Detecting Mental Disorders in Social Media Through Emotional Patterns - The Case of Anorexia and Depression

Mario Ezra Aragón, Adrian Pastor López-Monroy, Luis Carlos González-Gurrola, and Manuel Montes-y-Gómez

Abstract—Millions of people around the world are affected by one or more mental disorders that interfere in their thinking and behavior. A timely detection of these issues is challenging but crucial, since it could open the possibility to offer help to people before the illness gets worse. One alternative to accomplish this is to monitor how people express themselves, that is for example what and how they write, or even a step further, what emotions they express in their social media communications. In this article, we analyze two computational representations that aim to model the presence and changes of the emotions expressed by social media users. In our evaluation we use two recent public data sets for two important mental disorders: Depression and Anorexia. The obtained results suggest that the presence and variability of emotions, captured by the proposed representations, allow to highlight important information about social media users suffering from depression or anorexia. Furthermore, the fusion of both representations can boost the performance, equalling the best reported approach for depression and barely behind the top performer for anorexia by only 1 percent. Moreover, these representations open the possibility to add some interpretability to the results.

Index Terms—Mental disorders, emotional patterns, machine learning

1 Introduction

A mental disorder causes different interferences in the thinking and behavior of the affected person [1]. These interferences could vary from mild to severe, and could result in an inability to live routines in daily life and ordinary demands [2]. Common mental disorders such as depression and anorexia affect millions of people around the world. They may be related to a single incident causing excessive stress on the person or by a series of different stressful events. It is also well known that mental disorders tend to increase in countries experiencing generalized violence or recurrent natural disasters. For example, in 2018 a study of mental disorders in Mexicorevealed that 17 percent of its population has at least one mental disorder and one in four will suffer a mental disorder at least once in their life [3]. In another vein, in the modern world, we take for granted that social life could be experienced either in the physical world or in a virtual world created by social media platforms like Facebook, Twitter, Reddit, or similar platforms. This reality presents some challenges, but also great opportunities which, if properly addressed, could contribute to the understanding of *what* and *how* we communicate. In this regard, the goal of this study is to analyze, via the automatic identification of emotional patterns, social media documents[1] with the purpose of detecting the presence of signs of depression or anorexia in the population of that area [4], [5], [6]. Previous works have addressed the analysis of emotions of social media users by paying attention to their contrast and tone. They have mainly applied this analysis to predict users' age and gender as well as a range of sensitive personal attributes including sexual orientation, religion, political orientation [7], [8], income [9], and personality traits [10], [11]. According to these studies, the analysis of emotions in social media allows capturing important information related to users. This information presents an opportunity for us to extend the use of emotions in the detection of depression and anorexia in social media.

Former studies focused on the detection of depression and anorexia have mainly considered linguistic and sentiment analysis [12], [13], [14]. Note that the use of sentiments, i.e., polarity, was the preamble for the later use of emotions for the same task [15]. This line of thought exposed the potential of using emotions as features, such as "anger", "surprise" or "joy", instead of linguistic features or general sentiments like positive and negative. In this direction, in our previous work [16], we introduced a novel representation that was built using information extracted from emotions lexicons combined with word embeddings as a way to represent the information contained in users' documents. Then, using a clustering algorithm,

- *Mario Ezra Aragón and Manuel Montes-y-Gómez are with the National Institute of Astrophysics Optics and Electronics (INAOE), Puebla 72840, Mexico. E-mail: aragon.ezra@gmail.com, mmontesg@inaoep.mx.*
- *Adrian Pastor López-Monroy is with the Mathematics Research Center (CIMAT), Guanajuato 36023, Mexico. E-mail: beiceman@gmail.com.*
- *Luis Carlos González-Gurrola is with the School of Engineering, Autonomous University of Chihuahua, Chihuahua 22620, Mexico. E-mail: lcgonzalez@uach.mx.*

Manuscript received 21 May 2020; revised 13 Mar. 2021; accepted 14 Apr. 2021. Date of publication 27 Apr. 2021; date of current version 28 Feb. 2023. (Corresponding author: Mario Ezra Aragón.)
Recommended for acceptance by E. Mower Provost.
Digital Object Identifier no. 10.1109/TAFFC.2021.3075638

1. In this work, we refer as "document" to the concatenation of the posts of each user.

we created sub-groups of emotions, that conveniently we named as sub-emotions. These discovered sub-emotions provided a more flexible and fine-grained representation of users and a better performance for the detection of depression. In a few words, the idea behind this representation was to capture the presence of sub-emotions in users' posts. The intuition of our approach is that users suffering from depression would show a distribution of emotions different from healthy users. Motivated by the encouraging results of the representation based on sub-emotions, in this study we give a more complete treatment of the method. In particular, we propose a new representation that not only captures the presence of sub-emotions, but also models their changes over time. The intuition is to model emotional fluctuations that users with mental disorders could continuously present. This temporal information is later integrated to enrich the original approach. That is, we build a fusion of both representations, that at the end attains very competitive results, practically equal to those of the state-of-the-art approaches. Finally, we envision how these two representations can be applied beyond detecting depression to also detect other important mental disorder such as anorexia. Using this new representation we contrast emotional patterns between the two disorders, possibly finding what could be described as their emotional "silhouette".

The proposed static and dynamic representations, named as BoSE and Δ-BoSE respectively, are inspired in two hypotheses. The first one is that words assigned to coarse emotions in lexicons cannot capture subtle emotional differences, which in fact are what provide the most important insights into the mental health condition of users. For example, the lexicon associated with the anger emotion includes words such as *furious*, *angry* and *upset* that represent different degrees of anger, however, they are tagged with the same emotion. Thus, our proposal is to represent each user by a histogram of sub-emotions, which are discovered by clustering the embeddings of words inside coarse emotions. The second hypothesis is that people with depression and anorexia tend to expose greater emotional variability than a healthy person. In this case, the idea is to represent each user by a set of statistical values that describe the frequency changes of the sub-emotions over time. Based on these hypotheses, the contributions of this work for detecting people that have depression or anorexia are the following:

1) We further explore the BoSE representation and propose a new representation based on sub-emotions that allow capturing the emotional variability of social media users over time.
2) We propose an approach to combine both static and dynamic representations using early and late fusion strategies to improve the detection of depression.
3) We extend the use of these representations based on fine-grained emotions for the task of anorexia detection and contrast the discovered emotional patterns with those obtained from the task of depression detection.

The remainder of the paper is organized as follows: Section 2 presents a brief overview on the detection of mental health disorders using social media data. Section 3 describes in detail the creation of sub-emotions and how to convert text to these new sequences. Section 4 presents our emotion-based representations. Section 5 describes in detail our experiments, results, and their analysis. Finally, Section 6 presents our main conclusions.

2 Related Work

In this section we present an overview of previous works about the detection of depression and anorexia using social media data; we describe their strengths and opportunities, and contrast the strategies used to our proposal.

2.1 Depression Detection

Depression is a mental health disorder characterized by persistent loss of interest in activities, which can cause significant difficulties in everyday life [1], [17]. Studies focusing on the automatic detection of this disorder have used crowdsourcing as their main strategy to collect data from users who expressly have reported being diagnosed with clinical depression [18], [19]. Among these studies, the most popular approach considers words and word n-grams as features and employs traditional classification algorithms [13], [20], [21]. The main idea is to capture the most frequent words used by individuals suffering from depression and compare them against the most frequent words used by healthy users. This approach suffers because there is usually a high overlap in the vocabulary of users with and without depression.

Another group of works used a LIWC-based representation [22], aiming to represent users' posts by a set of psychologically meaningful categories like social relationships, thinking styles, or individual differences [18], [23]. These works have allowed a better characterization of the mental disorder conditions, nevertheless, they have only obtained moderately better results than using only the words. Recent works have considered ensemble approaches, which combine word and LIWC based representations with deep neural models such as LSTM and CNN networks [24], [25]. For example, in [25], [26], the combination of these models with features like the frequencies of words, user-level linguistic metadata, and neural word embeddings offered the best-reported result in the eRisk-2018 shared task on depression detection [27]. These works show that in social media texts exist useful information to determine if a person suffers from depression, but the results are sometimes hard to interpret. This is an important limitation since these types of tools are naturally aimed to support health professionals and not to take the final decisions. In [28], [29], the authors conduct studies to tackle this problem. They characterize users affected by mental disorders and provide methods for visualizing the data in order to provide useful insights to psychologists.

Lastly, some works have also considered representations based on sentiment analysis techniques [14], [30], [31]. These works have shown interesting results, indicating that negative comments are more abundant in people with depression than in users who do not suffer from this disorder. In a recent study [15], authors successfully proposed not only considering sentiments but also emotions to identify depression on Twitter users. This work was motivated by a psychological theory [32] that relates the manifestation of feelings and emotions with depression, and its objective

was to improve the interpretability of the results. In a previous work [16], we proposed to use a finer concept, called sub-emotions, reporting promising results to detect depression. Here, is where this study continues exploring this path, by proposing a new sub-emotion based representation, this time considering emotional changes through time, and also extending the potential use of this representation to detect anorexia.

2.2 Anorexia Detection

Anorexia nervosa is the most common eating disorder related to mental health. It is characterized by weight loss, difficulties maintaining appropriate body weight, and in many cases, a distorted body image. People with anorexia generally show abnormal attitudes towards food and unusual habits of eating. Also, they tend to exercise compulsively, purge via vomiting and laxatives, and binge eating.[2]

Some works have studied the manifestation of anorexia through social media content. For example, in [33], the authors proposed a method to automatically gather individuals who self-identified as eating disordered in their Twitter profile descriptions. They analyzed their social interactions and found that this kind of users has significant mixed patterns in tweeting preferences, language use, concerns of death, and emotions.

Regarding the automatic detection of anorexia in social media, several works have used syntactic and semantic features to characterize the structure and meaning of the posts [25], [34], [35], but these approaches suggest an overlap in the language used by users with anorexia and users without anorexia. Other works have applied sentiment analysis to study the emotional characteristics in the users' communications [12], [36]. Similar to depression, they mainly model the general sentiment (i.e., positive, negative, and neutral) that users express in their posts, and search for a relation between these sentiments and the signs of anorexia. Although this kind of methods have achieved some interesting results, they tend to fail in the classification of users without anorexia that usually express themselves in a negative way.

Recently, some works have also explored the use of deep learning techniques showing promising results [26], [37]. For example, in [38], the author proposed a solution based on neural networks, multi-task learning, domain adaptation, and Markov models. This work is still in its early stages, one of the issues is how to extend the mental health information resources. In a more recent work [39], the authors proposed a new neural network (NN) architecture consisting of 8 different neural sub-models, followed by a fusion component that concatenates the features and predict if a social media user has signs of anorexia. They concluded that the combination of the different models performs better than using them separately and that each kind of feature enriches the representation providing relevant information for the detection of anorexia. In line with these conclusions, the overview of the eRisk 2018 evaluation task [27] indicates that the best result was achieved by an approach that combines user-level linguistic metadata, frequencies of words, neural word embeddings, and a convolutional neural network. Notwithstanding the performance of these approaches, their complex designs and training frameworks make difficult to have good interpretability to better understand the severity of the disorder or support a preliminary diagnosis with newly discovered evidence. In [40], the authors develop a deep learning classifier that jointly models textual and visual characteristics that helps the detection of pro-eating disorder content that violates community guidelines. They used a million Tumblr posts to discovers deviant content in them. However, it is worth mentioning the work in [41], where the authors found that predicting a user with a mental disorder using their social media information although offers strong internal validity, suffers from external validity when tested on mental health patients; demonstrating that there is still work to be done in this area.

3 From Texts to Fine-Grained Emotions

Emotions are pervasive among humans and had widely been studied in different fields like psychology and neuroscience [42]. In particular, in psychology the correlation between emotions and mental disorders has been established, and how they manifest themselves in language through words is an active research area [14]. Supported by these insights is how we come to evaluate emotions, or more precisely, sub-emotions as an approach to effectively identify depression and anorexia on Reddit's users.

The proposed method for the detection of depression and anorexia considers representations of documents based on their expressed fine-grained emotions. In order to construct these representations, first, we generate groups of fine-grained emotions (referred as sub-emotions from here on) for each general emotion that belong to the EmoLEX lexicon [43]. This lexical resource indicates the association of words with eight emotions: Anger, Fear, Anticipation, Trust, Surprise, Sadness, Joy, and Disgust, and two sentiments: Negative and Positive.[3] The words were manually annotated and are available in 40 different languages. Then, we mask the text and represent each document using the sub-emotions labels instead of the original words. The following sections described in detail each step of this procedure.

3.1 Generating the Sub-Emotions

We represent the set of emotions within EmoLex in a formal way as $E = \{E_1, E_2, \ldots, E_{10}\}$, where $E_i = \{t_1, .., t_n\}$ is the set of words associated to emotion E_i.[4] We compute a vector for each word in the lexical resource using Wikipedia pre-trained sub-word embeddings of size 300 from FastText [44]. We empirically evaluated the vector size considering 100, 300, 500 as options, as well as word2vec[45], glove [46] word embeddings. After computing the vectors for each word (from each coarse emotion), we cluster them using the *Affinity Propagation (AP) algorithm*, which is a graph based clustering algorithm similar to k-means, but that does not require to establish the number of clusters in advance. This algorithm

2. https://www.nationaleatingdisorders.org/learn/by-eating-disorder/anorexia

3. In the rest of the paper we refer to these sentiments as emotions as well.

4. In the lexicon there are some words associated to more than one emotion.

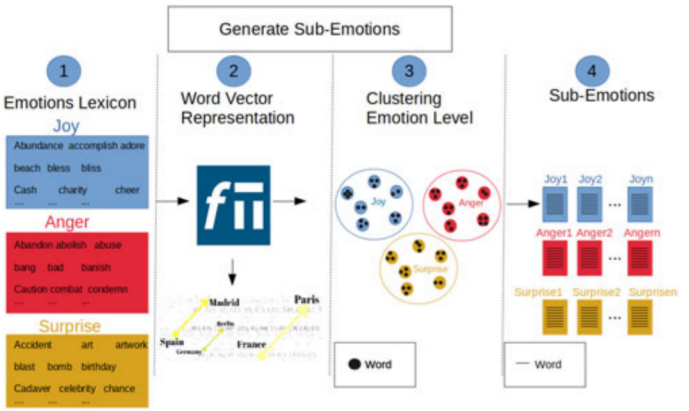

Fig. 1. Procedure to generate the sub-emotions for each emotion from the given lexical resource.

finds examples of members of the input set that are representative of clusters [47]. After the clustering, each centroid represents a different sub-emotion. That is, now each emotion is modeled as a set of sub-emotions, $E_i = \{S_1, \ldots, S_k\}$, where each S_j is a subset of the words from E_i. This process creates a set $\mathbb{S}$ with all computed sub-emotions. Fig. 1 depicts the whole process to generate the sub-emotions.

For the sake of completeness of how the vocabulary was distributed among emotions, and the number of generated clusters (sub-emotions) after applying the AP algorithm, Table 1 presents some statistics. It is interesting to notice that the average number of words per cluster (μW) is similar for all emotions, indicating that AP could find similar cluster distributions even for emotions with large vocabularies. For further analysis we also computed the average and standard deviation of the internal cohesion (μCoh and σCoh) for each emotion. The internal cohesion is a metric used to calculate how similar an object is to its own cluster. We calculated this value measuring the cosine similarity of each word with respect to the others in the same cluster. Based on this metric, we observe that some clusters present some cohesion, perhaps due to the lexicon containing words with similar contexts and topics.

It is noteworthy that some clusters provide an easy understanding and interpretability. As it can be observed, the obtained sub-groups of words allow separating each coarse emotion in different topics. These topics help to identify and capture more specific emotions used or expressed

TABLE 1
Size of the Vocabulary for Each Emotion Presented in the Lexical Resources, and Number of Generated Clusters (Cls)

Coarse Emotion Stats		Discovered Sub-Emotions Stats				
Emotion	Vocabulary	Cls	μW	σW	μCoh	σCoh
anger	6035	444	13.60	16.53	0.2932	0.1588
anticip	5837	393	14.77	20.53	0.2910	0.1549
disgust	5285	367	14.4	21.29	0.2812	0.1601
fear	7178	488	14.70	23.36	0.2983	0.1455
joy	4357	318	13.70	21.25	0.2928	0.1638
sadness	5837	395	14.78	20.48	0.2911	0.1549
surprise	3711	274	13.54	28.68	0.2874	0.1626
trust	5481	383	14.31	21.59	0.2993	0.1609
positive	11021	740	14.89	24.53	0.2967	0.1466
negative	12508	818	15.29	23.75	0.2867	0.1417

Anger			Joy		
anger1	***anger2***	***anger3***	***joy1***	***joy2***	***joy3***
abomination	growl	battle	accomplish	bounty	charity
fiend	growling	combat	achieve	cash	foundation
inhuman	thundering	fight	gain	money	trust
abominable	snarl	battler	reach	reward	humanitarian
unholy	snort	fists	goal	wealth	charitable
Surprise			**Disgust**		
surprise1	***surprise2***	***surprise3***	***disgust1***	***disgust2***	***disgust3***
accident	art	magician	accusation	criminal	cholera
crash	museum	wizard	suspicion	homicide	epidemic
disaster	artwork	magician	complaint	delinquency	malaria
incident	gallery	illusionist	accuse	crime	aids
collision	visual	sorcerer	slander	enforcement	polio

Fig. 2. Examples of words grouped in different sub-emotions.

by the users in their posts. For example, Fig. 2 shows some word examples of sub-emotions that were automatically obtained using this approach. It can be observed that words with similar context tend to group together. We can also notice that even for the same emotion each group of words shows different topics. For example, for the *Surprise* emotion one group expresses surprise related to art and museums, whereas other groups have words that are related to accidents and disasters, and magic and illusion respectively. In another example, the *Anger* emotion has one group that is related to topics of fighting and battles and another group with topics related to loud noises or growls.

3.2 Converting Text to Sub-Emotions Sequences

To follow with the procedure, we concatenate all the individual posts of a user and create a single document for each user. Then, we mask all users' documents replacing their words with a label that represents its closest sub-emotion. For this, after clustering word vectors of each coarse emotion, we compute prototypical sub-emotion vectors by averaging (column-wise) word embeddings in each cluster. We use these prototypes to count each word in text as an occurrence of a specific sub-emotion. For example, going back to Fig. 2, the sub-emotion *surprise2* is represented by the average of the vectors from the words: art, museum, artwork, gallery and visual that are presented column-wise. Once obtained these vectors, for each word t in a text sample we measure its cosine similarity with all sub-emotions vectors S, and substitute it by a label $\tau(t)$ related to its closest sub-emotion.[5] That is,

$$\tau(t) = S_j : \max_{\forall S_j \in \mathbb{S}} sim(\vec{t}, \vec{S_j}). \quad (1)$$

To illustrate this process consider the following two example sentences:

1) `The most important thing is to try and inspire people.`
2) `Im not good enough for the task.`

These sentences will be masked as:

1) `anticipation27 joy27 positive5 negative62 anticipation10 anticipation29 positive20 negative80 trust23 joy16`

5. We assigned the labels selecting the name of the emotion followed by the sequential number. For example, for anger the labels were assigned as: anger1, anger2, ... ,angerK. Where K is the number of clusters in that emotion.

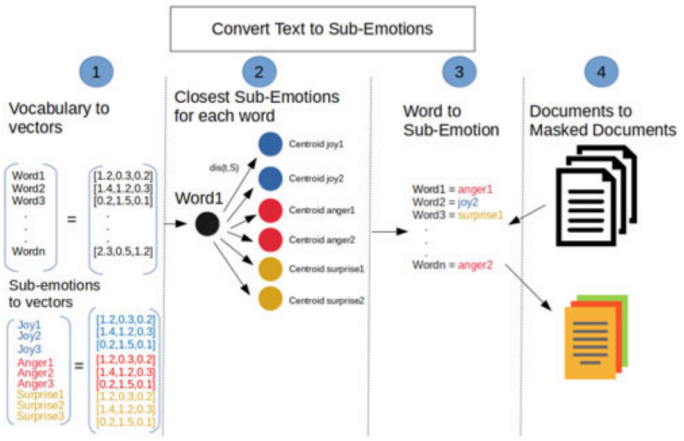

Fig. 3. Procedure to transform the texts to sub-emotions sequences.

(2) positive91 negative43 joy35 negative62 negative80 anticipation27 anticipation19

From these examples, it is possible to appreciate how different contexts are captured by different sub-emotions. It is important to mention that we replace the whole vocabulary of all users including stopwords with the closest sub-emotion. All this process is depicted in Fig. 3.

4 Emotion-Based Representations: BoSE and Δ-BoSE

4.1 Bag of Sub-Emotions: The BoSE Representation

After documents are masked, we build the *BoSE* representation by using histograms of sub-emotions. Basically, each document d is represented as a vector of weights associated to sub-emotions, $\vec{d} = \langle w_1, \ldots, w_m \rangle$, where m is the total number of generated sub-emotions and $0 \leq w_i \leq 1$ represents the relevance of sub-emotion S_i to the document d. This weight is computed in a *tf-idf* fashion as

$$w_i = freq(S_i, d) \cdot \log\left(\frac{|\mathcal{D}|}{\#_{\mathcal{D}}(S_i)}\right), \quad (2)$$

where $freq(S_i, d)$ represents a function that denotes the frequency of the sub-emotion S_i in the document d, $|\mathcal{D}|$ is the number of documents in the whole collection, and $\#_{\mathcal{D}}(S_i)$ is a function denoting the number of documents containing the sub-emotion S_i. As it can be seen, this representation only considers the presence of individual sub-emotions in the documents; thus we call it *BoSE-unigrams*. In the case of also considering the presence of sequences of sub-emotions, we named it *BoSE-ngrams*.

4.2 Δ-BoSE: A Dynamic Sub-Emotion Representation

One of the hypotheses of this work is that there exists some variability in how emotions are expressed by users with depression and anorexia. Following this intuition we propose a new representation to capture temporal emotional patterns; we named this representation Δ-*BoSE*.

To build the Δ-BoSE representation, we first divide the post history of each user in n parts or chunks.[6] Then, for each chunk we calculate its BoSE representation as described in Section 4.1. That is, we consider the chunks as individual but sequential documents. After this process, each one of the m sub-emotions is represented by a vector of n values, $\vec{S}_i = \langle w_{i,1}, \ldots, w_{i,n} \rangle$, where $w_{i,j}$ indicates the weight of sub-emotion S_i in the chunk j as determined by Formula (2).

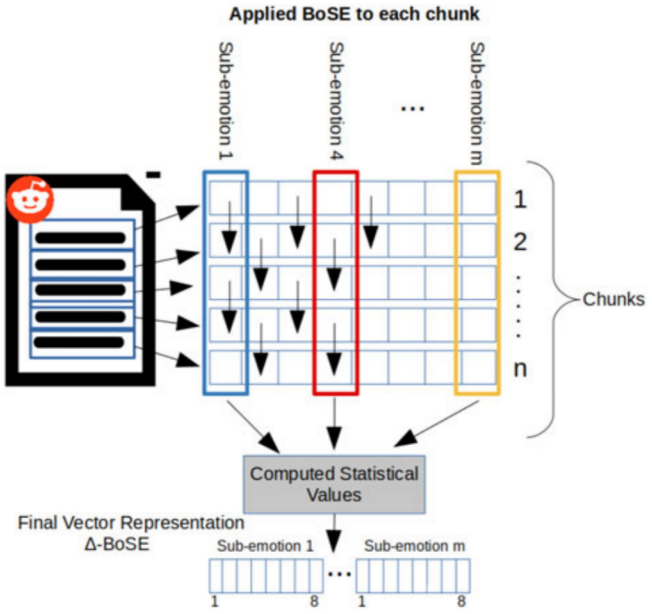

Fig. 4. Construction of the Δ-BoSE representation. First, BoSE is obtained for each part of the document; then, statistical values are calculated for each sub-emotion creating a new vector representation.

Given that our purpose is to model the temporal variability of the emotions, we decide to represent each sub-emotion by a Δ−vector of the following eight statistical values that capture its changes through the n-chunks sequence: mean(μ), sum ($\sum$), max-value(max), min-value(min), standard deviation (σ), variance(σ^2), average($\bar{x}$), and median($\tilde{x}$). This creates a new vector $\Delta\vec{S}_i = \langle \mu, \sum, max, min, \sigma, \sigma^2, \bar{x}, \tilde{x} \rangle$ that represents the changes of the sub-emotion S_i in the post history of the user. Finally, we concatenate the Δ-vectors from all sub-emotions in one single vector of size $8 \times m$, where m is the number of sub-emotions. Fig. 4 depicts this process.

5 Experiments and Results

5.1 Data Sets

To fully evaluate BoSE and Δ-BoSE we use the data sets from the eRisk 2018 evaluation tasks [27], [48]. These data sets contain the posts of several users from the Reddit platform. For each task, there are two categories of users: positive users, those who are affected by either anorexia or depression, and the control group, composed of people who do not suffer from any mental disorder. The positive class is composed by people who explicitly mentioned that they were diagnosed by a medical specialist with anorexia or depression, so users using vague expressions like "I think I have anorexia/depression" were discarded during the collection of the data. The control class is composed of random users from the Reddit platform.[7] It should be acknowledged that, to construct the positive group,

6. We consider 10 chunks similar to the e-Risk competition.

7. The creators of these data sets included in the control group users who often interact in the anorexia or depression threads to add more realism to the data and make it more realistic to detect positive users.

TABLE 2
Mental Dissorders Data Sets Used for Experimentation

Data set	Training		Test	
	P	C	P	C
Users dep eRisk'18	135	752	79	741
avg. num. posts	367.1	640.7	514.7	680.9
avg num. words per post	27.4	21.8	27.6	23.7
avg. activity period (days)	586.43	625.0	786.9	702.5
Users anor eRisk'18	20	132	41	279
avg. num. posts	372.6	587.2	424.9	542.5
avg num. words per post	41.2	20.9	35.7	20.9
avg. activity period (days)	803.3	641.5	798.9	670.6

(P = Positive, C = Control).

eRisk organizers first collected users using the specific searches previously mentioned. Through these searches they obtained self-expressions of depression or anorexia diagnosis, then, they manually reviewed the matched posts and verified if they were really genuine. This self-expression of depression or anorexia opens the possibility of having noise in both the control and positive group. This noise could also create some data bias in certain users of the data set that are more heavily represented than others. Table 2 shows how classes distributes within these data sets as well as some general information regarding the collections.

To offer a glimpse of the data sets, we present some examples of posts from the different classes of users. Our intention is to show that users who suffer from a mental illness as well as control users share personal experiences and their feelings about them, which for both can be positive and negative, making their identification a great challenge.

Depression

1) `After coming home from a road trip with a group of friends to celebrate my birthday.`
2) `Sometimes I can't help but think that they will be so much better off without me, and they know that they would be happier without me.`

Anorexia

1) `I'm happy to hear that you're okay with realizing you'll be on antidepressants for the rest of your life..`
2) `My coach looked over at me then muttered; "It's a shame. If she wasn't so BIG I'd consider her for the team.`

Control

1) `Nice job; it's not always easy with the clouds. I love the colors of those waters with the glacial moraine. Beautiful image.`
2) `It was difficult, I do not expect it to be well-received here, but even if one person find it useful i will continue.`

5.2 Experimental Settings

Preprocessing. The texts were normalized by lowercasing all words and removing special characters like urls, emoticons, and #; the stopwords were kept. Then, the preprocessed texts were masked using the created sub-emotions.

Classification. The main goal is to classify users into one of the two classes (Depressed/Control or Anorexia/Control). After building the *BoSE* representation, the most relevant features of the sequences of sub-emotions were selected using the term frequency – inverse document frequency (tf-idf) representation and chi^2 distribution X_k^2 [49]. With the selected features we fed a Support Vector Machine (SVM) with a linear kernel, $C = 1$, L2 normalization and weighted for class imbalance. We empirically search for the best number of features for each task, thus, we selected 3,000 features for depression and 1,500 for anorexia. We used the same number of features for BoSE and Δ-BoSE. The final prediction is using the whole post history of the users and we classify the user as positive if the SVM decides the example is closer to this class.

Baselines. Inspired in [15], we implemented a slightly different approach. That is, the original approach counts the exact presence of words from each emotion in each one of the posts, in our case we applied an approach similar to BoSE, masking the words with their more similar emotion. In other words, the original approach considers a hard matching of words from the lexicons, while ours considers a soft matching procedure. This approach is named Bag-of-Emotions (BoE). Also, the results are compared to the traditional Bag-of-Words representation. Both representations were created using word unigrams and n-grams; these are common baseline approaches for text classification. For both approaches, similar to BoSE and Δ-BoSE, we selected the same number of features using tf-idf representation and chi^2 distribution X_k^2, 3,000 for depression and 1,500 for anorexia. Similar to previous works in depression and anorexia detection, we add an LIWC-based representation using the categories as features. We also add some baselines based on deep learning approaches, using a CNN and a Bi-LSTM. The neural networks used 100 neurons, an adam optimizer, and word2vec and Glove embeddings with a dimension of 300. For the CNN we use 100 random filters of sizes 1, 2, and 3. Additionally, the obtained results are compared against the top-three participants of the eRisk 2018 evaluation tasks (these are explained in detail in sub-section E). For all this comparison we consider F_1 score over the positive class, which was suggested as the golden standard by the organizers of eRisk 2018 [27].

5.3 Evaluation of the BoSE Representation

In this study, we exhaustively evaluate BoSE-based representations, and we contrast them against the BoE and BoW schemes (using both unigrams and bigrams) and also against Deep Learning models (using Glove and word2vec) for the detection of Depression (eRisk '18) and Anorexia (eRisk '18). Table 3 presents the F_1 score over the positive class for this first evaluation. From this comparison, we appreciate that BoSE outperforms all baseline results, even in some cases with a good margin of difference (consider for instance the case of Anorexia). Surprisingly, the performance of deep learning models is remarkably low; to some extent this could be attributable to the small size of the employed data sets. Indeed, most participants of eRisk 2018 that employed this kind of models combined them with traditional approaches to leverage their results.

In order to analyze the obtained results, we plotted the users in a plane using both the BoW and the BoSE representations.