# Analysis and Experiments on Early Detection of Depression

Fidel Cacheda[1], Diego Fernández[1], Francisco J. Novoa[1], and Víctor Carneiro[1]

Telematics Research Group
Department of Computer Science
University of A Coruña, 15071, Spain
{fidel.cacheda, diego.fernandez, fjnovoa, victor.carneiro}@udc.es

**Abstract.** In this paper we present the participation of the Telematics Research group from the University of A Coruña at the eRisk 2018 task on early detection of signs of depression. We formalize the task as a classification problem and follow a machine learning approach. We perform an analysis of dataset and we propose three different feature types (textual, semantic and writing). We base our solution on using two independent models, one trained to predict depression cases and another one trained to predict non-depression cases, with two variants: Duplex Model Chunk Dependent (DMCD) and Duplex Model Writing Dependent (DMWD). Our results show how the DMWD model outperforms the DMCD on terms of $ERDE_5$, ranking in the top-10 submissions for this task.

**Keywords:** Depression · early risk detection · positive model · negative model.



## 1 Introduction

In this paper we present the participation of the Telematics Research group from the University of A Coruña (UDC) at the Conference and Labs of the Evaluation Forum (CLEF) in the eRisk 2018 lab. This lab is intended to explore the evaluation methodology, effectiveness metrics and practical applications of early risk detection on the Internet. This is the second year that this lab is run and it includes two main tasks: early detection of signs of depression and early detection of signs of anorexia. Our research group participated in the first task, presenting two different models that generated five separated runs.

The eRisk 2018 early detection of signs of depression was divided in two different stages: training and test [11]. The collection contains a sequence of writings in chronological order and for each user, her/his collection of writings has been divided into 10 chunks chronologically, each one containing a 10% of the user's writings. Initially the training set was provided with a whole history of writings for a set of training users, indicating explicitly the users that have been diagnosed with depression. The test stage consisted of 10 sequential releases of data throughout 10 consecutive weeks. The first release corresponded to the first chunk, containing the oldest writings for all test users. The second release consisted of the second chunk of data and so on so forth until the tenth chunk. After each release, for each user the systems provided one of three options: a) emit a decision of depression, b) emit a decision of non-depression, or c) delay the decision (that is, see more chunks of data). Once a decision is emitted, this decision is final and in the final chunk all users must have a decision.

The evaluation is done considering the ERDE metric [9]. In this way, the evaluation takes into account not only the correctness of the system's output (that is, the precision) but also the delay taken to emit its decision.

To deal with this problem we use a basic approach to measure the textual and semantic similarities between depressed and non-depressed users, but we also focus on other writing features, such as textual spreading (i.e. number of writings per user, number of words per writing), time elapsed between two consecutive writings and the moment when the writings were created. In our proposal we use two machine learning models, one trained to predict depression cases while the other is trained to predict non-depression cases and we provide two variants named Duplex Model Chunk Dependent and Duplex Model Writing Dependent.

The remaining of this article is organized as follows. In Section 2 we comment on related work. Section 3 provides a data analysis of the dataset used for this task. In Sections 4 and 5 we describe, respectively, the features selected and the model proposed. Section 6 presents our results for this task and, finally, Section 7 includes our conclusions and future work.

## 2 Related work

There are some previous publications that make use of social networks to identify and characterize the incidence of different diseases. For example, Chunara et al. analyzed cholera-related tweets published during the first 100 days of the 2010 Haitian cholera outbreak [6] or Prieto et al. in [14] propose a method to automatically measure the incidence of a set of health conditions in society just using Twitter. Also, Chew and Eysenbach use sentiment analysis on 2 million tweets to propose a complementary infoveillance approach [5].

Specifically related with depression, a few works intend to characterize depressed subjects from their social networking behavior. For example, at the CLPysch 2015 conference a task was organized to detected, among others, depressed subjects using Twitter posts [7]. Several groups participated in the task, with promising results, although none of them were focused on early detection.

Finally, it is fundamental to mention the CLEF Lab on Early Risk Prediction on the Internet 2017 [10]. In general, participants based their approaches on different features, such as lexical, linguistic, semantic or statistical. [19] followed a two-step classification, first post level and next user level, based on a Naives Bayes classifier. [2] proposed several supervised learning approaches and information retrieval models to estimate the risk of depression. In the work by Villegas et al. [20], the authors explicitly considered the partial information that is available in different chunks of data. In [8], they proposed a graph-based rep-

**Table 1. Analysis dataset statistics.**

| | Depressed | Control | Total |
|---|---|---|---|
| # subjects | 135 | 752 | 887 |
| # posts | $49,557$ | $481,837$ | $531,394$ |
| Avg. submissions per subject | 367.1 | 640.7 | 599.1 |
| Std. dev. submissions per subject | 420.5 | 629.6 | 610.3 |

resentation to capture some inherent characteristics of the documents and use a k-nearest neighbor classification system. The work in [16] considered a depression lexicon to extract features and used recurrent neural networks and support vector machines models for prediction. Trotzek et al. in [18] use linguistic meta information extracted from the users' texts and applied different models. In [12], the authors use a combination of lexical and statistics features to predict the risk of depression. Also, in our work [4] we compare a single machine learning model with a dual model, improving the results obtained on the lab by 10% using a dual model and following the eRisk time-aware evaluation methodology.

## 3 Data analysis

The dataset used in the eRisk 2018 early detection of signs of depression task consists of a set of text contents posted by users in Reddit [11]. Each individual has been previously classified as depressed or non-depressed. Additional information can be found in Table 1, where small differences between depressed and non-depressed users start to arise. For instance, as for the submissions, both average and standard deviation are significantly greater for control users.

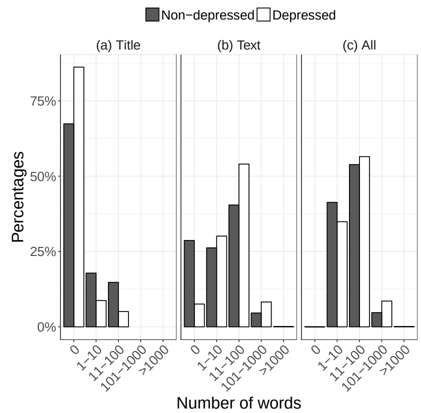
In order to detect depression in a user behavior it is essential to study the characteristics in the writings that determine this diagnosis for the subject. At this first stage, we have chosen several features easily measured, such as textual spreading, time gap and time span. We will assume that writing times are adapted to the user timezone.

**Textual spreading.** The textual spreading refers to the number of words composing a user writing. So, we are going to consider title length, text length, and the whole post length.

The comparison of those lengths according to the type of subject is depicted in Fig. 1. The first subfigure shows the number of words by title. Depressed and non-depressed users show a similar descending trend. As a consequence, those posts where users are commenting an existing reddit (so the title is not included) are more abundant. However, the difference between comments (the title length is 0) and new reddits (the title length is not 0) is higher for depressed subjects. Therefore, the latter are more prone to respond to an existing reddit than creating a new one (the title is mandatory in this case).

The second subfigure in Fig. 1 represents the number of words in the text, thus excluding the title. The percentages are higher for depressed users in all the intervals but the first one, which corresponds to writings without text. Apart

**Fig. 1.** Relative percentage for number of words used on title (a), text (b) and both fields (c) for depressed and non-depressed subjects.




from that, all users prefer short writings rather than long ones, since the number of writings with more than 100 words is lower than the others.

Regarding the third subfigure in Fig. 1, it compares the two types of subjects taking into account all the words in the writing. In this case, the differences between the subjects are smoothed and their results are similar. This is because differences in the first subfigure clearly compensate differences in the second one.

**Time gap.** Users present mainly two different behaviors with respect to the time gaps: having two consecutive writings or two writings separated by about one day. Non-depressed users follow a clearer routine, posting every day. Nevertheless, long gaps for depressed users are more sparse.

**Time span.** Regarding how users submit their writings during the week, non-depressed users tend to submit less writings at weekends and more in the middle of the week. This difference is more perceptible considering only new reddits. Moreover, depressed users behavior is more homogeneous, and it does not vary significantly at weekends. In spite of these differences, the behaviors for the two types of subjects are very similar when taking into account only comments.

Finally, depressed subjects send more posts and comments than non-depressed ones from midnight to midday, while the latter publish more in the afternoon. The main differences considering only comments appear six hours before midday (when depressed subjects are more active) and six hours after (when non-depressed subjects are more active). However, taking into account the new reddits, depressed users publish more from ten in the evening to six in the morning, but non-depressed subjects increase their difference in the afternoon.

In summary, regarding the aspects analyzed, subjects with or without depression tackle the submission of writings differently, in terms of number of words, gaps between writings, day of the week and hour.

## 4 Features

We formalize this problem as a binary classification problem using the presence or absence of a depression diagnosis as the label. However, if no strong evidence is found in one direction or another, the decision is delayed.

To address this machine learning problem, we resort to a features-based approach and design a collection of features that are expected to capture correlations between different aspects of the individual's writings and depression.

We propose three types of features: textual similarity, semantic similarity and writing features.

### 4.1 Textual similarity features

The training dataset is divided in two disjunctive sets: positive and negative subjects. Positive subjects refer to subjects diagnosed with depression, while negative subjects are those not diagnosed with depression. The main goal of these features is to estimate the degree of alignment of a subject's writings with positive or negative subjects measuring only the textual similarity between their writings. In this way we estimate the likeliness between a given subject versus positive and negative subjects.

We select a bag-of-words representation, ignoring words ordering and we employ two different measures extensively used in the literature: Cosine similarity [17] and Okapi BM25 [15].

For each subject we consider two fields with textual information: *title* and *text*. In order to capture potential differences between positive and negative subjects in the text, we measured the similarity in each field independently and concatenating all the textual information available for each writing. Therefore, in the same way as described in Section 3, for each subject we consider three textual scopes: *title*, *text* and *all*.

At the same time, for each active subject and scope we calculate the average, standard deviation, minimum, maximum and median of the scores obtained comparing this subject to every other positive subject. Then we repeat the same process for the scores with negative subjects. In both cases, the active subject is removed from the corresponding sample, whether positive or negative.

As a result we obtained 30 features for the Cosine similarity and another 30 features for the BM25 similarity.

### 4.2 Semantic similarity features

In order to capture semantic relationships among documents we apply Latent Semantic Analysis (LSA). LSA will explicitly learn the semantic word vectors by applying Singular Value Decomposition (SVD). This will project the input word-representation into a lower-dimensional space of dimensionality $k \ll V$, where semantically related words are closer than unrelated ones.

As in the previous case, each subject is represented as a document that aggregates all her/his writings but, in this case, no distinction is made between the different fields and all the textual information available is used to compute the singular values. Semantic similarity between two subjects is computed as the euclidean distance between the respective projections into the embedding space. This process is repeated for all positive and negative subjects and scores are aggregated calculating the average, standard deviation, minimum, maximum and median values. LSA is applied both following a full-text approach and removing stopwords and using Porter stemming [13]. Finally, we apply feature scaling to normalize the LSA scores computed following min-max normalization [1].

Overall, we have been able to capture 40 semantic features, considering two normalization options, two stopwords removal and stemming alternatives, five statistical measures, and positive and negative subjects.

### 4.3 Writing features

In order to complement the textual and semantic features, we also include a collection of features used to profile the characteristics of the subjects' writings. These features intend to capture part of the differences detected on Section 3, as we believe they may have an impact on the depression prediction.

From the data available on the dataset we extracted three main signals: textual spreading, time gap and time span.

**Textual spreading.** Textual spreading measures the amount of textual information provided by the subject in her/his writings. This set of features are intended to measure differences when users elaborate on their writings. To address this we introduce the following features:

- *NWritings*: the number of writings produced by the subject.
- *AvgWords*: the average number of words per writing. For each writing all the textual information available is considered.
- *DevWords*: standard deviation for the number of words per writing.
- *MinWords*: minimum number of words in the subject's writings.
- *MaxWords*: maximum number of words in the subject's writings.
- *MedWords*: median for the number of words in the subject's writing.

**Time gap.** Time gap intends to measure the frequency for a subject's writings by means of calculating the time spent between two consecutive writings. In this way, if a subject only has one writing in the time period considered, the time gap would be zero. Otherwise, the time gap will measure the number of milliseconds between two consecutive writings.

Also, a logarithmic transformation of the raw time gap values is considered. Therefore, the following two sets of features are considered:

- *TimeGap*: the aggregated information for the time lapse between two consecutive writings. These values are represented as the average, standard deviation, minimum, maximum and median.
- *LogTimeGap*: for the logarithmic transformation of the time gap values. The same aggregation values are computed for each subject.

**Time span.** This group of features is used to profile the moment when the writings were created by the subject. The following features are proposed:

- *Day*: percentage of writings corresponding to each day of the week.
- *Weekday*: accumulative percentage of writings created in a weekday.
- *Weekend*: accumulative percentage of writings posted during the weekend.
- *Hour*: percentage of writings corresponding to each hour of the day grouped into four homogeneous classes (0:00-5:59, 6:00-11:59, 12:00-17:59, 18:00-23:59).

## 5 Models

In order to learn a model from the previous features we employ a readily available machine learning toolkit. For the problem at hand we consider Random Forest (RF) [3] as the base machine learning model to solve the classification problem.

As this is not a traditional binary classification problem due to the delay option available when processing the different subjects' writings, our proposal is based on two Random Forest models, where each one has been trained with an independent set of features.

The positive model ($m_+$) is trained to predict depression cases, while the negative model ($m_-$) is trained to predict non-depression cases. For both models, and in order to make a firm decision, a threshold is set. If the probability is above the threshold then a diagnosis can be emitted and otherwise the decision must be delayed.

All models have been optimized on the training data using $ERDE_5$ or $ERDE_{50}$ scores before submitting the results for the first chunk and no modifications were applied later on. The same holds for the different thresholds utilized in each model.

Following, we describe the two variants for the prediction model proposed along with their configurations in the different submissions presented to the eRisk 2018 lab.

### 5.1 Duplex Model Chunk Dependent (DMCD)

This model applies the positive and negative models using a threshold function that depends on the number of chunks processed. The threshold function is used on the positive and negative models to determine if a firm decision (in any sense) can be provided. In this case, the threshold follows a decreasing step function being the same for both positive and negative models. It starts in 0.9 and decreases 0.1 every 2 chunks. For instance, after chunk 2 the value is 0.8, and after chunk 8 is 0.5.

More specifically, if the negative model probability is above the threshold a negative decision is emitted. Otherwise, if the positive model probability is above the threshold a positive decision is emitted and the decision is delayed in any other case.

**Table 2. Model configurations for the five submissions.**

| | UDCA | UDCB | UDCC | UDCD | UDCE |
|---|---|---|---|---|---|
| DMCD | ✓ | ✓ | | | |
| DMWD | | | ✓ | ✓ | ✓ |
| $th_w$ | | | 6 | 53 | 6 |
| Cosine Text | | $m_+$ | $m_+$ | $m_+$ | |
| Cosine All | $m_+$ | | | | $m_+$ |
| BM25 Text | | $m_+$ | $m_+$ | $m_+$ | |
| BM25 All | $m_+$ | | | | $m_+$ |
| LSA | $m_+$ | $m_+$ | $m_+$ | $m_+$ | $m_+$ |
| LSA Normalized | | | | | |
| LSA Stemming | | | $m_-$ | $m_-$ | |
| LSA Stemming Normalized | $m_-$ | $m_-$ | | | $m_-$ |
| Textual spreading | $m_+$ | $m_+$ | $m_+$ | $m_+$ | $m_+$ |
| Time gap | $m_+$ | $m_+$ | $m_+$ | $m_+$ | $m_+$ |
| Time span | $m_+$ | $m_+$ | $m_+$ | $m_+$ | $m_+$ |

### 5.2 Duplex Model Writing Dependent (DMWD)

In this case, the $m_+$ and $m_-$ are applied based exclusively on the number of writings generated for each subject. Therefore, we include a writings threshold, denoted as $th_w$, that is used to determine when the positive and negative models should be used. More specifically, if the number of writings is less than or equal to $th_w$, $m_+$ is applied, and otherwise $m_-$ is used.

In both cases, the model probability has to surpass a certain threshold to provide a definitive decision or else the decision is delayed. In case of the positive model the threshold was set at 0.9, while for the negative model was 0.5.

### 5.3 Model configurations

Table 2 provides the details for the five configurations submitted to the eRisk 2018 lab, indicating the features employed in each configuration both for the positive and negative models, as well as the variant utilized in each case.

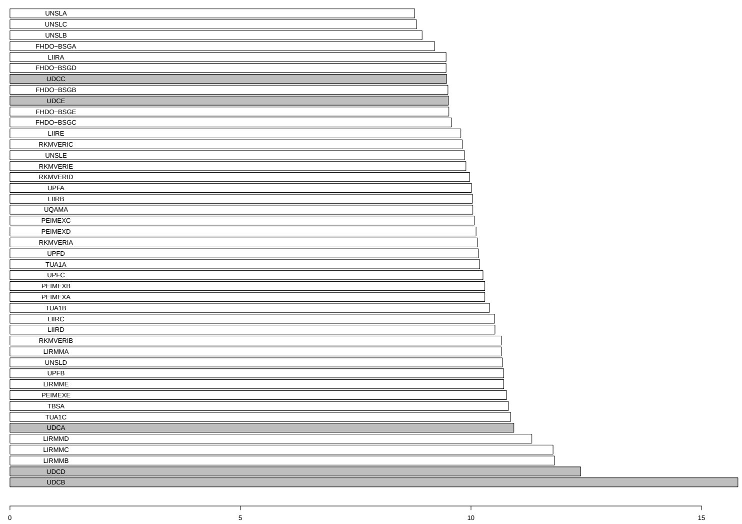
Note that the difference between configurations UDCC and UDCD relays on the $th_w$ that is set to 6 for UDCC and to 53 to UDCD in an attempt to optimize each configuration for $ERDE_5$ and $ERDE_{50}$, respectively.

## 6 Results

In the eRisk 2018 early detection of signs of depression a total amount of 11 institutions participated, submitting 45 results. The evaluation is focused not only on the correctness of the predictions, but also on the delay taken to emit the decision. In this sense, the organizers use $ERDE_5$ and $ERDE_{50}$, although also the traditional precision, recall and F1 are provided.

Figure 2 shows the official results for metric $ERDE_5$. Our best result is achieved with UDCC ranked in seventh place and followed closely by UDCE in

**Fig. 2.** $ERDE_5$ metric official results for eRisk 2018 early detection of signs of depression. Results for UDC are marked in grey.




ninth place. Both configurations follow the DMWD version and use $th_w = 6$ as writings threshold. Our best performing model uses cosine and BM25 metrics limited to the writings text field for the positive model and apply LSA with Porter stemming for the negative model, while our second best considers all textual fields for the cosine and BM25 metrics on the positive model, while a normalized LSA with stemming is used in the negative model.

Regarding $ERDE_{50}$ our results are more modest, with UDCA ranked twenty-third and UDCD ranked twenty-sixth. In this case, UDCA uses the DMCD alternative, while UDCD uses the DMWD ($th_w = 53$). Again, there are some differences in the features employed in each case. UDCA uses all textual fields to compute cosine and BM25 similarity metrics and normalized LSA with stemming for the negative model, while UDCD computes the cosine and BM25 metrics using exclusively the text field and the negative model is based on non-normalized LSA with stemming.

More interesting are the results regarding precision showed on Figure 3. In this case, our results are the worst from all participants. In fact, F1 metric shows the same results. From our point of view, this just highlights the fact that this task is not related with precision or F1 but, as the name states, with an early detection of the signs of depression.

The $ERDE$ metric penalizes the late detection of a positive case, being equivalent to a false negative. In this sense, our models try to focus on the early detection of positive cases, leaving in a second place precision and recall measures.

**Fig. 3.** Precision metric official results for eRisk 2018 early detection of signs of depression. Results for UDC are marked in grey.

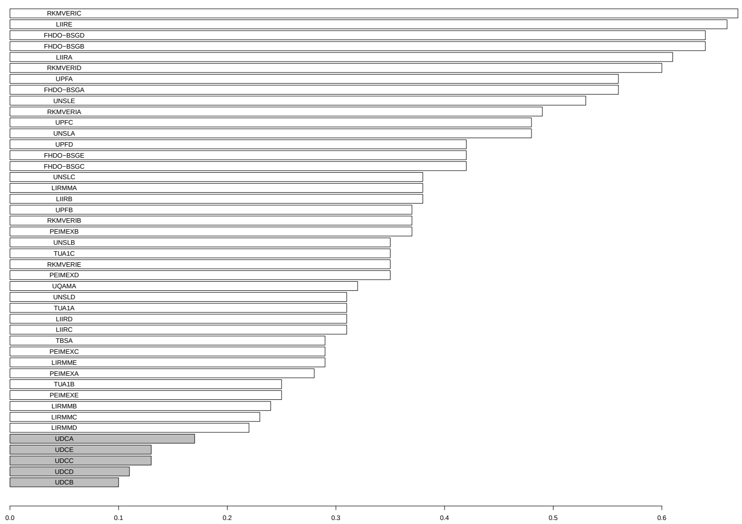



**Table 3. Oracle results for chunks 1 to 10.**

| | $ERDE_5$ | $ERDE_{50}$ | $F1$ | $P$ | $R$ |
|---|---|---|---|---|---|
| Oracle 1 | 7.62% | 3.79% | 1.0 | 1.0 | 1.0 |
| Oracle 2 | 8.47% | 4.70% | 1.0 | 1.0 | 1.0 |
| Oracle 3 | 8.90% | 5.30% | 1.0 | 1.0 | 1.0 |
| Oracle 4 | 9.15% | 5.99% | 1.0 | 1.0 | 1.0 |
| Oracle 5 | 9.32% | 6.34% | 1.0 | 1.0 | 1.0 |
| Oracle 6 | 9.47% | 6.70% | 1.0 | 1.0 | 1.0 |
| Oracle 7 | 9.56% | 6.74% | 1.0 | 1.0 | 1.0 |
| Oracle 8 | 9.61% | 6.92% | 1.0 | 1.0 | 1.0 |
| Oracle 9 | 9.62% | 7.55% | 1.0 | 1.0 | 1.0 |
| Oracle 10 | 9.63% | 7.80% | 1.0 | 1.0 | 1.0 |

The task proposed for the eRisk labs is interestingly difficult. As a matter of fact, using the golden truth provided by the organization we have calculated the performance obtained by an Oracle model that is able to predict perfectly all depression cases in all different chunks. Therefore, *Oracle 1* corresponds to an Oracle that is able to predict in the first chunk all depression cases, *Oracle 2* in the second chunk and so on.

Table 3 presents the results obtained for the *Oracle* at all chunks. As expected, precision, recall and F1 obtain perfect scores. However, $ERDE$ metrics are much more demanding. Specially $ERDE_5$, where any true positive decision that requires more than 5 writings would be penalized and soon become equivalent to a false negative.

Analyzing our results with respect to the Oracle models, we observe that our best performing model on the $ERDE_5$ score is equivalent to *Oracle 6*, while our best performing model on $ERDE_{50}$ is worse than *Oracle 10*. On the other side, analyzing the best results obtained by all participants in the depression task, the best performing model on $ERDE_5$ is slightly better than an *Oracle 3* while on $ERDE_{50}$ it is located between *Oracle 5* and *Oracle 6*.

In general, there seems to be more room for improvement on the $ERDE_{50}$ metric than on the $ERDE_5$ metric. This may be related with the fact that some users will have more than 5 writings in the first chunk, making an early prediction impossible.

## 7 Conclusions and future work

In this paper we have presented the participation of the Telematics Research group from the University of A Coruña at the eRisk 2018 task on early detection of signs of depression.

We have formalized the task as a classification problem and we have used a machine learning approach, designing three types of features in order to capture correlations between writings and depression signs: textual similarity, semantic similarity and writing features. Our proposal is based on two independent models, positive and negative, with two variants: DMCD and DMWD. Our results show how the DMWD model performs much better than the DMCD for $ERDE_5$ and it is among the top-10 submissions for the task. On the other side, our results on $ERDE_{50}$ are mediocre, with both variants performing below the average.

In the future, we expect to extend this work by studying other model combinations, with a focus on new machine learning algorithms and feature sets.

## Acknowledgments

This work was supported by the Ministry of Economy and Competitiveness of Spain and FEDER funds of the European Union (Project TIN2015-70648-P).

## References

1. Aksoy, S., Haralick, R.M.: Feature normalization and likelihood-based similarity measures for image retrieval. Pattern Recognition Letters **22**(5), 563–582 (2001)
2. Almeida, H., Briand, A., Meurs, M.J.: Detecting early risk of depression from social media user-generated content. In: Proceedings Conference and Labs of the Evaluation Forum CLEF (2017)
3. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
4. Cacheda, F., Fernández, D., Novoa, F., Carneiro, V.: Artificial intelligence and social networks for early detection of depression. Submitted for publication (2018)

5. Chew, C., Eysenbach, G.: Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. PloS one **5**(11), e14118 (2010)
6. Chunara, R., Andrews, J.R., Brownstein, J.S.: Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. The American journal of tropical medicine and hygiene **86**(1), 39–45 (2012)
7. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., Mitchell, M.: Clpsych 2015 shared task: Depression and ptsd on twitter. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. pp. 31–39 (2015)
8. Farías Anzaldúa, A.A., Montesy Gómez, M., López Monroy, A.P., González-Gurrola, L.C.: Uach-inaoe participation at erisk2017. In: Proceedings Conference and Labs of the Evaluation Forum CLEF. vol. 1866. NIH Public Access (2017)
9. Losada, D.E., Crestani, F.: A test collection for research on depression and language use. In: Conference Labs of the Evaluation Forum. pp. 28–39. Springer (2016)
10. Losada, D.E., Crestani, F., Parapar, J.: erisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 346–360. Springer (2017)
11. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk – Early Risk Prediction on the Internet. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). Avignon, France (2018)
12. Malam, I.A., Arziki, M., Bellazrak, M.N., Benamara, F., El Kaidi, A., Es-Saghir, B., He, Z., Housni, M., Moriceau, V., Mothe, J., et al.: Irit at e-risk. In: Proceedings Conference and Labs of the Evaluation Forum CLEF (2017)
13. Porter, M.F.: Readings in information retrieval. chap. An Algorithm for Suffix Stripping, pp. 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997)
14. Prieto, V.M., Matos, S., Alvarez, M., Cacheda, F., Oliveira, J.L.: Twitter: a good place to detect health conditions. PloS one **9**(1), e86191 (2014)
15. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr. **3**(4), 333–389 (Apr 2009)
16. Sadeque, F., Xu, D., Bethard, S.: Uarizona at the clef erisk 2017 pilot task: Linear and recurrent models for early depression detection. In: Proceedings Conference and Labs of the Evaluation Forum CLEF. vol. 1866. NIH Public Access (2017)
17. Singhal, A.: Modern information retrieval: A brief overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering **24**(4), 35–43 (2001)
18. Trotzek, M., Koitka, S., Friedrich, C.M.: Linguistic metadata augmented classifiers at the clef 2017 task for early detection of depression. In: Proceedings Conference and Labs of the Evaluation Forum CLEF (2017)
19. Villatoro-Tello, E., Ramírez-de-la Rosa, G., Jiménez-Salazar, H.: Uams participation at clef erisk 2017 task: Towards modelling depressed bloggers. In: Proceedings Conference and Labs of the Evaluation Forum CLEF (2017)
20. Villegas, M.P., Funez, D.G., Ucelay, M.J.G., Cagnina, L.C., Errecalde, M.L.: Lidic - unsl's participation at erisk 2017: Pilot task on early detection of depression. In: Proceedings Conference and Labs of the Evaluation Forum CLEF (2017)