Master thesis on Intelligent Interactive Systems

Universitat Pompeu Fabra

Early Detection of Eating Disorders in Reddit

Marc Mayans Yern

Supervisor: Ana Freire

Co-Supervisor: Diana Ramírez-Cifuentes

July 2018

Master thesis on Intelligent Interactive Systems

Universitat Pompeu Fabra

Early Detection of Eating Disorders in Reddit

Marc Mayans Yern

Supervisor: Ana Freire

Co-Supervisor: Diana Ramírez-Cifuentes

July 2018

Contents

Dedication

I would like to dedicate this project to everyone who has supported me directly and indirectly throughout the stage that has been the Master in Intelligent and Interactive Systems.

Acknowledgement

To Ana Freire for the given opportunity to collaborate in her projects, for her supervision and her guidance.

To Diana Ramírez-Cifuentes for helping me in the beginning of this thesis and for always being at my disposal to solve my doubts.

Last but not least, to my family and my girlfriend for all the support and patience that they have had with me.

This work was supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502).

Abstract

This thesis proposes an approach for the early detection of Anorexia Nervosa (AN) on social media. Our method is based on machine learning techniques using the processed texts written by social media users. This method relies on a set of features based on domain-specific vocabulary, topic modelling, psychological processes and linguistic information extracted from the users' writings. Moreover, other features are studied in order to exploit all the dataset resources. Additionally, we have compared some of the most known learning algorithms and we have introduced a minimum amount of information threshold to avoid some false positive predictions. This approach penalises the delay in the detection of positive cases in order to classify the users at risk as early as possible. By the early identification of anorexia, along with an appropriate treatment, the speed of recovery and the likelihood of staying free of the illness improves. The results of this thesis showed that our proposal is suitable for the early detection of AN symptoms in social media. Further research in this topic is needed to solve the problems stated in this project.

Keywords: Early risk detection; Eating disorders; Social media; Anorexia; Machine learning

Chapter 1

Introduction

1.1 Statement of the Problem

Eating Disorders (ED) are characterised by abnormal attitudes towards food and unusual eating habits [3]. Every 62 minutes, at least one person dies as a direct result from an eating disorder[1]. Anorexia Nervosa (AN) is an ED defined by the restriction in eating to keep a low weight [3]. With a mortality rate of 5% per decade, AN has the highest mortality rate of all mental disorders[1].

In 2003, eating disorders represented the third most common chronic illness in adolescent females worldwide. The prevalence of AN was about 0.3%, whereas Bulimia Nervosa (BN) was more common, with a prevalence of about 1% in young women and 0.1% in men [48]. In Europe, according to a more recent study conducted in 2016, AN was reported by 1-4% of women, and 0.3-0.7% of men. Among these people, only about one-third was detected by health-care [26]. Moreover, young people aged between 15 to 24 years old with anorexia have 10 times the risk of dying compared to their same age peers [31].

If left untreated, eating disorders tend to become more severe and less receptive to treatment [31]. The statistics mentioned above can provide an insight on how important is to detect their symptoms as soon as possible.

[1]Eating Disorders Coalition. Facts about eating disorders: What the research shows. (2016)

1.2 Motivation and Objectives

Early intervention in eating disorders is essential. According to the findings of Treasure et al. [44], when adolescents with AN are given family-based treatment within the first three years of the illness onset, they have a much greater likelihood of recovery.

Due to the fact that the symptoms associated with mental illnesses have been proved to be observable on social media [32], different automated methods to detect them are being designed. The review made by Guntuku's et al. [23] shows that most of these methods are based on the analysis of user-generated data present in online social networks, Web forums and blogs.

The current automated methods to detect eating disorders are based on machine learning techniques and do not consider the delay in detecting positive cases. We develop an approach suitable for the early detection of AN symptoms using a labelled dataset corresponding to the eRisk 2018[2] research collection [29], which contains writings posted in Reddit[3]. We aim to improve the state of art in the detection of AN and to open a way for researchers to create effective tools for the detection and prevention of anorexia.

1.3 Ethical Considerations

The aim of this thesis is to show the performance of the used techniques and to inspire social platforms to implement such methods in order to help people suffering from eating disorders.

Being aware that the applied procedure must respect the privacy of the sensitive community we are studying, it is very important to consider the ethical issues that arise from the proposed methods. Some important issues that need to be faced are related to the risk of false positives, the loss of methods accuracy and the user's confidentiality given the regularisation and the permanent surveillance present on

[2] http://early.irlab.org/
[3] http://www.reddit.com/

social media. Eating disorders are a controversial topic. For this reason, on the interventions aiming potential ill users, it is very important to assure a secure way to transmit and store the user's data and preserve their anonymity.

In general, online communities related to eating disorders prefer to remain hidden. Thus, we could ask ourselves: How these methods should be performed without making a negative impact in people's rights? In any case, they must be allowed to express their ideas on social media?

The platforms responsible of the design and the implementation of the interventions need to answer some questions. On this issue, is convenient to address the relevant stakeholders, such as designers, researchers, lawyers, politicians, individuals and clinicians in order to achieve a balance between the desire of helping individuals and the user's rights in social networks. For this task, could be interesting to consider the recommendations made by some authors for people interested in the usage of social media for psychologist purposes [25, 42].

1.4 Outline

We have organised this work in 6 different chapters. In the present Chapter 1, we introduce the motivations of this project and the research needs given the importance of the problem. In Chapter 2, we report the related work in detecting eating disorders on social media and the application of early risk measures. Chapter 3 shows our research proposal and hypotheses, focusing on describing the feature extraction process and the learning algorithms proposed. Furthermore, we describe how we are going to evaluate the performance of our built models. Chapter 4 explains our baselines and our experimental setup. In this chapter, we will expose our findings for the different experiments carried out and a brief discussion about the possible implications of the features extracted. Chapter 5 summarises our conclusions and Chapter 6 present some paths to follow in the future research in order to improve our work. Finally, Chapter 7 explains that some of the results of this thesis have been accepted for publication at International Conference on Internet Science 2018[4].

[4] `http://insci2018.org/`

Chapter 2

Related Work

2.1 Social Media and Anorexia

On the Web, the promotion of behaviours related to eating disorders is known as Pro-eating disorder site. The usage of these sites is prevalent among adolescents with these conditions [49]. Moreover, their engagement with this type of content has recently been suggested as a screening factor for these kind of illnesses [7]. On social media, people with eating disorders, such as anorexia and bulimia, can be identified by the usage of certain keywords that characterise and promote these conditions [2, 46]. In this sense, features or variables that have been extracted from labelled user-generated data [23] are used to build predictive models capable of doing an automated analysis of social media data. Based on the related work for detecting mental illnesses in online social platforms, the analysed data is obtained either by diagnosing participants with the usage of surveys [19, 37, 45], or by crawling directly the data from public online sources like Reddit, Twitter or Facebook [2, 11, 35]. The majority of mental illnesses' studies focus in the detection of depression and there are few dedicated to eating disorders. We assume that methods used for other mental illnesses could serve in our investigation.

In order to build predictive models, the most common features used for analysing and predicting mental illnesses are those extracted from the texts written by the users,

such as: topics [34,41,45] (topic modelling), frequencies of words or combinations of words (N-grams) [41,45], and features obtained using dictionaries like LIWC[1], which can provide an insight on the usage of self references, social words and emotions [18, 19]. Researchers in [10] show that eating disorders present a high level of quantifiable differences in terms of language usage with respect to other mental illness.

Related works also have studied the users' posting frequency in different periods of the day and year [2, 9, 18], and have also obtained features from the relationships between users, taking into account the number of friends, or followers [2, 17, 19] (egocentric social graphs). Additionally, some studies use features based on sentiment analysis, considering the subjectivity or polarity of a phrase [18,19,45]. Other researchers also analyse the use of emoticons [47] or have created a multimodal analysis taking into account text, emoticons and images from social media [24].

2.2 Early Risk Detection

To the extent of our knowledge, building predictive models to detect early risk of ED is not a widely explored task yet. For detecting other mental illnesses, such as depression, some works have attempted to do their analysis and build their models using only the data prior to the diagnosis [19, 45]. But the issue of the early risk detection is addressed by the work of Losada et al. [29], where the proposal of a temporal-aware risk detection benchmark, complements the evaluation on the accuracy of the decisions taken by the algorithms. In other words, this work proposes a new metric to measure the effectiveness of early alert systems. This metric known as Early Risk Detection Error (ERDE), which will be deeply detailed in Section 3.4, penalises the delay in detecting positive cases, and is suitable to evaluate our proposal.

[1] `http://liwc.wpengine.com/`

Chapter 3

Proposal

The main objective of our proposal is to detect early risk of anorexia in social media, minimising the ERDE measure and maximising the F_1 measure as we explain in Section 3.4. We use machine learning techniques that combine a set of features extracted from the concatenated writings of users on social media.

In the remainder of this section we are going to define the research hypothesis, how we proceed in order to extract the features from the writings and the learning algorithms that we will use to build our models. Besides, we explain the measures that we will use to evaluate the effectiveness of our proposal.

3.1 Research Hypothesis

We want to study the following hypothesis: The combination of machine learning techniques with the adequate feature extraction procedure is suitable for detecting early risk of AN in social media and, in particular, Reddit.

We aim to improve the state of art of early detection of anorexia on social media following the next steps:

- Extracting a diverse set of features from our dataset and testing their performance.

- Exploring the behaviour of the most powerful learning algorithms to better predict positive cases.
- Adding techniques to reduce false positive predictions.

An approach based on the *dynamic strategy* proposed in [29] is used. This method consists in building incrementally, writing per writing, a representation of each user (from the test set), and applying a classifier, which was previously trained with all the users' texts (from the train set). We consider the classification problem as a binary problem with the classes anorexic and non-anorexic. Notice that we process the title and the body of the users' posts to build their representation since we assume that the titles could contain relevant information. Following this approach, depending on the algorithm used, a decision is made if the classifier outputs a confidence value above a given threshold. The models with the highest F_1 score and the lowest value for the ERDE measure will be considered as the best.

3.2 Feature Extraction

We fed our models with features that characterise the content of the writings of each user. As we explain in Section 4.2.1, our dataset is more limited in resources than other datasets extracted from social networks such as Twitter. For this reason, we want to exploit all the possible features (to the extent of our knowledge) of our Reddit dataset. Further details of these features are explained below and summarised in Table 1.

Psychological and linguistic processes: We calculate features to characterise the users' writings. These features were calculated by taking into account the frequency of words belonging to the categories of the LIWC2007 dictionary [33], which has been previously used in detecting mental health issues [9, 12]. In this sense, scores that consider linguistic and psychological processes, as well as personal concerns and spoken categories were obtained. We consider a new feature value for each category defined in the LIWC2007 dictionary. The list and description of these categories can be found in [33]. The scores were calculated normalising the frequencies

of words by the total number of words in the writings of a user. Given that certain words could belong to multiple categories, the normalisation value was increased in one each time a word was part of more than one category.

Domain-related vocabulary: We defined 9 features by creating categories of words that belong to domains related to anorexia. The vocabulary for these categories was obtained from the codebook's domains and sample keywords defined in [2]. The domains are: anorexia, body image, food and meals, eating, caloric restriction, binge, compensatory behaviour, and exercises. These features were calculated in the same way as the psychological and linguistic processes features.

N-grams: They consist of sequences of contiguous words words within a given window (N). Studies have extensively used them in text mining and natural language processing tasks [21]. Since previous works have considered them as features for detecting depression and eating disorders [41, 45], we did a $tf \cdot idf$ vectorisation of the unigrams and bigrams of the training set writings. For this step, we used the *TfIdfVectorizer* from the *scikit-learn* Python library[1], with a stop-words list and the removal of the n-grams that appeared in less than 20 documents. The content of a document was defined by the concatenation of all the writings of a user.

Topic modelling: Topic modelling consists in automatically extracting and identifying topics that are present in documents in order to obtain hidden patterns of a corpus. A well-known method proposed by David Blei et al. [4] is the Latent Dirichlet Allocation (LDA), which is an unsupervised generative statistical model in which the topics are represented by a set of terms or words. Many authors show that, in tasks of prediction and classification, the use of this method is sounded. For instance, the authors in [38, 45, 50] conclude that features based on topic modelling are helpful in tasks for recognising depressive and suicidal users. Besides, this technique has been used in [8], combined with other features, to quantify and predict the mental illnesses severity in online pro-eating disorder communities.

To define the topics we used English stopwords and only considered the words that

[1] `http://scikit-learn.org/`

appeared at least in 10% of the training documents. The 50 features used by the model are given by the probabilistic distribution of 50 topics for each analysed text. This number was selected by the 10-fold cross validation technique in training tasks. The *LatentDirichletAllocation* module from the *scikit-learn* Python library was used to do this implementation.

Sentiments: We have added four features related to the sentiment polarity of the text. In order to do this, we have used the VADER Sentiment Analysis (Valence Aware Dictionary and sEntiment Reasoner) library[2], which is specifically attuned to sentiments expressed in social media [22]. The features correspond to the negative, positive and neutral percentages associated with each analysed text. The fourth feature corresponds to a compound percentage of the three previous features.

Time: Our dataset contains the time of publication from Reddit's server. Since we do not have any geographic information of the users, we are not able to know if their are posting during the day or at night. For this reason, we have added a feature that consists in the average of the publication interval between consecutive posts (in hours).

Removed posts: In the dataset, there are some posts where the body text appears as "[removed]". We have added a feature which is the ratio of deleted posts with respect to the number of posts by the user.

Pharmacological treatment: We have added a feature that is a ratio of medicines mentioned in the texts which are typically used in pharmacological treatment of eating disorders based on [20]. This feature is a ratio of the medicines mentioned in the texts over the total number of medicines examined. Also, we have considered another feature which is the ratio of words mentioned that refers to the type of these medicines, as could be antidepressant, anxiolytic or anti-emetic for instance.

[2] `https://github.com/cjhutto/vaderSentiment/`

Table 1: Features considered.

Feature Type	Details and resources	Number of features
Linguistic and psychological processes	LIWC	64
Domain-related vocabulary	Anorexia vocabulary	9
N-grams	Unigrams	4303
	Bigrams	667
Topic modelling	Topics using LDA	50
Sentiments	Sentiment Polarity	4
Time	Interval Post Frequency	1
Removed posts	Removed Posts Ratio	1
Pharmacological treatment	Medicines Ratio	2

3.3 Learning Algorithms

In this section we are going to describe briefly the algorithms used in this project. All this algorithms are supervised learning models and we have used the *scikit-learn* *Python* library for their implementation.

We explored four different prediction models, i.e., logistic regression, random forest, support vector machine, and multilayer perceptron since they have been used previously as classifiers for similar tasks [27, 29, 34]. They are explained below:

Logistic Regression (LR): Logistic regression is a common statistical method developed by statistician David Cox in 1958 [14] and used to predict a binary outcome given a set of independent variables. This algorithm fits data to a logistic function (see Figure 1) in order to predict the probability of occurrence for an event [1] (see Eq. 3.1).

$$p(C_1|\Phi) = \frac{1}{1 + e^{-\Phi \cdot x}} \tag{3.1}$$

Given that this algorithm is not a classifier since it outputs a probability in terms of the input, we need to choose a probability threshold to use it in a classification problem. If the probability of the input is higher than a predefined threshold the data belongs to a class, otherwise it belongs to another.

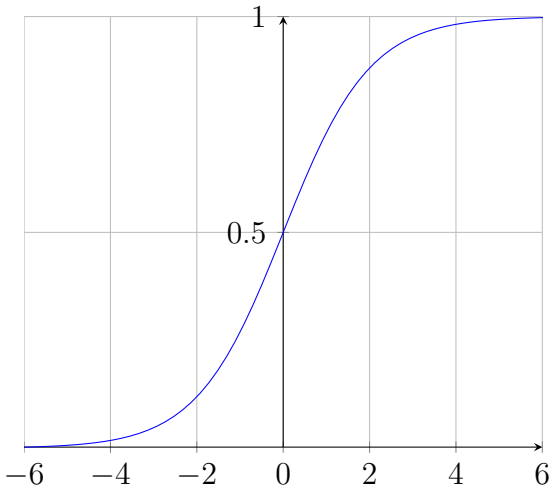

Figure 1: Logistic function representation in 2D.

Random Forest (RF): Random forest is an ensemble learning method for classification that works by building many decision trees at training time. Each tree depends on the values of a randomly generated vector that is tested independently and has the same distribution for all the trees in the forest. For the classification tasks, its output is the class that is the mode of the classes of the individual trees [6].

A decision tree is a well-known method for machine learning tasks which uses a tree-like graph of decisions and their possible consequences [6]. Since a decision tree tends easily to overfit, random forest is commonly used to average different trees in order to reach a trade-off between the variance and the bias. A simple example of a decision tree can be found in Figure 2.

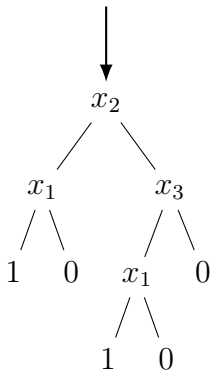

Figure 2: Example of a decision tree with three variables and two output classes. Each branch represents a decision over a variable.

Support Vector Machine (SVM): Support vector machine is an algorithm that finds a decision plane which maximises the distance between the classes to classify [13]. A graphical representation of this problem can be found in Figure 3.

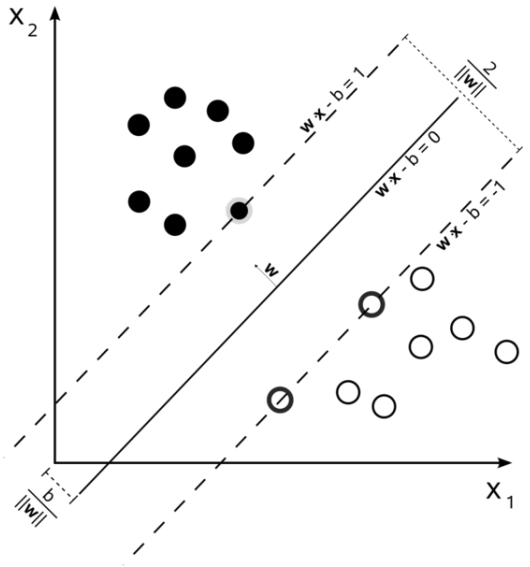

Figure 3: Graphic showing a hyperplane maximising the distance between two data classes, black and white dots, in 2D.

The algorithm seeks for the "maximum-margin hyperplane" that divides the data classes, but normally this data is not linearly separable. For this reason, SVM allows mislabelled examples in order to find a hyperplane that splits the data as best possible. This problem is solved by a technique called *Soft-Margin*.

However, in 1992 [5] suggested a way to create nonlinear classifiers using the kernel trick to maximum-margin hyperplanes. This enabled to operate in high-dimensional space to fit the maximum-margin hyperplane.

Multilayer Perceptron (MLP): A multilayer perceptron is a feed-forward artificial neural network [36]. An MLP contains at least three layers of nodes: an input layer, one or more hidden layers and an output layer. Each node uses a non-linear activation function except for the input nodes. The two more common activation functions are the hyperbolic tangent and the logistic function, both are sigmoids. For training, MLP commonly uses a technique called backpropagation [39,40], which is a generalization of the least mean squares algorithm in the linear perceptron. But

this type of minimisation present some cons. For instance the convergence is very slow and it may fall into a local minimum. Furthermore, this method is prone to overfitting, especially when number of nodes is large. An example of an MLP with 3 layers can be seen in Figure 4.

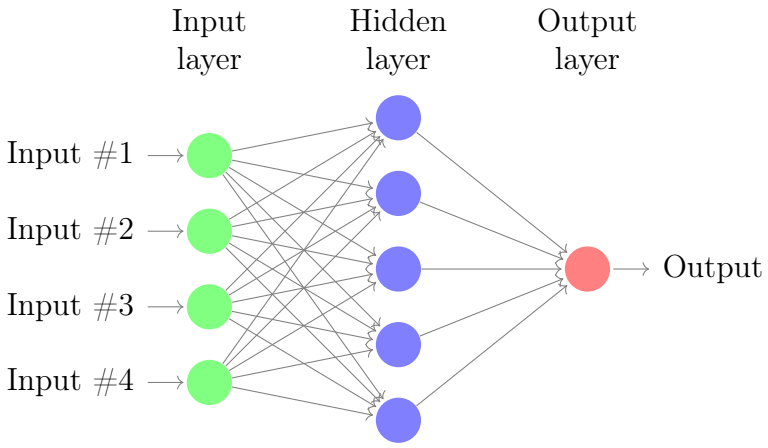

Figure 4: Example of a multilayer perceptron with a single hidden layer.

Unlike the linear perceptron, the combination of multiple layers and the non-linear activation allows to distinguish data that is not linearly separable [16]. The universal approximation theorem states that, under mild assumptions on the activation function, a feedforward neural network with a single hidden layer containing a finite number of neurons can approximate any continuous function on a compact subset of $\mathbb{R}^n$ [15].

3.4 Evaluation

Since we want to focus in the detection of positive examples (anorexic users), to evaluate the performance of our methods we report the Precision, Recall and F_1 measure. These metrics are based on the confusion matrix tool, which allows to a better visualisation of our models' performance. In Figure 5 is graphically represented a confusion matrix for a two class problem. Where TP is the number of correct predictions that are positive (true positive), FN is the number of incorrect predictions that are negative (false negative), FP is the number of incorrect predictions that are positive (false positive) and TN is the number of correct predictions that are negative (true negative).

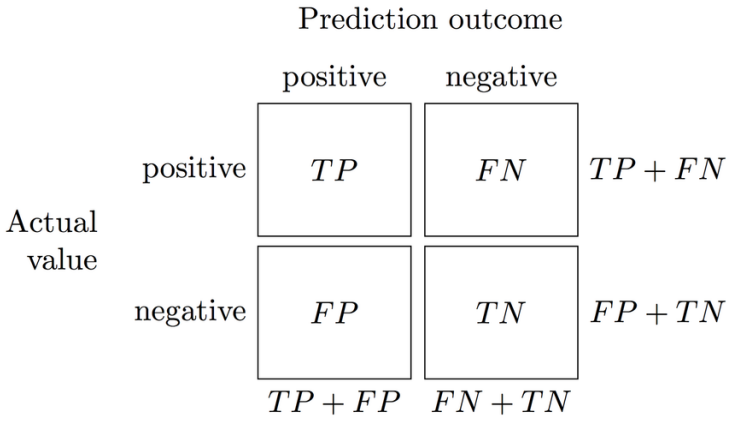

Figure 5: Confusion matrix for a two class problem [30].

The precision (P) measure indicates the fraction of anorexic users correctly classified from the examples that our model determines as positive (see Eq. 3.2). The recall (R) measure shows us the fraction of the total anorexic users that have been correctly classified (see Eq. 3.3).

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

We want to focus in the F_1 measure since it is the harmonic mean of the precision and recall metrics. This measure is shown in Eq. 3.4.

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3.4}$$

However, this commonly known measures do not indicate how early detection is made and for this we need to add some measures that are time-aware.

To accomplish this, we use a metric called Early Risk Detection Error (ERDE) proposed by David E. Losada and Fabio Crestani at their work [29] to evaluate our proposal. This measure gives a cost c to each binary decision d taken by the system at a number k of textual items seen before making a decision. In other words, it

penalises the delay taken by the model to emit a positive decision. This error is defined in our case as:

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{if } d = \text{False Positive (FP)} \\ c_{fn} & \text{if } d = \text{False Negative (FN)} \\ lc_o(k) \cdot c_{tp} & \text{if } d = \text{True Positive (TP)} \\ 0 & \text{if } d = \text{True Negative (TN)} \end{cases} \tag{3.5}$$

Where $c_{fp} = 0.13$, $c_{fn} = 1$, $c_{tp} = 1$, and the cost function $lc_o(k)$ is a sigmoid function to penalise the late decisions:

$$lc_o(k) = 1 - \frac{1}{1 + e^{k-o}} \tag{3.6}$$

Notice that the c_{fp} value was set according to the proportion of positive cases in the data and o is a parameter of (3.6) which defines the point at which the cost grows more quickly [29]. In Figure 6 we can see the behaviour of this function under two different values of o.

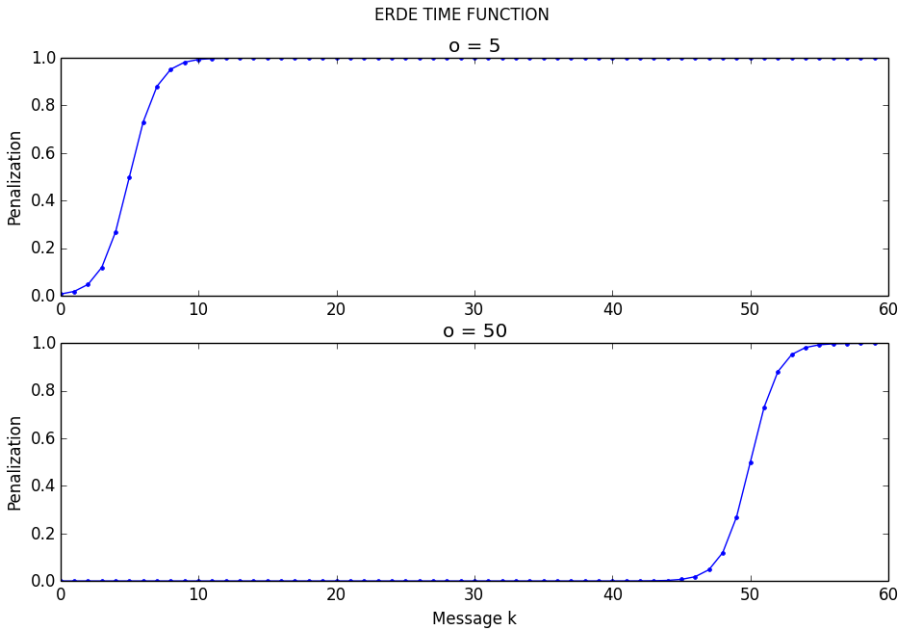

Figure 6: ERDE function representation in 2D.

The final ERDE value is the mean of the n ERDE values since we have n individuals in the collection with a decision taken in different k. In order to get the best model, we have to identify the ones with the highest F_1 measure and the lowest ERDE values.

Chapter 4

Results

4.1 Baselines

In order to test our approach we set four different baselines based on Losada et al. [29] since they have a similar objective and a dataset extracted from Reddit as well. The baselines are described as follows:

Random: This is a naive strategy that makes a random decision after seeing the first message. This decision could be positive ("anorexic") or negative ("non-anorexic"). This is a fast strategy, although we assume its effectiveness will be poor.

Minority: This is another naive strategy that predicts a positive decision ("anorexic") for each user right after processing the first message. Again, this strategy makes a quick decision but it is expected to have poor effectiveness.

First n: This strategy consists in concatenating the first n submissions from each user and getting the prediction provided by our model. If the user has less than n posts, all the data available for that user is used.

Dynamic: This strategy consists in analysing and building incrementally a representation of each user. Messages are concatenated one per one, and a prediction is made by our model each time a new message is added. This is done until a positive decision is reached. The system can emit a positive decision ("anorexic") only if the

classifier outputs a confidence value above a given threshold. As in the paper of Losada et al. [29], we tested three thresholds: 0.5, 0.75 and 0.9. This method does not work with a fixed number of messages. If the stream of texts is over, the method concludes with a negative decision ("non-anorexic").

For the *Dynamic* and the *First n* baselines we have employed a quite simple model, which consists in a Logistic Regression classifier trained and tested with features extracted from the posts. These features were based on the *tf-idf* vectorisation considering only unigrams.

4.2 Experimental Setup

Our experiments are conducted over the eRisk 2018 research collection [29], which contains a labelled dataset with writings of a control group and people diagnosed with anorexia that we will explain in detail in section 4.2.1. *Python 2.7.15*[1] and, in particular, the *scikit-learn Python* library was used for the implementation of the proposed methods.

In order to develop our final model we have followed the next steps described below:

1. We determine the baselines for our experiment in order to compare the performance of our model. These baselines are described in Section 4.3.1.

2. Using the best model of the mentioned baselines, we are going to add the new features explained in Section 3.2 separately, in order to know if they provide any interesting information to the learning algorithm.

3. We combine the best features resulting from the step 2 for testing the performance of the learning algorithms explained in Section 3.3.

On the evaluation of our model, we use the measures detailed in Section 3.4. We consider as the best models the ones that have the highest F_1 measure and the lowest ERDE measure.

[1] `https://docs.python.org/2/`

4.2.1 Dataset Description

Our method analysed a collection composed by chronologically ordered writings (posts or comments) from a set of Reddit users [29]. Users were labelled as anorexic and non-anorexic. The training set consists in a list of 20 users marked as anorexics and a control group of 132 users marked as non-anorexics. For the test set we have 41 positive users (anorexic) and 279 negative users (non-anorexic). The data provides more information besides the texts: the user id, the title of the message and the server time of the post. The dataset statistics are detailed in Table 2. On average anorexic users write fewer posts although these contain a larger number of words per submission rather than the posts written by non-anorexic users.

Table 2: Main statistics of the train and test collections.

	Train		**Test**	
	Anorexia	**Control**	**Anorexia**	**Control**
Num. subjects	20	132	41	279
Num. writings	7,452	77,514	17,422	151,364
Avg num. writings	372.6	587.2	424.9	542.5
Avg num. words per submission	41.2	20.9	35.7	20.9

Notice that we worked with an unbalanced dataset, which can negatively affect to the results due to the small number of positive training cases. Moreover, is important to stand out that our dataset is more limited in resources, have less features to exploit, respect to other social networks where there is more interactivity. For instance, Twitter can provide data about the followers, likes and comments of a post or a user.

4.2.2 Parameter Setting

By using the provided data we defined the training and testing sets. To train our models we applied 10-fold cross-validation and optimised the parameters through grid search. Each instance for our training task represented a user, and was defined by the features mentioned in Section 3.2. These features were extracted from all the sequentially-concatenated writings of each user (the title plus the body of the post). The test set allowed us to evaluate the behaviour of the *dynamic* method, where the

classifiers were applied each time a new writing was read.

Also, the 10-fold cross-validation process was used to define a threshold that represented the minimum probability value required by an instance to be classified as positive. After having tested different values for the LR and RF classifiers, this threshold was set to 0.75 and 0.55 respectively. For the SVM we have used a linear kernel and balanced class weights. We implemented an MLP with one hidden layer of 200 neurons using the logistic sigmoid function as the activation function. Finally, the solver for weight optimisation is the Limited-memory BFGS [28], which is an optimisation algorithm in the group of quasi-Newton methods that approximates the Broyden–Fletcher–Goldfarb–Shanno algorithm and gives us the best results in the cross-validation process.

Our approach modifies the dynamic strategy of [29] by defining a minimum amount of information that should be seen by the system before applying the classifier and emitting a positive decision. We include this threshold since we want to observe the progress of a user in the social media in order to predict an eating disorder. With the implementation of this threshold, we want to avoid some false positive predictions when the posts contain very few words. For instance, if the first post of a user is read and it contains five words with the word *laxative* mentioned twice, our classifier might give it a high score and classify the user as anorexic having seen only one short post. We assume that it is required to establish a minimum number of words to be seen before emitting a decision, since our approach is based on text analysis. We have understood that a temporal fixed threshold (based on time) is not efficient due to the variability of publications among users.

The threshold -number of words- is defined by the *text length threshold TLT* (see Eq. 4.1). To define the TLT we first assume that each user has a fixed number of words per post, denoted as *maxPostLenght*. To calculate this number we plotted a histogram to visualise the distribution of the number of words per post of all the users with anorexia (see Figure 7). We observe that 80% of the users wrote up to 90 words per post. Based on Pareto's principle we chose this number of words for the *maxPostLenght* value. We assumed that seeing just one post of a fixed

size was not enough, hence we considered that exploring more posts would reduce the amount of false positives. This number of posts, was defined by a percentage *selectionPercentage* of the *average number of posts per user*, 372.6 in our case, which was denoted as *avgPosts*. For our experiments we chose to work with 10% for the value of *selectionPercentage*. For the analysed dataset the TLT value resulted in 3353 words to be observed before processing the texts.

$$TLT = maxPostLenght \times avgPosts \times selectionPercentage \tag{4.1}$$

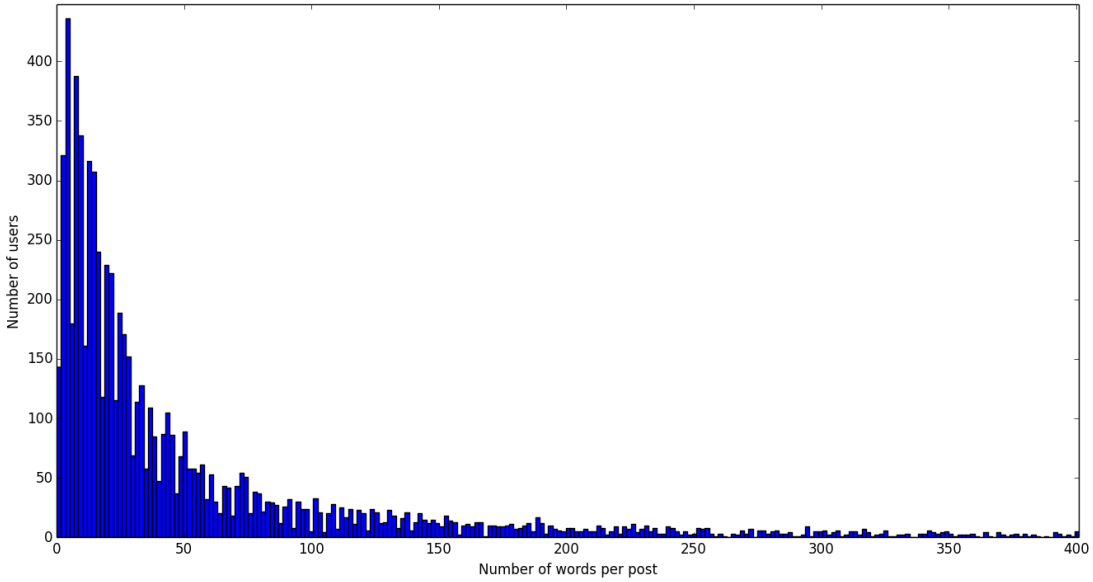

Figure 7: Histogram of the number of words per post.

4.3 Results Discussion

4.3.1 Baselines

The evaluation of the proposed baselines can be found in Table 3. The best results are obtained by the *Dynamic* strategy with a confidence threshold of 0.75. In terms of precision, the *First n* strategy offers better results. As we expect, the *Random* and *Minority* strategies have a poor performance. Also, it is noticeable that the best results of the ERDE metric are obtained by the *Dynamic* strategy. For this

reason, we choose this method (*Dynamic 0.75*) as our baseline to be compared with the models we propose.

Table 3: Baselines (linear regression). In bold, the best result for each evaluation metric.

	F1	**P**	**R**	**ERDE5**	**ERDE10**	**ERDE50**	**ERDE100**
Random	0.18	0.11	0.44	13.05%	12.95%	12.95%	12.95%
Minority	0.23	0.13	**1.00**	11.40%	11.17%	11.17%	11.17%
First 10	0.46	**0.81**	0.32	12.87%	11.17%	9.46%	9.46%
First 100	0.72	0.73	0.71	13.17%	13.15%	11.53%	8.80%
First 500	0.72	0.64	0.83	13.29%	13.27%	11.65%	11.42%
Dynamic 0.5	0.68	0.54	**0.93**	9.94%	8.81%	**5.68%**	**3.63**%
Dynamic 0.75	**0.72**	0.67	0.78	**9.89%**	**8.76%**	6.97%	5.80%
Dynamic 0.9	0.63	0.66	0.61	11.04%	10.19%	8.65%	7.40%

4.3.2 Features Individual Effect

In order to understand the performance of the proposed features in Section 3.2, we are going to add, separately, these features to our best model in the baselines. This model works with the dynamic strategy using logistic regression as learning algorithm with a probability threshold of 75%, using the *tf·idf* vectorisation (TFIDF). To choose the parameters for the LR algorithm we used the cross-validation technique as we explained in 4.2.1.

These features are: topics (LDA), sentiment polarity (SPOL), interval post frequency (IPF), removed posts ratio (REM), medicines ratio (MED), the LIWC dictionary categories and the special created anorexia category. The last two features are grouped as one type of feature since both of them are categories of words. We are going to refer them as LIWC. The results of these experiments are shown in Table 4.

Comparing with our baseline (see Table 3), the IPF and SPOL features have worsened the results (especially the SPOL) for all the measures except for the ERDE100 and the recall respectively. We will not include the IPF and SPOL features in our final model.

Table 4: Experiments to test the performance of the features extracted. In bold, the best result for each evaluation metric.

	F1	**P**	**R**	**ERDE5**	**ERDE10**	**ERDE50**	**ERDE100**
LR + TFIDF + IPF	0.70	0.65	0.76	11.35%	10.12%	7.54%	5.37%
LR + TFIDF + SPOL	0.28	0.16	**0.98**	13.58%	11.82%	8.84%	8.52%
LR + TFDIF + REM	0.72	0.67	0.78	10.40%	8.88%	7.05%	5.94%
LR + TFIDF + MED	0.72	0.67	0.78	10.40%	8.86%	7.05%	5.94%
LR + TFIDF + LIWC	**0.72**	**0.67**	0.78	10.40%	8.86%	7.05%	5.94%
LR + TFIDF + LDA	0.64	0.51	0.88	**8.19%**	**7.55%**	**4.21%**	**4.13%**

For the REM, MED and LIWC features, the results are very close to the baseline in terms of ERDE and exactly the same for the rest of measures. Given that these features do not improve neither worsen the results, we have tested these features together by adding them to our baseline (REM, MED and LIWC features). We obtained the same results for all the measures as when they are tested independently.

In any case, we are going to include the LIWC feature in our final models given its success in other studies based on the detection of mental illnesses through social media [9,12]. From these studies, we have seen that there exists significant differences on the language usage of diagnosed and control users. LIWC is a validated tool for the psychometric analysis of language data [43], which is able to reflect the language correlation of attentional focus, social relationships, emotional mood, individual differences and styles of thinking.

After analysing the training set, we observed that the 60% of the positive users and the 71% of the negative users have removed at least one post. These statistics tell us that there is no clear pattern among users that eliminate posts and the anorexia and we will not include the REM feature in our final models.

We have also observed that the 45% of the positive users and the 5% of the negative users have mentioned at least one medicine or a specific type of medicine. The differences are relevant and the MED feature will be considered in our final models.

The LDA feature has improved significantly the time-aware measures and the recall, but has decreased the precision and, in consequence, the F_1 score. This feature will be considered in our final models since it has improved some of the baseline results.

4.3.3 Proposed Methods

In Table 5 we report the results obtained after running the learning algorithms described in Section 3.3 in combination with best features concluded in Section 4.3.2. We can see that the usage of features based in the LIWC dictionary, the MED feature, the TLT value and the LDA features have improved the results of the baseline, in terms of the F1 score (0.78, increment of 0.06) and ERDE100 (4.99%, decrease of 0.81%), using the same learning algorithm (LR).

For the support vector machine classifier, we can see that the algorithm improves significantly the baseline in terms of the F1 score (0.85, increment of 0.13), and ERDE100 (4.22%, increment of 1.58%). We consider this as our best model.

In the case of the random forest algorithm, the results have worsened in terms of the F1 score having a low Recall value. In terms of the ERDE it does not get better results compared to our baseline.

With the multilayer perceptron algorithm, the F1 score has improved (0.78, increment of 0.06), with a higher precision (0.82, increment of 0.15), but the Recall value has decreased compared to the dynamic strategy of the baseline.

Table 5: Proposed classifiers with the modified Dynamic strategy. In bold, the best result for each evaluation measure.

	F1	**P**	**R**	**ERDE5**	**ERDE10**	**ERDE50**	**ERDE100**
LR + LDA + TFIDF + LIWC + MED	0.78	0.79	0.76	13.13%	13.12%	7.74%	4.99%
SVM + LDA + TFIDF + LIWC + MED	**0.85**	**0.85**	**0.85**	**13.05%**	**13.04%**	**7.26%**	**4.22%**
RF + LDA + TFIDF + LIWC + MED	0.62	0.70	0.56	13.21%	13.21%	9.65%	7.69%
MLP + LDA + TFIDF + LIWC + MED	0.78	0.82	0.76	13.09%	13.09%	8.01%	5.28%

Referring to our results, we can see that for all our models the ERDE5 and ERDE10 scores are high due to the TLT value that we have defined in order to decrease the number of false positives. In this sense, we have tested the best algorithm (SVM) results, with the same features but without the TLT restriction. Table 6 reports these results. The most remarkable thing is the improvement in all the ERDE measures, which is important considering our attempt to detect positive cases as early as possible.

Table 6: Comparison between the best model (SVM + LDA + TFIDF + LIWC + MED) with TLT and without TLT. In bold, the best result for each evaluation metric.

	F1	**P**	**R**	**ERDE5**	**ERDE10**	**ERDE50**	**ERDE100**
SVM (TLT)	**0.85**	**0.85**	0.85	13.05%	13.04%	7.26%	4.22%
SVM (NO TLT)	0.80	0.73	**0.88**	**8.44%**	**7.29%**	**4.19%**	**3.87%**

Chapter 5

Conclusion

In this work we proposed models for the early detection of cases of anorexia on social media. We presented a temporal-aware approach, which aims to penalise the delay in detecting positive cases.

Firstly, we established one initial hypothesis that we aimed to study. For this purpose, different machine learning models were built and tested. Concretely, we analysed the performance of the logistic regression, random forest, support vector machine and multilayer perceptron learning algorithms.

These models were fed with features based on linguistic information, domain-specific vocabulary, topics, psychological processes and analysis of medicines used in pharmacological treatment of AN, which prove to be the best. Additionally, we extracted other specific features for our dataset and we show how they do not contribute to a better performance of our models.

Finally, we defined a minimum amount of information, text length threshold (TLT), that should be seen by the system before applying the classifier and emitting a positive decision. This threshold helps to avoid some of the false positive predictions in our models but they have a later prediction, so the ERDE measure gets worse. This fact needs to be studied deeply in order to understand how much information is needed to predict optimally in the less time possible. Maybe, will be interesting

to search for a dynamic method that modifies the threshold depending on some variables or features that we could take into account.

We conclude that, in terms of the F_1 and ERDE, our best results have shown that the proposed approaches are suitable for the early detection of anorexia, as they clearly improve our baselines.

Chapter 6

Future Research

As a future work, new features and learning algorithms are going to be tested. For instance, other techniques should be explored in order to do the topic modelling, the use of more complex and specific dictionaries or the use of natural language techniques. Additionally, we will study in depth the introduction of voting methods and the use of genetic algorithms, as they have been previously applied with success in similar cases [27]. We also plan to investigate how to overcome the trade-off between avoiding the prediction of false positives and reducing the time needed to do the prediction. This means that we will explore a way to get a better precision without getting a high ERDE.

Finally, we will analyse other social media platforms providing different message formatting and information, such as Twitter.

Chapter 7

Publications

This thesis was summarised in a paper which is accepted for presentation in the 5th International Conference *Internet Science (INSCI 2018)*. The conference will take place in St. Petersburg, Russia, on October 24 to 26, 2018.

The conference, hosted by St. Petersburg University, the second largest university in Russia, is a focal point linking the academic and industrial communities in their evaluation of best Internet practices. The INSCI 2018 theme is *Internet in World Regions: Digital Freedoms and Citizen Empowerment*.

The reference of the paper is the following:

Diana Ramírez-Cifuentes, Marc Mayans and Ana Freire. Early Risk Detection of Anorexia in Social Media. International Conference on Internet Science – INSCI 2018. 2018.

List of Figures

List of Tables

Bibliography

[1] Agresti, A. Categorical Data Analysis. Wiley Series in Probability and Statistics. Wiley, 2013.

[2] Arseniev-Koehler, A., Lee, H., McCormick, T., and Moreno, M. Proana: Pro-eating disorder socialization on Twitter.

[3] Association., A. P., and Association., A. P. Diagnostic and statistical manual of mental disorders : DSM-5, 5th ed. ed. American Psychiatric Association Arlington, VA, 2013.

[4] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. J. Mach. Learn. Res. 3 (Mar. 2003), 993–1022.

[5] Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory (1992), ACM, pp. 144–152.

[6] Breiman, L. Random forests. Machine Learning 45, 1 (Oct 2001), 5–32.

[7] Campbell, K., and Peebles, R. Eating Disorders in Children and Adolescents: State of the Art Review.

[8] Chancellor, S., Lin, Z., Goodman, E. L., Zerwas, S., and De Choudhury, M. 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 (New York, NY, USA, 2016), CSCW '16, ACM, pp. 1171–1184.

[9] Coppersmith, G., Dredze, M., and Harman, C. Quantifying mental health signals in Twitter.

[10] Coppersmith, G., Dredze, M., Harman, C., and Hollingshead, K. From adhd to sad: Analyzing the language of mental health on Twitter through self-reported diagnoses. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (2015), pp. 1–10.

[11] Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., and Mitchell, M. Clpsych 2015 shared task: Depression and PTSD on Twitter. In CLPsych@HLT-NAACL (2015).

[12] Coppersmith, G., Harman, C., and Dredze, M. Measuring post traumatic stress disorder in Twitter. 579–582.

[13] Cortes, C., and Vapnik, V. Support-vector networks. Machine Learning 20, 3 (Sep 1995), 273–297.

[14] Cox, D. R. The regression analysis of binary sequences. Journal of the Royal Statistical Society. Series B (Methodological) (1958), 215–242.

[15] Csáji, B. C. Approximation with artificial neural networks. Faculty of Sciences, Etvs Lornd University, Hungary 24 (2001), 48.

[16] Cybenko, G. Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems 2, 4 (1989), 303–314.

[17] De Choudhury, M., Counts, S., and Horvitz, E. Social media as a measurement tool of depression in populations. In Proceedings of the 5th Annual ACM Web Science Conference (2013), ACM, pp. 47–56.

[18] De Choudhury, M., Counts, S., Horvitz, E. J., and Hoff, A. Characterizing and predicting postpartum depression from shared Facebook data. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (New York, NY, USA, 2014), CSCW '14, ACM, pp. 626–638.

[19] De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. Predicting depression via social media. AAAI.

[20] de Zwaan, M., and Roerig, J. Pharmacological treatment of eating disorders. Eating Disorders, Volume 6 (2003), 223–314.

[21] Elberrichi, Z. Text mining using n-grams, 01 2006.

[22] Gilbert, C. H. E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf (2014).

[23] Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., and Eichstaedt, J. C. Detecting depression and mental illness on social media: an integrative review. Current Opinion in Behavioral Sciences 18 (2017), 43 – 49. Big data in the behavioural sciences.

[24] Kang, K., Yoon, C., and Kim, E. Y. Identifying depressive users in Twitter using multimodal analysis. In Big Data and Smart Computing (BigComp), 2016 International Conference on (2016), IEEE, pp. 231–238.

[25] Kaslow, F. W., Patterson, T., and Gottlieb, M. Ethical dilemmas in psychologists accessing internet data: Is it justified? Professional Psychology: Research and Practice 42, 2 (2011), 105.

[26] Keski-Rahkonen, A., and Mustelin, L. Epidemiology of eating disorders in europe: prevalence, incidence, comorbidity, course, consequences, and risk factors.

[27] Leiva, V., and Freire, A. Towards suicide prevention: Early detection of depression on social media. In INSCI (2017).

[28] Liu, D. C., and Nocedal, J. On the limited memory BFGS method for large scale optimization. Mathematical programming 45, 1-3 (1989), 503–528.

[29] Losada, D., and Crestani, F. A test collection for research on depression and language use. In Proc. of Experimental IR Meets Multilinguality, Multimodality, and Interaction, 7th International Conference of the CLEF Association, CLEF 2016 (Evora, Portugal, September 2016), pp. 28–39.

[30] Masías, V. H., Valle, M., Morselli, C., Crespo, F., Vargas, A., and Laengle, S. Modeling verdict outcomes using social network measures: the Watergate and Caviar network cases. PloS one 11, 1 (2016), e0147248.

[31] Maximilian, F. M., and Norbert, Q. Mortality in eating disorders - results of a large prospective clinical longitudinal study. International Journal of Eating Disorders 49, 4, 391–401.

[32] Park, M., Cha, C., and Cha, M. Depressive moods of users portrayed in Twitter. 1–8.

[33] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. The Development and Psychometric Properties of LIWC2007. This article is published by LIWC Inc, Austin, Texas 78703 USA in conjunction with the LIWC2007 software program.

[34] Preoţiuc-Pietro, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., Schwartz, H. A., and Ungar, L. The role of personality, age, and gender in tweeting about mental illness. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (2015), Association for Computational Linguistics, pp. 21–30.

[35] Prieto, V. M., Matos, S., Alvarez, M., Cacheda, F., and Oliveira, J. L. Twitter: A good place to detect health conditions. In PloS one (2014).

[36] Ramchoun, H., Amine, M., Idrissi, J., Ghanou, Y., and Ettaouil, M. Multilayer perceptron: Architecture optimization and training. IJIMAI 4, 1 (2016), 26–30.

[37] Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., and Langer, E. J. Forecasting the onset and course of mental illness with Twitter data. In Scientific Reports (2017).

[38] Resnik, P., Garron, A., and Resnik, R. Using topic modeling to improve prediction of neuroticism and depression in college students. 1348–1353.

[39] Rosenblatt, F. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Tech. rep., CORNELL AERONAUTICAL LAB INC BUFFALO NY, 1961.

[40] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning internal representations by error propagation. Tech. rep., California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[41] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Park, G., Sap, M., Stillwell, D., Kosinski, M., and Ungar, L. H. Towards assessing changes in degree of depression through Facebook.

[42] Stern, S. R. Studying adolescents online: a consideration of ethical issues. In Readings in virtual research ethics: Issues and controversies. IGI Global, 2004, pp. 274–287.

[43] Tausczik, Y. R., and Pennebaker, J. W. The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology 29, 1 (2010), 24–54.

[44] Treasure, J., and Russell, G. The case for early intervention in anorexia nervosa: theoretical exploration of maintaining factors. British Journal of Psychiatry 199, 1 (2011), 5–7.

[45] Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., and Ohsaki, H. Recognizing depression from Twitter activity. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (New York, NY, USA, 2015), CHI '15, ACM, pp. 3187–3196.

[46] Wang, T., Brede, M., Ianni, A., and Mentzakis, E. Detecting and characterizing eating-disorder communities on social media. In WSDM (2017).

[47] Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., and Bao, Z. A depression detection model based on sentiment analysis in micro-blog social network. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (2013), Springer, pp. 201–213.

[48] Wijbrand, H. H., and van Hoeken Daphne. Review of the prevalence and incidence of eating disorders. International Journal of Eating Disorders 34, 4, 383–396.

[49] Wilson, J. L., Peebles, R., Hardy, K. K., and Litt, I. F. Surfing for thinness: a pilot study of pro-eating disorder web site usage in adolescents with eating disorders. Pediatrics 118 6 (2006), e1635–43.

[50] Zhang, L., Huang, X., Liu, T., Chen, Z., and Zhu, T. Using linguistic features to estimate suicide probability of chinese microblog users. CoRR abs/1411.0861 (2014).