self-harm (9-12%) [9]. This highlights the potential of the use of automatic means of detection on social media [20]. Such is the aim of the current task. We give hereafter a brief description of the corpus, the metrics, as well as our participation.

2.1 Task and Data

As previously mentioned, this task was first introduced in the previous iteration of eRisk (2019). In 2020, the dataset (training and test) consists of users exhibiting signs of self-harm and control users. For information regarding the labeling process, we direct the reader to [11]. Table 1 presents some statistics about the 2020 dataset. The test set is markedly different from the training set both in class proportions and user verbosity. Indeed, the ratio of positive subjects in the test is roughly double that of the training set. In addition, both positive and negative test users have fewer and shorter documents compared to their training counterparts. One of the chief concerns of the task is the *early* detection of

positive subjects. Therefore, during the test stage, a REST server² was set up by the task organizers to iteratively release user writings item-by-item during a limited period of time. Thereby, the participants had to send a GET request to retrieve the writing of each user. After each request, the processing/prediction pipeline runs and gives back to the server, via a POST request, the predictions about each individual. After each release of writings, a decision had to be emitted. Classifying a user as suffering from self-harm (decision: 1) was considered as final, while predicting the user not at risk (decision: 0) was open to updates in the following rounds. In order to evaluate the performance of the systems, and to explore ranking-based measures, the task organizers also asked participants to provide an estimated score of the level of self-harm with the decision.

2.2 Evaluation Metrics

Several metrics allow to evaluate the systems in T1: standard classification metrics like precision, recall and F_1 score as well as specific, time-aware classification metrics such as *ERDE*, $latency_{TP}$, *speed* and $F_{latency}$ (*i.e.* latency-weighted F_1). *ERDE* - *Early Risk Detection Error* - is a metric designed for eRisk tasks, taking into account the correctness of predictions and the delay taken by the system to make these predictions. The delay in decision for a given user is defined by k , the number of posts processed by the system before making a decision.

$latency_{TP}$ takes into account the latency for true positive predictions only because they represent users needing early intervention, as opposite to true negative predictions. This measure is based on the median number of posts the system has to analyze to detect true positives.

The last two metrics are *speed* and $F_{latency}$ [17]. Computing the *speed* requires a penalty factor, which takes into account the number of a user’s writings needed by the system to make a decision. $F_{latency}$ is a latency-weighted F_1 score, which combines the effectiveness of the system with a delay, by multiplying the F_1 score by the *speed* metric measuring the delay of the system.

Since the 2019 edition, the organizers have added a ranking-based evaluation process, which is not based on the binary predictions made but rather on their associated score. Two metrics are proposed for evaluating this ranking: the precision at k ($P@k$ - percentage of true positive users among the k users predicted by the system as presenting the highest risk); and the Normalized Discounted Cumulative Gain (*NDCG* - evaluates a system based on the relevance of the rankings built from its results). These metrics are computed after seeing k writings; this year, they were reported with four different k values: 1, 100, 500, 1000.

2.3 Related Work

In 2019, the best precision, F_1 and *ERDE* [7] were obtained by a system based on supervised learning for text classification called Sequential S3 - *SS3* for Smoothness, Significance, and Sanction - [6] submitted by the UNSL team. As for the

² <https://early.irlab.org/server.html>

Table 2. Summary of best results obtained on eRisk 2019 T2 (Self-harm)

System	Run	<i>precision</i>	<i>recall</i>	F_1	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	<i>speed</i>	$F_{latency}$
SS3 [7]	0	.710	.410	.520	.090	.073	2	1	.520
SS3 [7]	1	.310	.880	.460	.082	.049	3	.990	.450
LTL-INAOE [16]	0	.120	1	.220	.125	.106	1	1	.220

other evaluation metrics such as *recall*, $latency_{TP}$ and *speed*, the system submitted by LTL-INAOE achieved the best results with an approach based on the similarity between a given piece of text and a set of phrases potentially related to self-harm. Table 2 reports the best results obtained by the participating teams of eRisk 2019 T2. In the next Sections, some details are given about how we approached the problem of self-harm detection.

2.4 Topic Models

Topics discussed by users of social media could provide insight into their mental status [13]. We chose to explore this hypothesis by using Latent Dirichlet Allocation (LDA) [4], the most widely known topic modeling algorithm, as well as the Anchor variant [2]. For both topic modeling approaches, the best results were obtained when training the model using the entire textual production of each user as a single document, concatenating the posts together.

Latent Dirichlet Allocation. Two different LDA models were tested, one based on word stems and one based on word bigrams. Both models operate on documents with stop-words and short words (3 characters or fewer) removed. The first model further stems the remaining words. The second LDA model is trained instead on word bigrams. Once the LDA model is trained, users are mapped to a vector of topics. Finally, a logistic classifier is trained on these vectors. We use different training-validation splits to tune the hyper-parameters, namely the number of topics extracted. The best results are obtained by the stemmed unigram model, using a total of 14 topics.

Anchor Variant. This method tries to find a set of *anchor* words for each topic discovered. The anchor words will be assigned high probability in only one topic. The implementation of the Anchor-based system is similar to the LDA models previously discussed, using stop-word removal and stemming. Further, tokens used by over 60% of users are disregarded. As with the standard LDA approach, we find our best results in validation using 14 topics.

2.5 Neural Encoders

One of the principal challenges of the task is to combine the analyses of a user’s writings in order to arrive at a single prediction for said user. In this respect, the flexibility afforded by the back-propagation framework allowed us to explore several manners in which to structure prediction models. Broadly speaking, we

Table 3. Results on the test set of the RELAI systems in T1: Self-harm detection

System	precision	recall	F_1	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	speed	$F_{latency}$
Anchor (run 0)	.341	.865	.489	.188	.136	2	.996	.487
LDA (run 1)	.350	.885	.501	.190	.130	2	.996	.499
Contextualizer (run 2)	.438	.740	.550	.245	.132	8	.973	.535
IDA LSTM (run 3)	.291	.894	.439	.306	.168	7	.977	.428
DAN (run 4)	.381	.846	.525	.260	.141	7	.977	.513

distinguish two modes of aggregation encoding the documents making up a user into a single user encoding. The first mode, *nested* aggregation, uses two encoders. The first encoder encodes documents independently of each other. The second encoder aggregates these encoded documents together. The second mode of aggregation, *flat* aggregation, uses a single encoder combining the words from all documents simultaneously. We explored such aggregations with three different architectures as encoders: Deep Averaging Networks (DANs), Contextualizers, and Recurrent Neural Networks (RNNs). Contrary to the other two models, DANs [10] cannot account for the position of items so we only used nested aggregation with these models, having one DAN encode each of a user’s document independently, and a separate DAN aggregate these encoded documents. In the case of Contextualizers, a flat aggregation is more interesting, as even small parts of documents can be put into the context of other passages so we opted for a positional encoding consisting of a concatenation of three vectors of sinusoids [19] for each word: one of them corresponding to the position of the word in the document and the two others corresponding to the position of the document. Rather than simply providing the position of a document in the user’s history, we provide a position with two components: one counting the units of time elapsed since the writing of the post and a second one enumerates documents happening within the same unit of time (a day in our case). As for RNNs, we borrow the inter-document attention RNN approach described by [14] as the conditions are very similar.

2.6 Results and Discussion

Results of the RELAI systems on the test set are presented in Table 3 for the decision-based evaluation and Table 4 for the ranking-based evaluation. Overall, the proposed models appear to have erred on the side of caution, achieving high recall but relatively low precision.

In terms of ranking-based evaluation, precision increases from the measurement at 100 writings and remains high throughout for all models, suggesting issues with the policies mapping scores to decisions. The number of positive subjects being over 100, $P@10$ is less indicative of classification viability. While all models achieve high $NDCG@10$, neural models especially, $NDCG@100$ remains modest throughout. Contrasting this with the high recall and low precision achieved further illustrates the need for policy adjustments.

Table 4. Ranking-based evaluation on the test set of the RELAI systems in T1

Model	1 writing			100 writings		
	<i>P@10</i>	<i>NDCG@10</i>	<i>NDCG@100</i>	<i>P@10</i>	<i>NDCG@10</i>	<i>NDCG@100</i>
Anchor	.7	.80	.52	.8	.87	.52
LDA	.3	.28	.43	.6	.69	.47
Contextualizer	.2	.20	.27	.7	.81	.63
IDA LSTM	.2	.20	.27	.9	.94	.51
DAN	.2	.20	.27	.7	.68	.59
Model	500 writings			1000 writings		
	<i>P@10</i>	<i>NDCG@10</i>	<i>NDCG@100</i>	<i>P@10</i>	<i>NDCG@10</i>	<i>NDCG@100</i>
Anchor	.8	.87	.52	.8	.87	.50
LDA	.6	.69	.47	.7	.75	.47
Contextualizer	.8	.87	.70	.8	.87	.72
IDA LSTM	1	1	.59	1	1	.60
DAN	1	1	.71	.9	.81	.66

3 Early Signs of Depression (T2)

Given a user’s history writing and based on evidence found in it, T2 participants had to fill a standard depression questionnaire defined from Beck’s Depression Inventory (BDI) [3]. The questionnaire is composed of 21 questions with 4 possible answers (from 0 to 3), except for questions 16 and 18, where there are seven possible answers (0, 1a, 1b, 2a, 2b, 3a, 3b). The answers to each question represent an ordinal scale, each one associated to an integer value. The sum total of a subject’s answers is considered their score. Additionally, these scores are associated with the following categories: minimal depression (depression levels 0-9), mild depression (10-18), moderate depression (19-29) and severe depression (30-63). In T2, the proposed systems had to estimate each user’s response to each individual question. The predictions are therefore much more complex than those expected in T1. We give a brief description of the corpus, the metrics as well as our participation.

3.1 Task and Data

The second task of the eRisk 2020 lab was introduced in 2019, with the goal of exploring much finer-grained prediction of the severity of depression symptoms [12]. For this purpose, each subject was asked to fill the BDI questionnaire. The systems submitted by participants then had to estimate every user’s answer to each question given writing history of users. In order to assess the correctness of the responses provided by the participants, a number of metrics are used. Some are concerned with obtaining the exact answers whereas others are concerned with proximity in individual answers or overall BDI score. These metrics are described in the following Section. While participants in the previous iteration did not have training data at their disposal, 20 users were made available for eRisk 2020 participants, with 70 more used for evaluation. Their distributions according to the standard categorization used on BDI scores are shown in Table 5. As in T1, the training and test set differ in this respect.

Table 5. Summary statistics of the training and test sets for T2 (Depression)

	Training	Test
Total nb. users	20	70
Nb. minimal depression users	4 (20%)	10 (14%)
Nb. mild depression users	4 (20%)	23 (33%)
Nb. moderate depression users	4 (20%)	18 (26%)
Nb. severe depression users	8 (40%)	19 (27%)

3.2 Evaluation Metrics

Four metrics evaluate the systems trying to address T2. The first one is the *Average Hit Rate* (AHR). For a given user, the hit rate is simply the number of matches between the system automatic answers of the questionnaire and the user answers, i.e., the rate of correct guesses over the total. The AHR is then the mean hit rate across all users.

The second metric is the *Average Closeness Rate* (ACR). The closeness rate is a finer-grained measure of the disparity between the prediction and the ground truth for each answer, as defined by the ordinal scale on which the answers are placed. To calculate this, for each question, one takes the system and user’s answers and computes the absolute differences (ad) between them. The closeness rate for a user is the mean closeness rate for each question, and the ACR is the mean closeness rate across users.

The third metric is the *Average Difference between Overall Depression Levels* (ADODL), which is the mean over all users of the Difference between Overall Depression Levels (DODL), i.e. the absolute difference between the ground truth and the system predictions.

The last metric is the *Depression Category Hit Rate* (DCHR), which is the fraction of the cases where the system score and the user’s score fall in the same category.

3.3 Related Work

In eRisk 2019, the highest AHR, was achieved by the SS3 system trained using the dataset for the eRisk 2018 depression detection task [7]. Since the model was designed as a "yes or no" classifier, the authors had to make some modifications to return a depression level between 0 and 63 to be able to fill a BDI questionnaire. Additionally, a question-centered variant was built, achieving the aforementioned AHR. The best ACR (distance-based variant) was also achieved by a variant of the previous system using a probability distribution depending on the value of the expected answer. In terms of ADODL and DCHR, the best performances were reached with an unsupervised approach [1] using the distance between the answers and all the sentences of a user’s writing history. Table 6 reports the best results obtained by the participating teams of eRisk 2019

Table 6. Summary of best results obtained on eRisk 2019 T3 (Self-harm)

Run	AHR	ACR	ADODL	DCHR
CAMH GPT nearest unsupervised [1]	23.81%	57.06%	81.03%	45.00%
UNSLC [7]	41.43%	69.13%	78.02%	40.00%
UNSLE [7]	40.71%	71.27%	80.48%	35.00%

According to [11] these results show that it is possible to automatically extract some depression signals from social media activity. Although the performance is still modest and far from a really effective depression screening tool.

3.4 Approaching the Task as One of Authorship Attribution

The BDI was filled by only 20 users. Treating each of these users as one observation to be mapped to the answers they gave to the questionnaire would lead to a very limited number of examples. We hence approached the problem as one of authorship attribution by two different methods, which rely on decision models taking two documents and outputting the probability that both documents were written by the same user. The proposed systems exploit the decision models in different ways: one attempting to relate users to each other, and the other attempting to relate a user to the text contained in the BDI questionnaire itself. These methods are referred to as **user-based** and **answer-based** respectively.

One key advantage of this authorship framework is that the training of decision models does not require the annotation provided for the training set; the models can be trained on unannotated data from the same domain. The dataset from the eRisk 2018 depression risk detection task could hence be used for the training of the authorship attribution models, using the 2020 training data for validation. These models include LDA and a Contextualizer as well as a stylometry-based approach. As for validation, in the user-based approach, some users for whom the BDI is known were used as a knowledge base. The set of users was therefore split in half, using one half as a knowledge base and the other half for validation. On the contrary, the answer-based approach allows to validate on all 20 users.

3.5 Topic Models

One of the representation models used for this task was an LDA model. Using topic modeling, the strategy is to create topic vectors for users and then measure the distance between these in the user-based approach, or between these and the topic representation of answers for the answer-based approach. As previously mentioned, the LDA model is trained on the eRisk 2018 depression risk detection dataset. As in the Self-Harm task, each user’s posts are grouped together into larger documents. While the number of such groups will be the same for all users, the choice of it is set by observing its effect on validation results. The pre-processing then involves of stop-words, short words (3 letters and fewer)

and stemming. Using the pre-processed documents, a dictionary and a bag-of-words are created to train the LDA model. A filter is applied when creating the dictionary, removing words appearing in fewer than 20 documents or in over half of the documents. We find better results when requiring the model to find 30 topics. The trained LDA model is then used to create vectors for the documents from eRisk 2020 task 2 dataset. Each document is one of the reddit post included in the dataset. Finally, the distance between every pair of document vectors is measured using cosine similarity, which naturally falls in the unit interval, as the topic vectors are strictly positive. For both approaches, we aimed to maximize the ADODL metric. For the answer-based approach, the different experiments show that the best ADODL is reached when combining each user’s documents into 19 groups, with an LDA model trained for 30 topics. The ADODL attained by the user-based approach is approximately even when concatenating users’ posts into 10 to 19 groups. We opt to use 19 groups at test time as we posit this will allow for finer-grained predictions.

3.6 Contextualizer

Contextualizer encoders were also used for this task. This time, the aggregation considerations of the first task were no longer relevant. We tested two different approaches for the authorship decision task: encoding each document separately (parallel) or together (simultaneous). Both encoders were trained for this authorship task, ultimately using the depression questionnaire task as final validation, in both the user- and answer-based form. For the parallel encoder, the angular similarity between the document vectors is used. The simultaneous encoder, on the other hand, outputs the probability of the author being the same by design.

In order to prevent overfitting, we cease training of the authorship models by monitoring their accuracy on unseen pairs of documents, including unseen pairs of familiar documents, unseen documents by familiar users, as well as unseen users. After extensive testing, we select the parallel encoder for the user-based approach, and the simultaneous one for answer-based prediction.

3.7 Stylometry

This approach focuses on the *writing style* of a document in order to characterize its author. To this end, several linguistic features served as document representations, such as length of words and sentences, word and character frequencies and word and sentence lengths. These features were largely inspired by stylometric approaches to authorship attribution in instant messaging [15, 8] as well as legal proceedings and film reviews [21]. They are presented in Table 7. As with the LDA system, users’ documents have been concatenated together, in order to have the same number of documents per user while still accounting for all their production. Features are normalized with respect to the length of these groups, whether this length pertains to words, characters or sentences. These features result in document representations of size 585. These vector representations are

Table 7. Linguistic features and their associated number of dimensions used for the stylometry-based authorship model. The frequencies of the most frequent tokens are computed and compared irrespective of what these tokens are or whether they are the same for any two users.

Type	Feature
Syntactic	Frequency of select Parts of Speech (46)
Lexical	Frequency of most frequent word unigrams (100)
	Frequency of most frequent word bigrams (100)
	Frequency of most frequent character unigrams (100)
	Frequency of most frequent character bigrams (100)
	Frequency of most frequent character trigrams (100)
	Number of unique words (1)
	Number of alphanumeric characters (1)
	Number of digits (1)
	Number of non-ASCII characters (1)
Punctuation ratio (1)	
Morphological	Average length of words (1)
	Number of long words (1)
	Number of short words (1)
	Number of uppercase words (1)
	Number of uppercase characters (26)
Pragmatical	Average length of sentences (1)
	Number of hyperlinks (1)

Table 8. Results on the test set for T2 (Depression)

Model	AHR	ACR	ADODL	DCHR
LDA (answer-based)	28.50%	60.79%	79.07%	30.00%
LDA (user-based)	36.39%	68.32%	83.15%	34.29%
Contextualizer (answer-based, simultaneous)	21.16%	55.40%	73.76%	27.14%
Contextualizer (user-based, parallel)	36.80%	68.37%	80.84%	22.86%
Stylometry (user-based)	37.28%	68.37%	80.70%	20.00%

then compared using cosine similarity. As previously mentioned, validation was performed with the subjects for whom the BDI was available.

3.8 Results and Discussion

The results achieved on the test set are shown in Table 8. The more severe metrics, the hit rates, were fairly low for all five models. The user-based approach produced superior results across metrics and authorship models. This is unsurprising for LDA, where considerable parts of users’ activity will likely differ in subject matter from the BDI questionnaire. Given that the Contextualizer encoder matches documents individually to answers, there might be gains in performance to be obtained by considering the highest scoring document, rather than the average, for each answer. Nevertheless, although the answer-based approach

was outperformed by the user-based one, it has the very appealing advantage of not requiring annotated data, i.e. users with known BDIs.

4 Conclusion

This paper has described the experiments performed by the RELAI team from UQAM in the context of the eRisk 2020. Five models were submitted for each of the two tasks.

For the first task related to early detection of self-harm, two topic modeling systems were proposed, one using the standard LDA algorithm, and one relying on its Anchor variant. The three remaining systems were based on neural network, using three different architectures as encoders: Deep Averaging Networks (DANs), Contextualizers, and Recurrent Neural Networks (RNNs). All models are recall-oriented, which is arguably a safer decision policy. As evidenced by the ranking-based evaluations, however, tweaking this policy could result in greater precision. Globally, we achieved moderate results, the precision and recall obtained leads to a F_1 -score between 0.439 and 0.550 which is decent comparing to others systems. The Anchor model distinguished from our submitted models by its rapidity to provide fast predictions with little content. This could be explained by the presence of discriminative anchor words in provided user writings which allow to predict rapidly if a user is at risk or not.

For the second task related to early detection of depression, we approached the problem as one of authorship attribution by two different methods: **user-based** and **answer-based**. This approach affords the freedom to build decision models in a variety of ways. We relied again on LDA and the Contextualizer as well as a stylometry-based approach, achieving the best result among participants for ADODL (83.15%) with the LDA model with the user-based approach. This metric is arguably the most relevant when it comes to overall assessment of depression. Nonetheless, the ACR could be more interesting moving forward as it pertains to informing a clinician on the exact symptoms a patient is experiencing. Also, the LDA model shows a better balance between the different metrics. Almost all of the other approaches submitted achieved higher results than the average results for each metric. For example, the stylometric model user-based approach performed the second best AHR. We could also note some pertinent aspects. First, our systems are completely independent of the domain; they make decisions only on extracted features from the provided texts without requiring heavy processes of feature engineering or domain specific hand-crafted features. Also, the stark difference in the proportions in terms of number of users as well as the number of writings between the training set and the test set for the two tasks could impact the performance of submitted models. Overall, the test results show the promise of each approach. In future works, we will analyze in more detail the results obtained for each task. We plan to incorporate more carefully selected features to our decisions models which could grant a better ability to identify users at risk. Finally, given the unique nature of T2, we will explore different variations to improve predictions at a finer-grained level.

Reproducibility. The source code of the presented systems is available under GNU GPL v3 licence to ensure reproducibility. It can be found in the following repositories: <https://gitlab.ikb.info.uqam.ca/ikb-lab/nlp/eRisk2020>

References

1. Abed-Esfahani, P., Howard, D., Maslej, M., Patel, S., Mann, V., Goegan, S., French, L.: Transfer Learning for Depression: Early Detection and Severity Prediction from Social Media Postings. In: CLEF (Working Notes) (2019)
2. Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., Zhu, M.: A practical algorithm for topic modeling with provable guarantees. In: International Conference on Machine Learning. pp. 280–288 (2013)
3. Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J.: An Inventory for Measuring Depression. *Archives of General Psychiatry* **4**(6), 561–571 (06 1961). <https://doi.org/10.1001/archpsyc.1961.01710120031004>, <https://doi.org/10.1001/archpsyc.1961.01710120031004>
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *Journal of machine Learning research* **3**(Jan), 993–1022 (2003)
5. Brown, R.C., Plener, P.L.: Non-suicidal self-injury in adolescence. *Current psychiatry reports* **19**(3), 20 (2017)
6. Burdisso, S.G., Errecalde, M., y Gómez, M.M.: A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications* **133**, 182 – 197 (2019). <https://doi.org/https://doi.org/10.1016/j.eswa.2019.05.023>, <http://www.sciencedirect.com/science/article/pii/S0957417419303525>
7. Burdisso, S.G., Errecalde, M., Montes y Gómez, M.: UNSL at eRisk 2019: a unified approach for anorexia, self-harm and depression detection in social media. In: Working Notes of the Conference and Labs of the Evaluation Forum-CEUR Workshop Proceedings. vol. 2380 (2019)
8. Cristani, M., Roffo, G., Segalin, C., Bazzani, L., Vinciarelli, A., Murino, V.: Conversationally-inspired stylometric features for authorship attribution in instant messaging. In: Proceedings of the 20th ACM international conference on Multimedia. pp. 1121–1124 (2012)
9. Doyle, L., Treacy, M.P., Sheridan, A.: Self-harm in young people: Prevalence, associated factors, and help-seeking in school-going adolescents. *International journal of mental health nursing* **24**(6), 485–494 (2015)
10. Iyyer, M., Manjunatha, V., Boyd-Graber, J., Daumé III, H.: Deep unordered composition rivals syntactic methods for text classification. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1681–1691 (2015)
11. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2019 Early Risk Prediction on the Internet. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 340–357. Springer (2019)
12. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2020: Early Risk Prediction on the Internet. In: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névél, L. Cappellato, N. Ferro (eds) (ed.) Experimental IR Meets Multilinguality, Multimodality, and Interaction Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020). Springer International Publishing (2020)

13. Maupomé, D., Meurs, M.J.: Using Topic Extraction on Social Media Content for the Early Detection of Depression. *CLEF (Working Notes)* **2125** (2018)
14. Maupomé, D., Queudot, M., Meurs, M.J.: Inter and intra document attention for depression risk assessment. In: *Canadian Conference on Artificial Intelligence*. pp. 333–341. Springer (2019)
15. Mudit Bhargava, P.M., Asawa, K.: Stylometric Analysis for Authorship Attribution on Twitter. In: *Big Data Analytics: Second International Conference*. pp. 37–47 (2013)
16. Ortega-Mendoza, R.M., Fariás, D.I.H., Montes-y Gómez, M.: LTL-INAOE’s Participation at eRisk 2019: Detecting Anorexia in Social Media through Shared Personal Information. In: *CLEF (Working Notes)* (2019)
17. Sadeque, F., Xu, D., Bethard, S.: Measuring the latency of depression detection in social media. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. pp. 495–503 (2018)
18. Skinner, R., McFaul, S., Draca, J., Frechette, M., Kaur, J., Pearson, C., Thompson, W.: Suicide and self-inflicted injury hospitalizations in Canada (1979 to 2014/15). *Health promotion and chronic disease prevention in Canada: research, policy and practice* **36**(11), 243 (2016)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: *Advances in neural information processing systems*. pp. 5998–6008 (2017)
20. Xian, L., Vickers, S.D., Giordano, A.L., Lee, J., Kim, I.K., Ramaswamy, L.: # selfharm on Instagram: Quantitative Analysis and Classification of Non-Suicidal Self-Injury. In: *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)*. pp. 61–70. IEEE (2019)
21. Yunita Sari, Mark Stevenson, A.V.: Topic or Style? Exploring the Most Useful Features for Authorship Attribution. In: *Proceedings of the 27th International Conference on Computational Linguistics*. pp. 343–353 (2018)
22. Zetterqvist, M.: The DSM-5 diagnosis of nonsuicidal self-injury disorder: a review of the empirical literature. *Child and adolescent psychiatry and mental health* **9**(1), 1–13 (2015)