Early Detection of Depression and Anorexia from Social Media: A Machine Learning Approach

Faneva Ramiandrisoa
faneva.ramiandrisoa@@irit.fr
IRIT, Univ. de Toulouse
Toulouse, France
Univ. d'Antananarivo
Antananarivo, Madagascar

Josiane Mothe
Josiane.Mothe@irit.fr
IRIT, UMR5505 CNRS, INSPE, Université de Toulouse
Toulouse, France

ABSTRACT

In this paper, we present an approach on social media mining to help early detection of two mental illnesses: depression and anorexia. We aim at detecting users that are likely to be ill, by learning from annotated examples. We mine texts to extract features for text representation and also use word embedding representation. The machine learning based model we proposed uses these two types of text representation to predict the likelihood of each user to be ill. We use 58 features from state of the art and 198 features new in this domain that are part of our contribution. We evaluate our model on the CLEF eRisk 2018 reference collections. For depression detection, our model based on word embedding achieves the best performance according to the measure $ERDE_{50}$ and the model based on features only achieves the best performance according to precision. For anorexia detection, the model based on word embedding achieves the second-best results on $ERDE_{50}$ and recall. We also observed that many of the new features we added contribute to improve the results.

CCS CONCEPTS

• **Computing methodologies** → Supervised learning; • **Information systems** → Information retrieval.

KEYWORDS

Social Media Analysis, Text Mining, Depression Detection, Anorexia Detection, Early Risk Detection, weak signal detection

1 INTRODUCTION

Mental illness diagnosis has improved over decades [9]; however, it is acknowledged that early detection for early treatment is fundamental.Detection implies a medical consultation that sometimes takes time. While our aim is not to replace a medical diagnosis, we aim at studying whether social media analysis could help warning on some persons possibly suffering from a mental illness.

Indeed, in the last ten years, the use of social media platforms like Reddit, Facebook, or Twitter has increased and is still expected to grow in the next years [25] . Their users generate a lot of data that can be used to extract insights on users, on their communication practices [24], on location information [11] and on what they say. This information can also be used in medical-related applications, for example to help understanding consumer health information-seeking behavior [5], detecting mood [21], or sentiment about some diseases [22], for pharmacovigilance applications [14], or even for detecting depression [2] or suicidal ideation [3].

Our work is related to the latter applications. We aim at studying whether social media analysis and mining can help in mental illnesses detection. More specifically, we consider depression and anorexia detection tasks. We developed a machine learning model based on (a) a set of features that are extracted from users' writings and (b) vectors computed from users' writings (posts and comments). This model aims at predicting the likelihood for a user to be ill. While the principles we use for both depression and anorexia detection are the same, the main features used to detect one or the other illness are likely to differ. We thus analyze the differences on the two resulting models, specifically considering the important features in the users' writing representations. Results are based on two benchmark collections from the CLEF international forum [1].

With regard to automatic detection, these tasks can be considered as either a classification problem or a ranking problem. When considering depression detection for example, it can be considered as a binary classification problem : either the user is considered as (possibly) depressed or as non depressed. Alternatively, we can consider the depression detection as a ranking or a regression problem if the output is the likelihood for a user to be ill.

Supervised machine learning is the most common approach used in related work. The principle is that a model is trained on a set of annotated examples (training cases), then the trained model is used on cases for which the model has to make a decision (test cases). Evaluation considers ground truth on the test cases. Moreover, related work mainly consider a set of natural language processing (NLP) features extracted from texts [19] to represent items. While we re-use some features from the state of the art, in this paper, we also develop new features. In total we used 256 features from which 58 features are state of the art and 198 features are new for these tasks and part of our contribution. From the 198 features, 194 features are obtained from textual analysis across lexical categories and make use of the python library Empath [7]; the 4 remaining features are related to the text publication dates. Moreover, we combine these features with a word embedding content representation. We compare the resulting models, either combining representations or not, on two tasks in order to study which features are the most important and how much they differ from one task to the other. In this paper we

[1] Conference and Labs of the Evaluation Forum (CLEF) that promotes research, innovation, and development of information access systems www.clef-initiative.eu/

also investigate several machine learning algorithms and compare the results obtained.

This paper is organized as follows: Section 2 describes the tasks, data sets and evaluation measures, Section 3 overviews related work, Section 4 reports the model we propose, Section 5 reports the results,and finally Section 6 concludes this paper.

2 TASKS AND DATA SETS

The task and data used in this study are based on the CLEF Lab eRisk task [2] [16]. The main goal for both depression and anorexia detection tasks is to detect as early as possible some signs of depression/anorexia in texts. The detection is done on data collections composed of texts sorted in a chronological order and divided into 10 chunks. Chunk 1 contains the first 10% of each user's writings (the oldest), chunk 2 contains the second 10% and so forth.

Prediction for each user is to be given for each chunk when processed sequentially. The user has to be predicted as depressed/anorexic, as non depressed/non anorexic, or the system can postpone its decision waiting for the next data chunks. When a user receives a prediction, it is final and can not be reversed later. On the 10th and last chunk, the system has to make a decision for each user and the user has to be predicted either depressed/anorexic or not. More details about the tasks can be found in in [16].

As the problem is to detect as early as possible the sign of mental illnesses, a new measure named *ERDE* was defined in [16]. It takes into account the correctness of the system decision and the delay it took to emit its decision. *ERDE* is defined as follow:

$$ERDE_o d,k = \begin{cases} c_{fp} & if\ d\ is\ False\ Positive\ FP \\ c_{fn} & if\ d\ is\ False\ Negative\ FN \\ lc_o k \cdot c_{tp} & if\ d\ is\ True\ positive\ TP \\ 0 & if\ d\ is\ True\ Negative\ TN \end{cases} \quad (1)$$

Where d is the binary decision taken by the system with delay k for the user ; False (resp. True) Positive means d is positive and ground truth is negative (resp. positive); False (resp. True) Negative means d is negative and ground truth is positive (resp. negative); $c_{fn} = c_{tp} = 1$; c_{fp} is the proportion of positive cases in the test collection; $lc_o k = 1 - \frac{1}{1+e^{k-o}}$; o is a parameter and equal to 5 for $ERDE_5$ and equal to 50 for $ERDE_{50}$. The *ERDE* value of the model is the mean of the *ERDE* obtained for each user computed with Equation 1. For the ERDE measure, the smaller the value, the better. We also consider standard classification measures: precision, recall, and F-measure.

The *depression detection data set* is composed of chronological sequences of Reddit (www.reddit.com/) users' posts and comments. The CLEF eRisk data set was built by collecting submissions from any subreddits[3] for each user; those who had less than 10 submissions were excluded. Users were annotated as depressed (214 users) or non depressed (1,493 users). The training data set contains 135 depressed users and 752 non depressed, while the test data set contains 79 depressed users and 741 non depressed, for a total of 531,394 (resp. 545,188) posts/ comments in the training (resp. test) set. The *anorexia detection set* was built in the same way as the depression one but instead of searching for self-expressions of depression, self-expressions of anorexia were used. The training set contains 20 anorexic users and 132 non anorexic, while the test set contains 41 anorexic users and 279 non anorexic,for a total of 84,966 (resp. 168,786) posts or comments in the training (resp. test) set.

[2]https://early.irlab.org/2018/index.html, accessed on 2019-12-05

[3]Contents in Reddit platform are organized by areas of interest called "subreddits".

3 RELATED WORK

Many studies have investigated mental illness surveillance on social media such as depression detection [2], anxiety and OCD [10] or eating disorder detection [1]. There are also several evaluation frameworks related to social media analysis for mental illness detection such as eRisk [16] and CLPsych [17].

Mainly, the techniques used to detect illness on social media are supervised methods based on features extracted from texts. Many features have been defined in the literature, for example with regard to depression detection, we can quote : n-grams [2], key-phrases [15], the frequency of punctuation [19], word generalization/topic models [20], URL mentions [2], capitalized words [19] and word/paragraph embedding [19]), sentiment or emotion [2, 20], lexical resources such as antidepressant drugs name [2], linguistic features [19], activity or user behavior on the platform [2], Part-Of-Speech analysis [19], text readability [23], emoticons [19], *meta-information* [4], emotion [2], specific words [6]. In this section, we focus on related work that considers the same task and data as us. In their participation to eRisk 2018 challenge, Trotzek *et. al.* used four machine learning models [23]. While two of their machine learning models are based on CNN, the two others are based on features computed from user's text: a model based on user-level linguistic meta-data and Bags of Words (BoW), and a model based only on BoW. They also used a late fusion ensemble of three of these models: the one based on user-level linguistic meta-data and BoW, and the two based on CNN. The model based only on BoW achieved the top performance according to the measure $ERDE_{50}$ and F-measure in both tasks (depression and anorexia) at eRISK 2018. On the same task, Funez *et. al* [8] implemented two models: a model that uses Sequential Incremental Classification which classifies a user as risky based on the accumulated evidence, a model that uses a semantic representation of documents which considers the partial information available at a given time. The model based on semantic representation achieved the best results according to the measure $ERDE_5$ for depression and anorexia detection at eRisk 2018. The other model achieved the best precision for anorexia detection.

In this paper, we extend Ramiandrisoa *et. al.*' work [19] who built two machine learning models, one based on a set of features and the other based on a text representation using word embedding. Indeed, the models developed in [19] are simpler than the the ones from Funez *et. al* [8] or Trotzek *et. al.* [23] and are still very effective since the model based on a set of features achieved the second best precision at eRisk 2018. The Ramiandisoa's model uses several features from the literature of the domain, including some features from Trotzek *et. al.* [23]. We made the hypothesis that the best model of Ramiandrisoa *et. al.* (LIIRB) could have achieved better performance according to the measure $ERDE_{50}$ for depression detection if the prediction has started from chunk 1 while it started at chunk 3. We also made the hypothesis that having a richer text representation by adding more features could help the training and improve the results. For the study presented in this paper, we defined new features obtained from textual analysis across lexical categories.

Researchers found that users' mental health is correlated with the words they use [18]. Our hypothesis is the following: the writings of a user who suffers from depression or anorexia contain specific categories of words. For example texts are more likely to contain words belonging to sadness or fatigue related topics when written by depressed people; similarly food or weight related topics are more likely to be found in writings from anorexic people. While some of the features we used were specially designed for depression, we used them anyway for anorexia detection in order to study if they could be useful for other illness detection. Our results are compared to several baselines: Ramiandrisoa *et. al.* [19], Trotzek *et. al.* [23] and Funez *et. al* [8].

4 PROPOSED METHOD

We consider three models that we combine to detect depression/anorexia: (a) based on features extracted from users' writings (posts and comments), (b) based on vectors computed from users' writings and (c) combination of the two previous models.

To build the three models, we tested four classifiers which are often used in NLP and produced good results in the literature : SMO (Sequential Minimal Optimization), Random Forest, Logistic regression and Naive Bayes. We report the classifiers that gave the best results only. We found that on both depression and anorexia training data sets, Random Forest applied on the set of features achieves the best results. When using word embedding text representation, Logistic Regression achieves the best results. We report these models as ModRF and ModLR in this paper. For the combined model, we combine the output probability of the two models ModRF and ModLR. We report this later model as ModComb.

Feature-based text representation In total, we extracted 256 features; we used 58 features defined by the authors in [19] for their participation to eRisk, in which some features are specially designed for depression. 4 features are related to the text publication dates where we count the number of writings that a user has submitted in each season of a year, (one season[4] corresponds to 3 months), and 194 features are extracted using Empath tool[5] [7] that have never been used for this task in the past. These new features are very general and can be used for any text analyses and our contribution in this paper is to analyse their use for mental illnesses detection.

Even if we used the same features in both tasks, that does not mean that they are similarly important. In order to see what features are important for each task, we used χ^2 ranking[6] on the correspondent training data set. This method evaluates the importance of the feature by computing its χ^2 statistic value with respect to the target class (depressed/anorexic or non depressed/non anorexic).

The following features are the top ten according to χ^2 ranking for depression (Empath categories [7] are bold font) : *Frequency of "depress",* ***contentment, sadness, nervousness, shame****, Frequency of nouns (Part of speech frequency), Frequency of unigram feel (Bag of words), First person pronoun myself,* ***pain*** *and* ***love***. We observed that 154 features have a χ^2 statistic value higher than zero. Keeping these 154 features only in the model improves the results when training the model and it was also confirmed on the test collection ; 105 of these 154 features are from the new features we added where 104 are Empath categories and the last feature is the number of publications between June and August (season 3).

The top ten features for anorexia detection according to χ^2 are as follows : ***health, shame****, Depression symptoms and related drugs, First person pronoun myself,* ***nervousness, ugliness****, Frequency of nouns (Part of speech frequency),* ***body****, Frequency of unigram feel (Bag of words) and* ***sadness***. In that case, we observed that 57 features have χ^2 statistic value higher than zero; keeping these 57 features only, results are improved. 36 of these 57 features are new features we added where 35 are Empath categories and the last feature is the number of publications between March and May (season 2).

Considering the Empath categories [7], it seems that those related to sentiment are the most important for depression and those related to physical appearance and food are the most important for anorexia. A deeper analysis is needed to confirm this observation. Concerning the features related to the text publication dates, an analysis must be conducted in order to know why feature season 3 (resp. season 2) is important to our model for depression (resp. anorexia) detection.

We also observed that from features with $\chi^2 > 0$ on each task (154 for depression and 57 for anorexia), 48 features are common to both tasks. Three features from the 48 common features are very specific to depression but they are also useful for anorexia. These features are: drugs name, frequency of "depress" , and depression symptoms and related drugs. When we remove these three features, we observe F-measure decreases (from 0.71 to 0.67), as well as recall (from 0.60 to 0.55) and precision (from 0.86 to 0.84) (training with 10-folds cross-validation). Note that when identify the importance of features, training is based on gathering the 10 chunks users' writings.

Text representation based on text vectorization We also built a text representation based on text vectorization relying on *doc2vec* [13]. It represents users' writings as a vector. For this, we trained two separate *doc2vec* models on the training data. (a) Distributed Bag of Words model with 100 dimensional output [13]. (b) Distributed Memory model with 100 dimensional output which "ignore the context words in the input, but force the model to predict words randomly sampled from the paragraph in the output" [13].

Each user is represented by a vector. To compute the vector associated to a user, we computed first the vector of each of the user's writings and then averaged those vectors. At a given chunk, all the writings from this chunk and the writings from previous chunks were used to compute the vector. With regard to the training, we used all 10 training set chunks to represent the user by a vector. In the test stage, we represented the user by a vector computed with the available chunks. A user vector is a concatenation of the output of the distributed bag of words model and distributed memory model, resulting in a 200 dimensional vector.

5 RESULTS

In order to make a decision to annotate a user at a given chunk, we used a threshold that we set during the training stage, by testing different configurations. The way we defined the threshold is inspired form the work of [19]. We split the training data set into two subsets, one to train the model with the classifiers and one to test the model in order to define the threshold. As for depression, the training data set of eRisk 2018 is composed of training and test

[4] Season 1: December, January, and February; season 2: March, April, and May; etc.
[5] https://github.com/Ejhfast/empath-client, accessed on 2019-12-10
[6] We calculate χ^2 ranking by Weka tool

Table 1: $ERDE_5$ and $ERDE_{50}$ for detection of depression (left part) and anorexia (right part). The lower ERDE, the better.

	Depression					Anorexia				
Name	$ERDE_5$	$ERDE_{50}$	F	P	R	$ERDE_5$	$ERDE_{50}$	F	P	R
ModRF	9.62%	6.92%	0.58	**0.69**	0.51	12.40%	8.60%	0.71	0.89	0.59
ModLR	9.52%	**6.12**%	0.51	0.38	0.80	12.53%	6.27%	0.73	0.64	0.85
ModComb	9.52%	6.12%	0.51	0.38	0.80	12.34%	6.31%	0.72	0.62	0.85
UNLSA [8]	**8.78**%	7.39%	0.38	0.48	0.32	**11.40**%	7.82%	0.61	0.75	0.51
FHDO [23]	9.50%	6.44%	0.64	0.64	0.65	12.15%	**5.96**%	0.81	0.75	**0.88**
LIIRA [19]	9.46%	7.56%	0.50	0.61	0.42	12.78%	10.47%	0.71	0.81	0.63
LIIRB [19]	10.03%	7.09%	0.48	0.38	0.67	13.05%	10.33%	0.76	0.79	0.73

data sets of eRisk 2017; the splitting in eRisk 2017 is reused. For anorexia, we used the same threshold that we defined for depression as done by Funez *et. al.* [8] and *Trotzek et. al.* [23]. The idea behind this choice is to measure whether the models can perform well in detecting different mental diseases without changing the threshold. Our threshold is defined as follow: (a) For the model ModRF, a user is predicted as having the mental illness when the model predicts it with a probability higher than 0.5. (b) For model ModLR, a user is considered as depressed/anorexic if the model predicts it with a probability higher than 0.55 when using at least 20 of his writings, 0.7 when using at least 10 writings, 0.5 when using more than 200 writings and all probabilities above 0.9. All these values have been set using the training data sets only. (c) For the combined model ModComb, a user is considered as depressed/anorexic if model ModRF and model ModLR predict it. When the two models had different predictions, we gave priority to the prediction from model ModLR using the same threshold as depicted above. If the model ModLR does not predict the user as depressed/anorexic, then we considered the predictions from model ModRF. This priority was decided because model ModLR achieved better results than model ModRF on the training data set. In Table 1, the results with ModRF are obtained with the selected features.

The left side part of Table 1 presents the results of the three models on the depression test data set. We also report the best results from participants in eRisk 2018 when considering $ERDE_5$ and $ERDE_{50}$ namely UNLSA [8] and FHDO-BCSGB[23] and the best results of Ramiandrisoa *et. al* which are named LIIRA and LIIRB [19]. Other participants' results are details in [16]. We can see that there is no clear difference on results between the model ModLR and the combined model ModComb; however they achieve better results than model ModRF when considering $ERDE_5$, $ERDE_{50}$ and recall.

On eRisk 2018, compared to all the participants' results, our model ModLR achieves the best results according to $ERDE_{50}$; it is ranked 3rd according to recall (R). Our model ModRF achieves the best results according to precision (P).

Right side part of Table 1 reports the results of our three models on anorexia test data set. The best models when considering $ERDE_5$ and $ERDE_{50}$ from eRisk 2018, and the best results from Ramiandrisoa *et. al* [19]. When comparing our three models, we can see that the model ModLR gives the best results when considering $ERDE_{50}$, F-measure (F) and recall (R). Model ModRF gives the best results when considering precision (P) and the model ModComb gives the best results when considering $ERDE_5$.

When comparing to the other participants from eRisk 2018, model ModRF achieves the third-best results according to precision. It should be noted that the model ModRF is based on a set of features from which some are specially designed for depression detection. Using features that are designed for anorexia may improve the results of the model ModRF and ModComb. In short, our model ModLR achieved the second-best result according to the measure $ERDE_{50}$ and the measure recall (R).

6 CONCLUSION

This work aims at helping early detection of mental illness (depression and anorexia) by analyzing social media. We used machine learning approaches based on (a) features extracted from users' writings, (b) text representation using word embedding. We developed three models: one is based on features only, the second on word embedding text representation only and the lastest combines the two previous models. We used 58 features defined in [19] and 198 new features. The models are evaluated on two benchmark data sets provided at eRisk 2018 in CLEF international forum.

Our models can help to detect depression and anorexia. By adding new features, we outperformed the results of the authors in [19] and the results of the participants to the eRisk 2018 challenge according to two main evaluation measures ($ERDE_{50}$ and precision). For depression, when compared to other participants on the eRisk task, the model based on word embedding achieved the best performance according to the measure $ERDE_{50}$ (this measure evaluates both correctness of the decision and time used to take it) and third-best result according to recall. The model based on features only achieved the best performance according to precision. For anorexia, the word embedding models achieved the second-best result for $ERDE_{50}$ and recall and the feature model achieved the third-best precision. This result could be surprising since some of the features we used on anorexia detection task are specially designed for depression. This result leads us to think that there may be a link between depression and anorexia regarding the features that can help to detect them. We also observed that 105 of the 154 selected features for depression detection and 36 of the 57 selected features for anorexia that are selected are new features we added in this study.

For future work, we would like to investigate new features specifically designed for anorexia detection. On the other hand, we want to test different features selection such as the ones presented in [12]. Finally, we could analyze users' social signals such as the subject users leave comments on or like.

Ethical issue. While CLEF eRisk has its proper ethical policies, detecting depression, anorexia or any other human state or behavior raises ethical issues that are beyond the scope of the paper.

REFERENCES

[1] Stevie Chancellor, Zhiyuan Lin, Erica L. Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW 2016, San Francisco, CA, USA, February 27 - March 2, 2016*. 1169–1182.

[2] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. *ICWSM* (2013).

[3] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*. 2098–2110.

[4] Arman Cohan, Sydney Young, Andrew Yates, and Nazli Goharian. 2017. Triaging content severity in online mental health forums. *JASIST* 68, 11 (2017), 2675–2689.

[5] Zhaohua Deng and Shan Liu. 2017. Understanding consumer health information-seeking behavior from the perspective of the risk perception attitude framework and social support in mobile social media websites. *International journal of medical informatics* 105 (2017), 98–109.

[6] Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences* 115, 44 (2018), 11203–11208.

[7] Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding Topic Signals in Large-Scale Text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*. 4647–4657. https://doi.org/10.1145/2858036.2858535

[8] Dario G. Funez, Maria José Garciarena Ucelay, Maria Paula Villegas, Sergio Burdisso, Leticia C. Cagnina, Manuel Montes-y-Gómez, and Marcelo Errecalde. 2018. UNSL's participation at eRisk 2018 Lab. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*

[9] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* 18 (2017), 43–49.

[10] Bibo Hao, Lin Li, Ang Li, and Tingshao Zhu. 2013. Predicting Mental Health Status on Social Media - A Preliminary Study on Microblog. In *Cross-Cultural Design. Cultural Differences in Everyday Life - 5th International Conference, CCD 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part II*. 101–110.

[11] Thi Bich Ngoc Hoang and Josiane Mothe. 2018. Location extraction from tweets. *Information Processing & Management* 54, 2 (2018), 129–144.

[12] Léa Laporte, Rémi Flamary, Stéphane Canu, Sébastien Déjean, and Josiane Mothe. 2013. Nonconvex regularizations for feature selection in ranking with sparse SVM. *IEEE Transactions on Neural Networks and Learning Systems* 25, 6 (2013), 1118–1130.

[13] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. 1188–1196.

[14] Jing Liu and Gang Wang. 2018. Pharmacovigilance from social media: An improved random subspace method for identifying adverse drug events. *International Journal of Medical Informatics* 117 (2018), 33–43.

[15] Ning Liu, Zheng Zhou, Kang Xin, and Fuji Ren. 2018. TUA1 at eRisk 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*

[16] David E. Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of eRisk – Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*. Avignon, France.

[17] David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. 2016. CLPsych 2016 Shared Task: Triaging content in online peer-support forums. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA*. 118–127.

[18] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report. University of Texas at Austin.

[19] Faneva Ramiandrisoa, Josiane Mothe, Farah Benamara, and Véronique Moriceau. 2018. IRIT at e-Risk 2018. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*. Avignon, France.

[20] Philip Resnik, William Armstrong, Leonardo Max Batista Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan L. Boyd-Graber. 2015. Beyond LDA: Exploring Supervised Topic Modeling for Depression-Related Language in Twitter. In *Proceedings of CLPsych@NAACL-HLT*.

[21] Ramon Gouveia Rodrigues, Rafael Marques das Dores, Celso G Camilo-Junior, and Thierson Couto Rosa. 2016. SentiHealth-Cancer: a sentiment analysis tool to help detecting mood of patients in online social networks. *International journal of medical informatics* 85, 1 (2016), 80–95.

[22] María del Pilar Salas-Zárate, José Medina-Moreira, Katty Lagos-Ortiz, Harry Luna-Aveiga, Miguel Angel Rodriguez-Garcia, and Rafael Valencia-García. 2017. Sentiment analysis on tweets about diabetes: an aspect-level approach. *Computational and mathematical methods in medicine* 2017 (2017).

[23] Marcel Trotzek, Sven Koitka, and Christoph M. Friedrich. 2018. Word Embeddings and Linguistic Metadata at the CLEF 2018 Tasks for Early Detection of Depression and Anorexia. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*

[24] Rupa Sheth Valdez and Patricia Flatley Brennan. 2015. Exploring patientsâĂŹ health information communication practices with social network members as a foundation for consumer health IT design. *International journal of medical informatics* 84, 5 (2015), 363–374.

[25] Yu-Tseng Wang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2018. A Neural Network Approach to Early Risk Detection of Depression and Anorexia on Social Media Text. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*