IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 25, 2021, accepted August 18, 2021, date of publication September 10, 2021, date of current version September 28, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3112102*

# Enhanced Word Embedding Variations for the Detection of Substance Abuse and Mental Health Issues on Social Media Writings

**DIANA RAMÍREZ-CIFUENTES[1], CHRISTINE LARGERON[2], JULIEN TISSIER[2], RICARDO BAEZA-YATES[1,3], (Fellow, IEEE), AND ANA FREIRE[1,4]**

[1]Department of Information and Communication Technologies, Universitat Pompeu Fabra, 08018 Barcelona, Spain
[2]Laboratoire Hubert Curien UMR 5516, UJM-Saint-Etienne, CNRS, Université de Lyon, 42000 Saint-Etienne, France
[3]Institute for Experiential AI, Northeastern University, Boston, MA 02115, USA
[4]UPF Barcelona School of Management, 08008 Barcelona, Spain

Corresponding author: Diana Ramírez-Cifuentes (diana.ramirez@upf.edu)

This work was supported by the University of Lyon–IDEXLYON, the Auvergne-Rhône-Alpes Region, and the Spanish Ministry of Economy and Competitiveness through the Maria de Maeztu Units of Excellence Program under Grant MDM-2015-0502.

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the CIREP, which is the Institutional Review Board of the Pompeu Fabra University, as part of the ''Studying the Phenomena of Eating Disorders Through Social Media'' Project, under Approval No. 162.

**ABSTRACT** Substance abuse and mental health issues are severe conditions that affect millions. Signs of certain conditions have been traced on social media through the analysis of posts. In this paper we analyze textual cues that characterize and differentiate Reddit posts related to depression, eating disorders, suicidal ideation, and alcoholism, along with control posts. We also generate enhanced word embeddings for binary and multi-class classification tasks dedicated to the detection of these types of posts. Our enhancement method to generate word embeddings focuses on identifying terms that are predictive for a class and aims to move their vector representations close to each other while moving them away from the vectors of terms that are predictive for other classes. Variations of the embeddings are defined and evaluated through predictive tasks, a cosine similarity-based method, and a visual approach. We generate predictive models using variations of our enhanced representations with statistical and deep learning approaches. We also propose a method that leverages the properties of the enhanced embeddings in order to build features for predictive models. Results show that variations of our enhanced representations outperform in Recall, Accuracy, and F1-Score the embeddings learned with *Word2vec*, *DistilBERT*, *GloVe*'s fine-tuned pre-learned embeddings and other methods based on domain adapted embeddings. The approach presented has the potential to be used on similar binary or multi-class classification tasks that deal with small domain-specific textual corpora.

**INDEX TERMS** Classification algorithms, data mining, mental disorders, natural language processing, supervised learning.

## I. INTRODUCTION

Substance abuse and mental disorders are serious conditions that impact people's thinking, mood, feelings, and behavior. These conditions can also affect the dairy activities of people and the way they relate to others. This paper addresses the characterization of mental disorders and substance abuse conditions through the analysis of text cues on social media posts. The conditions considered are depression (DEP), eating disorders (ED), suicidal ideation (SUI) and alcoholism (ALC). Our main goals are twofold: first, to identify textual elements that characterize each of the conditions analyzed, and that distinguish these conditions from each other; including elements that differentiate mental conditions in general (MEN) from control cases (CON); second, to define automated methods capable to detect posts related to each

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar.

of the conditions addressed through the introduction of a word embedding generation model that identifies and takes advantage of the terms that are mostly used on the posts of users presenting a given condition. For that purpose, we built and evaluated different predictive models allowing to compare our enhanced embeddings with embeddings generated by other methods including domain adaptation approaches.

To address our first goal, we perform a comparative analysis of the posts to characterize the mental conditions studied using lexicons dedicated to five themes: affective processes and emotions, personal concerns and biological processes, linguistic elements, vocabulary related to risk factors and, topics of interest. We assume that the different groups (depression (DEP), eating disorders (ED), suicidal ideation (SUI) and alcoholism (ALC)) do not tackle the same topics and that they do not use the same vocabulary. We use statistical tests to check this hypothesis.

Regarding our second goal, we present a method for the generation of enhanced word embeddings for classification tasks on specialized domains. We formalize the problem in two ways: 1) as a binary task dedicated to the classification of posts (texts) of users with self-references related to substance abuse and mental health issues (MEN), and control posts (CON) which do not make reference to any of the prior conditions and, 2) as a multi-class classification task dedicated to the detection of posts related to depression (DEP), eating disorders (ED), suicidal ideation (SUI) and alcoholism (ALC). The particularity of these tasks is that texts are characterized by the usage of specific terms and expressions. This is the case of the eating disorders and alcoholism communities where it is common to find terms such as *thinspiration,* which refers to content that inspires a person to be thin; or *AA* that is used to refer to Alcoholics Anonymous. Consequently, we consider that this specific vocabulary must be exploited to solve in a more efficient way the classification task. Notably we assume that the classical embedding models learnt on large generic datasets are not suited and that they must be adapted.

Based on this hypothesis, our proposal takes advantage of the prior knowledge of terms that are predictive for each class and generates representations suited for the predictive task to address. It extends our previous work presented in [1], where enhanced word embeddings are generated for a binary classification task dedicated to Anorexia Nervosa screening.

The main new contributions of this paper are: 1) the generation of a Reddit dataset suitable for binary and multi-class classification tasks based on writings of users that state to have conditions such as depression, suicidal ideation, alcoholism and eating disorders, along with control cases; 2) a comparative analysis that characterizes the mental conditions addressed through the definition of textual features based on lexicons; 3) a method that improves and adapts the model presented in [1] to address a multi-class classification task; 4) a new approach to identify the most predictive terms for a given class taking into account binary and multi-class classification tasks; 5) the creation of predictive models based on deep learning approaches to compare our enhanced representations against other word embeddings' learning methods and domain adaptation approaches; 6) a type of feature designed for predictive models (named PSim) that leverages the properties of the embeddings generated with our method.

## II. RELATED WORK

Prior work has been dedicated to the detection of mental disorders on social media [2-8]. Most of it has been focused on the analysis of a single condition, which is usually compared to control cases [2-5]. Other studies have considered different risk levels over a single condition [6]; whereas only a few publications have been dedicated to the detection [7] and comparative analysis of multiple mental conditions, which are likely to be characterized by similar signs and symptoms [8]. Through our work, we do a further exploration of the linguistic dimensions, affective processes and emotions, personal concerns, vocabulary related to risk factors [9], and topics of interest linked to each condition, and define a method to identify the terms or n-grams that are highly related to them.

Researchers have created automated methods to detect mental disorders on social media by assuming that documents written by people presenting these disorders contain specific terms that describe signs and symptoms of a given condition [3]. However, before identifying these discriminant terms, it is necessary to find a suitable representation of the documents. Bag of Words (BoW) are among the most classical models considered. They allow to represent each text by a vector with components that are based on the number of times the terms of an index appear in the text [3]. More recently, word embedding models have been introduced, and they have proved to be very efficient for solving text mining tasks. In these models, terms are represented by vectors that are generated under the principle that words appearing in similar contexts are related, and they should have close representations in the vector space. Thus, one can compute a similarity score between two words by calculating the cosine value of their corresponding word vector and, a high value indicates that they are semantically related.

Examples of methods developed to generate word embeddings models are: *Word2vec* [10], where a vector is generated for each word in the corpus considering it as an atomic entity; *GloVe* [11] that defines a weighted least squares model for training on global word-word co-occurrence counts; or *fastText* [12] that addresses the morphology of words in a way such that a term is represented as a bag of character n-grams. More recent methods have addressed the issue of generating context aware representations, where polysemic terms are taken into account. Instances of these types of representations are *ELMo* [13] and *BERT* [14]. Among these methods, we consider *Word2vec* [10], and a distilled version of *BERT*: *DistilBERT* [15] to create our baseline models.

Embeddings that are generated through the prior approaches are often trained over large general corpora. However, when we consider their usage on domain specific classification tasks, in particular on the medical domain, it is common to have a reduced amount of labeled data to work with [1], [3]. Moreover, embeddings models learned exclusively in the domain corpus tend to not perform well on unseen cases with new vocabulary. Considering this issue, some methods have been developed to enhance the embeddings learned over small corpora. Those methods consist in incorporating external information [16], or adapting embeddings learned on large corpora to the task domain [2].

Within the enhancement methods, there is the work of [17] where approaches for combining different embedding sets to learn meta-embeddings are presented. Also, Faruqui *et al*. [16] propose a method that uses relational information from semantic lexicons for improving pre-built word vectors. Our approach surges as an alternative to handle small corpora and therefore some variations of these methods are considered as baselines to compare our model against other enhancement approaches.

In [1], we introduced a method based on *Dict2vec* [18], where in addition to the context defined by *Word2vec*, positive and negative sampling components are introduced. *Dict2vec* [18] works by using the lexical dictionary definitions of words to enrich the semantics of the embeddings generated over small corpora. This approach is based on the fact that all the words in the definition of a term from a dictionary are semantically related to the word they define, and therefore, the positive sampling component moves closer the vectors of words co-occurring in their mutual dictionary definitions, and the controlled negative sampling prevents to move these vectors apart.

According to our prior approach [1], dedicated to a binary document classification task, the positive sampling component consists in moving close to each other the vector representations of terms that are predictive for the main target class by defining a pivot vector $p$ towards which the vectors of predictive words are moved during the learning step. The negative sampling component, asides from preventing to move apart the vectors of words that are predictive for the target class, also puts apart from $p$ the vectors of the words that are the least predictive. In this paper, we extend this method to address multi-class classification tasks.

We also present a method that improves the performance of the embeddings generated. This is done through a modification of the objective function and the way to choose the set of words that are predictive for a given class, in a way such that enhanced embeddings can also be generated for multi-class classification tasks. We also introduce an alternative to identify predictive terms for a binary classification task. Moreover, for the development of predictive models, we propose the definition of a feature type named *PSim*, which leverages the properties of the embeddings generated through our approach.

## III. DATA COLLECTION

We collected our data to apply and evaluate our characterization and detection approaches. We defined two predictive tasks that are formalized as classification problems in two settings: each post can be classified as a document either related to depression (*DEP* class), eating disorders (*ED* class), suicidal ideation (*SUI* class) or alcoholism (*ALC* class) for *task 1* (multi-class task) and; as a document that is related to substance abuse or mental disorders (*MEN* class), or as a control document (*CON* class) for *task 2* (binary task).

The data collected for experimental purposes was gathered from a group of selected *subreddits*, which are forum communities of users on Reddit that are often focused on a specific subject of discussion. This is a suitable data source because it is likely to contain posts of users living with a given mental disorder [6], [19]. For instance, the *depression subreddit* contains posts of users with Depression, and from people that give advice and support to others.

We considered the following subreddits: for the suicide class: *Suicidewatch*; for the depression class: *depression*, for the alcoholism class: *alcoholism*, and for the eating disorders class: *eating_disorders, bulimia,* and *EatingDisorders*. As it was our intention to consider only posts of users living with the selected conditions and not control cases, we applied an automatic labelling approach where a post was first assigned the label of the *subreddit* it belonged to. Later, a first filtering approach was applied such that only posts with self-references were considered. With this purpose, we only kept posts containing keywords and phrases such as: *my alcoholism, I was diagnosed, I'm anorexic, etc*. From the starting 282,448 posts, with this filtering approach we kept only 13,174 posts.

For the multi-class task, we proceeded to discard posts of users with possible comorbidities. For the posts belonging to a given class, we did not keep posts with main general terms that describe other classes (*e.g.* for the *alcoholism* group we discarded posts containing the main terms: *depression, anorexia, bulimia, eating disorders and suicide*). We considered 9 main terms in total for this step.

After the filtering process, only 11,124 posts were kept. Finally, the keywords used for the first filtering approach were removed from 70% of the posts so that the keywords used during the data gathering would not interfere in the predictive models' behavior.

To collect control posts (CON), we took into account posts from subreddits where all types of posts were published. We considered posts from 18 randomly selected subreddits such as: *sports, celebs, books, fan theories, space, science, medical school, travel, history, economics, ask engineers, art fundamentals, lectures, unsolved mysteries, tales from call centers, law, legal advice and shower thoughts.* To discard posts that could be related to any of the issues studied, we deleted those containing self-references related to the mental conditions addressed. A total of 20,057 control posts were considered for our experimental approach.

For *task 1,* the posts considered are the ones that correspond to the depression (DEP), eating disorders (ED), suicidal ideation (SUI) and alcoholism (ALC) classes; all these posts constitute *dataset 1*. For task 2, the posts correspond to *Control* (*CON*) cases, and to those of the *MEN* class that consist of the union of all the posts of *task 1* (DEP + ED + SUI + ALC); they constitute *dataset 2*.

*Table 1* shows the details on the number of posts considered per class for both tasks. Python was selected as our main programming tool for the data collection and pre-processing approaches. Elements that could lead to the identification of users were removed (names, locations, and numbers).

**TABLE 1. Collection description.**

| Task | Class | Number of posts | Median number of terms per post |
|---|---|---|---|
| Task 1 (Dataset 1) | Suicide (SUI) | 7075 | 136 |
| | Depression (DEP) | 3015 | 177 |
| | Alcoholism (ALC) | 250 | 241 |
| | Eating disorders (ED) | 784 | 191 |
| Task 2 (Dataset 2) | Mental Conditions (MEN) | 11,124 | 152 |
| | Control (CON) | 20,057 | 141 |

Regarding the ethical concerns of the type of data addressed in this work, this research project has been approved by the ethical review board of the Pompeu Fabra University. A proper process of data transformation and anonymization has been followed to guarantee that no personal data is processed or stored. Only the transformed features extracted have been stored.

## IV. COMPARATIVE ANALYSIS OF MENTAL CONDITIONS

In order to meet our first goal, which consists in identifying the elements that characterize and differentiate each of the conditions considered, we perform a comparative analysis of the types of posts studied. With this purpose, we consider different psychological and linguistic perspectives that correspond to categories of lexicons, where each category is composed by a set of terms.

We generate numeric features for each of the categories analyzed within each perspective. To do so, for a given post we counted the frequency of terms belonging to each of the categories of the dictionaries, then the frequency was normalized by the size (in number of terms) of the full post. This approach was followed for all the lexicons' perspectives.

For the comparison of the groups analyzed (DEP, ED, SUI, and ALC), as in [20], we apply non-parametric tests after verifying that our features do not follow a normal distribution and that there is no homogeneity of variance for most of them. We first verified that there were features with significant differences among all the groups using Kruskal-Wallis' test [21]. Once we found there were features with significant differences, we performed Mann Whitney U's test [22] to check if the difference is significant for those features between certain pairs of groups. We also use Mann Whitney U's test to compare *mental conditions (MEN)* and *control (CON)* cases.

We analyzed 5 different perspectives as defined in the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary [23], which categorizes words in psychologically meaningful perspectives. We also consider other perspectives that are defined through related work addressing different mental conditions [9], [20]. The description of these perspectives and the results obtained for our comparative analysis are described as follows.

### a: AFFECTIVE PROCESSES AND EMOTIONS

To address these perspectives, we consider some of the LIWC's lexicon categories in addition to the categories described in EmoLex [24], which is a dictionary that associates terms to negative and positive sentiments, along with eight basic emotions: *anger, anticipation, trust, fear, surprise, sadness, disgust* and *joy*.

Table 2 shows the mean score values computed for each feature over the set of writings of each of the groups compared (MEN, CON, SUI, DEP, ED, ALC) and the P-values with the level of significance for each pair of classes compared using Mann-Whitney U's test. The averaged values per group are reported since the median values are equivalent to zero for a great number of features, as there are many categories with terms that can be found on very few writings.

Results show that there are features with high significant differences between the pairs of groups. As expected, this is notably true between the Mental Conditions (MEN) and Control (CON) groups, where the differences are significant for all the categories, confirming the quality of dataset 2.

Emotions such as *anger, fear, disgust* and *sadness* are expressed more on texts of users with suicidal ideation in comparison to texts of users with depression. These users (suicide risk) are the ones that express more negative emotions. Users with eating disorders are the most positive ones compared to the other groups within the conditions analyzed, and this can be due to the fact that eating disorders such as Anorexia and Bulimia can be characterized by the Transtheoretical Model of Health Behavior Change (TTM), where patients at the pre-contemplations stage are enthusiastic about their weight loss, and the social support they receive [20], [25]. The groups having the least difference between each other are the Depression and Alcoholism groups as shown by the non-significant results of the tests for several categories (*disgust, joy, surprise*, etc.). In Figure 1 (left), which presents a comparison of the emotions (Emolex) scores according to Plutchik's wheel [26], we can notice that *sadness* and *fear* are the emotions that mostly characterize users with mental conditions compared to the control group.

### b: PERSONAL CONCERNS AND BIOLOGICAL PROCESSES

Using LIWC, we explore also lexicon categories that are related to daily activities and concerns of users through

**TABLE 2.** Comparative results (means and p-values) between groups according to the Affective processes and emotions perspective.

| Categories | Mean values per group | | | | | | P- Values and significance level (Mann-Whitney U) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MEN | CON | SUI | DEP | ED | ALC | MEN - CON | SUI - DEP | SUI - ED | SUI - ALC | DEP - ED | DEP - ALC | ED - ALC |
| Fear | 3.30E-02 | 1.56E-02 | 3.67E-02 | 2.60E-02 | 2.78E-02 | 2.31E-02 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.002 ** | <0.001 *** |
| Disgust | 17.6E-03 | 7.74E-03 | 18.9E-03 | 15.0E-03 | 18.1E-03 | 15.2E-03 | <0.001 *** | <0.001 *** | 0.249 | 0.002 ** | <0.001 *** | 0.422 | <0.001 *** |
| Joy | 1.65E-02 | 1.46E-02 | 1.61E-02 | 1.69E-02 | 2.05E-02 | 1.62E-02 | <0.001 *** | <0.001 *** | <0.001 *** | 0.052 | <0.001 *** | 0.391 | 0.003 ** |
| Anticipation | 2.35E-02 | 2.19E-02 | 2.33E-02 | 2.36E-02 | 2.38E-02 | 2.10E-02 | <0.001 *** | <0.001 *** | 0.011 | 0.29 | 0.343 | 0.027 * | 0.027 * |
| Anger | 2.23E-02 | 1.21E-02 | 2.36E-02 | 2.11E-02 | 1.42E-02 | 1.79E-02 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.001 ** | <0.001 *** |
| Surprise | 10.1E-03 | 8.67E-03 | 10.2E-03 | 9.83E-03 | 12.2E-03 | 8.49E-03 | <0.001 *** | 0.005 ** | <0.001 *** | 0.46 | <0.001 *** | 0.165 | <0.001 *** |
| Trust | 2.48E-02 | 3.01E-02 | 2.41E-02 | 2.58E-02 | 3.02E-02 | 2.35E-02 | <0.001 *** | <0.001 *** | <0.001 *** | 0.34 | <0.001 *** | 0.045 * | <0.001 *** |
| Sadness | 3.71E-02 | 1.32E-02 | 3.92E-02 | 3.60E-02 | 2.55E-02 | 2.65E-02 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.395 |
| Positive emotions | 3.64E-02 | 4.64E-02 | 3.51E-02 | 3.78E-02 | 4.50E-02 | 3.55E-02 | <0.001 *** | <0.001 *** | <0.001 *** | 0.24 | <0.001 *** | 0.063 | <0.001 *** |
| Negative emotions | 5.24E-02 | 2.64E-02 | 5.43E-02 | 5.01E-02 | 4.35E-02 | 4.65E-02 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.030 * | 0.089 |
| Absolutist terms | 2.56E-02 | 1.30E-02 | 2.73E-02 | 2.38E-02 | 2.07E-02 | 1.64E-02 | <0.001 *** | 0.29 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.026 * |
| Cause and effect | 3.15E-02 | 3.28E-02 | 2.90E-02 | 3.58E-02 | 3.49E-02 | 2.90E-02 | 0.004 ** | <0.001 *** | <0.001 *** | 0.111 | 0.456 | <0.001 *** | 0.002 ** |
| Insight | 5.74E-02 | 5.89E-02 | 5.49E-02 | 6.43E-02 | 5.84E-02 | 5.50E-02 | 0.005 ** | <0.001 *** | <0.001 *** | 0.091 | 0.003 ** | 0.007 ** | 0.232 |
| Anxiety | 2.42E-02 | 9.32E-03 | 2.01E-02 | 3.00E-02 | 3.31E-02 | 1.98E-02 | <0.001 *** | <0.001 *** | <0.001 *** | 0.001 ** | <0.001 *** | 0.002 ** | <0.001 *** |
| Cognitive processes | 2.68E-01 | 2.59E-01 | 2.64E-01 | 2.80E-01 | 2.74E-01 | 2.57E-01 | <0.001 *** | <0.001 *** | <0.001 *** | 0.204 | 0.048 * | <0.001 *** | 0.003 ** |
| Certainty | 2.62E-02 | 2.09E-02 | 2.73E-02 | 2.62E-02 | 2.32E-02 | 2.04E-02 | <0.001 *** | 0.03 * | 0.067 | 0.03 * | 0.003 ** | 0.003 ** | 0.188 |
| Discrepancies | 3.68E-02 | 2.73E-02 | 4.02E-02 | 3.24E-02 | 2.92E-02 | 3.19E-02 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.491 | 0.026 * |
| Tentative | 4.85E-02 | 5.79E-02 | 4.70E-02 | 5.06E-02 | 4.90E-02 | 4.79E-02 | <0.001 *** | <0.001 *** | <0.001 *** | 0.049 * | 0.264 | 0.33 | 0.224 |
| See | 6.21E-03 | 11.6E-03 | 5.70E-03 | 6.58E-03 | 7.36E-03 | 6.67E-03 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.052 | 0.044 * | 0.28 |
| Listen | 10.7E-03 | 13.6E-03 | 9.79E-03 | 12.8E-03 | 8.90E-03 | 10.5E-03 | <0.001 *** | <0.001 *** | 0.08 | 0.045 * | 0.003 ** | 0.166 | 0.215 |
| Feel | 17.1E-03 | 7.71E-03 | 14.5E-03 | 22.9E-03 | 21.5E-03 | 14.5E-03 | <0.001 *** | <0.001 *** | <0.001 *** | 0.037 * | 0.373 | <0.001 *** | <0.001 *** |
| Social Processes | 1.37E-01 | 1.70E-01 | 1.33E-01 | 1.45E-01 | 1.25E-01 | 1.32E-01 | <0.001 *** | <0.001 *** | 0.023 * | 0.299 | <0.001 *** | 0.066 | 0.048 * |
| References to friends | 14.1E-03 | 5.98E-03 | 14.0E-03 | 16.1E-03 | 9.88E-03 | 12.3E-03 | <0.001 *** | <0.001 *** | <0.001 *** | 0.057 | <0.001 *** | 0.474 | <0.001 *** |
| References to family | 1.29E-02 | 1.00E-02 | 1.39E-02 | 1.17E-02 | 9.85E-03 | 1.20E-02 | <0.001 *** | 0.002 ** | <0.001 *** | 0.337 | 0.014 * | 0.076 | 0.005 ** |

(***P<.001, **P<.01, *P<.05)

general terms related to religion, work, leisure, money, health, and biological processes.

Table 3 reports the comparative results obtained for these categories. Control writings obtain the highest scores for the topics *work, money*, and *home* and the lowest for *biological processes, body*, or *health*. We can notice a high mean value for the usage of terms related to *death* and *sexuality* for the Suicide class and, for the topics: *body, ingest* and *biological processes* in the eating disorders class. These last two topics also characterize the alcoholism class, which obtains the highest mean scores for both of them, along with the *leisure* category. We can also see that the ED class obtains the highest score for the achievement category and the lowest score for the religion and death categories in comparison to the other conditions. Concerning the SUI group, it obtains the lowest scores in the *work, achievement*, and *leisure* categories whereas the depression class has the second highest value for the usage of terms related to *death*, and this score is

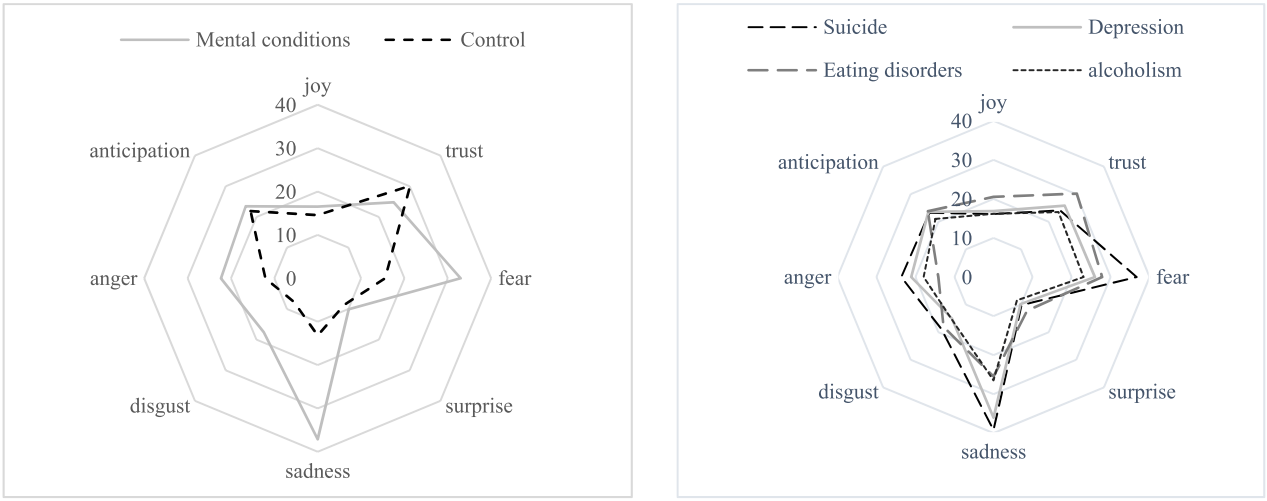



**FIGURE 1.** Emotions (Emolex) scores according to the basic emotions of Plutchik's wheel. The scores from Table 3 were multiplied by 1000 to ease the visualization.

**TABLE 3.** Comparative results (means and p-values) between groups according to the personal concerns and biological processes perspective.

| Categories | Mean values per group | | | | | | P- Values and significance level (Mann-Whitney U) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MEN | CON | SUI | DEP | ED | ALC | MEN – CON | SUI – DEP | SUI – ED | SUI – ALC | DEP – ED | DEP – ALC | ED – ALC |
| Work | 3.87E-02 | 10.8E-02 | 3.33E-02 | 5.03E-02 | 3.41E-02 | 4.12E-02 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.437 | <0.001 *** |
| Achievement | 3.72E-02 | 4.32E-02 | 3.47E-02 | 3.72E-02 | 5.47E-02 | 3.96E-02 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.003 ** | <0.001 *** |
| Leisure | 1.79E-02 | 3.46E-02 | 1.43E-02 | 1.81E-02 | 1.70E-02 | 9.70E-02 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.425 | <0.001 *** | <0.001 *** |
| Home | 8.24E-03 | 14.1E-03 | 7.71E-03 | 9.46E-03 | 7.76E-03 | 8.40E-03 | <0.001 *** | <0.001 *** | 0.299 | <0.001 *** | 0.001 ** | 0.169 | 0.004 ** |
| Money | 8.33E-03 | 40.8E-03 | 8.61E-03 | 7.47E-03 | 5.84E-03 | 13.6E-03 | <0.001 *** | 0.007 ** | 0.408 | <0.001 *** | 0.053 | 0.003 ** | <0.001 *** |
| Religion | 3.61E-03 | 4.54E-03 | 3.61E-03 | 3.40E-03 | 1.58E-03 | 5.90E-03 | <0.001 *** | 0.466 | <0.001 *** | 0.008 ** | <0.001 *** | 0.011 * | <0.001 *** |
| Sexual | 14.1E-03 | 4.19E-03 | 18.4E-03 | 8.12E-03 | 4.28E-03 | 4.83E-03 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.147 | 0.006 ** |
| Death | 33.9E-03 | 5.84E-03 | 50.1E-03 | 7.02E-03 | 1.34E-03 | 3.18E-03 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** |
| Biological processes | 8.16E-02 | 3.04E-02 | 7.20E-02 | 7.61E-02 | 15.3E-02 | 18.7E-02 | <0.001 *** | 0.003 ** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** |
| Body | 17.9E-03 | 8.37E-03 | 18.5E-03 | 16.7E-03 | 20.6E-03 | 14.6E-03 | <0.001 *** | 0.158 | <0.001 *** | 0.166 | 0.005 ** | 0.243 | 0.169 |
| Ingest | 13.4E-03 | 4.61E-03 | 4.37E-03 | 6.13E-03 | 85.6E-03 | 122E-03 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** |
| Health | 4.09E-02 | 1.42E-02 | 3.24E-02 | 4.60E-02 | 7.69E-02 | 8.40E-02 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.124 |

(***P<.001, **P<.01, *P<.05)

significantly higher in comparison to the eating disorders and alcoholism classes as confirmed by the p-value.

### c: LINGUISTIC ELEMENTS

This perspective addresses the usage of grammatical and syntactical elements such as verbs, adverbs, pronouns, articles, and prepositions. It also considers the different verbal times and pronoun types. We use LIWC for this perspective as well.

In Table 4, we can observe that the writings of users of the MEN group, in comparison to the CON group, tend to have more first-person singular pronouns, use more negations, adverbs, verbs, and past and present verb tenses. In comparison to the other conditions, the suicide group is characterized mainly by the usage of pronouns, especially first-person singular pronouns. It is also characterized by the reduced usage of second person pronouns, past verb tenses and articles; and the high usage of negations, and present and future verb tenses. A characteristic of the depression class is the high usage of third person plural pronouns in comparison to the other conditions. It also gets scores significantly lower than the suicide class but also significantly higher than the ED and ALC classes in the following categories: verbs, pronouns, and present verb tense. The ED group is characterized by the

**TABLE 4.** Comparative results (means and p-values) between groups according to the linguistic elements perspective.

| Categories | Mean values per group | | | | | | P- Values and significance level (Mann-Whitney U) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MEN | CON | SUI | DEP | ED | ALC | MEN - CON | SUI - DEP | SUI - ED | SUI - ALC | DEP - ED | DEP - ALC | ED - ALC |
| First person singular pronouns | 13.4E-02 | 5.18E-02 | 14.3E-02 | 12.0E-02 | 11.7E-02 | 11.4E-02 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.091 | 0.012 * | 0.085 |
| First person plural pronouns | 1.70E-03 | 6.71E-03 | 1.47E-03 | 2.29E-03 | 1.60E-03 | 2.54E-03 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.017 * | <0.001 *** |
| Second person pronouns | 4.68E-03 | 11.0E-03 | 4.08E-03 | 5.21E-03 | 4.81E-03 | 5.21E-03 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.303 | 0.022 | 0.017 * |
| Third person singular pronouns | 7.45E-03 | 16.6E-03 | 7.04E-03 | 8.64E-03 | 8.53E-03 | 8.76E-03 | <0.001 *** | <0.001 *** | 0.009 ** | 0.018 * | 0.124 | 0.459 | 0.289 |
| Third person plural pronouns | 7.61E-03 | 13.7E-03 | 7.56E-03 | 8.23E-03 | 5.92E-03 | 4.81E-03 | <0.001 *** | <0.001 *** | 0.027 * | 0.046 * | <0.001 *** | 0.001 ** | 0.315 |
| Negations | 1.86E-02 | 1.03E-02 | 2.09E-02 | 1.60E-02 | 1.35E-02 | 1.28E-02 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.002 ** | 0.239 |
| Affirmations | 1.65E-03 | 2.16E-03 | 1.77E-03 | 1.79E-03 | 1.84E-03 | 1.29E-03 | <0.001 *** | <0.001 *** | <0.001 *** | 0.026 * | 0.497 | 0.466 | 0.477 |
| Adverbs | 6.57E-02 | 4.69E-02 | 6.65E-02 | 6.64E-02 | 6.64E-02 | 5.96E-02 | <0.001 *** | 0.108 | 0.136 | <0.001 *** | 0.337 | <0.001 *** | <0.001 *** |
| Articles | 3.79E-02 | 6.79E-02 | 3.68E-02 | 3.87E-02 | 4.03E-02 | 5.11E-02 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.030 * | <0.001 *** | <0.001 *** |
| Verbs | 1.52E-01 | 1.27E-01 | 1.56E-01 | 1.50E-01 | 1.37E-01 | 1.35E-01 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.346 |
| Personal pronouns | 15.6E-02 | 9.98E-02 | 16.3E-02 | 14.5E-02 | 13.8E-02 | 13.5E-02 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.15 |
| Total pronouns | 2.50E-01 | 1.75E-01 | 2.59E-01 | 2.42E-01 | 2.19E-01 | 2.07E-01 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.014 * |
| Prepositions | 1.22E-01 | 1.28E-01 | 1.19E-01 | 1.23E-01 | 1.25E-01 | 1.36E-01 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.107 | <0.001 *** | <0.001 *** |
| Past verb tense | 3.38E-02 | 3.22E-02 | 3.22E-02 | 3.78E-02 | 3.87E-02 | 4.04E-02 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.114 | 0.028 * | 0.135 |
| Present verb tense | 9.67E-02 | 7.46E-02 | 10.0E-02 | 9.35E-02 | 8.32E-02 | 7.80E-02 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.046 * |
| Future verb tense | 9.77E-03 | 9.09E-03 | 11.7E-03 | 6.60E-03 | 6.05E-03 | 6.58E-03 | 0.366 | <0.001 *** | <0.001 *** | <0.001 *** | 0.3 | 0.085 | 0.05 |

(***P<.001, **P<.01, *P<.05)

usage of first-person plural pronouns which is significantly higher than the SUI class but significantly lower than the DEP and ALC classes. Finally, the ALC group is characterized by a low usage of adverbs, and a high usage of articles and prepositions.

### *d: MENTAL DISORDERS, SUBSTANCE ABUSE RELATED VOCABULARY AND RISK FACTORS*

We study terms related to eating disorders, self-loathing, self-injuries, explicit suicidal ideation references, substance abuse, lack of social support, and discrimination or abuse. These categories were taken from those analyzed in [9]. Additionally, we considered the categories defined by Arseniev *et al*. [26] with terms related to Anorexia Nervosa and its symptoms. These categories refer to topics such as: anorexia promotion, body image, body weight, caloric restrictions, compensatory behaviors, and exercise. We also consider names of antidepressants.

Results regarding this perspective can be seen in Table 5. We can notice that there are significant differences for most categories between the MEN and CON groups, with higher mean values for the former. Evidently, when compared to the other mental conditions' groups, the SUI group obtains very significantly high score for the *explicit suicide* category; it also obtains the lowest mean value for *the food and meals* category and, the second lowest score for the *explicit depression* category with highly significant differences with the remaining classes. The categories that characterize the DEP group are the *explicit depression* and *antidepressants*, while for the ED group are those related with *food and meals*, *caloric restriction, anorexia promotion, eat verb, body image, binge eating, body weight, compensatory behavior* and *laxatives*. Regarding the ALC group, one can notice a high value for the *substance abuse* category, as expected, but also the lowest mean value for the *hate* category when compared with the other conditions.

### *e: TOPICS OF INTEREST*

Thanks to Empath [28], which generates and validates lexical categories using a corpus with 1.8 billion words, we retain 200 prebuilt topics such as sports, social media, music, and politics, among others. Fig. 2 shows only the top 20 Empath topics having the most significantly different values (P<.05) [20] between each pair of classes compared, including the mental conditions and control classes. The mean value for each class compared for each topic is shown.

**TABLE 5. Comparative results (means and p-values) between groups according to the mental disorders, substance abuse related vocabulary and risk factors perspective.**

| Categories | Mean values per group | | | | | | P- Values and significance level (Mann-Whitney U) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MEN | CON | SUI | DEP | ED | ALC | MEN - CON | SUI - DEP | SUI - ED | SUI - ALC | DEP - ED | DEP - ALC | ED - ALC |
| Abuse | 6.20E-04 | 5.78E-04 | 6.50E-04 | 7.69E-04 | 3.22E-04 | 3.51E-04 | <0.001 *** | 0.002 ** | 0.195 | 0.027 * | 0.012 * | 0.219 | 0.015 * |
| Explicit suicide | 29.3E-04 | 1.34E-04 | 43.5E-04 | 7.18E-04 | 1.39E-04 | 7.39E-04 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.017 * | 0.043 * |
| Food and meals | 15.7E-04 | 5.52E-04 | 4.08E-04 | 9.49E-04 | 142E-04 | 10.2E-04 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.059 | <0.001 *** |
| Exercise | 36.4E-05 | 14.7E-05 | 8.80E-05 | 73.8E-05 | 118E-05 | 23.2E-05 | <0.001 *** | <0.001 *** | <0.001 *** | 0.027 * | 0.007 ** | 0.151 | 0.023 * |
| Caloric restriction | 6.10E-04 | 1.10E-04 | 1.75E-04 | 3.81E-04 | 51.4E-04 | 12.1E-04 | <0.001 *** | <0.001 *** | <0.001 *** | 0.087 | <0.001 *** | 0.448 | <0.001 *** |
| Anorexia promotion | 89.5E-05 | 3.94E-05 | 14.9E-05 | 34.8E-05 | 944E-05 | 13.1E-05 | <0.001 *** | <0.001 *** | <0.001 *** | 0.264 | <0.001 *** | 0.058 | <0.001 *** |
| Eat verb | 62.7E-05 | 5.08E-05 | 9.80E-05 | 23.1E-05 | 676E-05 | 12.3E-05 | <0.001 *** | <0.001 *** | <0.001 *** | 0.044 * | <0.001 *** | 0.32 | <0.001 *** |
| Suicide methods | 25.5E-04 | 12.3E-04 | 35.1E-04 | 9.55E-04 | 13.4E-04 | 7.41E-04 | <0.001 *** | <0.001 *** | <0.001 *** | 0.011 * | 0.351 | 0.39 | 0.326 |
| Explicit depression | 156E-05 | 3.47E-07 | 21.7E-05 | 522E-05 | 85.0E-05 | 18.5E-05 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | 0.008 ** |
| Hate | 18.3E-04 | 1.47E-04 | 20.4E-04 | 13.1E-04 | 11.7E-04 | 2.31E-04 | <0.001 *** | 0.017 * | 0.032 * | <0.001 *** | 0.263 | <0.001 *** | <0.001 *** |
| No social support | 12.3E-04 | 2.25E-04 | 13.5E-04 | 13.5E-04 | 4.22E-04 | 4.98E-04 | <0.001 *** | 0.001 ** | <0.001 *** | 0.152 | <0.001 *** | 0.022 * | 0.045 * |
| Self-harm | 47.7E-04 | 4.67E-04 | 53.6E-04 | 37.2E-04 | 19.7E-04 | 27.9E-04 | <0.001 *** | <0.001 *** | <0.001 *** | 0.046 * | <0.001 *** | 0.297 | 0.056 |
| Bullying | 6.55E-04 | 5.34E-04 | 6.11E-04 | 6.37E-04 | 6.91E-04 | 38.0E-04 | <0.001 *** | 0.29 | 0.412 | 0.093 | 0.463 | 0.122 | 0.128 |
| Substance abuse | 52.7E-04 | 7.74E-04 | 27.2E-04 | 27.7E-04 | 24.8E-04 | 114E-03 | <0.001 *** | 0.020 * | 0.341 | <0.001 *** | 0.073 | <0.001 *** | <0.001 *** |
| Self-hatred | 11.5E-04 | 1.74E-04 | 14.3E-04 | 7.68E-04 | 5.73E-04 | 1.87E-04 | <0.001 *** | 0.007 ** | 0.020 * | 0.005 ** | 0.251 | 0.029 * | 0.074 |
| Helplessness | 14.4E-04 | 1.73E-04 | 14.7E-04 | 15.6E-04 | 3.21E-04 | 2.31E-04 | <0.001 *** | 0.046 * | <0.001 *** | 0.004 ** | <0.001 *** | 0.001 ** | 0.223 |
| Insomnia | 56.0E-06 | 15.4E-06 | 38.0E-06 | 103E-06 | 5.00E-06 | 84.0E-06 | <0.001 *** | 0.06 | 0.197 | 0.385 | 0.077 | 0.415 | 0.197 |
| Sexual orientation | 10.8E-05 | 7.59E-05 | 17.6E-05 | 6.80E-05 | 2.10E-05 | 0.000 | <0.001 *** | 0.007 ** | 0.004 ** | 0.03 * | 0.051 | 0.078 | 0.212 |
| Body image | 23.7E-05 | 2.48E-05 | 14.2E-05 | 3.10E-05 | 223E-05 | 1.60E-05 | <0.001 *** | 0.083 | <0.001 *** | 0.231 | <0.001 *** | 0.367 | <0.001 *** |
| Relationship issues | 24.1E-05 | 7.24E-05 | 25.1E-05 | 28.0E-05 | 7.10E-05 | 24.3E-05 | <0.001 *** | 0.155 | 0.001 ** | 0.209 | <0.001 *** | 0.335 | 0.001 ** |
| Binge eating | 72.1E-05 | 3.12E-05 | 5.60E-05 | 7.00E-05 | 822E-05 | 153E-05 | <0.001 *** | 0.172 | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** | <0.001 *** |
| Body weight | 46.8E-05 | 12.0E-05 | 9.90E-05 | 26.4E-05 | 468E-05 | 19.7E-05 | <0.001 *** | <0.001 *** | <0.001 *** | 0.002 ** | <0.001 *** | 0.448 | <0.001 *** |
| Antidepressants | 770E-06 | 5.25E-06 | 268E-06 | 200E-05 | 363E-06 | 0.000 | <0.001 *** | <0.001 *** | 0.459 | 0.041 * | <0.001 *** | <0.001 *** | 0.044* |
| Compensatory behavior and laxatives | 31.3E-05 | 1.08E-05 | 13.1E-05 | 4.30E-05 | 355E-05 | 9.40E-05 | <0.001 *** | 0.062 | <0.001 *** | 0.317 | <0.001 *** | 0.12 | <0.001 *** |
| No spirituality | 7.00E-06 | 2.53E-06 | 10.0E-06 | 3.00E-06 | 0.000 | 22.0E-06 | <0.001 *** | 0.201 | 0.124 | 0.013 * | 0.189 | 0.003 ** | 0.006 ** |

(***P<.001, **P<.01, *P<.05)

Regarding all the pairs of classes compared, we can observe that the SUI group, compared to all the other conditions groups is characterized by addressing topics such as: *kill, crime, prison, weapon, war, fight, aggression, negative emotions,* and *hate*. The ED group is characterized by topics as *food, eating, cooking, restaurant, shopping,* and *strength*, which can easily be linked to the condition. The topics that characterize the ALC group are *alcohol, liquid, party, smell,* and *poor*, which is a category that normally implies the usage of terms related to economic issues, but in this case the category is likely to be representative because within its terms the word *alcoholism* can be found. The DEP

group is characterized only by the *neglect* topic compared to all the other conditions, this topic considers terms such as: *depressed, loneliness, fear, depression, loathing, hopelessness* and *suffer*.

When the DEP group is compared to the ED and ALC groups, we can observe that *suffering, emotional, shame* and *negative emotion* are topics that obtain significantly higher scores for the DEP group. Regarding the SUI vs. DEP class, we can see that the depression group expresses more feelings of *contentment, love* and *zest*, and it also addresses more topics related to daily activities such as *white-collar job, occupation,* and *sports*. Notice too that when the ED and ALC groups are compared, the ALC group addresses more leisure related topics such as *party, night, car,* and *vacation*. Finally, compared to the control group, as shown at Fig. 3, the mental conditions group obtains higher mean values on topics that address feelings and emotions.

These findings confirm our hypothesis according to which the vocabulary used by the different groups is not the same. Specific topics, with their corresponding terms, are highly addressed by a given group, such as caloric restriction by the eating disorders' group. They also reveal less obvious associations, which suggest that such terms could be efficiently exploited as predictive features to automatically determine the belonging of a user to a group depending on their writings. That is what we propose to do by generating enhanced word embeddings dedicated to classification tasks over specialized and small corpora.

## V. EMBEDDINGS GENERATION

We introduce a method to generate word embeddings enhanced for both binary and multi-class classification tasks that involve small corpora. This is a method that can also be applied to other domain-specific classification tasks.

As in *Dict2Vec* [18], we consider positive and negative components but, unlike it, our model learns enhanced vectors dedicated to a particular classification task. Its interest is to represent the vectors of terms (word level n-grams) that are predictive for a given class close to each other, and far from those terms that are predictive for the remaining classes. For that, we define positive and negative predictive pairs, which are based on the definition of a list of terms (word level 1-2grams) which are themselves predictive for each class. These pairs are later used as inputs of the embedding learning model.

### A. PREDICTIVE PAIRS GENERATION

To provide an appropriate input to our learning model, we define a set of positive predictive pairs and a set of negative predictive pairs, which are built as follows for each type of task (multi-class or binary).

#### 1) MULTI-CLASS CLASSIFICATION TASK

To address a multi-class classification task, for each class $c_n$, a list of *positive predictive terms* denoted as $c_n$*_predictive* is built. The process to generate these lists is described in *Algorithm 1*.

We summarize the process in 4 main steps: *first*, based on $X^2$[27], we aim to identify the classes for which each unique (1-2)-gram of the corpus is more relevant. Given the labeled documents as input, for the $X^2$ definition of relevant terms for each class, a BoW model with a Boolean representation denoting the existence of a term in a given document is generated (Boolean_matrix), along with the classes (labels) to which each document belongs. Then, we proceed to calculate the $X^2$ scores for each term and class. As for a given term $t$, a $X^2$ score is obtained for each class $c_n$ in the list of existing classes $C$ and stored ($X^2scores_t$), we choose the class $c_n$ for which $t$ obtains the highest $X^2$ score (max among the scores for each class in $X^2scores_t$) and we add $t$ to the list of relevant terms of $c_n$ ($rel\_c_n$) according to the $X^2$ test (Steps 1 to 7 in *Algorithm 1*). By this way, a list of relevant terms is generated for each of the classes in $C$ and, each term is relevant for one single class.

As every (1-2)gram of the vocabulary is defined as relevant for a given class regardless of having a very low $X^2$ score, it is important to select only the most relevant terms for a class *i.e.* a subset of all the terms relevant to $c_n$. In this case, differing from our first proposal [1], where a $X^2$ score threshold is selected based on the distribution of the scores, in the *second* step we proceed to create a Tf.Idf representation of the posts (documents) for all the terms (1-2grams) of the corpus. A Tf.Idf model provides a weight for a term in a document. Then, for each class $c_n$ in $C$ we apply Mann Whitney U's test [22] for each term $t$ belonging to $rel\_c_n$ in order to compare the Tf.Idf scores for $t$ of all the documents that correspond to $c_n$ and the Tf.Idf scores for $t$ of the documents belonging to each one of the remaining classes in C ($C\backslash\{c_n\}$). This step corresponds to the statements 8 to 13 in *Algorithm 1*.

In the *third* main step, for those pairs of classes where the P-value obtained for a term $t$ by the Mann Whitney U's test is lower than a given threshold (0.001 in our use case), we calculate the mean Tf.Idf score obtained by $t$ for each class of the pair, and then we pick the class for which the mean Tf.Idf value is the highest as the one for which $t$ is relevant (steps 13 to 25 in *Algorithm 1*).

At the *fourth* main step, if $t$ is relevant for the same class $c_n$ on all its comparisons with the remaining classes in $C$, then it is added to the list of positive predictive terms for this class ($c_n$*_predictive)* (Steps 26 to 28 in *Algorithm 1*). The list of *negative predictive terms* of $c_n$, which is denoted as $c_n$*_negative_predictive,* it contains all the terms that are part of the list of positive predictive terms of every other class. Once the lists of positive and negative predictive terms (1-2 grams) of each class are defined, we proceed to generate the inputs required for our embeddings learning approach, which consist of two lists of *predictive pairs*: the *positive predictive pairs'* list and the *negative predictive pairs'* list. To generate the list of positive predictive pairs, we select one pivot term for each class in $C$. The pivot term for a class

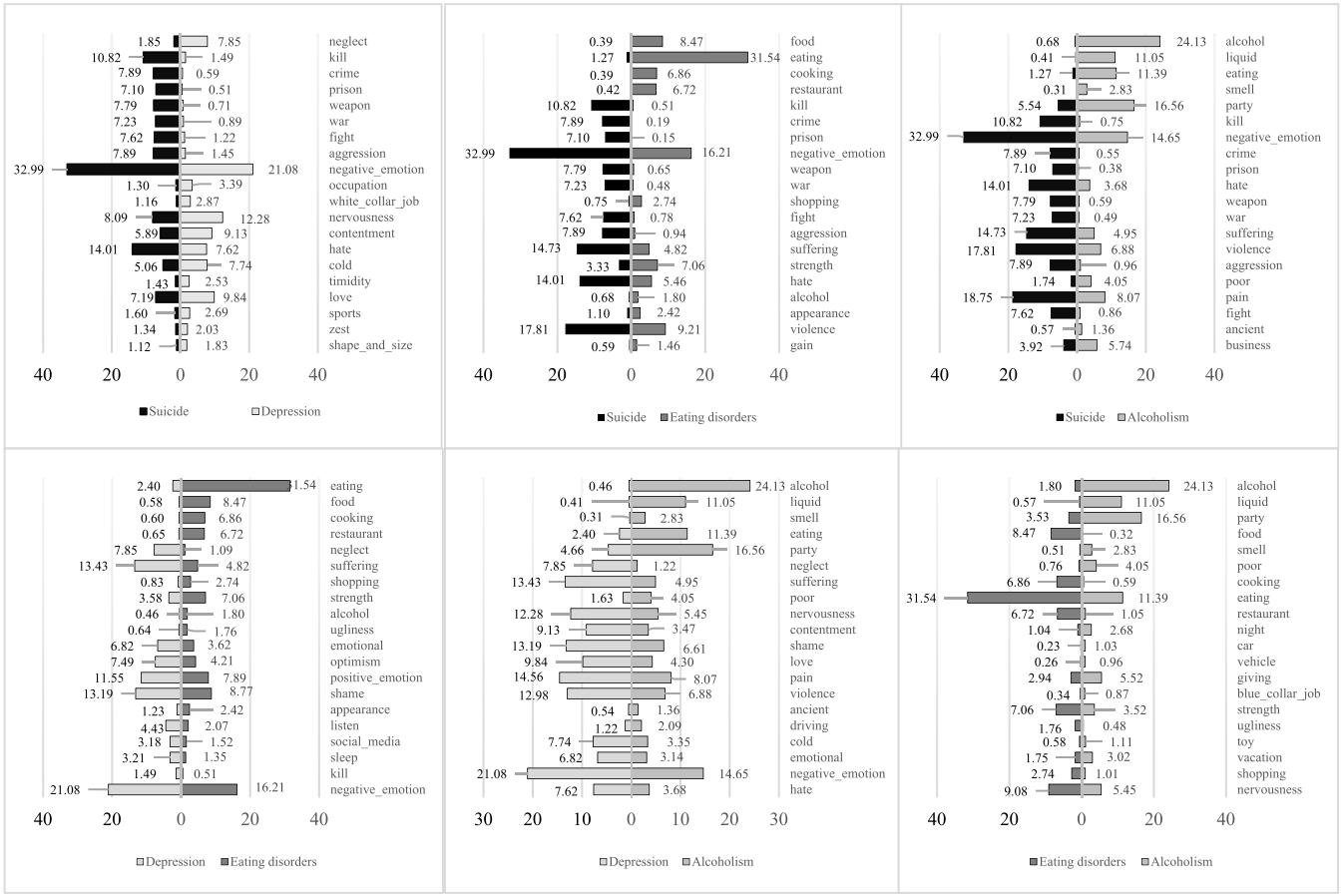



**FIGURE 2.** Top 20 Empath topics with most significantly different values (P<.05) between each pair of classes compared (multi-class task). The mean value for each class compared and topic is shown.

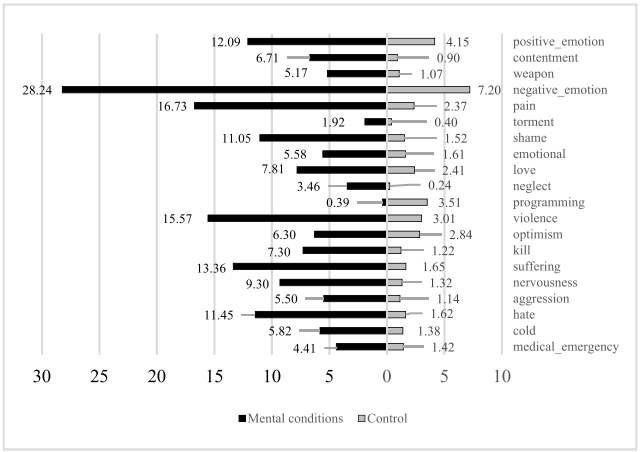



**FIGURE 3.** Top 20 Empath topics with most significantly different values (P<.05) between the classes compared (binary task). The mean value for each class and topic is shown.

$c_n$ is given by the term with the highest $X^2$ score within the terms in $c_n$ _*predictive*. This will be considered as a pivot term and, the vectors of the predictive terms of $c_n$ will be moved towards the vector of this term. Considering our use case as an instance (*task 1*), the pivot terms for the suicide, depression, eating disorders, and alcoholism classes were respectively: *kill*, *depression*, *eating*, and *alcoholism*. A positive predictive pair is then composed by a pivot term and a term that is part of the list of positive predictive terms of the class for which the pivot term belongs to. Considering our use case (*task 1*), positive predictive pairs instances are *(eating, anorexia)* and *(alcoholism, beer)*. This approach consists in pairing with their pivot term all the terms of the list of positive predictive terms of each class to compose the corresponding positive predictive pairs list.

For generating the list of negative predictive pairs, each pair is given by a pivot term, and a term that belongs to the list of negative predictive terms of the class for which the pivot term belongs to. In our use case, examples of negative predictive pairs' instances are *(eating, beer)*, and *(alcoholism, anorexia)*, as this pairing approach consists in pairing with their pivot term all the terms of the list of negative predictive terms for each class in $C$. Each of these pairs are added to the negative predictive pairs list.

### 2) BINARY CLASSIFICATION TASK

For the case where there are only two classes ($c_1$ and $c_2$), we follow a similar approach as for the multi-class task, but we consider that for binary tasks, the $X^2$ resulting predictive terms are the same for both classes. Here, we differ from

**Algorithm 1** Positive Predictive Terms' Lists Generation for Multiclass Classification Tasks

**Input:** labeled_documents, Classes list $C$
**Output:** positive predictive terms lists $c_n$_predictive

1. $Boolean\_matrix \leftarrow$ generate_Boolean_matrix ($labeled\_documents$)
2. $X^2$_scores $\leftarrow$ calculate_$X^2$_scores($Boolean\_matrix$)
3. for every term $t$ in $X^2\_scores$
4. $max\_score \leftarrow \max(X^2\_scores_t)$
5. $c_n \leftarrow$ class to which $max\_score$ corresponds
6. append $t$ to $rel\_c_n$
7. end for
8. $Tf.Idf\_model \leftarrow$ generate_Tf.Idf_representation ($labeled\_documents$)
9. for every class $c_n$ in $C$
10. for every term $t$ in $rel\_c_n$
11. $is\_relevant \leftarrow$ True
12. for every class $c_m$ in $C \backslash \{c_n\}$
13. $Pval \leftarrow$ get_P_value_using_Mann_Whitney_U's_test ($Tf.Idf_{model\,t,c_n}$, $Tf.Idf_{model\,t,c_m}$)
14. if $Pval$ <0.001
15. $mean_{t,c_n} \leftarrow \text{mean}(Tf.Id\_model_{t,c_n})$
16. $mean_{t,c_m} \leftarrow \text{mean}(Tf.Id\_model_{t,c_m})$
17. if $mean_{t,c_n} < mean_{t,c_m}$
18. $is\_relevant \leftarrow$ False
19. break
20. end if
21. else
22. $is\_relevant \leftarrow$ False
23. break
24. end if
25. end for
26. if $is\_relevant$
27. add $t$ to $c_n\_predictive$
28. end if
29. end for
30. end for

our prior approach [1], which assigns the class for which a term is relevant based on the ratio of documents that contain the term, and on the class they belong to. The steps to obtain the predictive terms for this task type are shown in *Algorithm 2*.

In this paper, we first consider the same initial main step as for multi-class tasks except that for the binary case, we define a $X^2$ score threshold based on the distribution of the scores of all the terms in order to keep only relevant terms. Then, these terms, regardless of the class they are relevant for (as it is not known through the $X^2$ test) are added to a list of binary relevant terms (*binary_rel_terms*) (steps 1 to 3 of *Algorithm 2*). Later, to identify the class for which the terms in the *binary_rel_terms* list are predictive, and to discard terms that are not relevant enough, we execute the main steps 2 to 4 of the approach for the multi-class task considering that for the second main step, Mann Whitney U's test is applied for each term in the *binary_rel_terms* list and that the comparison is done between the Tf.Idf scores of the documents according to the respective class they belong to (steps 4 to 6 of *Algorithm 2*). In this sense, if the p-value threshold is met for a given term, then it is directly added to the list of predictive terms of the class for which it obtains the greatest mean Tf.Idf score ($c_n$_predictive). This corresponds to steps 5 to 13 in *Algorithm 2*.

Later, considering that we only address two classes, we define a single pivot term, which is given by the term that obtains the highest $X^2$ score, and that is part of the list of predictive terms of one of the classes to predict. With our use case (*task 2*) as an instance, and considering its nature, where control cases are characterized by terms that are not related to mental disorders but that can be related to many other types of topics, we choose our pivot term, which corresponds to the unigram *feel* as it belongs to the main class to predict (*MEN*).

For this case, a positive predictive pair is composed by the pivot term, and a term that is part of the list of predictive terms of the pivot term's class. In our use case, instances of positive predictive pairs for *task 2* are (*feel, abused*), (*feel, antidepressants*) and (*feel, attempted suicide*). Finally, each negative predictive pair is composed by the same single pivot term (*feel*), and a term that is part of the predictive terms list of the remaining class. For our use case,

**Algorithm 2** Predictive Terms' Lists Generation for Binary Classification Tasks

**Input:** labeled_documents
**Output:** predictive terms lists $c_n$*_predictive*

```
Boolean_matrix ← generate_Boolean_matrix(labeled_documents)
X^2_scores ← calculate_X^2_scores(Boolean_matrix)
binary_rel_terms ← terms that obtain X^2 scores over a threshold
Tf.Idf_model ← generate_Tf.Idf_representation (labeled_documents)
for every term t in binary_rel_terms
        Pval ← get_P_value_using_Mann_Whitney_U's_test (Tf.Idf_model_t,c1 , Tf.Idf_model_t,c2)
        if Pval < 0.001
            mean_t,c1 ← mean(Tf.Id_model_t,c1)
            mean_t,c2 ← mean(Tf.Id_model_t,c2)
             if mean_t,c1 > mean_t,c2
                        add t to c1_predictive
             else
                        add t to c2_predictive
             end if
        end if
end for
```

instances of negative predictive pairs are (*feel, account*), (*feel, mechanical*), or (*feel, agent*).

### B. LEARNING APPROACH

Given the positive and negative pairs, the aim of this method consists in determining a vector representation of the terms in such a way that the vectors of positive predictive terms are represented close to their corresponding pivot vector and far from the pivot vectors of the remaining classes. These embedding representations are obtained by optimizing a global objective function.

Adopting the notation in [1], the objective function for a target term $\omega_t$ (1) is given by the aggregation of *Word2vec*'s (target term $\omega_t$, context term $\omega_c$) pair cost, a positive sampling cost (2) and a negative sampling cost (3). *Word2vec*'s cost is given by $\mathrm{l}(v_t, v_c)$ where l corresponds to the logistic loss function, and $(v_t)$ and $(v_c)$ are the vectors of $\omega_t$ and $\omega_c$ respectively.

$$J(\omega_t, \omega_c) = \ell(v_t, v_c) + J_{pos}(\omega_t) + J_{neg}(\omega_t) \tag{1}$$

The positive sampling component $J_{pos}$ is calculated for each target term according to Equation 2. $P(\omega_t)$ represents the set of n-grams that form a positive predictive pair with the n-gram $\omega_t$. The vectors $v_t$ and $v_i$ represent $\omega_t$ and $\omega_i$ respectively. Like in *Dict2vec*, a weight $\beta_P$ represents the importance of the positive sampling component during the learning phase. As it is our goal to keep the vector of the pivot term as a fixed element towards which other predictive terms get close to, whenever a pivot term happens to be the target term, the positive and negative sampling values are null. In this sense the positive sampling cost is zero if $\omega_t$ is a pivot term, otherwise its value is calculated according to (2). This represents a modification of the cost considered in our prior approach [1] where the same issue of keeping the pivot term as fixed as possible was addressed by normalizing the cost with the size of the predictive pairs set of the term ($|P(\omega_t)|$).

$$J_{pos}(\omega_t) = \beta_P \sum_{\omega_t \epsilon P(\omega_t)} \ell(v_t \cdot v_i) \tag{2}$$

For the negative sampling cost $J_{neg}$ defined in Equation (3), according to the first component, the vectors of the terms forming a positive predictive pair with $\omega_t$ are not moved away from $\omega_t$ thanks to the modification of the negative random sampling cost of *Word2vec,* where a set $F(\omega_t)$ of $k$ random terms from the vocabulary are moved away from the vector of $\omega_t$ considering that those random terms are not likely to be semantically related. The second component corresponds to the negative predictive pairs cost, which is the cost of putting apart from the vectors of the pivot terms the representations of the most predictive terms of other classes. $N(\omega_t)$ represents the set of all the terms (n-grams) that form a negative predictive pair with the term $\omega_t$, while $\beta_N$ represents the weight that defines the importance of the negative predictive pairs' component. Again, for not affecting the position of the vectors of pivot terms, whenever $\omega_t$ is a pivot term, the cost of the second component in (3) is zero.

$$J_{neg}(\omega_t) = \sum_{\substack{\omega_t \epsilon F(\omega_t) \\ \omega_t \notin P(\omega_t)}} \ell(-v_t \cdot v_i) + \beta_N \sum_{\omega_J \epsilon P(\omega_t)} \ell(-v_t \cdot v_J) \tag{3}$$

Finally, the sum of the cost of every (target, context) pair is what defines the global objective function (4) where $n$ is the size of the window and $C$ is the corpus size.

$$J = \sum_{t=1}^{C} \sum_{c=-n}^{n} J(\omega_t, \omega_{t+c}) \tag{4}$$

### C. ENHANCED EMBEDDINGS VARIATIONS

For both use cases, we define 4 variations of our embeddings with the aim of improving the representations obtained.

To address this goal we use related approaches [2], [16], [17] with which our method is compatible.

We label the proposed model described in the prior subsection (*V.B.*) as *Embedding model 0*. Thus, the first variation (*Embedding model 1*) consists in learning embeddings with our model after using *GloVe*'s pre-learned embeddings to define the starting weights of the vectors of terms. This is a popular transfer learning approach, which introduces information from a bigger corpus and enhances the performance of the embeddings that are being learned over the domain corpus. This technique has been applied on similar tasks such as the detection of suicide related writings [2]. The second variation (*Embedding model 2*) consists in applying Faruqui's retrofitting method [16] over the representations of the *Embedding model 0*. For the third variation, given a pre-learned embedding that associates for a term $\omega$ a pre-learned vector $v_{pr}$, and a vector $v$ learned through our approach for the same term $\omega$ with the same length $n$ as $v_{pr}$, an embedding of the *Embedding model 3* is defined by the concatenation of both representations $(v_{pr}+v)$ and the application of truncated SVD as a dimensionality reduction method so that the new vector of $\omega$ is given by $SVD(v_{pr}+v)$ [17], this variation is considered because it obtained the best results for the binary classification task addressed in [1]. Finally, the *Embedding model 4* corresponds to the retrofitting approach applied over the *Embedding model 1*.

### D. FEATURES FOR PREDICTIVE MODELS BASED ON ENHANCED EMBEDDINGS

In this section, we propose a feature generation method that leverages the properties of the embeddings generated through our method. The obtained features can be used for machine learning models. Our proposal takes into account that in our embedding model, the predictive terms of a class $c$ are represented close to its pivot term in the vector space. Thus, if we define a vector representation of a writing (document) that corresponds to $c$ and consider that predictive terms of $c$ are likely to be found in the writing, we can assume that their presence is likely to influence the placement of the vector that represents the whole writing. In this sense, the vector that represents the document (a writing/post) should be closer to the vector of the pivot term of $c$ in comparison to the vectors of documents that do not contain the predictive terms.

Based on our prior statement, we define the *pivot similarity (PSim)*, which is calculated for each document and for each class to predict. Considering $c$ as a class from the set of classes to predict $C$, $t$ a term belonging to the set $T$ of $n$ terms composing the document $D$, $v_t$ being the vector representation of $t$ and, $vp_c$ representing the vector of the pivot term of $c$, the value of *PSim* for $D$ and $c$ is defined by the cosine similarity between $vp_c$ and the average of the vectors associated to the terms belonging to D. It is given by (5).

$$PSim\,(D, c) = \cos\_sim\left(\frac{\sum_{t\in T} v_t}{n}, vp_c\right) \tag{5}$$

In a document classification task, for each document there will be as many features as classes to predict, except for a binary task where, as there is a single pivot term there is only one feature to define. Each feature corresponds to the *PSim* value between the document and the pivot term of a class.

## VI. EMBEDDINGS EXPERIMENTAL AND EVALUATION FRAMEWORK

This section explains the embedding generation process in our use cases dedicated to the detection of writings related to mental disorders. It also describes the methods adopted to evaluate the embeddings generated as well as their variations and, it also reports the results of these evaluations.

### A. EMBEDDINGS GENERATION PROCESS

In order to generate the embeddings and to evaluate their performance, *dataset 1* and *dataset 2* were split into training (70%) and test sets (30%). The distribution of the instances on each split was proportional for each class. Table 6 gives the details of the datasets for both tasks. Two corpora were defined, a corpus corresponding to *dataset 1* and a corpus that corresponds to *dataset 2*.

**TABLE 6. Train and test sets description.**

| Task | Class | Train set Number of posts | Test set Number of posts |
|---|---|---|---|
| **Task 1 (dataset 1)** | SUI | 5,306 | 1,769 |
| | DEP | 2,261 | 754 |
| | ALC | 188 | 62 |
| | ED | 588 | 196 |
| **Task 2 (dataset 2)** | MEN | 8,343 | 2,781 |
| | CON | 15,042 | 5,015 |

The process defined to generate the predictive pairs was applied over the training set, where each post was represented by a document and its label, which corresponds to the document class (SUI, DEP, ED, ALC, MEN, or CON). Then, we followed the process described in the *Predictive Pairs Generation* section.

Table 7 shows the list of the top 15 most predictive terms for each class. For the alcoholism class, we observed that despite having a reduced number of posts (see Table 1) it is a class that can be characterized by a large number of terms, whereas the suicide class, despite having the largest number of writings, does not have a large amount of unique distinguishable terms. For this same task, the list of negative predictive terms for each class was given by the list of terms that were predictive for all the other classes. For *task 2,* the number of positive predictive pairs obtained was 351, and the number of negative predictive pairs was 202. Table 8 shows the top 15 most predictive terms for each class. These findings contribute to achieve our first objective as these are terms that characterize the mental conditions studied.

From the pre-learned embeddings of *GloVe* [11], we also consider the top 20 terms with the highest similarity to

**TABLE 7.** Pivots and list of the top 15 most predictive terms for each class (task 1).

| Class | SUI | DEP | ED | ALC |
|---|---|---|---|---|
| **Pivot terms** | *Kill* | *Depression* | *Eating* | *Alcoholism* |
| **Terms' number** | 11 | 6 | 45 | 56 |
| **Terms** | *Suicide* | *Anxiety* | *Eating disorder* | *Alcohol* |
| | *Die* | *Depressed* | *Bulimia* | *Alcoholic* |
| | *Want die* | *Depression anxiety* | *Purging* | *Drinking* |
| | *Killing* | *Energy* | *Ed* | *Drink* |
| | *Live* | *Mental health* | *Purge* | *Sober* |
| | *Dead* | *Sad* | *Weight* | *Aa* |
| | *Anymore* | - | *Recovery* | *Beer* |
| | *Just want* | - | *Food* | *Sobriety* |
| | *Cares* | - | *Anorexia* | *Drank* |
| | *Care* | - | *Eat* | *Liquor* |
| | *Kill_myself* | - | *Binge* | *Drinks* |
| | - | - | *Calories* | *Drunk* |
| | - | - | *Bulimic* | *Stop drinking* |
| | - | - | *Binging* | *Beers* |
| | - | - | *Restricting* | *Drinking problem* |

**TABLE 8.** Pivot and list of the top 15 most predictive terms for each class (task 2).

| MEN (feel) | CON |
|---|---|
| *life* | *company* |
| *kill* | *customer* |
| *depression* | *calls* |
| *die* | *theory* |
| *friends* | *engineering* |
| *suicide* | *information* |
| *depressed* | *service* |
| *suicidal* | *book* |
| *feeling* | *legal* |
| *mental* | *number* |
| *anxiety* | *center* |
| *hate* | *question* |
| *pain* | *phone* |
| *shit* | *engineer* |
| *scared* | *account* |

each pivot term (*eating, kill, depression and alcoholism* for *task 1,* and *feel* for *task 2*) and terms highly related to the conditions such as *anorexia, suicide, bulimia, die, anxiety, eating disorder, alcohol and alcoholic*. We add the terms to the list of predictive terms of the respective class only if they are relevant for the class according to the $X^2$ score, or if they are not already part of the vocabulary and are semantically related to the pivot term considering the context of the task. This last aspect is considered as for terms as *die*, for instance, most of the vectors of terms with the highest cosine similarity are words in German. In the case where a term was not part of the corpus vocabulary, prior to learning, it was added to the corpus at the end of the last document belonging to the class. The *GloVe*'s pre-learned vectors were the 100 dimensions embeddings learned over 2B tweets with 27B tokens, and with 1.2M vocabulary terms. These embeddings are also the ones used for the baselines and some of our embedding variations.

After having the predictive pairs defined, in order to generate the embeddings, for its corresponding task, each corpus considered for training purposes consisted of the concatenation of all the texts from all the training posts. Stop words were removed. This resulted in a training corpus with a size of 800,319 tokens and a vocabulary size of 23,450 unique terms for dataset 1, and a corpus with a size of 2,230,423 tokens, with a vocabulary of 55,620 unique terms for dataset 2. For both datasets, to consider the predictive bigrams on the learning process, the words forming a predictive bigram were represented as a single term in the corpus. To learn our embeddings, we used as hyper parameters a window size of 5 with 5 random negative pairs chosen for negative sampling. We trained with one thread per worker and 5 epochs. Different values for $\beta_P$ and $\beta_N$ were tested.

### B. EVALUATION APPROACHES

We adapt the evaluation approaches of [1] that consist in a cosine-similarity-based evaluation, an evaluation based on visualization, and a predictive task evaluation. These different approaches are described in the following sections.

#### 1) AVERAGE COSINE SIMILARITY EVALUATION

This first evaluation method is applied over the *Embedding model 0*. We adopt the approach of [1] in the binary task evaluation (*task 2*) and adapt it for the existence of multiple classes for *task 1*.

For the case of *task 1*, for each class $c$ we average the cosine similarities between the vector of the pivot term of $c$ and each of the vectors of the remaining positive predictive terms of $c$ to obtain a positive score $P$ for the class. We also calculate a negative score $N$ for each class, which is given by the average of the cosine similarities between the vector of the pivot term of $c$ and each of the vectors of the remaining negative predictive terms of $c$. To address *task 2, P* and $N$ are calculated considering the pivot term (*feel*) of the main class to predict, which is the *MEN* class in this case.

For this evaluation approach we also choose as baselines: the embedding model that we introduced in [1] (*Baseline 1*) where the positive and negative components of the objective function were normalized considering the size of the list of predictive terms for the target term; and an embedding model generated using *Word2vec* exclusively (*Baseline 2)*, which corresponds to the case where $\beta_P = 0$ and $\beta_N = 0$. For the proposed models and, in comparison with the baselines, we expect to obtain better representations with our enhanced

**TABLE 9.** Task 1 (multi-class) – Average cosine similarity evaluation results.

| Values for $\beta_P$ and $\beta_N$ | P Scores | | | | N Scores | | | |
|---|---|---|---|---|---|---|---|---|
| | ALC | DEP | ED | SUI | ALC | DEP | ED | SUI |
| $\beta_P = 10$ and $\beta_N = 10$ (Prior method – Baseline 1) | 0.88 | 0.85 | 0.94 | 0.93 | 0.37 | 0.51 | 0.25 | 0.53 |
| $\beta_P = 0$ and $\beta_N = 0$ (*Word2vec* – Baseline 2) | *0.96* | *0.96* | *0.96* | 0.94 | 0.96 | 0.93 | 0.95 | 0.75 |
| $\beta_P = 0.01$ and $\beta_N = 0.01$ | *0.96* | *0.96* | *0.96* | 0.95 | 0.95 | 0.92 | 0.95 | 0.74 |
| $\beta_P = 0.05$ and $\beta_N = 0.05$ | 0.92 | 0.84 | 0.94 | 0.81 | 0.8 | 0.78 | 0.59 | 0.55 |
| $\beta_P = 0.1$ and $\beta_N = 0.1$ | 0.91 | 0.83 | 0.95 | *0.96* | 0.59 | 0.66 | 0.48 | 0.53 |
| $\beta_P = 0.5$ and $\beta_N = 0.5$ | 0.9 | 0.88 | 0.94 | 0.95 | 0.41 | 0.52 | 0.19 | 0.57 |
| $\beta_P = 1$ and $\beta_N = 1$ | 0.89 | 0.89 | 0.94 | *0.96* | 0.39 | 0.51 | 0.18 | 0.52 |
| $\beta_P = 5$ and $\beta_N = 5$ | 0.87 | 0.86 | 0.94 | 0.95 | 0.39 | 0.51 | 0.16 | 0.54 |
| $\beta_P = 10$ and $\beta_N = 10$ | 0.88 | 0.86 | 0.94 | 0.94 | 0.34 | 0.48 | 0.1 | 0.52 |
| $\beta_P = 25$ and $\beta_N = 25$ | 0.86 | 0.82 | 0.94 | 0.93 | 0.25 | 0.39 | 0.04 | 0.45 |
| $\beta_P = 35$ and $\beta_N = 35$ | 0.84 | 0.79 | 0.93 | 0.93 | 0.2 | 0.34 | 0.00 | 0.39 |
| $\beta_P = 50$ and $\beta_N = 50$ | 0.82 | 0.76 | 0.93 | 0.93 | *0.15* | *0.3* | *-0.04* | *0.33* |

The best results obtained by the configurations are in cursive.

embeddings in such way that $P$ keeps a high value while $N$ is kept as low as possible. We also study the impact of the parameters by assigning different values to $\beta_P$ and $\beta_N$.

For both tasks, the best results were obtained with equal values for $\beta_P$ and $\beta_N$. For *Task 1,* they are described in Table 9. Remember that for P the higher the score the better, while for N the lower the better. Even though the values for P for most of the models are lower compared to *Baseline 2*, a good balance is obtained considering how the value for N decreases for all the classes. This means that the method has managed to obtain representations where the vectors of predictive terms for a class are represented far enough from the vectors of terms that are predictive for other classes, while keeping a high cosine similarity with the vectors of the terms that are predictive for their own class. Notice that for *Baseline 1*, which corresponds to our prior method [1], we present the configuration for $\beta_P = 10$ and $\beta_N = 10$ as it obtained the best balance between the P Scores and N Scores for the approach. We observe that the results for the same configuration with the new approach are particularly better, especially considering the N Score and the ED class.

Results obtained for *task 2* are displayed in Table 10. Embeddings generated through our method obtained better results in comparison to the baselines as for P the scores are higher, whereas for N the scores are lower. For *Baseline 1*, corresponding to our prior method [1], we present the configuration: $\beta_P = 1$ and $\beta_N = 1$ as it obtained the best P Score while keeping a good balance with the N Score. Again, better results are obtained by the new approach.

### 2) VISUALIZATION EVALUATION

This second evaluation approach allows us to visually observe how some of the predictive terms for each class (top 15 terms according to the $X^2$ score) are distributed in the vector space without applying our embeddings' generation method (*Word2vec* - baseline), and how they are distributed after its application (*Embedding model 0*). We also consider the enhanced representation provided by *Embedding model 2* [16]. To generate the plots, Principal Component Analysis (PCA) is used as the dimensionality reduction method to reduce the vectors' dimensions from 100 to 2. For each plot we report PCA's Total Explained Variance Percentage (TEVP), which is an indicator of the percentage of information retained by the two resulting components, and that is given by the aggregation of the Explained Variance Ratio of each component. The high rates reported confirm the global quality of the representation. For each task, we retain the configuration which led to the best results according to the average cosine similarity evaluation.

**TABLE 10.** Task 2 (binary) – Average cosine similarity evaluation results.

| Values for $\beta_P$ and $\beta_N$ | P Score | N Score |
|---|---|---|
| $\beta_P = 1$ and $\beta_N = 1$ (prior method – Baseline 1) | 0.68 | 0.44 |
| $\beta_P = 0$ and $\beta_N = 0$ (*Word2vec* – Baseline 2) | 0.66 | 0.41 |
| $\beta_P = 0.01$ and $\beta_N = 0.01$ | 0.72 | 0.41 |
| $\beta_P = 0.05$ and $\beta_N = 0.05$ | 0.73 | 0.42 |
| $\beta_P = 0.1$ and $\beta_N = 0.1$ | 0.74 | 0.43 |
| $\beta_P = 0.5$ and $\beta_N = 0.5$ | *0.75* | 0.42 |
| $\beta_P 1$ and $\beta_N = 1$ | *0.75* | 0.42 |
| $\beta_P = 5$ and $\beta_N = 5$ | 0.74 | 0.42 |
| $\beta_P = 10$ and $\beta_N = 10$ | 0.74 | 0.42 |
| $\beta_P = 25$ and $\beta_N = 25$ | 0.74 | 0.41 |
| $\beta_P = 35$ and $\beta_N = 35$ | 0.73 | 0.40 |
| $\beta_P = 50$ and $\beta_N = 50$ | 0.73 | *0.37* |

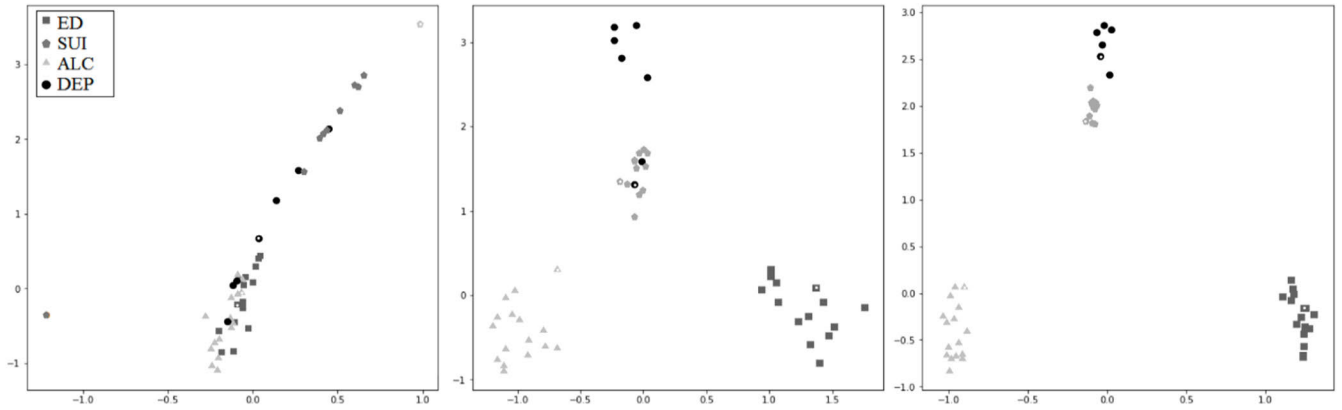



**FIGURE 4.** Task 1 - Vectors in two dimensions of the top 15 predictive terms of each class. The representations correspond to the 1) Word2vec baseline model (TEVP = 97%), 2) Embedding model 0 (TEVP = 72%), and 3) Embedding model 2 (TEVP = 91%). White dots are placed over pivot terms.

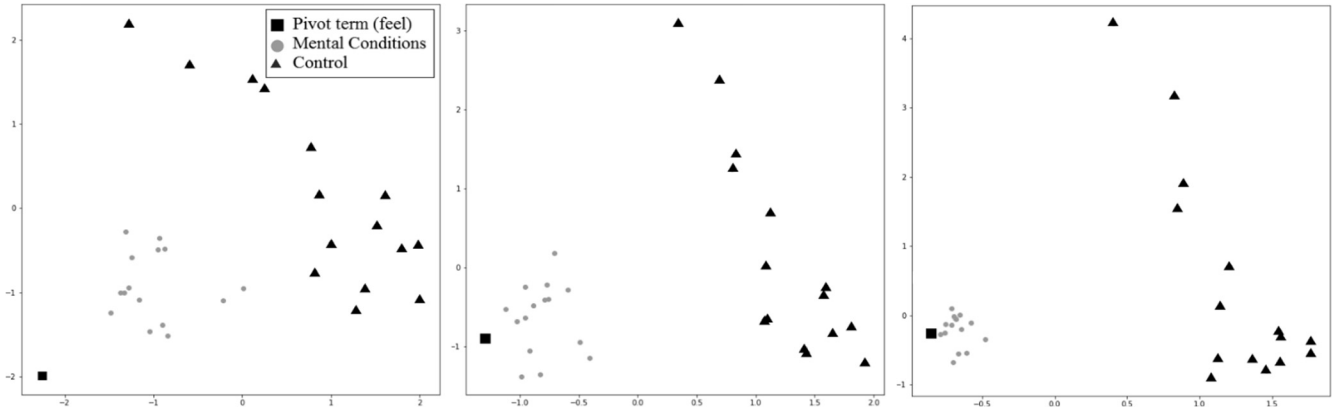



**FIGURE 5.** Task 2- Vectors in two dimensions of the top 15 predictive terms of each class. The representations correspond to the 1) Word2vec baseline model (TEVP = 36%), 2) Embedding model 0 (TEVP = 48%), and 3) Embedding model 2 (TEVP = 60%). The pivot term is represented by a square.

The results of this evaluation approach for *task 1*, obtained with $\beta_P = 10$ and $\beta_N = 10$, are displayed in Fig. 4. This clearly shows that better representations are obtained by *Embedding model 0* and *Embedding model 2* [16], where the vectors of predictive terms from any given class are represented close to each other, while making themselves distinguishable from the vectors of predictive terms for other classes. Suicide and depression related terms cannot be easily separated, which is consistent with the fact that both of these conditions tend to be closely related [6].

It is important to mention that while the closeness between terms for our approach is given by the cosine distance, the retrofitting approach [16] converges to changes in the Euclidean distance of adjacent vertices. Notice that for *Embedding model 2,* in order to apply Faruqui's approach to the *Embedding model 0*, as required by the input format of the retrofitting approach, a word has to be linked to all the words that we want it to be represented close to. For our tasks, each predictive term was associated to its pivot term and to each pivot term all its predictive terms were linked.

Fig. 5 shows the results for *task 2*. We consider the $\beta_P = 1$ and $\beta_N = 1$ configuration as the best one according to the P score, and it also obtains a reduction in the N score in comparison to *Baseline 1* according to the average cosine similarity evaluation. As for *task 1*, we can notice that the vectors of terms that are predictive for the main class (*MEN*) are placed closer to the pivot term and thus far from the terms that are predictive for the *Control (CON)* class with the proposed models.

### 3) PREDICTIVE TASK EVALUATION

Finally, for accomplishing our second main goal and to evaluate the performance of our embeddings generation approach, we process *task 1* (*dataset 1*) and *task 2* (*dataset 2*). We define a set of baselines based on state-of-the-art approaches. The same training and test sets exploited in the embeddings' generation process are used for this evaluation.

#### *a: BASELINES*

For both tasks considered for evaluation, we define 9 baselines: *Baseline 0* corresponds to a BoW model based on term level (1-2) grams. More than a baseline, this is a model kept as a reference as we are mainly focused on the evaluation of the models that make use of word embeddings on predictive tasks with small corpora. *Baseline 1*, kept as a reference model as well, consists of a model based on

the features extracted using the lexicons described in the *Data Analysis Approach* section. *Baseline 2* corresponds to a model that uses *DistilBERT* context aware pre-trained embeddings with the goal of building a deep learning model with transfer learning. *Baseline 3* consists of using *GloVe*'s pre-trained embeddings without any fine-tuning approach on the domain corpus. *Baseline 4* corresponds to a model where the word embeddings are learned on the training set using the classic *Word2vec* approach. *Baseline 5* is given by an enhanced version of *Baseline's 4* embeddings, using Faruqui's *et al.* [16] retrofitting method. *Baseline 6* applies the retrofitting method over *GloVe*'s pre-learned embeddings, while *Baseline 7* corresponds to an embedding model where *GloVe's* pre-learned embeddings provide the starting weights for learning embeddings on the training set with *Word2vec.* Finally, Baseline 8 is a model that uses the embeddings generated using our prior approach presented in [1]. Table 11 shows the baseline models and the proposed embedding variations that they can be compared to.

**TABLE 11. Baselines and proposed embedding models (variations) to compare.**

| Baselines | Embedding models (variations) |
|---|---|
| Baseline 0 (BoW) | All |
| Baseline 1 (lexicon) | All |
| Baseline 2 (distilBERT) | All |
| Baseline 3 (GloVe) | All |
| Baseline 4 (Word2vec) | Embedding Model 0 (predictive terms) |
| Baseline 5 (Word2vec + retrofitting) | Embedding model 2 (predictive terms + retrofitting) |
| Baseline 6 (GloVe + retrofitting) | Embedding model 4 (GloVe's initial weights + predictive terms + retrofitting) |
| Baseline 7 (GloVe's initial weights + Word2vec) | Embedding model 1 (GloVe's initial weights + predictive terms) |
| Baseline 8 (prior approach predictive terms) | Embedding model 0 (predictive terms) |

### b: CLASSIFICATION METHODS

We use as classifiers the Scikit Learn Python library implementations for Logistic regression (LR) and Random Forest (RF). These classifiers are trained using a parameter grid search, with a 5-fold cross validation performed for each parameter combination. The model with the best results is kept for its evaluation later over the test set.

We also consider a deep learning approach previously tested on a similar task [6] consisting of a Convolutional Neural Networks (CNN) model. This is denoted as our first deep learning approach *DL1*. In order to train this model, posts were represented as sequences of terms, and these terms were represented by word embeddings. For the CNN, the embeddings sequences instances were given as the model input. We used a filter window ({2,3,5} terms). We then applied max pooling and passed the output to either a SoftMax (multi-class task) or Sigmoid (binary task) layer to generate the final output. For *Baseline 2*, *DistilBERT*"s output is computed into a single vector with average pooling, and later two dense layers are added to predict the probability of each class; the classifier thus obtained is denoted *DL2*.

For the deep learning models, 75% of the training instances (posts) were selected for training the model and 25% were considered for validation. Notice that for presenting the results of the deep learning models, we average the results obtained by 5 runs over the test sets (with unseen cases).

For all the classifiers, we defined class weights' parameters for addressing the reduced amount of training samples for certain classes. This was done in such way that all the classes were considered equally important.

### c: EMBEDDING BASED INPUTS FOR PREDICTIVE METHODS

For the inputs, each instance is represented by an individual post (document) to which a class is assigned. For *Baseline 0* a Tf.Idf vectorization of the documents has been applied, considering a list of stop-words and the removal of the n-grams that appeared in less than 20 documents. For *Baseline 1*, we considered as features all the scores obtained for the lexicons categories of *Section IV*. For this baseline, each of the categories in tables 2, 3, 4 and 5 were considered as features, along with the 200 prebuilt categories of the Empath tool. To get the score for a category (feature), we consider the frequency of terms belonging to it, then the frequency is normalized by the size (in number of terms) of the full post.

Later, we consider approaches for using embeddings as inputs depending on the classification method selected. The first input, named *aggregation input*, used for testing machine learning approaches, such as Logistic Regression and Random Forest, consisted in representing a document through the aggregation of the vector representations of the terms in the document, normalized by the size (words count) of the document. Within this same method, a $L_2$ normalization was applied to all the instances. The other input approaches were suitable for generating deep learning models, which require the input data to be integer encoded, so that each term is represented by a unique integer, we denote this input as the *Emb. sequence input*. Notice that *distilBERT*"s input uses a different tokenization approach for which a proper input structure should be provided.

We also build models that use features created through the *PSim* approach as defined in Equation (5). For *task 1,* a predictive model with 4 features, one per class, was built with this method; each feature corresponds to the *PSim* value between the document and the pivot term of a class. For *task 2*, a model with only one feature was built as there is only one pivot term belonging to the main class to predict. These features are referred as the *PSim input* in our results section.

### d: EVALUATION MEASURES

For both tasks our evaluation measures are Precision (*P*), Recall (*R*), F1-Score (*F1*) and Accuracy (*A*). The results for *task 1* (multi-class) correspond to the macro average

**TABLE 12.** Task 1 (multi-class) – Predictive task evaluation results.

| Result type | Models | Input approach | Classifier | P | R | F1 | A |
|---|---|---|---|---|---|---|---|
| Reference baselines | Baseline 0 (BoW) | BoW | LR (C = 10) | *93.66%* | *89.49%* | *91.47%* | *92.48%* |
| | Baseline 1 (lexicon) | Lexicon scores | LR (C = 100) | 37.12% | 37.98% | 37.27% | 63.43% |
| Embeddings Baselines | Baseline 2 (distilBERT) | Embeddings sequence | DL2 | 54.00% | 69.00% | 58.00% | 71.00% |
| | Baseline 3 (GloVe) | Embeddings sequence | DL1 | 86.67% | 83.06% | 84.66% | 87.17% |
| | Baseline 4 (Word2vec) | Aggregation | LR (C = 100) | 70.11% | 62.98% | 65.23% | 81.77% |
| | Baseline 5 (Word2vec + retrofitting) | Aggregation | LR (C = 100) | 72.27% | 66.25% | 68.70% | 82.27% |
| | Baseline 6 (GloVe + retrofitting) | Embeddings sequence | DL1 | 87.65% | 83.16% | 85.12% | 87.84% |
| | Baseline 7 (GloVe's initial weights + Word2vec) | Embeddings sequence | DL1 | *88.82%* | 81.26% | 84.31% | 88.03% |
| | Baseline 8 (prior approach predictive terms) | Aggregation | DL1 | 84.09% | 75.42% | 78.93% | 84.67% |
| Best results for the enhanced embeddings models | Embedding model 0 (predictive terms) | Aggregation | LR (C = 100) | 79.69% | 78.73% | 79.20% | 84.29% |
| | Embedding model 1 (GloVe's initial weights + predictive terms) | Embeddings sequence | DL1 | 87.97% | 84.49% | 86.00% | 87.38% |
| | Embedding model 2 (predictive terms + retrofitting) | Aggregation | LR (C = 100) | 79.67% | 75.58% | 77.42% | 83.64% |
| | Embedding model 3 (SVD combination) | Embeddings sequence | DL1 | 87.78% | 82.49% | 84.74% | 87.24% |
| | Embedding model 4 (GloVe's initial weights + predictive terms + retrofitting) | Embeddings sequence | DL1 | 87.74% | *84.64%* | *86.03%* | *88.56%* |
| Best results for each input approach | Embedding model 4 (GloVe's initial weights + predictive terms + retrofitting) | PSim | LR (C = 5) | 76.39% | 72.09% | 73.99% | 80.55% |
| | Embedding model 1 (GloVe's initial weights) | Aggregation | LR (C = 100) | 83.12% | 81.06% | 82.03% | 86.19% |
| | Embedding model 4 (GloVe's initial weights + predictive terms + retrofitting) | Embeddings sequence | DL1 | 87.74% | *84.64%* | *86.03%* | *88.56%* |

Best baselines and embedding models' variations results for Precision (P), Recall (R), F1-Score (F1) and Accuracy (A). The best results obtained by the configurations are in cursive. For all the enhanced embedding models and *Baseline 8*: $\beta_P = 10$ and $\beta_N = 10$.

scores for *P*, *R* and *F1*, while their micro average scores are equivalent to the Accuracy. The results for *task 2* (binary) for *P*, *R* and *F1* correspond to the main class to predict (*MEN*), and we consider the Accuracy to evaluate the performance of the models for both classes. All the evaluation results correspond to those obtained over the test sets defined for each task, which correspond to cases that have not been seen before by the models, nor they have been used for tuning parameters.

### e: RESULTS

For both tasks, we report the best results obtained for each embedding model, including the baselines. We also report those embeddings models that obtained the best results for each input approach (*PSim, Aggregation and Embeddings sequence*); and to exclusively compare the embeddings models regardless of the input approach, we also present the results of a single input method (*aggregation input*) for all the embeddings. This last input approach also corresponds to the method used in [1].

Regarding the parameters of the LR models, for both tasks' models we used a *one vs. rest* approach with *Scikit Learn's liblinear* solver. The values of the *C* parameter are defined through a grid search, and its value for each model is mentioned next to the classifier type in each results table.

For the RF classifiers we use *Scikit Learn's* default parameters except for the number of trees in the forest (*n_estimators*).

For *task 1*, according to the results presented in Table 12, the type of classification method that obtains the best results for the embedding based inputs is the deep learning model *DL1*, which obtains the best results for 7 models. Notice that the BoW reference model obtains the best results for the task, which is consistent with the findings in related work addressing similar tasks for the detection of mental health issues [19]. Regardless of this, considering exclusively the approaches based on word embeddings, we observe that the *Embedding Model 4* is the one that obtains the best results for recall, F1-score, and accuracy. Moreover, we can see better results (F1) when the enhanced embeddings models (that

**TABLE 13. Task 1 (multi-class) – Predictive task evaluation results – aggregation input for the enhanced embeddings.**

| Result type | Models | Input approach | Classifier | P | R | F1 | A |
|---|---|---|---|---|---|---|---|
| Reference baselines | Baseline 0 (BoW) | BoW | LR (C=10) | *93.66%* | *89.49%* | *91.47%* | *92.48%* |
| | Baseline 1 (lexicon) | Lexicon scores | LR (C=100) | 37.12% | 37.98% | 37.27% | 63.43% |
| Embeddings Baselines | Baseline 2 (distilBERT) | Embeddings sequence | DL2 | 54.00% | 69.00% | 58.00% | 71.00% |
| | Baseline 3 (GloVe) | Aggregation | LR (C = 100) | 80.17% | 79.15% | 79.63% | 85.04% |
| | Baseline 4 (Word2vec) | Aggregation | LR (C = 100) | 70.11% | 62.98% | 65.23% | 81.77% |
| | Baseline 5 (Word2vec + retrofitting) | Aggregation | LR (C = 100) | 72.27% | 66.25% | 68.70% | 82.27% |
| | Baseline 6 (GloVe + retrofitting) | Aggregation | LR (C = 100) | 80.50% | 79.70% | 80.07% | 84.32% |
| | Baseline 7 (GloVe's initial weights + Word2vec) | Aggregation | LR (C = 100) | 79.55% | 80.08% | 79.71% | 85.87% |
| Best results for each Model | Embedding model 0 (predictive terms) | Aggregation | LR (C = 100) | 79.69% | 78.73% | 79.20% | 84.29% |
| | Embedding model 1 (GloVe's initial weights + predictive terms) | Aggregation | LR (C = 100) | *83.12%* | *81.06%* | *82.03%* | *86.19%* |
| | Embedding model 2 (predictive terms + retrofitting) | Aggregation | LR (C = 100) | 79.67% | 75.58% | 77.42% | 83.64% |
| | Embedding model 3 (SVD combination) | Aggregation | LR (C = 100) | 81.16% | 78.36% | 79.69% | 85.15% |
| | Embedding model 4 (GloVe's initial weights + predictive terms + retrofitting) | Aggregation | LR (C = 100) | 81.07% | 79.97% | 80.50% | 85.69% |

Baselines and embedding models' variations results for Precision (P), Recall (R), F1-Score (F1) and Accuracy (A), using the *Aggregation input* for all the enhanced embeddings models. The best results obtained by the configurations are in cursive. For all the enhanced embedding models and *Baseline 8*: $\beta_P = 10$ and $\beta_N = 10$.

use our learning approach) are compared with embedding models (*Baselines 4,5* and *7*) that use *Word2vec* 's learning approach (*i.e.* $\beta_P = 0$ and $\beta_N = 0$). This can be seen when comparing *Baseline 4* (F1 = 65.23%) vs *Embedding Model 0* (F1 = 79.20%); *Baseline 5* (F1 = 68.70%) vs. *Embedding model 2* (F1 = 77.42%); and *Baseline 7* (F1 = 84.31%) vs. Embedding *model 1* (F1 = 86%).

Remember that we consider the Recall and F1-Scores as our most relevant measures given the nature of the task, where a misclassification error would imply a mistaken diagnosis. Furthermore, the accuracy of the system cannot be very reliable given the limited number of instances existing for the alcoholism and eating disorders classes.

For the embeddings baselines, the best results (F1) were obtained by *Baseline 6*, which corresponds to the CNN model that considers a retrofitted [16] version of pre-trained *GloVe* embeddings. Regarding the enhanced embedding models, we can observe that embeddings learned exclusively through our approach (*Embedding model 0*) provide better results (F1) compared to *Baselines 1, 2, 4, 5* and *8*. *Embedding model 4* is also the best model for generating the *PSim input.* This is a promising result for this approach as with only 4 features the Accuracy achieved is only 8.01% lower than the one of the best embedding model (*Embedding model 4 − Embeddings sequence input)*, and only 11,93% lower than the BoW model.

Table 13 shows the results obtained by a single input type (Aggregation input) and classification method (LR) for *task 1*. Among the embedding models, we observe that the best results in Precision, Recall, F1 Score and Accuracy are given by the *Embedding model 1*. Based on the F1 score, we can see that the embedding models 1 and 4, enhanced through our approach, outperform all the embeddings baselines. Notably, we can observe a 13.97% increase in the F1 Score when comparing the embeddings learned through our approach (*Embedding model 0*) vs. *Baseline 4 (Word2vec)*. Moreover, we have proved the usefulness of the predictive terms defined through our method (Section V.A) for their usage on similar approaches, such as Faruqui's *et al.* retrofitting method [16], where their usage as semantically related terms implied obtaining better results for *Baseline 5* (F1 = 68.70%) vs. *Baseline 4* (F1 = 65.23%); and for *Baseline 6* (F1 = 80.07%) vs. *Baseline 3* (F1 = 79.63%). Also, considering our learning approach combined with the retrofitting method, better results were obtained for the *Embedding model 2* (F1 = 77.42%) *vs. Baseline 5* (F1 = 68.70%) and, the *Embedding model 4* (F1 = 80.50%) *vs Baseline 6* (F1 = 80.07%) cases.

For task 2 the results for the predictive task are presented in Table 14. We can see that as for the prior task, the BoW reference model obtains the best results. Addressing the embeddings models, which are our main point of interest,

**TABLE 14.** Task 2 (binary) – Predictive task evaluation results.

| Result type | Models | Input approach | Classifier | P | R | F1 | A |
|---|---|---|---|---|---|---|---|
| Reference baselines | Baseline 0 (BoW) | BoW | LR (C = 10) | *98.05%* | *97.45%* | *97.75%* | *98.40%* |
| | Baseline 1 (lexicon) | Lexicon scores | RF (n_estimators = 100) | 68.24% | 89.54% | 77.45% | 81.40% |
| Embeddings Baselines | Baseline 2 (distilBERT) | Embeddings sequence | DL2 | 90.87% | 93.82% | 92.32% | 94.43% |
| | Baseline 3 (GloVe) | Embeddings sequence | DL1 | 95.96% | 96.31% | 96.12% | 97.23% |
| | Baseline 4 (Word2vec) | Embeddings sequence | DL1 | 96.82% | 94.26% | 95.49% | 96.83% |
| | Baseline 5 (Word2vec + retrofitting) | Embeddings sequence | RF (n_estimators = 1000) | 94.60% | 96.44% | 95.51% | 96.77% |
| | Baseline 6 (GloVe + retrofitting) | Embeddings sequence | DL1 | 96.09% | 96.23% | 96.15% | 97.25% |
| | Baseline 7 (GloVe's initial weights + Word2vec) | Embeddings sequence | RF (n_estimators = 1000) | 96.89% | 94.17% | 95.51% | 96.84% |
| | Baseline 8 (prior approach predictive terms) | Embeddings sequence | DL1 | 95.77% | 95.97% | 95.82% | 97.01% |
| Best results for the enhanced embeddings models | Embedding model 0 (predictive terms) | Embeddings sequence | RF (n_estimators = 1000) | 95.85% | 96.40% | 96.13% | 97.23% |
| | Embedding model 1 (GloVe's initial weights + predictive terms) | Embeddings sequence | DL1 | 96.77% | 96.19% | *96.46%* | *97.48%* |
| | Embedding model 2 (predictive terms + retrofitting) | Aggregation | RF (n_estimators = 1000) | 95.42% | *96.58%* | 96.00% | 97.13% |
| | Embedding model 3 (SVD combination) | Embeddings sequence | DL1 | 96.05% | 95.94% | 95.96% | 97.11% |
| | Embedding model 4 (GloVe's initial weights + predictive terms + retrofitting) | Embeddings sequence | DL1 | *97.37%* | 95.41% | 96.37% | 97.44% |
| Best results for each input approach | Embedding model 4 (retrofitting + GloVe's initial weights + predictive terms) | PSim | LR (C = 150) | 86.93% | 94.50% | 90.56% | 92.97% |
| | Embedding model 0 (Predictive terms) | Aggregation | RF (n_estimators = 1000) | 95.85% | 96.40% | 96.13% | 97.23% |
| | Embedding model 1 (GloVe's initial weights + predictive terms) | Embeddings sequence | DL1 | 96.77% | 96.19% | *96.46%* | *97.48%* |

Best baselines and embedding models' variations results for Precision (P), Recall (R), F1-Score (F1) and Accuracy (A). The best results obtained by the configurations are in cursive. For all the enhanced embedding models: $\beta_P = 1$ and $\beta_N = 1$.

we can see that despite being small, there is an improvement in the results obtained by the enhanced embeddings. This is confirmed by the results showed in Table 15 where, as for the prior task, a single input approach is considered (*Aggregation input*). According to Table 14, we can observe that the baseline with the best result (F1) is *Baseline 6*. We can also see that the best results (P, R, F1 and A) for all the embedding models are given by variations of the enhanced embeddings. In addition, the best *PSim* result (F1) is only 5.9% lower than the best embedding model score (*Embedding model 1 – Embeddings sequence input*). For this particular task, we can notice that despite obtaining better results with the enhanced embeddings, the differences with the baselines are minimal.

## VII. DISCUSSION

One of the first aims of this work was to determine lexical features characterizing each of the mental conditions (classes) studied. Results show that there are many elements that distinguish writings of control users (CON) from those of people with certain mental disorders and substance abuse conditions (MEN), while there are fewer elements that distinguish the conditions analyzed (SUI, DEP, ED and ALC) from each other. This is corroborated by the predictive task evaluation as well, where better scores are obtained on the binary task (*Task 2*) that classifies CON and MEN writings compared to the multi class task dedicated to the detection of writings of multiple conditions (*Task 1*). Notably, terms related to emotions and feelings are expressed more in writings of the MEN class, while words concerning topics such as work, money, and home are more frequent in the CON class writings. We can also observe that, as expected, the categories that imply risk factors for mental disorders such as self-harm, suicidal ideation references, self-hatred, substances abuse, lack of social support, bullying or other types of abuse, obtained higher scores for the MEN group, providing evidence of the fact that the aspects that are considered on screening processes [28] can also be identified on social media posts.

We see that some of the categories addressed can characterize exclusively certain conditions such as the highly

**TABLE 15.** Task 2 (binary) – Predictive task evaluation results – aggregation input for the enhanced embeddings.

| Result type | Models | Input approach | Classifier | P | R | F1 | A |
|---|---|---|---|---|---|---|---|
| Reference baselines | Baseline 0 (BoW) | BoW | LR (C = 10) | *98.05%* | 97.45% | *97.75%* | *98.40%* |
| | Baseline 1 (lexicon) | Lexicon scores | RF (n_estimators = 100) | 68.24% | 89.54% | 77.45% | 81.40% |
| Embeddings Baselines | Baseline 2 (distilBERT) | Embeddings sequence | DL2 | 90.87% | 93.82% | 92.32% | 94.43% |
| | Baseline 3 (GloVe) | Aggregation | LR (C = 100) | 90.99% | 97.34% | 94.06% | 95.61% |
| | Baseline 4 (Word2vec) | Aggregation | RF (n_estimators = 1000) | 94.57% | 95.90% | 95.23% | 96.58% |
| | Baseline 5 (Word2vec + retrofitting) | Aggregation | RF (n_estimators = 1000) | 94.60% | 96.44% | 95.51% | 96.77% |
| | Baseline 6 (GloVe + retrofitting) | Aggregation | RF (n_estimators = 1000) | 95.11% | 93.64% | 94.36% | 96.01% |
| | Baseline 7 (GloVe's initial weights + Word2vec) | Aggregation | RF (n_estimators = 1000) | 96.89% | 94.17% | 95.51% | 96.84% |
| Best results for the enhanced embeddings models | Embedding model 0 (predictive terms) | Aggregation | RF (n_estimators = 1000) | 95.85% | 96.40% | *96.13%* | *97.23%* |
| | Embedding model 1 (GloVe's initial weights + predictive terms) | Aggregation | RF (n_estimators = 1000) | *97.23%* | 94.61% | 95.90% | 97.11% |
| | Embedding model 2 (predictive terms + retrofitting) | Aggregation | RF (n_estimators = 1000) | 95.42% | 96.58% | 96.00% | 97.13% |
| | Embedding model 3 (SVD combination) | Aggregation | LR (C = 100) | 91.35% | *97.55%* | 94.35% | 95.83% |
| | Embedding model 4 (GloVe's initial weights + predictive terms + retrofitting) | Aggregation | RF (n_estimators = 1000) | 96.90% | 94.46% | 95.67% | 96.95% |

Baselines and embedding models' variations results for Precision (P), Recall (R), F1-Score (F1) and Accuracy (A), using the Aggregation input for all the enhanced embeddings models. The best results obtained by the configurations are in cursive. For all the enhanced embedding models: $\beta_P = 1$ and $\beta_N = 1$.

significant expression of negative emotions by the SUI class; the references to caloric restrictions, body image, laxatives, and body weight of the ED class; the references to antidepressants of the DEP class; and the reference to topics related to leisure activities, which often involve drinking, for the ALC class.

The second element to consider in this discussion is the performance of the predictive models based on the embeddings generated through the approach proposed in this manuscript. Our evaluation shows that the proposal is suitable for addressing domain specific document classification tasks with small corpora as some of the enhanced variations obtain the best results in Recall, F1-Score and Accuracy compared to the embeddings-based baselines for both tasks. Results also show that for the tasks addressed, word embedding based models are less accurate compared to BoW models. This matches the findings of related work dedicated to the detection of depression [19].

Another interesting aspect of this work concerns the performance of the predictive models that use the PSim features, as with only 4 features the accuracy achieved is only 8.01% lower than the one of the best embedding model (embeddings sequence input) for the multi-class task and, only 4.51% lower for the binary task.

As for the limitations of this work, it is important to mention the constrains imposed by the social platform, and the biases introduced by it as we cannot obtain information that clinicians normally get during the screening process, an instance of this is demographic data (location, gender, or age). There are also limitations given by the data collection, and the labelling processes, which did not involve the intervention of human annotators.

Finally, regarding the reproducibility of this work, even if we do not store any personal data or information that can allow the identification of the posts' authors, for ethical reasons, the resulting embeddings' models and the lexicon-based features calculated will be available only under request and justification of their usage purpose [29].

## VIII. CONCLUSION AND FUTURE WORK

We have analyzed Reddit posts related to conditions such as: suicidal ideation, depression, eating disorders and alcoholism; along with control cases, providing a study of lexicon-based features that characterize each condition. We have considered affective, emotional, and linguistic aspects, including psychological risk factors and signs that characterize these conditions. Results show that we can also trace on social media the aspects that (according to specialists) characterize the conditions studied.

We have also proposed an approach for generating enhanced word embeddings for binary and multi-class classification tasks on small and domain specific corpora. The method includes a process to identify predictive terms for each given class. This way, the embedding vectors associated to predictive terms are represented close to each other and far from the representations of terms that are predictive for the remaining classes. These vectors are obtained by extending *Word2vec*'s objective function, which takes those predictive terms as inputs. We evaluated the enhanced embeddings and defined several variations of these embeddings through their

combination with transfer learning and domain adaptation approaches. Furthermore, we defined a feature type denoted as PSim which leverages the properties of the embeddings learned with our approach and can be used as input for any classifier.

Results show that our enhanced embeddings and some of their variations outperform similar embeddings-based methods at least in recall, F1-score and accuracy for both tasks. These findings are promising, and therefore future work involves evaluating the performance of embeddings generated with this method on similar textual multi-class classification tasks with domains that can be characterized by the usage of specialized vocabulary.

It is also important to recall that even if our findings suggest that the detection of mental conditions on social media is possible, a further implementation and deployment of detection tools shall require a proper study of the legal frameworks that regulate such types of tools along with a risk-benefit assessment that addresses the capability of such tools to be misused [3], [29].

## REFERENCES

[1] D. Ramírez-Cifuentes, C. Largeron, J. Tissier, A. Freire, and R. Baeza-Yates, "Enhanced word embeddings for anorexia nervosa detection on social media," in *Advances in Intelligent Data Analysis XVIII. IDA* (Lecture Notes in Computer Science), vol. 12080. Cham, Switzerland: Springer, 2020, pp. 404–417.

[2] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, "Natural language processing of social media as screening for suicide risk," *Biomed. Informat. Insights*, vol. 10, Jan. 2018, Art. no. 117822261879286.

[3] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: An integrative review," *Current Opinion Behav. Sci.*, vol. 18, pp. 43–49, Dec. 2017.

[4] X. Huang, L. Zhang, D. Chiu, T. Liu, X. Li, and T. Zhu, "Detecting suicidal ideation in Chinese microblogs with psychological lexicons," in *Proc. IEEE 11th Int. Conf. Ubiquitous Intell. Comput. IEEE 11th Int. Conf. Autonomic Trusted Comput. IEEE 14th Int. Conf. Scalable Comput. Commun. Associated Workshops*, Dec. 2014, pp. 844–849.

[5] P. L. Úbeda, F. M. Plaza-Del-Arco, M. C. Díaz-Galiano, L. A. U. Lopez, and M.-T. Martín-Valdivia, "Detecting anorexia in Spanish tweets," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process. (RANLP)*, 2019, pp. 655–663.

[6] H.-C. Shing, S. Nair, A. Zirikly, M. Friedenberg, H. Daumé, III., and P. Resnik, "Expert, crowdsourced, and machine assessment of suicide risk via online postings," in *Proc. 5th Workshop Comput. Linguistics Clin. Psychol.: Keyboard Clinic*, 2018, pp. 25–36.

[7] A. Benton, M. Mitchell, and D. Hovy, "Multitask learning for mental health conditions with limited social media data," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 152–162.

[8] E. A. Ríssola, M. Aliannejadi, and F. Crestani, "Beyond modelling: Understanding mental disorders in online social media," in *Proc. Adv. Inf. Retr.: 42nd Eur. Conf. IR Res. (ECIR)*, vol. 12035, 2020, pp. 296–310.

[9] D. Ramírez-Cifuentes, A. Freire, R. Baeza-Yates, J. Puntí, P. Medina-Bravo, D. A. Velazquez, J. M. Gonfaus, and J. González, "Detection of suicidal ideation on social media: Multimodal, relational, and behavioral analysis," *J. Med. Internet Res.*, vol. 22, no. 7, Jul. 2020, Art. no. e17758.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: https://arxiv.org/abs/1301.3781

[11] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[12] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.

[13] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, vol. 1, 2018, pp. 2227–2237.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, vol. 1, 2019, pp. 4171–4186.

[15] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*. [Online]. Available: https://arxiv.org/abs/1910.01108

[16] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith, "Retrofitting word vectors to semantic lexicons," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2015, pp. 1606–1615.

[17] W. Yin and H. Schütze, "Learning word meta-embeddings," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2016, pp. 1351–1360.

[18] J. Tissier, C. Gravier, and A. Habrard, "Dict2Vec: Learning word embeddings using lexical dictionaries," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 254–263.

[19] M. Trotzek, S. Koitka, and C. Friedrich, "Word embeddings and linguistic metadata at the CLEF 2018 tasks for early detection of depression and anorexia," in *Proc. Working Notes CLEF Conf. Labs Eval. Forum*, Avignon, France 2018.

[20] D. Ramírez-Cifuentes, A. Freire, R. Baeza-Yates, N. Sanz Lamora, A. Álvarez, A. González-Rodríguez, M. Lozano Rochel, R. Llobet Vives, D. A. Velazquez, J. M. Gonfaus, and J. Gonzàlez, "Characterization of anorexia nervosa on social media: Textual, visual, relational, behavioral, and demographical analysis," *J. Med. Internet Res.*, vol. 23, no. 7, Jul. 2021, Art. no. e25925.

[21] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *J. Amer. Stat. Assoc.*, vol. 47, no. 260, p. 583, 1952.

[22] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, no. 1, pp. 50–60, 1947.

[23] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Lang. Social Psychol.*, vol. 29, no. 1, pp. 24–54, Mar. 2010.

[24] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Comput. Intell.*, vol. 29, no. 3, pp. 436–465, Aug. 2013.

[25] P. Leichner, "A new treatment approach to eating disorders in youth," *Brit. Columbia Med. J.*, vol. 47, no. 1, pp. 23–27, 2005.

[26] A. Arseniev, H. Lee, T. McCormick, and M. Moreno, "#Proana: Pro-eating disorder socialization on Twitter," *J. Adolescent Health*, vol. 58, pp. 659–664, Jun. 2016.

[27] M. Mowafy, A. Rezk, and H. El-Bakry, "An efficient classification model for unstructured text document," *Amer. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 1, p. 16, 2018.

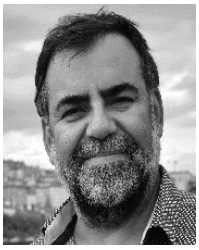
[28] Institute of Medicine (US) Committee on Prevention of Mental Disorders, P. J. Mrazek, and R. J. Haggerty, Eds., "Risk and protective factors for the onset of mental disorders," in *Reducing Risks for Mental Disorders: Frontiers for Preventive Intervention Research*. Washington, DC, USA: National Academies Press, Jun. 1994.

[29] S. Chancellor, M. L. Birnbaum, E. D. Caine, V. M. B. Silenzio, and M. De Choudhury, "A taxonomy of ethical tensions in inferring mental health states from social media," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2019, pp. 79–88.

**DIANA RAMÍREZ-CIFUENTES** studied information systems and computer science engineering from the National Polytechnic School, Quito, Ecuador. She received the master's degree in data and linked systems from Lyon University–UJM-Saint-Etienne, France. She is currently pursuing the Ph.D. degree with the Web Science and Social Computing Research Group, Pompeu Fabra University. Her research interest includes application of computational techniques to address mental health issues on social media.

**CHRISTINE LARGERON** received the Ph.D. degree in computer science from Claude Bernard University, Lyon, France, in 1991, and the HDR degree from Jean Monnet University, in 2004. She is currently a Professor with Jean Monnet University, where she is also the Head of the Complex Data Analysis Team, Hubert Curien Laboratory. She has published more than 100 papers in refereed international conferences and journals. Her main interests include data mining and information retrieval, and her current research focuses on developing methods to efficiently deal with textual and relational data.

**JULIEN TISSIER** received the Ph.D. degree in computer science, in 2020. He is currently a Postdoctoral Researcher with Lyon University–UJM-Saint-Etienne, France. The topic of his thesis was to develop new methods to learn word embeddings to improve the linguistic information contained in these representations, as well as to increase the speed of computation of semantic similarities by using binary vectors. His current research work is focused on learning embeddings for different applications, and specially for temporal dynamic graphs.



**RICARDO BAEZA-YATES** (Fellow, IEEE) received the Ph.D. degree in CS from the University of Waterloo, Canada, in 1989. He is currently a Research Professor with the Institute for Experiential AI, Northeastern University. He is also a part-time Professor with Universitat Pompeu Fabra, Barcelona, Spain, and the Universidad de Chile, Santiago. Before he was the CTO of NTENT, a semantic search technology company-based in California and prior to these roles, he was the VP of Research at Yahoo Labs, based in Barcelona, and later in Sunnyvale, CA, USA, from 2006 to 2016. He is the coauthor of the best-seller *Modern Information Retrieval* textbook (Addison-Wesley, 2nd edition, in 1999 and 2011), that won the ASIST 2012 Book of the Year Award. His research interests include web search and data mining, information retrieval, bias and ethics on AI, data science, and algorithms in general. In 2009, he was named as an ACM Fellow and an IEEE Fellow, in 2011, among other awards and distinctions. From 2002 to 2004, he was elected to the Board of Governors of the IEEE Computer Society. From 2012 to 2016, he was elected to the ACM Council. Since 2010, he has been a Founding Member of the Chilean Academy of Engineering.

**ANA FREIRE** received the Ph.D. degree in computer science. She is currently a Senior Lecturer at UPF Barcelona School of Management. She leads Suicide prevenTion in sOcial Platforms (STOP), a multidisciplinary project to study mental health issues in social media through artificial intelligence. She contributed with more than 40 publications and several patents. Her research has been published in top international journals, such as *JMIR* and conferences, such as ACM SIGIR and ACM WSDM. Her research interests include computational techniques for social good, including applications in sustainability and health.