mathematics

MDPI

Article

Cognitive Emotional Embedded Representations of Text to Predict Suicidal Ideation and Psychiatric Symptoms

Mauricio Toledo-Acosta [1], Talin Barreiro [1], Asela Reig-Alamillo [2], Markus Müller [3], Fuensanta Aroca Bisquert [4,5], Maria Luisa Barrigon [6], Enrique Baca-Garcia [7,8,9,10] and Jorge Hermosillo-Valadez [1,*]

[1] Computational Semantics Laboratory, Centro de Investigación en Ciencias, Universidad Autónoma del Estado de Morelos, Cuernavaca 62209, Morelos, Mexico; mauricio.toledo@uaem.mx (M.T.-A.); tlnedworld@gmail.com (T.B.)
[2] Cognitive Linguistics Laboratory, Centro de Investigación en Ciencias Cognitivas, Universidad Autónoma del Estado de Morelos, Cuernavaca 62209, Morelos, Mexico; assela.reig@uaem.mx
[3] Complex Systems Laboratory, Centro de Investigación en Ciencias, Universidad Autónoma del Estado de Morelos, Cuernavaca 62209, Morelos, Mexico; muellerm@uaem.mx
[4] Instituto de Matemáticas, Unidad de Cuernavaca, Universidad Nacional Autónoma de México, Cuernavaca 62209, Morelos, Mexico; fuen@im.unam.mx
[5] CNRS-UMI 4584–LaSoL Laboratorio Internacional Solomon Lefschetz, Cuernavaca 62210, Morelos, Mexico
[6] Department of Psychiatry, University Hospital Jimenez Diaz Foundation, 28050 Madrid, Spain; luisa.barrigon@quironsalud.es
[7] Department of Psychiatry, University Hospital Rey Juan Carlos, 28933 Mostoles, Spain; EBaca@quironsalud.es
[8] Department of Psychiatry, General Hospital of Villalba, 28400 Madrid, Spain
[9] Department of Psychiatry, University Hospital Infanta Elena, 28342 Valdemoro, Spain
[10] Research Department, Universidad Católica del Maule, Talca 03550, Maule, Chile
* Correspondence: jhermosillo@uaem.mx; Tel.: +52-777-329-7040

Received: 10 September 2020; Accepted: 16 November 2020; Published: 23 November 2020

Abstract: Mathematical modeling of language in Artificial Intelligence is of the utmost importance for many research areas and technological applications. Over the last decade, research on text representation has been directed towards the investigation of dense vectors popularly known as word embeddings. In this paper, we propose a cognitive-emotional scoring and representation framework for text based on word embeddings. This representation framework aims to mathematically model the emotional content of words in short free-form text messages, produced by adults in follow-up due to any mental health condition in the outpatient facilities within the Psychiatry Department of Hospital Fundación Jiménez Díaz in Madrid, Spain. Our contribution is a geometrical-topological framework for Sentiment Analysis, that includes a hybrid method that uses a cognitively-based lexicon together with word embeddings to generate graded sentiment scores for words, and a new topological method for clustering dense vector representations in high-dimensional spaces, where points are very sparsely distributed. Our framework is useful in detecting word association topics, emotional scoring patterns, and embedded vectors' geometrical behavior, which might be useful in understanding language use in this kind of texts. Our proposed scoring system and representation framework might be helpful in studying relations between language and behavior and their use might have a predictive potential to prevent suicide.

Keywords: cognitive-emotional embedded representations; topological-geometrical clustering; suicide ideation prediction

1. Introduction

Suicide is a major public health concern in modern society as it is one of the leading causes of death worldwide [1]. As such, researchers from many areas worldwide have dedicated great effort to preventing suicide [2]. Our approach is to mathematically represent and analyze the linguistic patterns of texts written by subjects under psychiatric treatment who tend to express death wishes. Natural Language Processing (NLP) based on semantic text similarity methods has been applied to this and other similar tasks such as Sentiment Analysis [3,4].

One of the most effective methods used in NLP consists in computing dense vector representations of words by means of neural networks. A telling example is the acclaimed `word2vec` model [5]. This model captures semantic relations between words when trained in large enough corpora possessing a rich grammatical structure and coherence (typical examples of such corpora are the Google News data set, Wikipedia, or the British National Corpus). Thus, nowadays, one can find pretrained embedded representations of words that were trained in this kind of corpora and that are already packaged within some programming languages' NLP libraries (e.g., gensim in Python). However, the writing of subjects suffering from depression may lack the congruence and structure of texts written in healthier emotional conditions. Moreover, people suffering from depression may produce lexical associations that may not be the usual (or "standard") collocations. As a consequence, using pretrained word embeddings may induce lexical similarities that might not reflect the actual usage of language under depression. Moreover, any of such methods fall short in grasping the emotional weight carried by words. Thus, we propose to incorporate cognitive tools in the task.

The relationship between language and emotions has received attention from linguists, psychologists, and those interested, in general, in language and cognition. Among other questions, scholars have investigated the ways in which certain linguistic units, and the choice between available structural or lexical linguistic resources in a communicative event, express emotions and, more specifically, express the producer's (speaker or writer) emotional state [6].

There is consensus that the expression of the internal emotional states can be achieved by different means in oral linguistic communication: prosody, para-linguistic communication, and the lexicon, where lexical units conveying emotions constitute an important semantic domain. In written communication, para-linguistic information and prosody are missing and the search for indices of the writer's emotional state should focus on the semantic analysis of content words.

The emotional state of the speaker can be reflected in the use of different word classes: adjectives, nouns, verbs, and adverbs, among others [7] can convey emotional meanings and, more interesting and challenging, this can be done both because of their literal meaning or by means of indirect, non-literal expressions, such as metaphorical semantic extensions. The complex nature of emotional language offers, therefore, a challenge for its mathematical formalization and processing [8].

The problem of automatically determining the emotional content of a written message falls under the umbrella term of Sentiment Analysis [3,4]. There are countless contributions in the literature dealing with this problem. Our work is a multidisciplinary effort that seeks to mitigate the impact of suicide in contemporary society, by providing computational tools that could eventually lead to a better understanding of those language patterns in this phenomenon. Our aim is to provide new tools for analyzing patterns of language that might underlie suicide and death thoughts. In this paper, we provide a mathematically sound approach to compute cognitive embedded representations of words, grasping and embedding their emotional content.

Our contribution is a geometrical-topological framework for Sentiment Analysis, that includes a hybrid method that uses a cognitively-based lexicon together with word embeddings to generate graded sentiment scores for words, and a new topological method for clustering dense vector representations in high-dimensional spaces. On the one hand, it has been empirically shown that unsupervised dense vector representations of words have a significant potential to discover semantic correlations [9], and even capture latent knowledge [10] in written language. Our hypothesis is that if we appropriately score every word in a corpus according to its cognitive-emotional content, then this score might allow forming new word embeddings, whose properties and geometric behavior may allow the discovery of possibly new semantic associations in depressive states. The quality of these new representations may be assessed via their ability to predict suicidal ideation and psychiatric symptoms in written language. On the other hand, we propose a novel approach to find clusters of word embeddings. The proposed technique is a partitioning method that clusters sparsely distributed dense vector representations in high dimensional spaces. Contrary to subspace clustering methods [11] and density-based methods [12], our approach searches for clusters in the full set of features and performs well in high-dimensional settings. Furthermore, our method does not need to know the number of clusters a priori, as it happens with other partitioning techniques such as those based on K-Means [13–16]. Our framework is useful in detecting word association topics, emotional scoring patterns, and embedded vectors' geometrical behavior, which might be useful in understanding language use in this kind of texts. As other authors have pointed out [17,18], we believe that constructing meaningful models of the emotional content of a text message is of primary importance for exploring and understanding the influence of the multiple linguistic and extra-linguistic factors involved in the analysis of psychiatric symptoms.

The rest of this paper is organized as follows. In Section 2, we review the related work. In Section 3, we describe the corpus, the proposed cognitive emotional framework, and the experimental settings to validate our approach. In Section 4, we present the results obtained. In Section 5, we discuss their interpretation. Finally, in Section 6, we summarize the findings of our work and present some ideas for future research.

2. Related Work

Our work is closely related to the research area of Sentiment Analysis, which has shown an impressive growth over the last two decades with the outbreak of computer-based analysis and the availability of subjective texts on the Web [19]. In this section, we review the articles related to our work in order to put forward our contribution.

Sentiment analysis is concerned with the automatic extraction of sentiment-related information from text. Traditionally, sentiment analysis has been about opinion polarity (positive, negative, and neutral), but in recent years there has been an increasing interest in the affective dimension (angry, happy, sad, etc.) [19]. The main types of sentiment analysis algorithms belong to one of these three classes: Knowledge-based, Machine Learning, and Hybrid systems. We now make a review of related works under these approaches.

Knowledge-based approaches perform sentiment analysis based on lexicons. Lexicons are collections of known terms that, in the case of sentiment analysis, contain sentiment-bearing words or phrases (including adjectives, verbs, nouns, and adverbs) [20] and typically incorporate sentiment word lists from many resources [21,22]. To construct lexicons, one can use or compile dictionaries (usually annotated by humans) or create lists of prototypical words that are further enriched with corpus data by seeking syntactic-semantic similarities. Lexical approaches have been preferred over text classifiers mainly because they are less sensitive to domain adaptation problems and are more flexible to parameterize with lexical, syntactical or semantic cues that depend on context (e.g., political or social contexts) [22].

Under the label of Machine Learning, we can find two kinds of approaches. The first can be viewed as a feature engineering problem, in which the objective is to find a suitable set of affect features in combination with an appropriate Machine Learning technique (e.g., Support Vector Machines, Classification Trees, Probabilistic models, etc.) [23–25]. The second approach concerns the use of neural networks or deep learning architectures to learn *sentiment-specific word embeddings* [26,27]. The core idea in Tang et al. [26] was to automatically rank contexts of words by predicting their polarity and then use the informative contexts for learning the sentiment embeddings of tweets in a novel architecture. Li et al. [27] use prior knowledge available both at word level and document level for training word embeddings for the specific task of sentiment analysis. The main idea was to leverage the intuition behind GloVe [28] (another popular model to compute word embeddings), which is that good word vectors can be learned from the ratios of co-occurrence probabilities. Under a different perspective, Jabreel and Moreno [29] proposed a deep learning architecture to transform the sentiment multi-label classification problem into a single-label (i.e., binary or multi-class) problem.

Hybrid systems combine both approaches. Our work belongs to this category. Here, we can cite Appel et al. [30], whose method uses basic NLP tools, a sentiment lexicon enhanced with the assistance of SentiWordNet [21], and fuzzy sets to estimate the semantic orientation polarity and its intensity for sentences. More recently, Zainuddin et al. [31] proposed a hybrid sentiment classification method for Twitter by embedding a feature selection method. The authors used principal component analysis (PCA), latent semantic analysis (LSA), and random projection (RP) as feature extraction methods. They presented a comparison of the accuracy of the classification process using Support Vector Machine, Naïve Bayes, and Random Forest classifiers. They achieve performance rates of the order of 76%. Wu et al. [32] used text mining techniques to extract symptom profiles and functional impairment diagnoses of major depressive disorders from electronic health records of the Taiwan's National Health Insurance Research database. Wang et al. [33] proposed a Mental Disorder Identification Model by detecting emotions as a sequential pattern, based on a sequential emotion analysis algorithm to identify the intensity of the mental disorder. To this end, they used authoritative criteria of mental disorders and professional blogs about mental health from the psychology community [33].

In relation with our approach, Xue et al. [34] use word embeddings from `word2vec` to compute the Sentiment Orientation [35] of a Weibo (Tweet). In order to attain their goal, the objective of Xue et al. [34] was to construct a Sentiment Dictionary based on a basic dictionary for which each word was previously annotated by humans with its polarity and intensity. The Sentiment Dictionary was constructed by extracting the 100 most similar words to every word in a Weibo, using the cosine similarity measure over the embeddings of both the words of the Weibo and the words of the annotated basic dictionary. Then, the authors proposed scoring methods for computing the Sentiment Orientation for a Weibo. Another example using `word2vec` is Al-Amin et al. [36], who analyze short texts from Bengali micro-blogging websites by using a tagged corpus of comments and sentiment scoring formulae based on empirically (trial and error) tuned parameters to achieve the highest performance.

In medicine, NLP (text mining) and machine learning techniques have been mainly used for analysis of Electronic Health Records, and so assist clinicians in their work [37,38]. Nevertheless, to directly study patients' language is a promising field, especially in psychiatry, as language is the expression of thought, and, consequently, a window into the mind and emotions [39]. NLP techniques have been applied in different psychiatric fields. In schizophrenia, NLP has shown tangential and concrete speech and reductions in semantic coherence and syntactic complexity [40]. In elder people, NLP could predict loneliness with high precision by analyzing audio-taped interviews [41]. NLP has also been used to predict therapeutic alliance analyzing psychotherapy sessions recordings [42], for detecting risk of school violence [43], or for screening suicide risk assessing social media posts [44]. Machine learning techniques for classification (predictive) tasks have also been applied successfully in psychiatry. In [45], the authors compared NLP-based models analyzing free text answering to an open question, against logistic regression prediction models analyzing answers to structured psychiatric questionnaires. In [46], the authors explore some applications of word embeddings to the characterization of some psychiatric symptoms. In [47], adapted word embeddings are used to recognize psychiatric symptoms in psychiatrist texts. Nevertheless, there are challenges to solve when using NLP and machine learning methods in health. Useful patterns should be extracted, since an accurate classification and interpretability of the results is not easy, and human-understandable explanations for classification results are crucial [17,18]. This is the main purpose of our work.

In this context, we would like to stress that our work does not intend to learn sentiment-specific word embeddings such as in Tang et al. [26] or Li et al. [27]. In our approach, word embeddings are instrumental in achieving our sentiment word and message representations. However, in contrast to the above approaches, we use a cognitively-based lexicon together with word embeddings to generate graded sentiment scores for words. Using cognitive information in sentiment analysis is not new. SenticNet 3 [48], for instance, is inspired by neuroscience and cognitive psychology theories on associative memory, in order to build a publicly available semantic and affective resource for concept-level sentiment analysis. Mishra et al. [49] proposed to augment traditional features used for sentiment analysis and sarcasm detection, with cognitive features derived from the eye movement patterns of human annotators recorded while they annotated short text with sentiment labels. The authors used subjective words (Positive words, Negative words) computed against a lexicon [50], obtained from the Multi-perspective Question Answering (MPQA) Opinion Corpus [51]. Zucco et al. [52] proposed a multimodal system based on the acquisition and analysis of vocal signals, facial expression images, and textual posts extracted from social networks.

These works suggest that the problem of grasping emotional content only from text is difficult to tackle. In a recent survey, Hussein [53] concluded that domain dependency is an essential factor related to sentiment evaluation. In this respect, our aim is to provide new tools for analyzing patterns of language that might underlie suicide and death thoughts. Thus, we propose a cognitively-based lexicon suited for the context of the corpus we are dealing with. Therefore, our contribution in this context, is a hybrid method that computes sentiment scores from semantic correlations of words in the corpus with our cognitive-linguistic knowledge base. Moreover, we propose a new topological method for clustering sparsely distributed dense vector representations in high-dimensional spaces. The proposed method bears some similarities to subspace clustering [11] and density-based techniques [12]. The approach of subspace clustering is to search for clusters of points using proper subsets of features [11,16,54]. These methods aim at removing noise or non-relevant features that can mask the structure of a sparse group of points [16,55]; thus, overcoming some of the problems arising from the so-called curse of dimensionality [11,56,57]. Contrary to these approaches, our method searches for clusters in the full set of features, assuming that all features bear similar significance, while keeping in mind the sparsity of the data. When working with word embeddings, the search for clusters in the full set of features is important, as it has been empirically shown that word embeddings might be grouped by semantic similarity (topical similarity), which is a property captured by the whole vector [5]. Similarly to subspace clustering methods, our method performs well in high-dimensional settings. In contrast, some density-based clustering methods such as DBSCAN [58] usually do not perform well with high dimensional data [59] and therefore are hard to use to discover groups of dense word embeddings sharing similarities. On the other hand, one of the most used clustering algorithm in text mining is K-means [13] and its variants, such as Spherical K-Means [14], Multi-Cluster Spherical K-Means [15], and Robust and Sparse K-Means [16]. However, in these methods the number of clusters is a hyperparameter which might not be known a priori. Our method is also a partitioning approach, but does not need to know the number of clusters beforehand as it discovers them automatically. We now proceed to explain our approach in detail.

3. Materials and Methods

Our topological-geometric framework aims at providing representation and analysis tools to grasp and embed the emotional weight carried by each word in the corpus described in Section 3.1. In Section 3.2, we define the scoring and representation system, while in Section 3.3, we define the proposed topological clustering method and state some important properties of the representation transformation. Finally, in Section 3.4 we describe the experimental settings to validate the approach.

3.1. Sampling Methodology and Corpus Description

The corpus in which we test our system consists of short free-form text documents, produced by adults in follow-up due to any mental health condition [45,60,61] .

The project included 20,000 adult outpatients that had attended any of the psychiatric services within the Psychiatry Department of Hospital Fundación Jiménez Díaz in Madrid, Spain. This Department comprises six community mental health centers and is part of the Spanish National

Health Service, providing tax-funded medical care to a catchment area of around 850,000 people. The study was carried out in compliance with the Declaration of Helsinki and approved by the Comité de Ética de la Investigación Clínica de la Fundación Jiménez Díaz (CEIC-FJD), under the license number PIC 76-2013_FJD-HIE-HRJC, approved on 01/28/2014 in act number 01/14.

From May 2014, all clinicians (doctors, clinical psychologists, and nurses) of the Psychiatry Department used the MEmind Wellness Tracker systematically in their clinical activity after receiving specific training, see the works in [45,60,61] for details. All patients in follow-up in the aforementioned facilities were eligible for the study. Inclusion criteria were either male or female outpatients, aged 18 or older, who gave written informed consent approved by the Comité de Ética de la Investigación Clínica de la Fundación Jiménez Díaz (CEIC-FJD), under the license number PIC 76-2013_FJD-HIE-HRJC, approved on 01/28/2014 in act number 01/14. Exclusion criteria were illiteracy, refusal to participate, current imprisonment, being under guardianship and emergency situations in which the patient's state of health did not allow for a written informed consent.

The documents used in this paper are anonymous unstructured answers to an open-ended question related to the participant's current mental state ("how are you feeling today?"). Participants were able to enter responses to questionnaires up to once per day and were instructed to answer as often as they wished. Along with the free text message, participants also had to report how they were thinking about suicide or death in that particular moment.

Out of the total sample, 20% of the patients used the free text field of the MEmind Wellness Tracker [62]. Therefore, the corpus consists of 5489 short free-form text documents and 12,256 tokenized words (i.e., unique words). Previous to the word embedding training phase, the corpus was curated manually in order to avoid misspellings. The minimum length of messages is 1 (i.e., a one-word message), the maximum length is 77 and the average number of words per message is 21.

3.2. Cognitive-Emotional Scoring System and Representation

In this section, we define the Cognitive-Emotional scoring system, our embedded representation of words using this scoring system, and the representation of messages together with their classification. Let the corpus be the set $\mathcal{M} := \{m_1, ..., m_M\}$ of M messages, and let $\mathcal{V}$ be the vocabulary of size N. Let a message m be a sequence of $k \leq N$ words $w_1, ..., w_k$ (not necessarily different from each other), thus we write $m = \{w_1, ..., w_k\}$ to describe the content of message m. We denote by $\mathcal{M}_w$ the multi-set of all messages containing the word w; i.e., if a word occurs more than once in one message, then this message will appear in $\mathcal{M}_w$ the same number of times. Let $F(w)$ be the term frequency of w in the corpus, i.e., the total number of occurrences of w. For the sake of clarity, we will denote by the same letter w the word in the corpus and its respective word embedding $w \in \mathbb{R}^d$.

The semantic similarity between two words w_1 and w_2 will be represented by their cosine similarity:

$$\mathrm{sim}(w_1, w_2) = \frac{w_1 \cdot w_2}{\|w_1\| \|w_2\|}, \tag{1}$$

where $w_1 \cdot w_2$ is the usual inner product in $\mathbb{R}^d$, and $\| \cdot \|$ is the Euclidean norm. As the cosine similarity is the cosine of the angle between w_1 and w_2, we will assume, for the rest of this work, that $\|w\| = 1$ for all word embeddings. Therefore, all word embeddings live in the unit sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$.

The cosine similarity is not a metric in this vector space $\mathbb{R}^d$. Therefore, we will be using the *angular distance* related to the cosine similarity and given by

$$d(w_1, w_2) = \arccos\left(\mathrm{sim}(w_1, w_2)\right), \tag{2}$$

which is a metric on the unit sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ [63]. We will consider the branch of the arccos function given by $\arccos : [-1, 1] \to [0, \pi]$.

It has been empirically shown that word embeddings are able to model syntactic and semantic aspects of language, by leveraging their statistical regularities during the learning process [5,28,64,65].

Their computation from text corpora has been an active research field over the last decade with many applications in healthcare [66–68]. Although many such computational methods now exist, any of them is useful for our purposes, see the work in [69] for an interesting comparative study, and therefore the actual method used for computing the embeddings is not a concern in our approach. We have chosen the model proposed by Mikolov et al. [5], known as `word2vec`, because it is computationally efficient in building high-performance representations of topically related words (e.g., cities and countries) and semantic relations (e.g., analogies of the kind "*man is to woman as king is to queen*") [5,9].

However, semantic similarity alone does not fully account for the emotional weight of words, as word embeddings for positive and negative words can be really close in $\mathbb{R}^d$. The graph of Figure 1 shows an example of this situation; it depicts the nearest neighbors of the word *alegre* (cheerful). The words alegre (cheerful), relajada (relaxed), contenta (glad), and mejores (better) represent positive emotional states, however the word tensa (tense) hints an opposite emotional state.

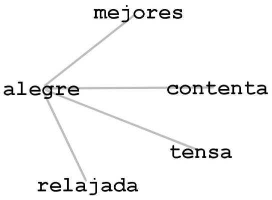

Figure 1. Nearest neighbors of the positive word *alegre* (cheerful). The negative word tensa (tense) is close in the embedding space to the rest of positive words *relajada* (relaxed), *contenta* (glad), and *mejores* (better).

The relationship between language and emotions has received attention from linguists, psychologists, and those interested, in general, in language and cognition. Among other questions, scholars have investigated the ways in which language use, and linguistic communication in general, expresses the producer's (speaker or writer) emotional state [6].

There is consensus that the expression of the internal emotional states can be achieved by different means in oral linguistic communication: prosody, para-linguistic communication, and the lexicon, where lexical units conveying emotions constitute an important semantic domain. In written communication, para-linguistic information and prosody are missing and the search for indices of the writer's emotional state should focus on the semantic analysis of content words.

Based on these facts, the following list represents a simple yet representative *knowledge base* of positive words, denoted by $\mathcal{P}$, i.e., words conveying a cognitively known state of wellness.

$\mathcal{P} :=$ [acompañada (accompanied), adelante (forward), alegre (cheerful), bien (fine), bienestar (well-being), buena (good), bueno (good), capaz (capable), contenta (happy), descansada (rested), descansado (rested), dormida (asleep), dormido (asleep), durmiendo (sleeping), día (day), futuro (future), fácil (easy), mejor (better), mejorar (to get better), mejores (better), positivo (positive), principio (beginning), relajada (relaxed), relajado (relaxed), tranquila (calm), tranquilo (calm), vida (life), vivir (to live), útil (helpful) .]

Similarly, we also consider the following list, which represents the *knowledge base* of negative words, denoted by $\mathcal{N}$, i.e., words conveying a cognitively known state of stress.

$\mathcal{N} :=$ [atrás (backwards), cansada (tired), cansado (tired), despierta (awake), despierto (awake), difícil (difficult), duro (rough), empeorar (to worsen), fin (end), incapaz (helpless), inútil (useless), mal (bad), mala (bad), malestar (discomfort), malo (bad), morir (to die), muerte (death), negativo (negative), nerviosa (nervous), nervioso (nervous), noche (night), pasado (past), peor (worst), peores (worst), sola (alone), tensa (tense), tenso (tense), triste (sad).]

These lists include straightforwardly positive and negative words, such as adjectives, adverbs, and verbs semantically associated with life *vs.* death and wellness *vs.* discomfort. They include as well other words which maintain with these semantic fields a cognitive metaphorical relation, culturally determined but reflected in language use [70]: *day* and *night*, and *beginning* and *end* are conceptual metaphors for life and death; likewise, *to live* is associated with *moving* or *looking forward*, *towards the future*, unlike their opposites (*backwards* and *past*).

Our scoring system intends to account for the cognitive-emotional weight carried by each word in the corpus. On the one hand, the cognitive aspect is implicit in the tag of each message, which reflects the state of anxiety that each subject recognizes in herself/himself. We now define and describe the implementation of this scoring system in detail. We will assign to each message a weight

$$\theta_m \in [0,1],$$

which is the weight of a message related to its tag. Each message has two self-reported tags given by the subject (see the work in [61] for details).

During data collection, each participant answered the question *"Have you ever felt that you had no desire to live?"* using a 6-point Likert scale offering the following options; 0: "all the time", 1: "most of the time", 2: "more than half of the time", 3: "less than half of the time", 4: "occasionally", and 5: "never". From the point of view of a classification task, this labeling scheme produced very unbalanced data sets. Thus, we combined these labels in what we called the cognitive confidence of a message $P(m)$, in order to deploy 3 balanced data sets. Thus, the cognitive confidence of a message is described by

$$P(m) = \begin{cases} 0, & \text{corresponding to score 5 in the Likert scale} \\ 1, & \text{corresponding to scores 3 or 4 in the Likert scale} \\ 2, & \text{corresponding to scores 1 or 2 in the Likert scale.} \end{cases}$$

These will be the classes we want to predict, using a classifier **C** as will be described later in this subsection. It is now possible to provide concrete values for θ_m, the weight of a message related to its tag.

Conventionally, for each message m, we will say that the cognitive weight is $\theta_m = 1$, if the message has a cognitive confidence ($P(m)$) value of 0 or 2. With this, we suggest that the message has a "maximum" cognitive weight, i.e., the certainty or level of confidence of the subject about his/her emotional state is the highest. On the other hand, if the message m has a $P(m)$ value of 1, we will say that $\theta_m = 0.5$. Therefore, we define θ_m as follows.

$$\theta_m = \begin{cases} 1, & \text{if } P(m) = 0 \text{ or } P(m) = 2 \\ 0.5, & \text{if } P(m) = 1 \end{cases} \tag{3}$$

Intuitively, the emotional weight of each word must depend also on the lexical collocations in each message (i.e., the association of the word with positive or negative words). As we have said, the sets $\mathcal{P}$ and $\mathcal{N}$ each represent a knowledge base about the possible emotional state of the subjects. Therefore, we can assign a *positive* weight (θ_m) to the words of $\mathcal{P}$ and a *negative* weight ($-\theta_m$) to the words of $\mathcal{N}$. If the word does not belong to any of the previous sets, we will consider its semantic relatedness to any of the terms in those sets as computed by Equation (1).

Before defining the scoring system, we need to state some definitions. First, we denote by $\overline{\mathcal{P}}$ and $\overline{\mathcal{N}}$ the sets containing up to k nearest neighbors, having strictly positive cosine similarity with respect to every word in $\mathcal{P}$ and $\mathcal{N}$, respectively. The number k is a hyperparameter of this model.

Now, we define a directed graph $G = (V, E)$ such that the vertices are given by

$$V = \mathcal{P} \cup \mathcal{N} \cup \overline{\mathcal{P}} \cup \overline{\mathcal{N}}.$$

The directed edges of G are determined by the following condition. If $z \in \mathcal{P} \cup \mathcal{N}$ and $w \in \overline{\mathcal{P}} \cup \overline{\mathcal{N}}$, then $\overrightarrow{z,w} \in E$ if and only if w is one of the k-nearest neighbors of z.

Finally, for every word $w \in \overline{\mathcal{P}} \cup \overline{\mathcal{N}}$, we define

$$\begin{aligned} \sigma(w) &= \max_{\overrightarrow{z,w} \in E} \operatorname{sim}(w,z) \\ \eta(w) &= \operatorname*{argmax}_{\overrightarrow{z,w} \in E} \operatorname{sim}(w,z) \end{aligned} \tag{4}$$

The purpose of Equation (4) is to formalize the association between every word $w \in \overline{\mathcal{P}} \cup \overline{\mathcal{N}}$ and its nearest neighbor in $\mathcal{P} \cup \mathcal{N}$, given by $\eta(w)$. In Figure 2, we show an example of the graph G.

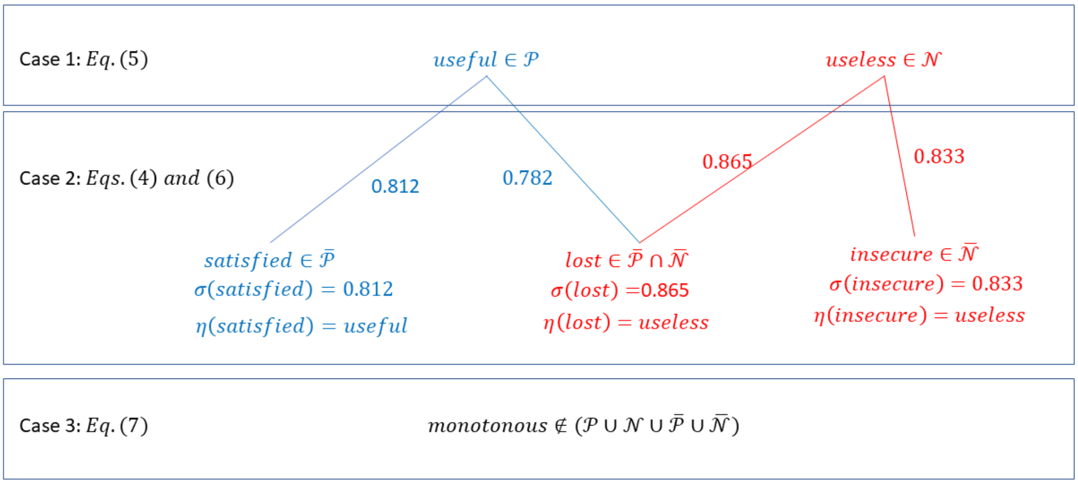
Case 1: *Eq.* (5)
useful ∈ $\mathcal{P}$
useless ∈ $\mathcal{N}$
Case 2: *Eqs.* (4) *and* (6)
0.812
0.782
0.865
0.833
satisfied ∈ $\bar{\mathcal{P}}$
$\sigma(satisfied) = 0.812$
$\eta(satisfied) = useful$
lost ∈ $\bar{\mathcal{P}} \cap \bar{\mathcal{N}}$
$\sigma(lost) = 0.865$
$\eta(lost) = useless$
insecure ∈ $\bar{\mathcal{N}}$
$\sigma(insecure) = 0.833$
$\eta(insecure) = useless$
Case 3: *Eq.* (7)
monotonous ∉ ($\mathcal{P} \cup \mathcal{N} \cup \bar{\mathcal{P}} \cup \bar{\mathcal{N}}$)

Figure 2. Part of the graph G showing only two of the nearest neighbours of *útil (useful)* and *inútil (useless)*. The words *satisfecha (satisfied)* and *perdida (lost)* are two of the nearest neighbours of *útil (useful)*. On the other hand, the words *perdida (lost)* and *insegura (insecure)* are two of the nearest neighbours of *inútil (useless)*. The similarities between pairs of neighbours are shown along the respective edges. At the bottom of the Figure, we show the word *monotono (monotonous)* that does not belong to either $\mathcal{P} \cup \mathcal{N}$ or $\overline{\mathcal{P}} \cup \overline{\mathcal{N}}$, and therefore is not a vertex of G.

The score of the word w will be denoted by $s(w)$. In order to compute $s(w)$, we will first introduce a pre-scoring function $\tilde{s}_m(w)$ for each message m where the word w appears. We define $\tilde{s}_m(w)$ using three rules as follows.

1. If $w \in \mathcal{P} \cup \mathcal{N}$ then:

$$\tilde{s}_m(w) = \begin{cases} \theta_m, & \text{if } w \in \mathcal{P} \\ -\theta_m, & \text{if } w \in \mathcal{N} \end{cases} \tag{5}$$

2. If $w \notin \mathcal{P} \cup \mathcal{N}$ and $w \in \overline{\mathcal{P}} \cup \overline{\mathcal{N}}$ then:

$$\tilde{s}_m(w) = \begin{cases} \theta_m \cdot \sigma(w), & \text{if } \eta(w) \in \mathcal{P} \\ -\theta_m \cdot \sigma(w), & \text{if } \eta(w) \in \mathcal{N} \end{cases} \tag{6}$$

3. If $w \notin \mathcal{P} \cup \mathcal{N}$ and $w \notin \overline{\mathcal{P}} \cup \overline{\mathcal{N}}$ then:

$$\tilde{s}_m(w) = 0. \tag{7}$$

Thus, the cognitive-emotional score of each word can be calculated thanks to the cognitive confidence of each message and to its semantic relatedness with the terms of the knowledge base.

In order to illustrate the workings of Equations (4)–(7), we propose to observe Figure 2 which gives an example of the directed graph G and how these equations operate. The figure shows the three possible scenarios that we might encounter when computing $\tilde{s}_m(w)$ as we go through the messages where the word w appears. The first case is when the analyzed word in the message is actually one of the prototypical words, either belonging to $\mathcal{P}$ or $\mathcal{N}$. The example in Figure 2 shows an hypothetical message m containing the words w_1 = useful $\in \mathcal{P}$ and w_2 = useless $\in \mathcal{N}$, which are nodes of the directed graph G (see top of the Figure). In this case, Equation (5) assigns the values $\tilde{s}_m(w_1) = \theta_m$ and $\tilde{s}_m(w_2) = -\theta_m$, respectively, in this particular message. The second case is when some of the words contained by m are in $\overline{\mathcal{P}} \cup \overline{\mathcal{N}}$. In this situation, it is also possible that a word belongs to both sets as is illustrated by the word *lost* in the middle part of Figure 2. In this case, Equation (4) solves the conundrum as it retrieves the closest neighbor from either list. Then, Equation (6) computes the value $\tilde{s}_m(w)$ for each word. Finally, if the word does not belong to any set, Equation (7) assigns zero to $\tilde{s}_m(w)$.

Our final score for each word $s(w)$ is the sum of weighted scores $\tilde{s}_m(w)$ normalized by the term frequency:

$$s(w) = \frac{1}{F(w)} \sum_{m \in \mathcal{M}_w} \tilde{s}_m(w). \tag{8}$$

It is straightforward to verify that $-1 \leq s(w) \leq 1$.

Finally, we describe the proposed Cognitive-Semantic representation of words, based on this scoring system. This representation is a transformation of the original `word2vec` embeddings. More precisely, by a Cognitive-Semantic representation of a word w, we mean a vector representation $R(w)$ given by

$$R(w) = s(w) \cdot w \in \mathbb{R}^d \tag{9}$$

where $s(w)$ is the score $s(w)$ defined by Equation (8). These points are inside the closed unit ball $B^d \subset \mathbb{R}^d$, since $\|w\| = 1$ and $|s(w)| \leq 1$. This fact is illustrated in Figure 3.

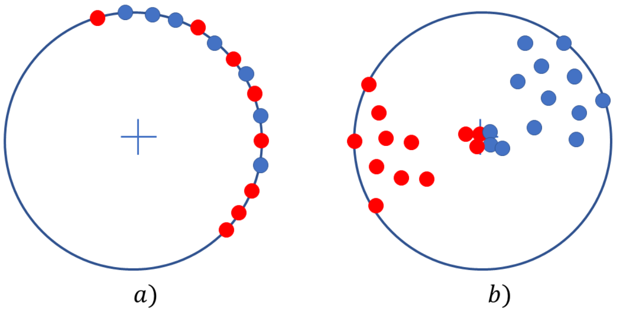

Figure 3. (**a**) Points in the original embedding space lie in the unit hypersphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$, represented by the circumference, words are not grouped by polarity. (**b**) After applying the transformation R, the points $R(w)$ lie inside the closed unit ball B^d, represented by the closed circle, points are now grouped by polarity.

Using (9), we compute a representation for each message in the corpus. This representation is obtained as the mean of the representations of the words in the message. Thus, for each message $m = \{w_1, ..., w_n\}$ we compute the vector $\mathbf{R}(m) \in \mathbb{R}^d$ defined as

$$\mathbf{R}(m) = \frac{1}{n} \sum_{i=1}^{n} R(w_i). \tag{10}$$

In order to validate these representations, we propose to assess their relevance in predicting the messages' tags by means of a classification task.

To this end, we consider the matrix $A \in \mathcal{M}_{M\times d}$ of message representations and the column vector $b \in \mathcal{M}_{M\times 1}$ of message tags given by

$$A = \begin{pmatrix} \mathbf{R}(m_1) \\ \dots \\ \mathbf{R}(m_M) \end{pmatrix}, \quad b = \begin{pmatrix} P(m_1) \\ \dots \\ P(m_M) \end{pmatrix}, \tag{11}$$

where $P(m_i)$ is the the cognitive confidence of the message m_i as described above.

3.3. *Topological-Geometrical Clustering and Properties of the Transformation R(w)*

In this section, we first describe the proposed clustering method to discover groups of words in low-density embedding spaces such as the representation space of this corpus. Then, we prove an important geometrical property of the transformation $R(w)$ (Equation (9)). Both the clustering method and the property of $R(w)$ stem from the fact that we are using the angular metric, instead of the usual Euclidean distance over the embedding space.

As we will argue (see Discussion), the corpus we are mining is neither *rich* nor *complex* enough to produce the semantic regularities expected when applying `word2vec`. When trained in large corpora (millions of words), `word2vec` tends to group words together by topic forming clusters of points in the embedding space. In our case (few thousands of words), this is not totally true. After applying current clustering techniques, such as K-means [13], DBSCAN [58], or Agglomerative Clustering [71], we obtained 1 or 2 clusters depending on the hyperparameters, even if we know there are much more topics in the corpus (e.g., Health: pain, headache, illness, etc.; Relationships: mother/father, brother/sister, son/daughter, etc.). However, after inspecting the nearest neighbors of words related to each other by some topics, specific to the type of this corpus (e.g., Drugs), it was evident that some of their embeddings were close to each other. On the other hand, words related to the topic of *Places*, for instance, were not very close to each other, even when `word2vec` usually groups words from this kind of topic together. Therefore, the motivation for investigating a new method to discover groups of words plausibly having a topical relation.

These observations are the main motivation behind our clustering algorithm. The proposed method is a contribution at a nascent state. Our clustering technique bears some similarities to subspace clustering, which has been proposed to overcome high-dimensional clustering problems [11,59,72,73], it also bears some similarities to density-based methods such as DBSCAN [58]. However, to the best of our knowledge, it is a novel algorithm.

Although proving the formal properties of our method is out of the scope of this paper, we provide insight into its functioning and its empirical implementation. Future work concerns the formal analysis of the method under a solid and sufficiently exhaustive experimental setting. Before we describe our clustering technique, we need to define and recall some geometric observations of the space $\mathbf{R}^d$.

3.3.1. Definitions and Preliminary Observations

The diameter of a finite set $D \subset \mathbb{R}^d$ is defined as

$$\operatorname{diam}(D) = \max\{d(x,y) \mid x,y \in D\}, \tag{12}$$

where d is the angular metric defined in Equation (2). By a set of clusters we mean a collection of sets of words whose vectors form a set of diameter less than or equal to a certain diameter ρ. Intuitively, this diameter ρ has to be small enough such that the similarity between any pair of words in a same set is positive. In the following definition, we formalize this intuition.

Definition 1. *Let ρ be a small enough diameter, we denote by $\mathcal{D}$ a collection of pairwise disjoint finite subsets of $\mathbb{R}^d$ such that each element $D \in \mathcal{D}$ satisfies $diam(D) \leq \rho$ and $sim(x, y) > 0$ for all $x, y \in D$.*

It is worth noting that Definition 1 says nothing about the distance between clusters or about the quality or validation of consistency of the clustering method. Moreover, the reader must bear in mind that if we relax the constraint of $d(x, y)$ being a metric, to become just a distance or a semi-metric [63], nothing changes for Definition 1. Therefore, the collection $\mathcal{D}$ can be assumed to be either a collection of well-formed clusters, in the sense that they show good separability, or a mere partition of the set of word embeddings.

We will now divide the space $\mathbb{R}^d$ into 2^d open regions, called hyperoctants (this is the general term; in $\mathbb{R}^2$ they are called quadrants, and in $\mathbb{R}^3$ octants.) . A hyperoctant is a region of space defined by one of the 2^d possible combinations of coordinate signs $(\pm, ..., \pm)$. For example, in $\mathbb{R}^3$, the following combinations of signs are two hyperoctants; $+++$ and $-+-$, the former denotes the set of points with three positive coordinates, and the latter the set of points with the first and third coordinates negative and the second positive. As it can be readily seen, in low-dimensional spaces (e.g., $\mathbb{R}^2$), there would be only four regions. However, in general, the number of words in a corpus is significantly smaller than the number of hyperoctants in the embedding space. The typical embbeding space has dimension $d = 100$, and therefore has a number of hyperoctants in the order of 10^{32}; the size of a typical large vocabulary is in the order of 10^6 words. This observation means that the hyperoctant of each point may be a useful label for the location of the point in space, as the number of labels (hyperoctants) is much larger than the number of points. This is the main intuition behind our clustering method.

The method consists in searching for clusters by searching through a graph whose nodes are connected by minimal Levenshtein distances ([74]) between hyperoctants. Instead of searching groups of points verifying some angular distance criterion, we search for points that have *close* hyperoctants, where closeness is defined by a hyperparameter.

Now, we construct a graph $G_d = (V, E)$ associated to the set of hyperoctants of $\mathbb{R}^d$. The set V is the set of hyperoctants labeled by their respective sign combination ϕ. Two sign labels ϕ_1 and ϕ_2 are connected by an edge in E if the Levenshtein distance between ϕ_1 and ϕ_2 is 1. In Figure 4 we show an example of the graph G_d for $d = 3$.

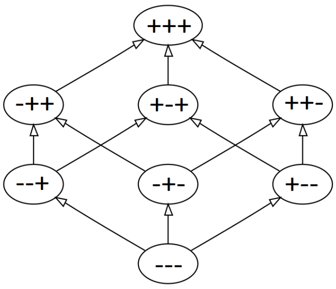

Figure 4. The graph G_d for $d = 3$ representing the spatial location of the 8 octants of $\mathbb{R}^3$.

Now, we define the density of a finite set $A \subset \mathbb{R}^d$ as

$$\delta(A) = \frac{|A|}{\text{diam}(A)}.$$

This function accounts for how dense the set A is, in terms of how many elements it contains with respect to the maximum distance between the elements.

3.3.2. Hyperoctant Tree Search Clustering Method

We can now describe the clustering method proposed in this paper. Consider the graph G_d and denote by U the set of vertices of G_d such that their respective hyperoctants contain one word embedding. The hypothesis is that the sparsity of the points in the embedding space is such that every hyperoctant contains at most one word embedding. Although this requires a deeper analysis, we have verified that this is the case in this corpus. The clustering method consists of applying the following steps.

1. We first choose a density δ_0, which will be the minimum density per cluster.
2. Now, we perform a Depth-first search on the entire graph starting at the bottom vertex $-\ldots-$.
3. The first time we encounter a vertex in U, we add the word embeddings w in this vertex to a set D, and we write $D = \{z_1\}$ as a short notation meaning that the embedding in that node was added.
4. We continue with the search until we encounter another vertex $z_2 \in U$. We add this vertex to D only if
$$\operatorname{diam}(D) \leq \frac{|D|}{\delta_0}$$
5. We now search for each nearest neighbor y of each word in D and add it to D if
$$\operatorname{diam}(D \cup \{y\}) = \operatorname{diam}(D).$$
6. We recursively apply steps 4 and 5, until we encounter a vertex such that its inclusion in D violates the condition in step 4, or a neighbor increases $\operatorname{diam}(D)$ in step 5. In such a case, we create a new cluster and restart the process at step 3 until all vertices and neighbors have been explored.

Thanks to Definition 1, we may now state and prove an important property of transformation $R(w)$ (Equation (9)), regarding its geometric behavior.

Proposition 1. *Let $\mathcal{D}$ be a collection of clusters complying to Definition 1. For any $D \in \mathcal{D}$, it holds $\operatorname{diam}(D) \leq \operatorname{diam}\left(R(D) \setminus \{\mathbf{0}\}\right)$, where $\mathbf{0} \in \mathbb{R}^d$ is the origin.*

Proof. For any $D \in \mathcal{D}$ and any $x, y \in D$ such that $R(x), R(y) \neq \mathbf{0}$, we have $\operatorname{sim}(x,y) > 0$ (Definition 1). Using the properties of the dot product, it follows that

$$\operatorname{sim}\left(R(x), R(y)\right) = \frac{s(x)s(y)}{|s(x)| \cdot |s(y)|}\operatorname{sim}(x,y) = \begin{cases} \operatorname{sim}(x,y) & \text{, if } s(x)s(y) \geq 0 \\ -\operatorname{sim}(x,y) & \text{, if } s(x)s(y) < 0. \end{cases}$$

As $|s(x)|, |s(y)| \leq 1$, it follows that $\operatorname{sim}(x,y) \geq \operatorname{sim}\left(R(x), R(y)\right)$. Now, the function arc-cosine is decreasing in its domain $[-1, 1]$; therefore, we have

$$d(x,y) \leq d\left(R(x), R(y)\right).$$

Therefore, $\operatorname{diam}(D) \leq \operatorname{diam}\left(R(D)\right)$. □

We would like to stress that Proposition 1 holds even if distance (2) may no longer be a metric in the transformed space $R(\mathbb{R}^d)$.

3.4. Experimental Settings

In order to obtain the word embeddings, we used the model `word2vec` from the Python module `gensim`. The hyperparameters used were

$$\text{min\_count} = 1, \quad \text{size} = 100, \quad \text{workers} = 4, \quad \text{window} = 5, \quad \text{iter} = 30.$$

Thus, for every word w, we have a vector in $w \in \mathbb{R}^d$, with $d = 100$. The hyperparameter k of the scoring system was set to $k = 20$. The lists $\mathcal{P}$ and $\mathcal{N}$ have sizes 29 and 28, respectively.

The classifier **C** used on the system A, b defined in (11) was a boosted decision tree implemented with `xgboost` (Extreme Gradient Boosting). We split the dataset of messages into training, validation and testing subsets with the respective ratios 18:2:5. The hyperparameters of the classifier were tuned using cross-validation.

We compare the classification task implementing Cognitive-Semantic representations against using only word2vec embedded vectors. Namely, we use the classifier **C** on the system A, b described in (11), we call this system $\mathcal{S}_1$. On the other hand, we use the same classifier **C** on the system A', b, where $A' \in \mathcal{M}_{M\times d}$ is defined as follows,

$$A' = \begin{pmatrix} \mathbf{R}'(m_1) \\ \dots \\ \mathbf{R}'(m_M) \end{pmatrix},$$

where $\mathbf{R}'(m)$ is a vector representation of each message m, given as the average of the word embeddings contained in the message. Specifically, if $m = \{w_1, ..., w_n\}$, then

$$\mathbf{R}'(m) = \frac{1}{n}\sum_{i=1}^{n} w_i.$$

We denote by $\mathcal{S}_2$ the system A', b.

We perform this comparison between both systems $\mathcal{S}_1$ and $\mathcal{S}_2$ to assess the effect of transforming the `word2vec` embeddings.

4. Results

In this section, we present the results obtained with the described methods of Section 3 applied to this particular corpus. We first show the results on clusters constructed with our method described in Section 3.3. We then present our results using the scoring and representation system. Finally, we present the results on the classification task.

After applying the `word2vec` model to the corpus, word embeddings in $\mathbb{R}^d$ are obtained. The graph depicted in Figure 5 represents the set of clusters $\mathcal{D}$ in $\mathbb{R}^d$ with $d = 100$, obtained with our clustering method described in Section 3.3 and which comply to Definition 1 with $\rho = 1.35$ rad and $\delta_0 = 29$. These clusters were obtained before we apply the transformation $R(w)$ (Equation (9)). The clusters have been labeled by topic. The size of the node is proportional to the diameter of the cluster. The size of the edges are proportional to the similarity between the centroids of the clusters and the labels on the edges are the angular distances between the centroids of the clusters. The edges were drawn only if the angular distance between these centroids was smaller than an arbitrary threshold 0.8. It is worth noting that this graph shows that some clusters are very close to each other. The reader must bear in mind that our definition of clusters only asks for sets to be pairwise disjoint (Definition 1). Consider the convex hull of a cluster in the embedding space, then a pair of convex hulls may overlap (e.g., see Figure 6), but the sets by themselves do not overlap. On the other hand, recall that distances are in radians and that our proposed clustering method ensures that the sets of points (clusters) are disjoint in the sense that every point in one cluster remains within the diameter of that cluster.

In order to visualize these clusters, Figure 6 shows a bidimensional reduction of the `word2vec` word embeddings, using PCA and keeping the first two principal components. We discuss these clusters in the next section.

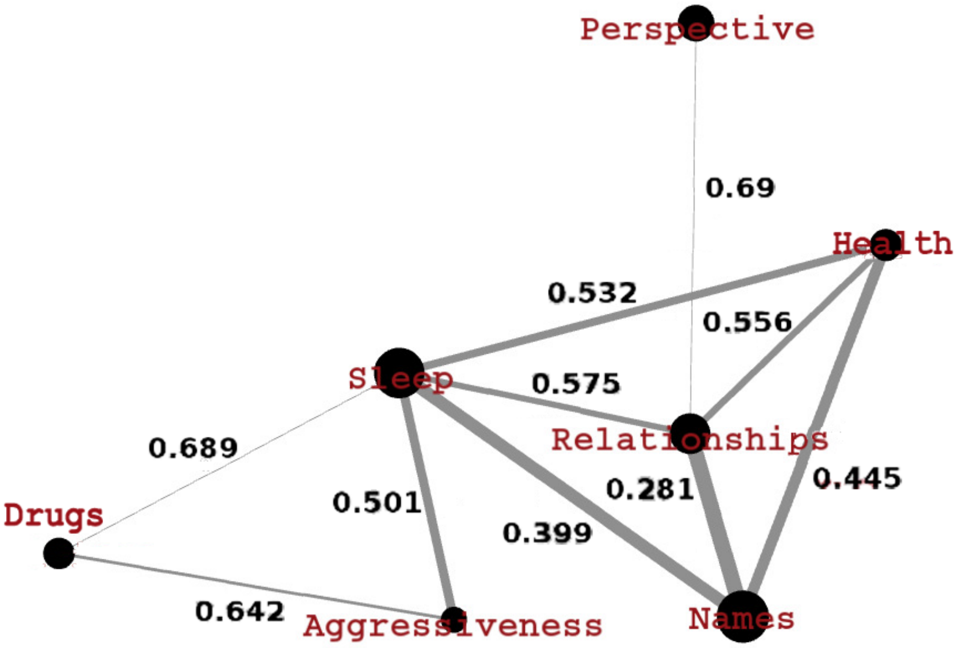

Figure 5. Clusters of words in $\mathbb{R}^d$, depicted as a graph. The nodes have been labeled by topic. The clusters were obtained after applying `word2vec` and before transformation $R(w)$.

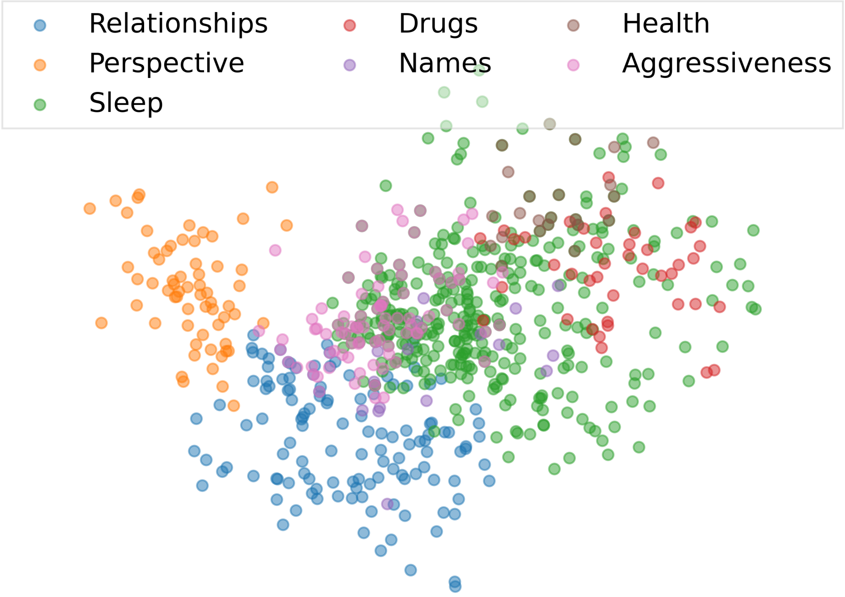

Figure 6. Clusters of word embeddings by topic, depicted using principal component analysis (PCA).

Table 1 describes these clusters and their topics, including some of their members. The name of the topics were arbitrarily chosen in function of the semantic contents that we inspected after their construction.

Table 1. Description of clusters depicted in Figure 5 including some members.

Cluster/Topic	**Topic Description**	**Some Members**
Relationships	Relationships between members of a family, friends or acquaintances.	brother, sister, mother, grandmother, daughter, son, neighbour.
Perspective	Words related to the future or the attitude towards life in general	To live, peace, to abandon, to face, perspective, to hate, frustration.
Sleep	Words related to sleep or rest.	sleep, sleepless, asleep, incubate, inactive, rested.
Names	Names or last names of people	Luisa, Juan, Yolanda, Teresa, Sánchez.
Health	Words regarding injuries, sickness or ailments.	sclerosis, throat, exhaustion, muscular, ocular.
Aggressiveness	Words bearing various levels of aggressiveness.	To warn, to trample, excuses, wall, break up, to cheat, to demand, slander, bruise.
Drugs	Names and terms related to medications, mainly antidepressants.	Escitalopram, lormetazepam, lorazepam, ibuprofen, orfidal, paroxetine, sycrest.

After applying our scoring system, we obtained 922 words with non-zero score, 384 positive words (words whose score is greater than 0), and 468 negative words (words whose score is less than 0). Within the positive words, 47% comes from the list $\overline{\mathcal{N}}$, 46% from the list $\overline{\mathcal{P}}$, and 7% from $\mathcal{P}$. Within the negative words, 55% comes from the list $\overline{\mathcal{N}}$, 38% from the list $\overline{\mathcal{P}}$, and 7% from $\mathcal{N}$. The sizes of the lists $\overline{\mathcal{P}}$ and $\overline{\mathcal{N}}$ were 384 and 461, respectively. In Table 2, we show the top 20 words with the most positive score and the top 20 words with the most negative score, along with its English translation. Figure 7 shows the histogram of the non-zero scores of words in the corpus.

Table 2. Top 20 most positive and most negative words, respectively.

Word	Score	Translation	Word	Score	Translation
descansada	1.0	rested	tenso	−1.0	tense
mejorar	1.0	to get better	empeorar	−1.0	to get worse
acompañada	1.0	accompanied	muerte	−1.0	death
principio	1.0	beginning	peores	−1.0	worst
mejores	1.0	best	atrás	−0.9783	backwards
descansado	1.0	rested	malo	−0.9726	bad
bienestar	1.0	well−being	fin	−0.9655	end
durmiendo	0.9821	sleeping	mala	−0.96	bad
adelante	0.9818	forward	duro	−0.9583	rough
positivo	0.9787	positive	negativo	−0.95	negative
útil	0.9706	useful	nervioso	−0.949	nervous
dormido	0.9684	asleep	sol	−0.9474	sun
relajada	0.9661	relaxed	pasado	−0.947	past
tranquilo	0.965	serene	cansado	−0.9423	tired
fácil	0.96	easy	triste	−0.9364	sad
tranquila	0.9583	serene	librarme	−0.934	freeing my self
contenta	0.9516	content	arrepentirme	−0.92	to regret
alegre	0.95	happy	cansada	−0.9194	tired
buena	0.9494	good	nerviosa	−0.9159	nervous
relajado	0.9286	relaxed	concentrándome	−0.915	focusing

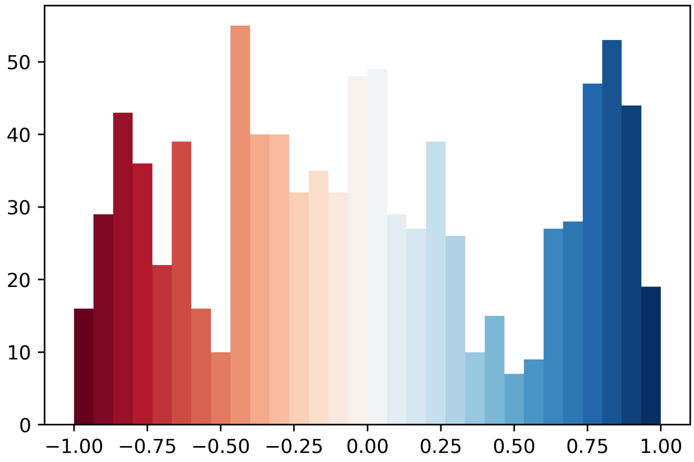

Figure 7. Histogram of the non-zero scores of words in the corpus. The colors are related to the score.

Once we have the scores of each word, we computed representations $R(w)$. The resulting representations are the original embeddings scaled and polarized by the scores (Equation (9)). In order to visualize the resulting vectors, Figure 8 shows a bidimensional reduction of the representations $R(w)$ using PCA.

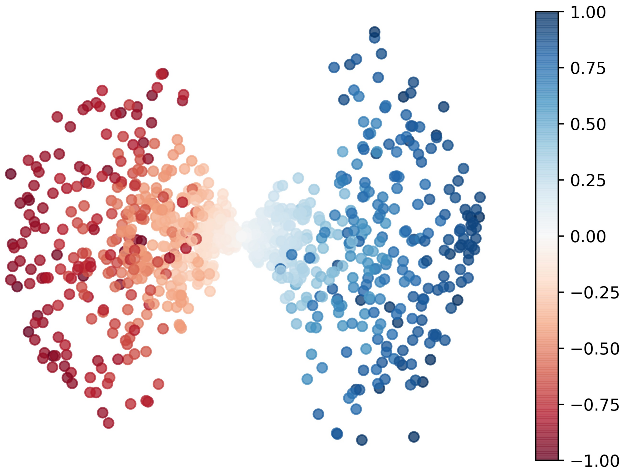

Figure 8. Bidimensional reduction of the points $R(w)$, using PCA. The color denotes the score $s(w)$.

If we visualize the clusters of Figure 6, after substituting in each cluster the original word embeddings by the corresponding transformations $R(w)$, we obtain Figure 9.

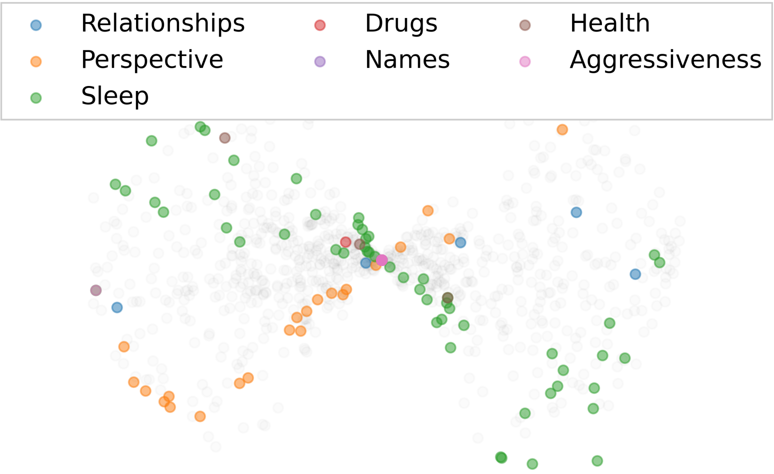

Figure 9. Bidimensional reduction of the points $R(w)$ in each cluster of $\mathcal{D}$, using PCA. The gray points are the rest of the words in the corpus.

Regarding the classification task, the accuracy metrics of the implementation of the classifier **C** on $\mathcal{S}_1$ and $\mathcal{S}_2$ were 62.99% and 62.13%, respectively.

5. Discussion

Our proposed clustering method shows good qualitative results as shown by Figures 5 and 6. These figures show that our method was able to find pairwise disjoint sets of words with a topical relation between the elements of each set. The clustering method is applied in the original embedding space, before we apply our transformation $R(w)$. In this sense, the clusters only reflect how the words are grouped by topic in Figure 6.

After applying the transformation $R(w)$, Figure 8 shows a net difference between what we called *positive* and *negative* words (in a cognitive sense). To the authors' knowledge, there has not been such a clear distinction between the sentiment of words in a corpus in the literature. This net division between polarity assessments brings new insights about the use of language, which can be leveraged with topic similarity issues. In these respects, the clusters of Figure 6 are observed differently after applying $R(w)$ as shown in Figure 9. This is a direct consequence of $R(w)$. Figure 9 shows the sets $R(D)$ for each D in our cluster collection $\mathcal{D}$; in other words, they are not new sets computed with our method. In these respects, these transformed sets do not have to satisfy Definition 1. Therefore, the transformation of the clusters, previously obtained by our clustering method, does not yield clusters in the transformed space $R(w)$ in general. This is where the interest of our scoring system takes on all its relevance as we will explain now.

Figure 9 shows two things: On the one hand, it shows that members in a same cluster, related by a single topic, may be differentiated by emotional polarity. On the other hand, it also shows that some topics have a "neutral" polarity in the sense that their members distribute more densely around the origin when using representations $R(w)$; in other words, members of these topics do not neatly separate in this polarity space, or may even collapse to the origin. This is particularly the case of *Drugs*, *Names*, and *Aggressiveness* topics.

This empirical phenomenon is the manifestation of Proposition 1, which states the following idea. As we discussed before, each cluster $D \in \mathcal{D}$ is a set of points that are close together in $\mathbb{R}^d$. After applying the transformation $R(w)$, some points in each cluster might be moved away far from their former neighbors, thus increasing the diameter of the set $R(D)$, i.e., members in a same cluster, may be differentiated by emotional polarity after using the transformation $R(w)$. In geometric terms, some clusters expand in diameter when they are represented by coordinates $R(w)$. It is important to note again that these sets $R(D)$ might no longer be a cluster (under the original criterion of Definition 1) in the new *polarity* space $R(\mathbb{R}^d)$. Moreover, new clusters might form with the union of subsets from the original clusters; in other words, there might be a reconfiguration of the original clusters. We did not apply our clustering method to this new polarity space, as it was out of the scope of this paper.

It is worth observing the particular case of *Sleep* and *Perspective* topics, as Figure 9 shows that some members separate in positive/negative subsets. This might certainly imply semantic nuances that could lead to a more in-depth analysis. Still, it could also imply that certain words that we would consider as positive or negative appear on the opposite side of the polarity space, because this representation does not allow us to identify meanings of the "negation" type, which could explain this phenomenon. This might be illustrated by the word "sol" (sun), that appears (perhaps unexpectedly) on the negative side of Table 2. This is a research direction in the future. Notwithstanding, these results allow us to hypothesize that the proposed scoring and representation system might be helpful in studying relations between language and emotions. This is certainly a future research vein that might produce interesting and useful results.

Observing now the scoring results, there is an intersection of the lists reported in Table 2 with the respective lists $\mathcal{P}$ and $\mathcal{N}$. However, there are positive words in $\mathcal{P}$ that do not appear in the most positive words of Table 2. For instance, the word *bien* (fine) is in the knowledge base of positive words, but not among the most positive words, according to the scoring system. A similar situation occurs with some negative words in $\mathcal{N}$.

This phenomenon is explained by the fact that there are words that seem to be positive/negative, on its own, but they may appear in many messages with distinct θ_m weights (i.e., in messages expressing a more or less level of anxiety), which makes that when their scores are normalized by their frequency, their overall score lowers down. However, any word in $\mathcal{P}$ has a higher score than any of the words in $\mathcal{N}$ as expected. This is shown in Table 2. We believe this is an interesting outcome of the scoring system: it allows to identify words out of the prototypical lists that might reflect a positive/negative emotional content, because they were collocated with positive/negative prototypical words in messages with low/high rates of anxiety. In other words, the scoring system allowed to qualify the emotional polarity of the words, yielding scores congruent both with plausible cognitive interpretations and with the way in which words co-occurred in the text.

It is worth to mention that the size of the sets $\mathcal{P}$ or $\mathcal{N}$ does not impact on the number of neighbors, because that number depends only on a distance criteria over the embedding space. Nevertheless, the size of these lists does have an impact on the affective nuances that can be grasped by the model. Moreover, the criteria with which the contents of $\mathcal{P}$ or $\mathcal{N}$ are chosen is of the upmost importance. To illustrate this, consider the family of words associated with "sleep". The word sleep was put into the list $\mathcal{P}$ from a cognitive-linguistic point of view. However, in a psychiatric interpretation, the word sleep may have a negative connotation as this word might reflect a desire of suicide. We performed another experiment in order to observe the effects in the resulting scores of changing the family of words associated with sleep to the list $\mathcal{N}$. Table 3 shows the effects.

Table 3. Top 20 most positive and most negative words after putting the family of words corresponding to "sleep" in the negative knowledge base (list $\mathcal{N}$).

Word	Score	Translation	Word	Score	Translation
bienestar	1.0	well−being	empeorar	−1.0	to worsen
acompañada	1.0	accompanied	muerte	−1.0	death
descansada	1.0	rested	peores	−1.0	worst
mejores	1.0	better	tenso	−1.0	tense
descansado	1.0	rested	durmiendo	−0.9821	sleeping
principio	1.0	beginning	atrás	−0.9783	backwards
mejorar	1.0	to get better	despierto	−0.9747	awake
adelante	0.9818	forward	malo	−0.9726	bad
positivo	0.9787	positive	dormido	−0.9684	asleep
útil	0.9706	useful	fin	−0.9655	end
relajada	0.9661	relaxed	mala	−0.96	bad
tranquilo	0.965	peaceful	duro	−0.9583	tough
fácil	0.96	easy	negativo	−0.95	negative
tranquila	0.9583	peaceful	nervioso	−0.949	nervous
contenta	0.9516	happy	sol	−0.9474	sun
alegre	0.95	cheerful	pasado	−0.947	past
buena	0.9494	good	cansado	−0.9423	tired
relajado	0.9286	relaxed	triste	−0.9364	sad
franca	0.9209	frank	librarme	−0.9342	to get rid of
pintando	0.9195	painting	despierta	−0.9286	awake

Observing Table 3, we see now that the words associated with sleep are among the most negative. However, it is interesting to note that the word *despierto* (awake), which is its antonym, also appears among the most negative words, while it never showed up in Table 2. This example shows interesting phenomena with respect to our scoring system. The first observation is that the words of the family "awake" have a similarity score around 0.7 with the words of the family "sleep". When "sleep" was put in list $\mathcal{P}$, the word "awake" got a score near 0, meaning that it received almost the same number of times positive scores and negatives scores. This is indicative that the word "awake" was a neighbor of other positive and negative words in our lists in more or less the same amount. However, when "sleep" was put in list $\mathcal{N}$, the word "awake" receives a score of -0.9821, meaning that the contribution of neighboring negative words was higher this time. As the neighboring coefficient is always the same (i.e., the cosine similarity remains the same in both experiments), the only possibility is that the negative score of "awake" was due to the effect of "sleep" and possibly other words accompanying it when it passed to the negative list. This is interesting because it could imply that if we focus the attention on sleep, we should also focus the attention on awake. It is worth to mention how this finding in fact has a clinical sense, as it is known how sleep disturbances and suicide are related [75]. Interestingly, our findings of how words related to sleep can be positive or negative could add a refined interpretation of this phenomena in clinical settings and be personalized. Thus, these results highlight the need to establish extra-linguistic criteria and human supervision for labeling words, as the implications of semantic associations might differ depending on the context. In these respects, we must stress that in order to tag words with a clinical sense, the contribution of specialists in the field of psychiatry is essential.

A last comment on words with a score value of 0, as there are indeed many such words. This situation is due to either of the following reasons. (1) The word is not in $\mathcal{P} \cup \mathcal{N} \cup \overline{\mathcal{P}} \cup \overline{\mathcal{N}}$. This happens when the word in question was not a prototypical word and it was not retrieved as one of the closest k neighbors of words in either $\mathcal{P}$ or $\mathcal{N}$, which suggests a relatively poor semantic relation between the word and the prototypical lists. However, this situation may change by increasing the hyperparameter k. (2) The final score of the word, as calculated by Equation (8), turns out to be 0. This occurs when the word has been used in messages with opposite polarities and with balanced weights, suggesting that the word bears no semantic content associated with a particular sentiment (polarity value).

Turning now to the classification task, it is worth noting that the slightly above average performance of the classifier is, perhaps, to be expected, as a consequence of the relatively small size of the corpus. Thus, we believe these representations will increase their impact with a bigger corpus. Indeed, most word embedding algorithms, like `word2vec`, are based on word co-occurrence statistics. In order for such algorithms to uncover an underlying Euclidean semantic space, it has to be shown that the co-occurrences are related to a semantic similarity assessment and that they can be embedded into a Euclidean space. In [76], the authors describe two statistics to empirically test this agreement: *centrality* and *reciprocity factor* R_f. The latter should be greater than $\frac{1}{2}$; however, in our model we have $R_f = 0.257$. This supports our claim that this corpus does not provide an ideal framework for `word2vec` to produce the expected regularities in the word embeddings. More empirical evidence of this proposition is given by the following facts.

- 48.87% of words appear only once in the corpus.
- The vocabulary is rather limited compared to the size of the corpora where `word2vec` is usually trained. The usual size of such vocabularies is in the order of millions of tokens.
- 7% of the messages consist of only a few words lacking a coherent grammatical structure.

Despite these facts, the classification performance was 30% above chance, and as expected, our cognitive-semantic representation is seemingly preserving word similarity.

As a whole, our results show that embedded word representations made it possible to discern sets of lexical association in the text messages. However, by themselves, these representations do not allow discerning emotional polarity or strength. Our cognitively-based lists of prototypical words allowed to compute scores which we used to create new embedded representations with a positive or negative connotation from a cognitive-linguistic point of view. However, as we discussed, extra-linguistic criteria such as clinical considerations might be paramount in order to better assess these emotional states. The scoring method achieves this distinction very clearly at the word level, which allows us to validate the hypothesis that the proposed new representation using this scoring system enables the discovery of semantic correlations that could be reflecting a depressive state.

Another important contribution is our clustering approach. Although in its nascent state, the method produced nice sets of words that we could relate them by arbitrary topics. The new representations of words, $R(w)$, expand the clusters of lexical association in the vector space, potentially modifying the association sets. This could be leveraged by the clustering method, as new clusters might be discovered by taking into account the emotional connotation of words. Using our message representation $R(m)$, we obtained relatively low classification performance, even though rates nearly $1/3$ above chance were achieved. Further research on this phenomenon needs to be undertaken in order to better understand the impact of our representations on the classification task.

6. Conclusions

Helping to prevent suicide is the major motivation of our work. Our aim was to provide tools for constructing meaningful models of the emotional content of a text message, in order to explore and understand the influence of the multiple linguistic and extra-linguistic factors involved in the analysis of psychiatric symptoms.

In this paper, we proposed a hybrid method that uses a cognitively-based lexicon together with word embeddings to generate graded sentiment scores for words. The word embeddings were computed using the `word2vec` model. The scoring system consists in a set of rules that account for the emotional weight of each word in the corpus. The score for each word is computed as a function of the tag of each message containing the word and its semantic relatedness with the terms of a small cognitive knowledge base. The scoring system allowed to qualify the emotional polarity of words, yielding scores between -1 and 1 congruent with plausible cognitive interpretations and the way in which the words co-occurred in the text.

Using the scores as scaling factors of the word embeddings, we proposed new embedded representations for words, called *cognitive-semantic representations*. These new representations neatly separated the embeddings into positive and negative words. The cognitive-semantic representations preserve the semantic similarities derived from `word2vec` while capturing the emotional weight behind the words as shown by classification experiments.

We also proposed a new topological method for clustering dense vector representations in high-dimensional spaces. The proposed clustering method performs the cluster search in a graph which divides and organizes the embedding space into pairwise disjoint open regions, called hyperoctants. These hyperoctants provide a simple and useful label for each word embedding. The criteria for clustering is based on a compromise between the size of the diameter and the density of each cluster. We applied the clustering method to the word embeddings computed by `word2vec`, and found several pairwise disjoint sets of words with a semantic relation between them. As this clustering method is still in its initial state, a formal analysis and improvement of this method is in progress.

The analysis of the results of our scoring and representation framework revealed interesting phenomena. First, we observed that the choice of the words for our knowledge base may have semantic implications with important clinical interpretations. Second, the cognitive-semantic representations reconfigure the clusters of the word embeddings by expanding or collapsing them. This cluster reconfiguration shall be further investigated in the future, as it might reveal new semantic associations related to emotional states. Finally, we propose two possible veins of research for future work: On the one hand, it would be interesting to delve into new algorithms to estimate the cognitive-emotional scores automatically, in order to construct better predictors and improve the classification performance. To achieve this, further data to train the models with shall allow us to predict, in real-time with better accuracy, the emotional state of a person through his or her writing, in order to deploy effective protocols of helpful interventions. On the other hand, an alternative perspective is to investigate the sequential structure of our texts [77] by using techniques from statistical physics which could capture nontrivial interrelations between syntagmatic sequences, by means of entropy based measures or visibility graphs, that have been recently successfully applied to musical pieces [78].

Our work brings together cognitive linguists, psychiatrists, mathematicians, and computer scientists. The results presented herein are encouraging. We hypothesize that our cognitive-based scoring and representation framework might be helpful in studying relations between language and behavior and that their use might have a predictive potential to prevent suicide.

Author Contributions: Conceptualization, J.H.-V. and E.B.-G.; methodology, J.H.-V., A.R.-A., and M.T.-A.; software, T.B. and M.T.-A.; validation, J.H.-V. and E.B.-G.; formal analysis, M.T.-A., Markus-Müller, and T.B.; investigation, M.T.-A. and T.B.; validation, J.H.-V., M.M. and M.L.B., F.A.B., and E.B.-G.; resources, E.B.-G. and J.H.-V.; data curation, T.B., M.T.-A., and J.H.-V.; writing–original draft preparation, M.T.-A., T.B., A.R.-A., and J.H.-V.; writing–review and editing, M.T.-A., A.R.-A., M.M., J.H.-V., M.L.B., F.A.B., and E.B.-G.; visualization, M.T.-A. and T.B.; supervision, J.H.-V. and E.B.-G.; project administration, J.H.-V.; funding acquisition, J.H.-V. and E.B.-G. All authors have read and agreed to the published version of the manuscript.

Funding: Research was partially funded by CONACYT Projects A1-S-24213 of Basic Science, 263377 of Frontiers of Science and CONACYT grants 28268 and 30053, by Carlos III (ISCIII PI16/01852), American Foundation for Suicide Prevention (LSRG-1-005-16), the Madrid Regional Government (B2017/BMD-3740 AGES-CM 2CM; Y2018/TCS-4705 PRACTICO-CM), MINECO/FEDER ('ADVENTURE', id. TEC2015-69868-C2-1-R), MCIU Explora Grant 'aMBITION' (id. TEC2017-92552-EXP) and MICINN ('CLARA', id. RTI2018-099655-B-I00).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript.

NLP	Natural language processing
MPQA	Multi-Perspective Question Answering
CEIC-FJD	Comité de Ética de la Investigación Clínica de la Fundación Jiménez Díaz
PCA	Principal component analysis
DBSCAN	Density-based spatial clustering of applications with noise.

References

1. Baca García, E.; Aroca, F. Factores de riesgo de la conducta suicida asociados a trastornos depresivos y ansiedad. *Salud Ment.* **2014**, *37*, 373–380. [CrossRef]
2. Turecki, G. Preventing suicide: where are we? *Lancet. Psychiatry* **2016**, *3*, 597. [CrossRef]
3. Ge, J.; Vazquez, M.; Gretzel, U. Sentiment analysis: a review. In *Advances in Social Media for Travel, Tourism and Hospitality: New Perspectives, Practice and Cases*; Routledge: London, UK, 2018; pp. 243–261.
4. Serrano-Guerrero, J.; Olivas, J.A.; Romero, F.P.; Herrera-Viedma, E. Sentiment analysis: A review and comparative analysis of web services. *Inform. Sci.* **2015**, *311*, 18–38. [CrossRef]
5. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 3111–3119.
6. Out, C.; Goudbeek, M.; Krahmer, E. Do Speaker's emotions influence their language production? Studying the influence of disgust and amusement on alignment in interactive reference. *Lang. Sci.* **2020**, *78*, 101255. [CrossRef]
7. Foolen, A. The relevance of emotion for language and linguistics. *Moving Ourselves, Moving Others: Motion and Emotion in Intersubjectivity, Consciousness and Language*; John Benjamins Publishing Company: Amsterdam, The Netherlands, 2012; pp. 349–369.
8. Benamara, F.; Taboada, M.; Mathieu, Y. Evaluative Language Beyond Bags of Words: Linguistic Insights and Computational Applications. *Comput. Linguist.* **2017**, *43*, 201–264. [CrossRef]
9. Jatnika, D.; Bijaksana, M.A.; Suryani, A.A. Word2Vec Model Analysis for Semantic Similarities in English Words. *Procedia Comput. Sci.* **2019**, *157*, 160–167. [CrossRef]
10. Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K.A.; Ceder, G.; Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95–98. [CrossRef]
11. Sim, K.; Gopalkrishnan, V.; Zimek, A.; Cong, G. A survey on enhanced subspace clustering. *Data Min. Knowl. Discov.* **2013**, *26*, 332–397. [CrossRef]
12. Bhattacharjee, P.; Mitra, P. A survey of density based clustering algorithms. *Front. Comp. Sci.* **2020**, *15*, 151308. [CrossRef]
13. MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, Statistical Laboratory of the University of California, Berkeley, CA, USA, 21 June–18 July 1965 and 27 December 1965–7 January 1966; University of California Press: Berkeley, CA, USA, 1967; pp. 281–297.

14. Dhillon, I.S.; Modha, D.S. Concept decompositions for large sparse text data using clustering. *Mach. Learn.* **2001**, *42*, 143–175. [CrossRef]
15. Tunali, V.; Bilgin, T.; Camurcu, A. An Improved Clustering Algorithm for Text Mining: Multi-Cluster Spherical K-Means. *Int. Arab J. Inform. Technol.* **2016**, *13*, 12–19.
16. Brodinová, Š.; Filzmoser, P.; Ortner, T.; Breiteneder, C.; Rohm, M. Robust and sparse k-means clustering for high-dimensional data. *Adv. Data Anal. Classif.* **2019**, *13*, 905–932. [CrossRef]
17. Gao, J.; Liu, N.; Lawley, M.; Hu, X. An interpretable classification framework for information extraction from online healthcare forums. *J. Healthc. Eng.* **2017**, *2017*, 2460174 . [CrossRef] [PubMed]
18. Stewart, R.; Velupillai, S. Applied natural language processing in mental health big data. *Neuropsychopharmacology* **2020**, *46*, 252–253.
19. Mäntylä, M.V.; Graziotin, D.; Kuutila, M. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Comp. Sci. Rev.* **2018**, *27*, 16–32. [CrossRef]
20. Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-Based Methods for Sentiment Analysis. *Comput. Linguist.* **2011**, *37*, 267–307. [CrossRef]
21. Esuli, A.; Sebastiani, F. SentiWordNet: A high-coverage lexical resource for opinion mining. *Evaluation* **2007**, *17*, 26.
22. Thelwall, M.; Buckley, K.; Paltoglou, G. Sentiment strength detection for the social web. *J. Am. Soc. Inform. Sci. Technol.* **2012**, *63*, 163–173. [CrossRef]
23. Abbasi, A.; Chen, H. Affect Intensity Analysis of Dark Web Forums. In Proceedings of the 2007 IEEE Intelligence and Security Informatics, New Brunswick, NJ, USA, 23–24 May 2007; pp. 282–288.
24. Agarwal, A.; Xie, B.; Vovsha, I.; Rambow, O.; Passonneau, R. Sentiment Analysis of Twitter Data. In Proceedings of the Workshop on Languages in Social Media, LSM '11, Portland, OR, USA, 23 June 2011; pp. 30–38.
25. Gautam, G.; Yadav, D. Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In Proceedings of the 2014 Seventh International Conference on Contemporary Computing (IC3), Noida, India, 7–9 August 2014; pp. 437–442.
26. Tang, D.; Wei, F.; Qin, B.; Yang, N.; Liu, T.; Zhou, M. Sentiment embeddings with applications to sentiment analysis. *IEEE Trans. Knowl. Data Eng.* **2015**, *28*, 496–509. [CrossRef]
27. Li, Y.; Pan, Q.; Yang, T.; Wang, S.; Tang, J.; Cambria, E. Learning word representations for sentiment analysis. *Cogn. Comput.* **2017**, *9*, 843–851. [CrossRef]
28. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global Vectors for Word Representation. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Volume 14, pp. 1532–1543.
29. Jabreel, M.; Moreno, A. A deep learning-based approach for multi-label emotion classification in tweets. *Appl. Sci.* **2019**, *9*, 1123. [CrossRef]
30. Appel, O.; Chiclana, F.; Carter, J.; Fujita, H. A hybrid approach to the sentiment analysis problem at the sentence level. *Knowl.-Based Syst.* **2016**, *108*, 110–124. [CrossRef]
31. Zainuddin, N.; Selamat, A.; Ibrahim, R. Hybrid sentiment classification on twitter aspect-based sentiment analysis. *Appl. Intell.* **2018**, *48*, 1218–1232. [CrossRef]
32. Wu, C.S.; Kuo, C.; Su, C.H.; Wang, S.; Dai, H.J. Using text mining to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records. *J. Affect. disord.* **2019**, *260*, 617–623. [CrossRef] [PubMed]
33. Wang, L.; Liu, H.; Zhou, T. A Sequential Emotion Approach for Diagnosing Mental Disorder on Social Media. *Appl. Sci.* **2020**, *10*, 1647. [CrossRef]
34. Xue, B.; Fu, C.; Shaobin, Z. A study on sentiment computing and classification of sina weibo with word2vec. In Proceedings of the 2014 IEEE International Congress on Big Data, Anchorage, AK, USA, 27 June–2 July 2014; pp. 358–363.
35. Turney, P.D.; Littman, M.L. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inform. Syst. (TOIS)* **2003**, *21*, 315–346. [CrossRef]

36. Al-Amin, M.; Islam, M.S.; Uzzal, S.D. Sentiment analysis of bengali comments with word2vec and sentiment information of words. In Proceedings of the 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox's Bazar, Bangladesh, 16–18 February 2017; pp. 186–190.
37. Velupillai, S.; Hadlaczky, G.; Baca-Garcia, E.; Gorrell, G.M.; Werbeloff, N.; Nguyen, D.; Patel, R.; Leightley, D.; Downs, J.; Hotopf, M.; et al. Risk assessment tools and data-driven approaches for predicting and preventing suicidal behavior. *Front. Psychiatry* **2019**, *10*, 36. [CrossRef] [PubMed]
38. Corcoran, C.M.; Benavides, C.; Cecchi, G. Natural language processing: opportunities and challenges for patients, providers, and hospital systems. *Psychiatr. Ann.* **2019**, *49*, 202–208. [CrossRef]
39. Pinker, S. *The Stuff of Thought: Language as a Window into Human Nature*; Penguin: London, UK, 2007.
40. Corcoran, C.M.; Cecchi, G. Using language processing and speech analysis for the identification of psychosis and other disorders. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **2020**, *5*, 770–779. [CrossRef] [PubMed]
41. Badal, V.D.; Graham, S.A.; Depp, C.A.; Shinkawa, K.; Yamada, Y.; Palinkas, L.A.; Kim, H.C.; Jeste, D.V.; Lee, E.E. Prediction of Loneliness in Older Adults Using Natural Language Processing: Exploring Sex Differences in Speech. *Am. J. Geriatr. Psychiatry* **2020**. [CrossRef] [PubMed]
42. Goldberg, S.B.; Flemotomos, N.; Martinez, V.R.; Tanana, M.J.; Kuo, P.B.; Pace, B.T.; Villatte, J.L.; Georgiou, P.G.; Van Epps, J.; Imel, Z.E.; et al. Machine learning and natural language processing in psychotherapy research: Alliance as example use case. *J. Couns. Psychol.* **2020**, *67*, 438. [CrossRef] [PubMed]
43. Ni, Y.; Barzman, D.; Bachtel, A.; Griffey, M.; Osborn, A.; Sorter, M. Finding warning markers: Leveraging natural language processing and machine learning technologies to detect risk of school violence. *Int. J. Med. Inform.* **2020**, *139*, 104137. [CrossRef] [PubMed]
44. Coppersmith, G.; Leary, R.; Crutchley, P.; Fine, A. Natural language processing of social media as screening for suicide risk. *Biomed. Inform. Insights* **2018**, *10*. [CrossRef] [PubMed]
45. Cook, B.L.; Progovac, A.M.; Chen, P.; Mullin, B.; Hou, S.; Baca-Garcia, E. Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. *Comput. Math. Methods Med.* **2016**, *2016*, 8708434. [CrossRef] [PubMed]
46. Pellegrini, A.M.; Chan, S.; Brown, H.E.; Rosenquist, J.N.; Vuijk, P.J.; Doyle, A.E.; Perlis, R.H.; Cai, T. Integrating questionnaire measures for transdiagnostic psychiatric phenotyping using word2vec. *PLoS ONE* **2020**, *15*, e0230663.
47. Zhang, Y.; Li, H.J.; Wang, J.; Cohen, T.; Roberts, K.; Xu, H. Adapting word embeddings from multiple domains to symptom recognition from psychiatric notes. *AMIA Summits Transl. Sci. Proc.* **2018**, *2018*, 281.
48. Cambria, E.; Olsher, D.; Rajagopal, D. SenticNet 3: A Common and Common-Sense Knowledge Base for Cognition-Driven Sentiment Analysis. In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI'14, Québec City, QC, Canada, 27–31 July 2014; pp. 1515–1521.
49. Mishra, A.; Kanojia, D.; Nagar, S.; Dey, K.; Bhattacharyya, P. Leveraging Cognitive Features for Sentiment Analysis. In Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, Berlin, Germany, 11–12 August 2016; Association for Computational Linguistics: Berlin, Germany, 2016; pp. 156–166.
50. Wilson, T.; Wiebe, J.; Hoffmann, P. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, Vancouver, BC, Canada, 6–8 October 2005; pp. 347–354.
51. Wiebe, J.; Wilson, T.; Cardie, C. Annotating Expressions of Opinions and Emotions in Language. *Lang. Resour. Eval.* **2005**, *39*, 164–210. [CrossRef]
52. Zucco, C.; Calabrese, B.; Cannataro, M. Sentiment analysis and affective computing for depression monitoring. In Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, USA, 13–16 November 2017; pp. 1988–1995.
53. Hussein, D.M.E.D.M. A survey on sentiment analysis challenges. *J. King Saud Univ. Eng. Sci.* **2018**, *30*, 330–338. [CrossRef]
54. Witten, D.M.; Tibshirani, R. A framework for feature selection in clustering. *J. Am. Stat. Assoc.* **2010**, *105*, 713–726. [CrossRef]

55. Galimberti, G.; Manisi, A.; Soffritti, G. Modelling the role of variables in model-based cluster analysis. *Stat. Comp.* **2018**, *28*, 145–169. [CrossRef]
56. Houle, M.E.; Kriegel, H.P.; Kröger, P.; Schubert, E.; Zimek, A. Can shared-neighbor distances defeat the curse of dimensionality? In Proceedings of the International Conference on Scientific and Statistical Database Management, Heidelberg, Germany, 31 June–2 July 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 482–500.
57. Beyer, K.; Goldstein, J.; Ramakrishnan, R.; Shaft, U. When is "nearest neighbor" meaningful? In Proceedings of the International conference on database theory, Jerusalem, Israel, 10–12 January 1999; Springer: Berlin/Heidelberg, Germany, 1999; pp. 217–235.
58. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, Portland, OR, USA, 2–4 August 1996; pp. 226–231.
59. Xu, R.; Wunsch, D. Survey of clustering algorithms. *IEEE Trans. Neural Netw.* **2005**, *16*, 645–678. [CrossRef] [PubMed]
60. Porras-Segovia, A.; Molina-Madueño, R.M.; Berrouiguet, S.; López-Castroman, J.; Barrigón, M.L.; Pérez-Rodríguez, M.S.; Marco, J.H.; Díaz-Oliván, I.; de León, S.; Courtet, P.; et al. Smartphone-based ecological momentary assessment (EMA) in psychiatric patients and student controls: A real-world feasibility study. *J. Affect. Disord.* **2020**, *274*, 733–741. [CrossRef] [PubMed]
61. Berrouiguet, S.; Barrigón, M.L.; Castroman, J.L.; Courtet, P.; Artés-Rodríguez, A.; Baca-García, E. Combining mobile-health (mHealth) and artificial intelligence (AI) methods to avoid suicide attempts: the Smartcrises study protocol. *BMC Psychiatry* **2019**, *19*, 277. [CrossRef] [PubMed]
62. Barrigón, M.L.; Berrouiguet, S.; Carballo, J.J.; Bonal-Giménez, C.; Fernández-Navarro, P.; Pfang, B.; Delgado-Gómez, D.; Courtet, P.; Aroca, F.; Lopez-Castroman, J.; et al. User profiles of an electronic mental health tool for ecological momentary assessment: MEmind. *Int. J. Methods Psychiatr. Res.* **2017**, *26*, e1554. [CrossRef] [PubMed]
63. Deza, M.M.; Deza, E. Encyclopedia of distances. In *Encyclopedia of Distances*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–583.
64. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
65. Mnih, A.; Kavukcuoglu, K. Learning word embeddings efficiently with noise-contrastive estimation. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 2265–2273.
66. Wang, Y.; Rastegar-Mojarad, M.; Komandur-Elayavilli, R.; Liu, H. Leveraging word embeddings and medical entity extraction for biomedical dataset retrieval using unstructured texts. *Database* **2017**, *2017*, bax091. [CrossRef]
67. Wang, Y.; Liu, S.; Afzal, N.; Rastegar-Mojarad, M.; Wang, L.; Shen, F.; Kingsbury, P.; Liu, H. A comparison of word embeddings for the biomedical natural language processing. *J. Biomed. Inform.* **2018**, *87*, 12–20. [CrossRef]
68. Chen, Q.; Peng, Y.; Lu, Z. BioSentVec: Creating sentence embeddings for biomedical texts. In Proceedings of the 2019 IEEE International Conference on Healthcare Informatics (ICHI), Xi'an, China, 10–13 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
69. Levy, O.; Goldberg, Y.; Dagan, I. Improving distributional similarity with lessons learned from word embeddings. *Trans. Assoc. Comput. Linguist.* **2015**, *3*, 211–225. [CrossRef]
70. Gathigia, M.G.; Wang, R.; Shen, M.; Tirado, C.; Tsaregorodtseva, O.; Khatin-Zadeh, O.; Minervino, R.; Marmolejo-Ramos, F. A cross-linguistic study of metaphors of death. *Cogn. Linguist. Stud.* **2018**, *5*, 359–375. [CrossRef]
71. Zepeda-Mendoza, M.L.; Resendis-Antonio, O. Hierarchical Agglomerative Clustering. In *Encyclopedia of Systems Biology*; Springer: New York, NY, USA, 2013; pp. 886–887.
72. Kriegel, H.P.; Kröger, P.; Zimek, A. Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering. *ACM Trans. Knowl. Discov. Data* **2009**, *3*, 1. [CrossRef]
73. Müller, E.; Günnemann, S.; Assent, I.; Seidl, T. Evaluating Clustering in Subspace Projections of High Dimensional Data. *Proc. VLDB Endow.* **2009**, *2*, 1270–1281. [CrossRef]
74. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* **1966**, *10*, 707–710.

75. Porras-Segovia, A.; Pérez-Rodríguez, M.M.; López-Esteban, P.; Courtet, P.; López-Castromán, J.; Cervilla, J.A.; Baca-García, E.; Barrigón M, M.L. Contribution of sleep deprivation to suicidal behaviour: A systematic review. *Sleep Med. Rev.* **2019**, *44*, 37–47. [CrossRef] [PubMed]
76. Hashimoto, T.B.; Alvarez-Melis, D.; Jaakkola, T.S. Word embeddings as metric recovery in semantic spaces. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 273–286. [CrossRef]
77. Mikros, G.K.; Macutek, J. *Sequences in Language and Text*; De Gruyter Mouton: Berlin, Germany; Boston, MA, USA, 24 April 2015.
78. González-Espinoza, A.; Martínez-Mekler, G.; Lacasa, L. Arrow of time across five centuries of classical music. *Phys. Rev. Res.* **2020**, *2*, 033166. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.