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

Mario Ezra Aragón, A. Pastor López-Monroy, Luis C. González, 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 study, 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%. Moreover, these representations open the possibility to add some interpretability to the results.**

***Index Terms*—Mental Disorders, Emotional Patterns, Machine Learning**

I. 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% 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]–[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]–[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, 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

M.E. Aragón and M. Montes-y-Gómez are with National Institute of Astrophysics Optics and Electronics, Puebla, Mexico.

A.P. López-Monroy is with Mathematics Research Center, Guanajuato, Mexico.

L.C. González is with Autonomous University of Chihuahua, School of Engineering, Chihuahua, Mexico.(e-mail: lcgonzalez@uach.mx)

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3075638, IEEE Transactions on Affective Computing

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.

II. 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.

A. 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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3075638, IEEE Transactions on Affective Computing

B. 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.

III. 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.

A. Generating the sub-emotions

We represent the set of emotions within EmoLex in a formal way as $E = \{E_1, E_2, ..., E_{10}\}$, 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 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, $E_i = \{S_1, ..., S_k\}$, where each S_j is a subset of the words from E_i. This process creates a set S with all computed sub-emotions. Figure 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

[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.

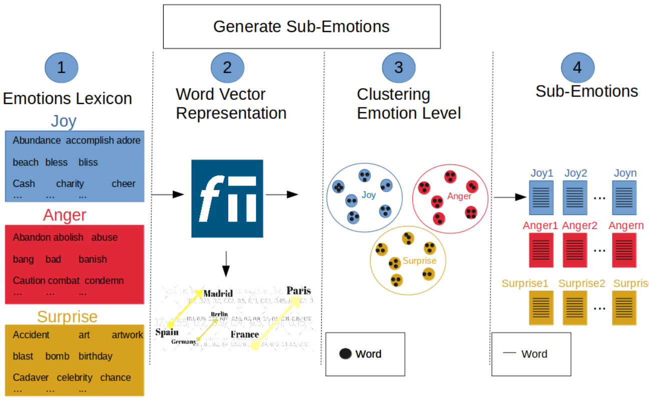


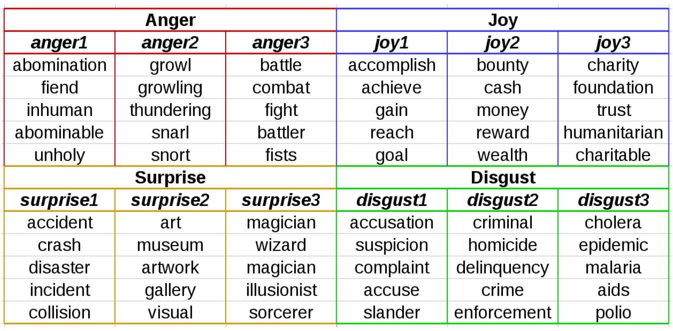
Figure 1. Procedure to generate the sub-emotions for each emotion from the given lexical resource.

clusters (sub-emotions) after applying the AP algorithm, Table I 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.

Table I
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

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 by the users in their posts. For example, Figure 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.



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

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

B. 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 Figure 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,

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

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`
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 Figure 3.

[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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3075638, IEEE Transactions on Affective Computing

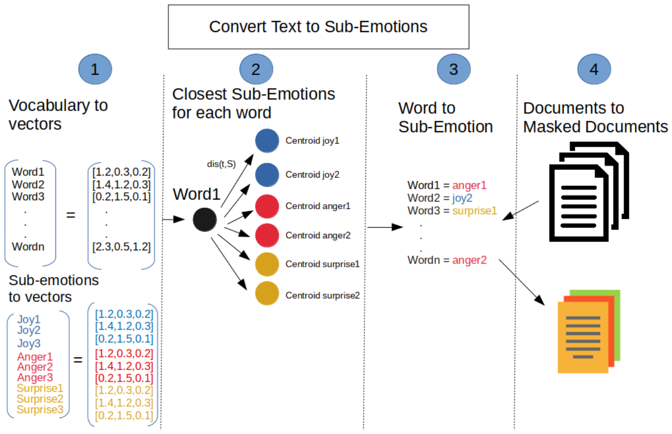


Figure 3. Procedure to transform the texts to sub-emotions sequences.

IV. Emotion-based Representations: BoSE and Δ-BoSE

A. 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, $\vec{d} = \langle w_1, \cdots, w_m \rangle$, 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**.

B. 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 IV-A. 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, $\vec{S_i} = \langle w_{i,1}, \cdots, w_{i,n} \rangle$, where $w_{i,j}$ indicates the weight of sub-emotion S_i in the chunk j as determined by Formula 2.

Given that our purpose is to model the temporal variability of the emotions, we decide to represent each sub-emotion by a $\Delta-$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. Figure 4 depicts this process.

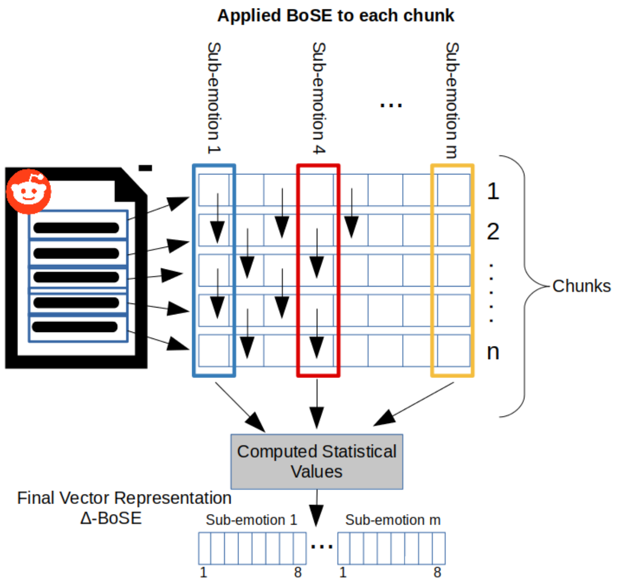


Figure 4. Construction of the Δ-BoSE representation. First, BoSE is obtained for each part of the document; then, statistical values are calculated for each sub-emotion creating a new vector representation.

V. Experiments and Results

A. 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, eRisk organizers first collected users using the specific searches previously mentioned. Through these searches they obtained self-expressions of depression or anorexia diagnosis,

[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 II
MENTAL DISSORDERS DATA SETS USED FOR EXPERIMENTATION. (P = POSITIVE, C = CONTROL)

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

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 II 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.

B. 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 3000 features for depression and 1500 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, 3000 for depression and 1500 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].

C. 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 III 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. To generate these visualizations we used the T-distributed Stochastic Neighbor Embedding (t-SNE) algorithm

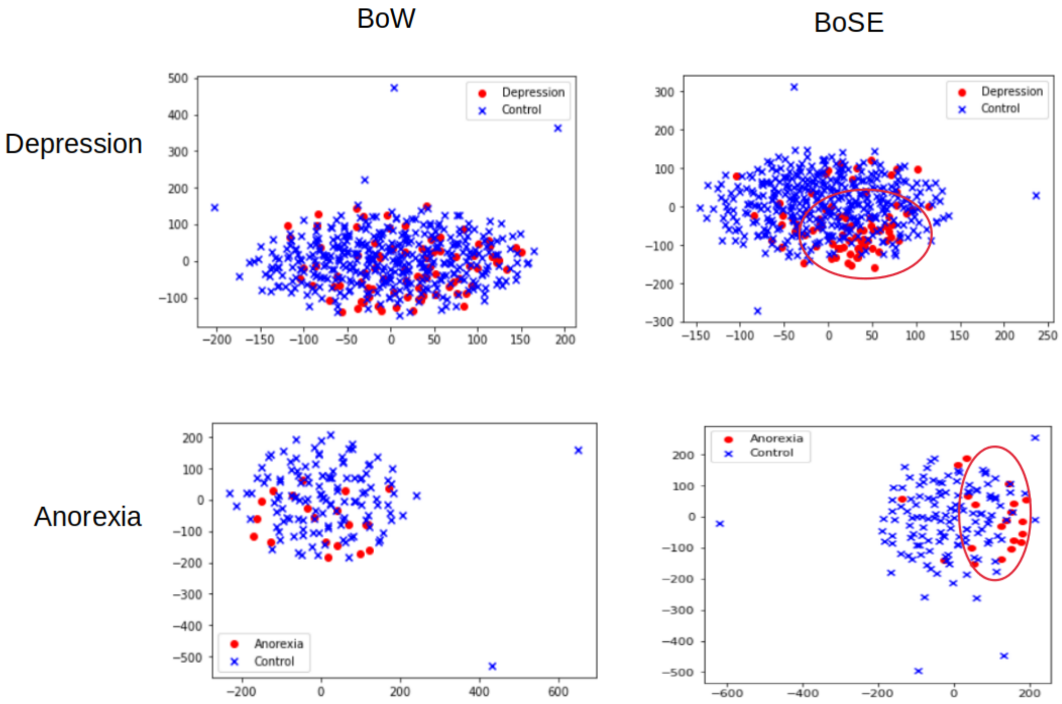


Figure 5. t-SNE visualization for BoW and BoSE representation in both tasks.

Table III
F_1 RESULTS OVER THE POSITIVE CLASS:
BOSE AND BASELINE METHODS

Method	Dep'18	Anor'18
BoW-unigrams	0.54	0.69
BoE-unigrams	0.60	0.50
BoSE-unigrams	0.61	**0.82**
BoW-ngrams	0.54	0.69
BoE-ngrams	0.58	0.58
BoSE-ngrams	**0.63**	0.81
LIWC	0.38	0.54
BiLSTM-Glove	0.46	0.46
BiLSTM-word2vec	0.48	0.56
CNN-Glove	0.51	0.54
CNN-word2vec	0.48	0.57

[50], which is a nonlinear dimensionality reduction technique well-suited for plotting high-dimensional spaces in a low-dimensional space. For this analysis we used the vector representation of 3000/1500 features obtained using tf-idf with chi^2 distribution for both BoSE and BoW (previously mentioned in Section V-B). Figure 5 offers an interesting perspective of the advantage of using BoSE over BoW to allow the classifier to build a better classification function. We even analyzed the boundary cases and found similar distribution in the sub-emotions, this could be due to the similarity in the topics captured by the sub-emotions that users posted and shared. For example, we have a user with the following paragraph: *"This philosophy is suicidal. And that's not bashing it, this is exactly what it is. Let's not forget Gandhi. "Whenever you are confronted with an opponent. Conquer him with love." Who would tell the Jews that suicide is more heroic than fighting for your life: "But the Jews should have offered themselves to the butcher's knife. They should have thrown themselves into the sea from cliffs."* This user is in the control group, where the user explicitly refers to suicide and violence, but in this context is referring to a philosophy that could be not his own ideals. Nevertheless, these examples present a challenge for the classifier."

Beyond classification performance over the complete user history, eRisk workshop also considers their early prediction. We perform an additional experiment to explore the effect of data size over our prediction. For this, the eRisk workshop offers the data in parts (chunks) and evaluates not only the effectiveness but the anticipation of the prediction. To assess how well BoSE performs in giving early predictions, we predict the users at each chunk and compare the results against baseline approaches. For each chunk, we perform a classification using the accumulative information of the users at that moment. The classifier assigns the label of positive (depression or anorexia) if their probability of being of that class is higher than 50%. Therefore, for each chunk, we have more evidence and the classifier made better decisions. Figure 6 presents a plot of the results from BoSE and baselines over each chunk of data. From this plot it can be observed that BoSE achieves an outstanding performance for the Anorexia data set, even with $F_1 = 0.56$ when only using the first chunk available, while the second best approach only reaches $F_1 = 0.34$. For the case of Depression, in most cases BoSE achieves the best prediction, making this more evident towards the inclusion of the latest chunks of data.

From the first round of experiments, we highlight the following observations:

1) BoSE outperformed the BOW representation in both

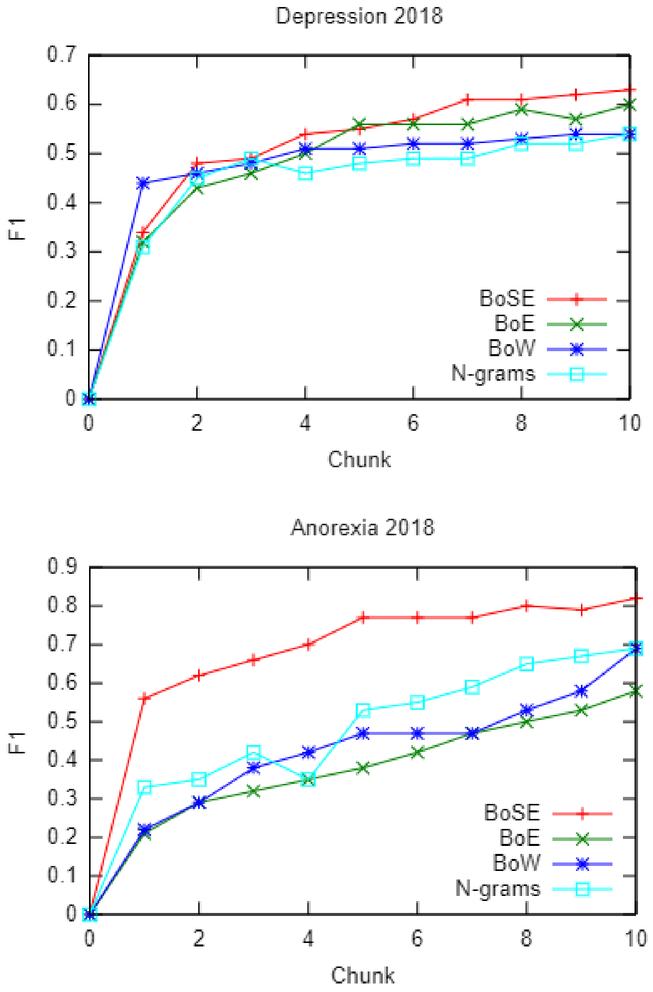


Figure 6. Results by chunk in the data sets. X-axis represent the chunks and Y-axis the F1 result.

tasks, indicating that considering emotional information is more relevant for the detection of depression and anorexia in online communications than only considering the use of words.

2) The use of sub-emotions as features helps to improve the results from a representation that only considers coarse emotions. This result confirms our hypothesis that such approach is more effective to capture subtle changes of emotions in users with depression and anorexia.
3) BoSE obtained competitive results in terms of F_1 over the positive class, even without considering all users' posts (with around 70% of the information), suggesting that useful emotional patterns can be detected from a few texts from the users.

D. Evaluation of the $\Delta-$BoSE representation

To evaluate the hypothesis that users with mental disorders experience more variability in their expressed emotions we evaluate different ways to enrich the BoSE representation to include this information. To do this we test Early and Late fusion schemes to incorporate $\Delta-$BoSE and BoSE within the same traditional classification process. Table IV presents the results of this experiment. We observe that only using BoSE is more informative than capturing sub-emotion changes over time. Nonetheless, when we fusion both representations, the performance of the classification improves, as it can be observed by the *Late Fusion* approach, where we combined the decision of two classifiers (one classifier using BoSE while the other uses $\Delta-$BoSE) by means of an OR gate over their corresponding decisions[8].

Table IV
F_1-SCORES FOR BOSE, Δ-BOSE AND THEIR COMBINATIONS

	Depression'18	Anorexia'18
BoSE	0.63	0.82
Δ-BoSE	0.53	0.79
Early Fusion	0.62	0.77
Late Fusion	**0.64**	**0.84**

E. Comparison against the eRisk participants

As it is described in the overview of the eRisk-2018 shared task [27], a total of 35 models were submitted to the anorexia detection task and 45 to the depression detection task, ranging from simple to advanced deep learning models. Particularly, the first place submitted results from four machine learning models and an ensemble model that combines the predictions of the four previous models. They employed user-level linguistic metadata, a bag of words representation, neural word embeddings from Glove, and a convolutional neural network [25]. The team in second place implemented a system based on two different models, one that considers the temporal variation of terms, and other that carries out an incremental classification. The first model uses a semantic representation of documents considering the explicit information available at each chunk, whereas, the second model does an incremental estimation of the association of each user to each class based on the accumulated information at each chunk [51] [9]. Table V shows how our approach (i.e., the fusion of BoSE and Δ-BoSE) compares against the top places at the eRisk 2018 evaluation tasks. We can observe that our approach achieves competitive results in both tasks. Nonetheless, it is important to mention that the participants focused on obtaining early and accurate predictions of the users, while our approach focuses exclusively on determining accurate classifications. On the other hand, in our previous experiment (refer to Figure 6), BoSE obtained competitive results in terms of F_1 over the positive class, even without considering all users' posts (with around 70% of the information). Beyond these results, the presented approach seems simpler and even opens the possibility to add some degree of understanding or interpretability to what has been captured by the classification model.

For a further analysis of these results, Figure 7 presents a boxplot of the F_1, precision, and recall scores of all participants from both tasks. The green **X** represents the results

[8] For Early Fusion we concatenate both representations and employed only one classifier.

[9] The team in third place did not send a report describing their approach.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3075638, IEEE Transactions on Affective Computing

Table V
F_1, *Precision* AND *Recall* RESULTS OVER THE POSITIVE CLASS: BOSE FUSION APPROACH AND THE TOP PERFORMERS AT ERISK

Task	Depression 2018			Anorexia 2018		
Metric	F1	P	R	F1	P	R
first place	**0.64**	0.64	0.65	**0.85**	0.87	**0.83**
second place	0.60	0.53	**0.70**	0.79	**0.91**	0.71
third place	0.58	0.60	0.56	0.76	0.79	0.73
Late Fusion	**0.64**	**0.67**	0.61	0.84	0.87	0.80

of the BoSE late fusion approach. We can appreciate that our results are in the highest quartile (except for recall in depression) for both tasks, indicating that the representation of the presence and changes of fine-grained emotions obtains competitive results in depression and anorexia detection.

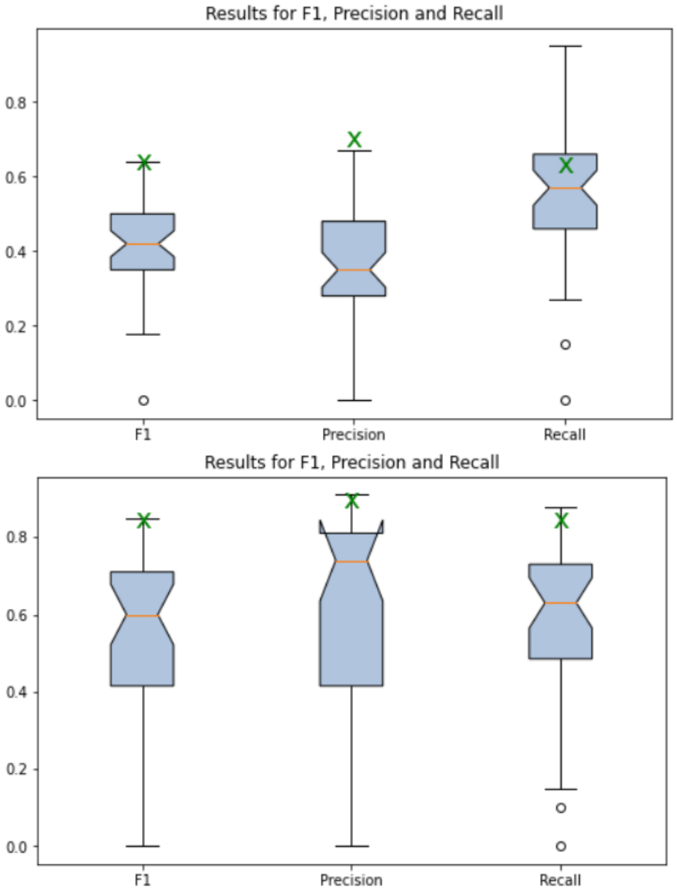


Figure 7. Boxplot of the F1 scores for depression (upper part) and anorexia (bottom part), where the green X represents our BoSE late fusion approach.

F. Is there a sub-emotion pattern for depression or anorexia?

Table VI shows some of the top relevant sub-emotions, according to the chi^2 distribution, as well as some examples of the words that correspond to these sub-emotions in the depression and anorexia tasks. Most of the sub-emotions that present high relevance for the detection of depression are related to negative topics, for example, the anger sub-emotions are associated with the feeling of abandonment or unsociability, and the disgust sub-emotions are related to delusion, insecurity, and desolation. These sub-emotions capture the way a depressed user expresses their opinions about the world. In contrast, for the anorexia detection task, the sub-emotions that present higher relevance are related to embarrassment, self-harm, and eating topics. For example, the anger sub-emotions are related to victimization, bleeding, or bruises. The disgust emotions are associated with mental states of defeat and internal organs related to eating. The latter clearly showed that these sub-emotions capture the essence of the problems of a person that suffers from anorexia. On the other hand, it is interesting to see that the intersection of anger is in sub-emotions that refer to lying or rejection.

Intending to recognize a possible emotional silhouette of both tasks, in Figure 8 we present a comparative histogram of their emotions computed from their 100 most frequent sub-emotions. We can observe that the distribution of the emotions, as captured from the sub-emotions, presents some differences between anorexia and depression. For example, we can appreciate the emotion *anger*, even when it is present in both tasks, users suffering from depression show more types of anger in their posts, as it is appreciated in Table VI. On the contrary, *fear* related sub-emotions are more frequent in users suffering from anorexia. There are also some emotions with similar frequencies in both groups of users, such as the case of the *disgust* emotion; nevertheless, their particular sub-emotions, as shown in Table VI, cover very different topics. We can conclude that the sub-emotions help us to find sub-groups of topics related to the different mental disorders and obtain the themes of the problems related to each task.

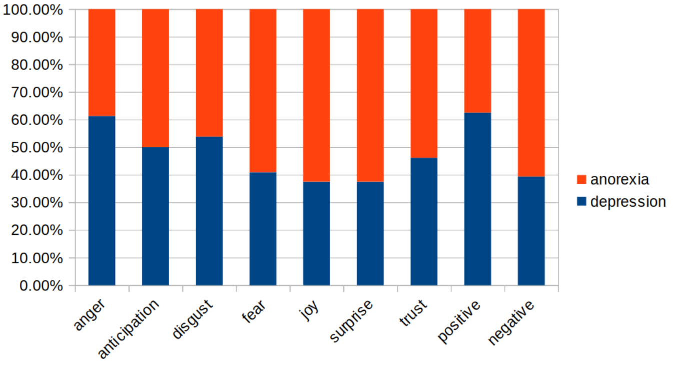


Figure 8. Emotions distribution for each task.

But, how these differences could be captured by Δ-BoSE?. In Figure 9 we compare the occurrences of certain sub-emotions over time (i.e., through the 10 chunks) in the control group (colored in orange) and the mental disorder group (colored in blue). We selected some of the top sub-emotions based on their chi-squared value for each task. These signals indicate the average occurrence of the sub-emotions in all users from each group. We observe that the control group, in both disorders, presents fewer changes or peaks through time than the mental disorder group, thus suggesting that emotional variability may exist and that it could be further exploited through emotion-based representations and Machine Learning methods to identify certain type of users.

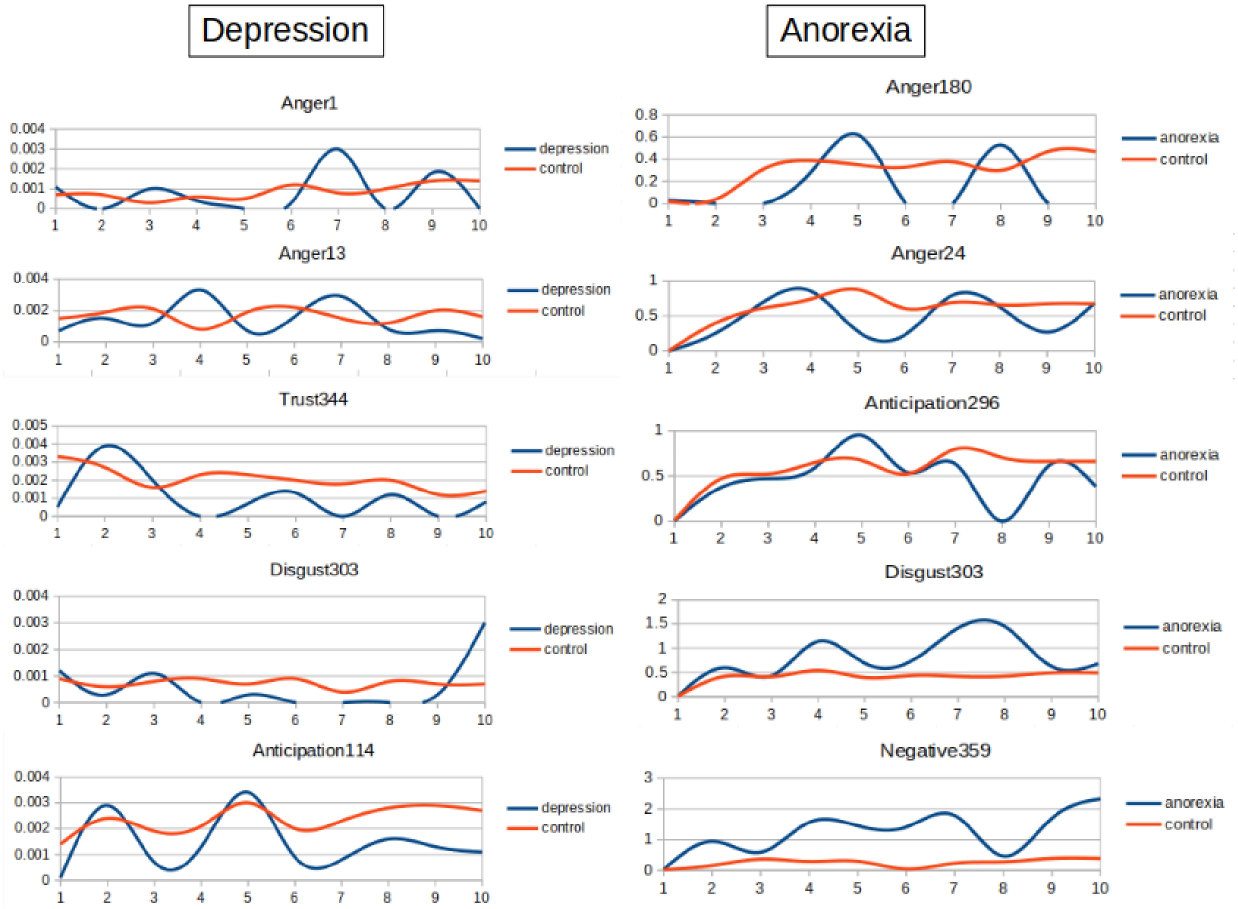


Figure 9. Comparison of the emotional signals between control and mental-disorder groups. X-axis represents the chunks (time span) and Y-axis represents the average value of the sub-emotion at each chunk.

VI. Conclusion

In this work, we showed that representations based on fine-grained emotions can capture more specific topics and issues that are expressed in social media documents by users that unfortunately experience depression or anorexia. That is, the automatically extracted sub-emotions present useful information that helps the detection of these two mental disorders. On the one hand, the BoSE representation obtained better results than the proposed baselines, including some deep learning approaches, and also improved the results of only using broad emotions as features. On the other hand, the inclusion of a dynamic analysis over the sub-emotions, called Δ-BoSE, improved the detection of users that presents signs of anorexia and depression, showing the usefulness of considering the changes of sub-emotions over time. It is worth mentioning the simplicity and interpretability of both representations, then creating a more straightforward analysis of the results. Finally, the capability to model the emotional behavior of users using their social media data presents an opportunity for future wellness facilitating technologies. This kind of technology can serve as warning systems that provide wide-area analysis and information related to a mental disorder respecting user privacy. This information could include the presence of mental disorders in certain areas, and the authorities could decide to create professional assistance or emotional support, that the users will decide whether to take or not. We believe that it is important to mention when we analyze social media content, we may have concerns regarding individual privacy or certain ethical considerations. These concerns appear due to the usage of information that could be sensitive, given the personal behavior and emotional health of the users. The experiments and usage of this data are for research and analysis only, and the misuse or mishandling of the information is prohibited.

Acknowledgment

The first author thanks for doctoral scholarship CONACyT-Mexico 654803.

References

[1] R. Kessler, E. Bromet, P. Jonge, V. Shahly, and Marsha., "The burden of depressive illness," *Public Health Perspectives on Depressive Disorders*, 2017.

[2] W. H. Organisation, "Mental health: Fact sheet," *https://www.euro.who.int/en/health-topics/noncommunicable-diseases/mental-health*, 2019.

[3] M. Renteria-Rodriguez, "Salud mental en mexico," *NOTA-INCyTU NÚMERO 007*, 2018.

Table VI
EXAMPLES OF RELEVANT SUB-EMOTIONS FOR ANOREXIA AND DEPRESSION DETECTION

Depression	
anger1	abandoned, deserted, unattended
anger11	unsociable, crowd, mischievous
anticip10	disappointed, inequality, infidelity
anticip99	desolate, seclude, inhospitable
disgust16	unsatisfactory, delusion, influence
disgust11	insecurity, desolation, incursion
fear17	hysterical, immaturity, injury
fear20	suffer, unhealed, sickness
joy1	abundant, abundance, plentiful
joy14	life, time, moment
negative10	accident, incident, fatal
negative11	accuse, complaint, wrongdoing
positive10	approval, authorize, approved
positive100	attention, notoriety, publicity
surprise8	anxious, desire, wanted
surprise20	distress, embarrassment, shame
trust9	completion, continuation, progress
trust20	ambition, desire, wanted
Anorexia	
anger4	bruising, contusion, bleeding, fracture
anger15	delinquent, victimized, victimization
anticip10	hurting, refused, anxious, afraid
anticip12	ashamed, embarrass, upset, disgust
disgust32	breakdown, fight, crushed, abandoned
disgust21	stomach, intestinal, bile, esophagus
fear19	food, eating, eat, consume
fear101	illness, sickness, suffer
joy10	gain, attain, surpass
joy13	age, young, youngster
negative65	bathroom, toilet, washroom
negative105	bread, cake, tart
positive10	feeding, nutriment, nutrient
positive19	helpful, information, relevant
surprise18	oddness, quirkiness, strangeness
surprise28	betrayal, deceiver, desolation
trust23	admiration, fondness, esteem
trust25	hunger, thirst, solace
Intersection	
anger108	traumatic, trauma, traumatize
anger120	lie, rejection, restriction
anticip104	attack, offensive , harass
anticip107	alcoholism, alcoholic, drink
disgust15	aggravation, distress, traumatic
disgust17	criticize, condemn, repudiate
fear114	aggression, hostile, discord
fear15	cautious, wary, pessimistic
joy100	completion, progression, succeeding
joy104	effort, chance, opportune
negative115	conceal, hide, concealment
negative200	abstain, discourage, discouragement
positive108	sociable, approachable, dependable
positive110	protect, safeguard, protection
surprise106	increase, growth, increasing
surprise108	oddness, quirkiness, strangeness
trust18	game, scoreboard, gaming
trust110	house, residence, building

[4] S. Guntuku, D. Yaden, M. Kern, L. Ungar, and J. Eichstaedt, "Detecting depression and mental illness on social media: an integrative review," *Current Opinion in Behavioral Sciences*, 2017.

[5] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett, and A. Leenaars, "Suicide note classification using natural language processing: A content analysislin heidelberg," *Biomed Inform Insights*, 2010.

[6] P. Chikersal, D. Belgrave, G. Doherty, A. Enrique, J. E. Palacios, D. Richards, and A. Thieme, "Understanding client support strategies to improve clinical outcomes in an online mental health intervention," *In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[7] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the national academy of sciences*, 2013.

[8] M. Kosinski, Y. Bachrach, P. Kohli, D. Stillwell, and T. Graepel, "Manifestations of user personality in website choice and behaviour on online social networks," *Machine learning*, 2014.

[9] D. Preoţiuc-Pietro, S. Volkova, V. Lampos, Y. Bachrach, and N. Aletras, "Studying user income through language, behaviour and affect in social media.," *PloS one 10.9*, 2015.

[10] T. Correa, A. Willard Hinsley, and H. G. De Zuniga, "Who interacts on the web?: The intersection of users' personality and social media use.," *Computers in human behavior 26.2*, 2010.

[11] S. Volkova and Y. Bachrach, "Inferring perceived demographics from user emotional tone and user-environment emotional contrast," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016.

[12] D. Ramírez-Cifuentes and A. Freire, "Upf's participation at the clef erisk 2018: Early risk prediction on the internet," *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France*, 2018.

[13] H. Schwartz, J. Eichstaedt, M. Kern, G. Park, M. Sap, D. Stillwell, M. Kosinski, and L. Ungar, "Towards assessing changes in degree of depression through facebook," *In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014.

[14] G. Coopersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in twitter," *Workshop on Computational Linguistics and Clinical Psychology*, 2014.

[15] C. Xuetong, D. Martin, W. Thomas, and E. Suzanne, "What about mood swings? identifying depression on twitter with temporal measures of emotions," *Companion Proceedings of the The Web Conference 2018, International World Wide Web Conferences Steering Committee*, pp. 1653–1660, 2018.

[16] M. Aragón, A. López-Monroy, L. González-Gurrola, and M. Montes-y Gómez, "Detecting depression in social media using fine-grained emotions," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

[17] C. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLOS Medicine, Public Library of Science*, 2006.

[18] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," *In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, 2013.

[19] M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations.," *In Proceedings of the 5th Annual ACM Web Science Conference*, 2013.

[20] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, "Recognizing depression from twitter activity," *In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015.

[21] G. Coppersmith, C. Harman, and M. Dredze, "Measuring post traumatic stress disorder in twitter," *In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[22] Y. Tausczik and J. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of Language and Social Psychology*, 2010.

[23] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, "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*, 2015.

[24] M. Trotzek, S. Koitka, and C. Friedrich, "Linguistic metadata augmented classifiers at the clef 2017 task for early detection of depression," *Proceedings of the 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland*, 2017.

[25] M. Trotzek, S. Koitka, and C. Friedrich, "Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia," *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France*, 2018.

[26] N. Liu, Z. Zhou, K. Xin, and F. Ren, "Tua1 at erisk 2018," *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France*, 2018.

[27] D. Losada, F. Crestani, and J. Parapar, "Overview of erisk 2018: Early risk prediction on the internet (extended lab overview)," *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France*, 2018.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3075638, IEEE Transactions on Affective Computing

[28] E. A. Ríssola, M. Aliannejadi, and F. Crestani, "Beyond modelling: Understanding mental disorders in online social media," *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal*, 2020.

[29] S. Burdisso, M. Errecalde, and M. Montes-y Gómez, "A text classification framework for simple and effective early depression detection over social media streams," *Expert Systems With Applications, Vol. 133*, 2019.

[30] D. Preotiuc-Pietro, J. Eichstaedt, G. Park, M. Sap, L. Smith, V. Tobolsky, H. Schwartz, and L. Ungar, "The role of personality, age and gender in tweeting about mental illnesses," *In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*, 2015.

[31] G. Coppersmith, K. Ngo, R. Leary, and A. Wood, "Exploratory analysis of social media prior to a suicide attempt," *In Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, 2016.

[32] A. P. Association, "Diagnostic and statistical manual of mental disorders (5th ed.)," *American Psychiatric Association*, 2013.

[33] T. Wang, M. Brede, A. Ianni, and E. Mentzakis, "Detecting and characterizing eating-disorder communities on social media," *In Proceedings of the Tenth ACM International conference on web search and data mining*, 2017.

[34] F. Ramiandrisoa, J. Mothe, B. Farah, and V. Moriceau, "Irit at e-risk 2018," *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France*, 2018.

[35] R. Ortega-Mendoza, A. Lopez-Monroy, A. Franco-Arcega, and M. Montes-Y-Gómez, "Peimex at erisk2018: Emphasizing personal information for depression and anorexia detection," *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France*, 2018.

[36] W. Ragheb, B. Moulahi, J. Aze, S. Bringay, and M. Servajean, "Temporal mood variation: at the clef erisk-2018 tasks for early risk detection on the internet," *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France*, 2018.

[37] Y. Wang, H. Huang, and H. Chen, "A neural network approach to early risk detection of depression and anorexia on social media text," *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France*, 2018.

[38] R. Masood, "Adapting models for the case of early risk prediction on the internet," *Advances in Information Retrieval. ECIR 2019. Lecture Notes in Computer Science, vol 11438.*, 2019.

[39] E. Mohammadi, H. Amini, and L. Kosseim, "Quick and (maybe not so) easy detection of anorexia in social media posts," *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland.*, 2019.

[40] S. Chancellor, Y. Kalantidis, J. A. Pater, M. De Choudhury, and D. A. Shamma, "Multimodal classification of moderated online pro-eating disorder content," *In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017.

[41] S. K. Ernala, M. L. Birnbaum, K. A. Candan, A. F. Rizvi, W. A. Sterling, J. M. Kane, and M. De Choudhury, "Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals," *In Proceedings of the 2019 chi conference on human factors in computing systems*, 2019.

[42] L. Canales and P. Martínez-Barco, "Emotion detection from text: A survey," *Processing in the 5th Information Systems Research Working Days (JISIC)*, 2014.

[43] S. Mohammad and P. Turney, "Crowdsourcing a word-emotion association lexicon," *Computational Intelligence*, pp. 436–465, 2013.

[44] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, 2016.

[45] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *ICLR*, 2013.

[46] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

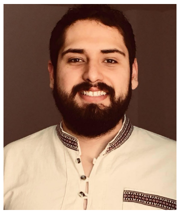
[47] P. Thavikulwat, "Affinity propagation: A clustering algorithm for computer-assisted business simulation and experimental exercises," *Developments in Business Simulation and Experiential Learning*, 2008.

[48] D. E. Losada, F. Crestani, and J. Parapar, "erisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations," *Proceedings of the 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland*, 2017.

[49] C. Walck, "Hand-book on statistical distributions for experimentalists," *University of Stockholm, Internal Report SUF–PFY/96–01*, 2007.

[50] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, 2008.

[51] D. Funez, M. Garciarena-Ucelay, M. Villegas, S. Burdisso, L. Cagnina, M. Montes-Y-Gómez, and M. Errecalde, "Unsl's participation at erisk 2018 lab," *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France*, 2018.



Mario Ezra Aragón is currently a Ph.D. student in Computing Science at the National Institute of Astrophysics Optics and Electronics, Puebla, Mexico. He received a B.Eng. and M.Eng in Computer Engineering from Universidad Autónoma de Chihuahua, in 2015 and 2017 respectively. His current research interests include natural language processing, pattern recognition, machine learning and deep learning.

Adrian Pastor López-Monroy obtained the MS and PhD degrees on Computer Science from National Institute for Astrophysics, Optics and Electronics (INAOE, Mexico) in 2012 and 2017 respectively. He received the best PhD thesis on AI award from the Mexican Society on AI in 2017. In 2017 and 2018 he was Postdoctoral Fellow and Lecturer at the University of Houston, in Texas USA. Since 2018, Dr. López-Monroy is full time researcher at Department of Computer Science of the Mathematics Research Center (CIMAT, Mexico). Among his research topics of interest are deep learning, natural language processing, and language & vision. He serves as regular reviewer on many journals and conferences including: ACL, NAACL, EMNLP, IP&M, KNOSYS, ESWA, among others.

Luis C. González obtained the PhD in Information Technology from the University of North Carolina at Charlotte (USA) in 2011. In 2017, he received a Google Research Award for his work in the intersection of Machine Learning and Intelligent Transportation Systems. Since 2011 he is a full time professor at the School of Engineering in The Autonomous University of Chihuahua (Northern Mexico). Dr. González interests are on the application of Machine Learning models in Intelligent Transportation Systems, Industrial-related problems and Natural Language Processing Tasks.

Manuel Montes-y-Gómez is researcher at the National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico. His research is on automatic text processing. He is author of more than 200 journal and conference papers in the fields of information retrieval, text mining and authorship analysis. He has been visiting professor at the Polytechnic University of Valencia, the University of Geneva, and the University of Alabama at Birmingham. He is a founding member of the Mexican Association of Natural Language Processing, and has been organizer of the National Workshop on Language Technologies, the Mexican Workshop on Plagiarism Detection and Authorship Analysis, and the Mexican Autumn School on Language Technologies.