Approximating the Mental Lexicon from Clinical Interviews as a Support Tool for Depression Detection

Esaú Villatoro-Tello[*]
Universidad Autónoma Metropolitana
Mexico
evillatoro@cua.uam.mx

Gabriela Ramírez-de-la-Rosa
Universidad Autónoma Metropolitana
Mexico
gramirez@cua.uam.mx

Daniel Gática-Pérez[†]
Idiap Research Institute
Switzerland
daniel.gatica-perez@idiap.ch

Mathew Magimai.- Doss
Idiap Research Institute
Switzerland
mathew.magimaidoss@idiap.ch

Héctor Jiménez-Salazar
Universidad Autónoma Metropolitana
Mexico
hjimenez@cua.uam.mx

ABSTRACT

Depression disorder is one of the major causes of disability in the world that can lead to tragic outcomes. In this paper, we propose a method for using an approximation to a *mental lexicon* to model the communication process of depressed and non-depressed participants in spontaneous North American English clinical interviews. Our approach, inspired by the Lexical Availability theory, identifies the most relevant vocabulary of the interviewed participant, and use it as features in a classification process. We performed an in-depth evaluation on the DAIC-WOZ [20] and the E-DAIC [11] clinical datasets. Obtained results indicate that our approach can compete against recent contextual embeddings when modeling and identifying depression. We show the generalization capabilities of our algorithm using outside data, reaching a macro $F1 = 0.83$ and $F1 = 0.80$ in the DAIC-WOZ and E-DAIC datasets respectively. An analysis of our method's interpretability allows understanding how the classifier is making its decisions. During this process, we observed strong connections between our obtained results and previous research from the psychological field.

CCS CONCEPTS

• **Applied computing** → Life and medical sciences; • **Human-centered computing**;

KEYWORDS

Depression Detection, Mental Lexicon, Language Production, Interpretable Models, Text Classification

ACM Reference Format:
Esaú Villatoro-Tello, Gabriela Ramírez-de-la-Rosa, Daniel Gática-Pérez, Mathew Magimai.- Doss, and Héctor Jiménez-Salazar. 2021. Approximating the Mental Lexicon from Clinical Interviews as a Support Tool for Depression Detection. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3462244.3479896

1 INTRODUCTION

Nowadays, depression disorders are among the principal mental illness problems present in many people around the world. According to the World Health Organization (WHO), more than 300 million people suffer from depressive disorders [58]. This type of mental illness is a leading cause of the person inability to function and represents a significant health and economic burden [31]. At worst, depression can lead to suicide if not diagnosed on time or if not treated adequately. Statistics from the WHO indicate that during the year 2015, near 788 thousand people died due to suicide [58]. Depression is an extremely heterogeneous disorder that can be difficult to detect given it can occur at any age, and is not limited to any specific life situations [32]. Additionally, considering that many people have low levels of mental health literacy, recognizing the symptoms of depression is a complicated task.

Although the severity of suffering a mental illness is well known by psychologists, there is an acknowledged necessity for digital solutions for addressing the burden of mental illness diagnosis and treatment. It is recognized that won't be possible to treat people by professionals alone, and even if possible, some people might require to use alternative modalities to receive mental health support [59]. Examples of recent efforts building technology towards this direction are the SimSensei Kiosk [11] a virtual human interviewer designed to create an engaging face-to-face interaction with patients; Woebot [17] and Wysa [24] dialog systems for health and therapy support for patients that have depressive symptoms; Expressive Interviewing [57], a conversational agent aiming at support users to cope with COVID-19 issues. Although the task of automatic diagnosis needs further research before its practical deployment, the vast majority of these technologies are designed to provide support both to users and experts.

Accordingly, during the last decade the Natural Language Processing (NLP) research community has been interested in making first steps towards computer-supported detection of mental disorders. As described in [37], language is a powerful indicator of

[*] Also with Idiap Research Institute (esau.villatoro@idiap.ch).
[†] Also with Ecole Polytechnique Fédérale de Lausanne (EPFL).

ICMI '21, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8481-0/21/10…$15.00
https://doi.org/10.1145/3462244.3479896

personality, social or emotional status, but also mental health. Consequently, researchers have attempted to predict or analyze depression in the social media domain [9, 34, 60], and more recently, early depression detection has also been investigated [7, 27–29, 52, 56]. Contrary to previous work, our scenario is not social media but clinical-interviews mediated by an automatic agent; thus, we aim at detecting when a interviewee has (or not) depressive symptoms by analyzing their interventions in such scenario. We consider this as an important and necessary step forward in the automatic detection of depressive symptoms in the context of digital health support technologies. For this, our central hypothesis is that participants suffering a mental disorder will have a distinctive *available lexicon*, which can be further modeled as features for training an automatic classifier.

Therefore, inspired by the *Lexical Availability* (LA) theory from psycholinguistics [19], in this paper, we propose a novel linguistically motivated approach for identifying the vocabulary flow used by a group of people in a given communicative situation. To the best of our knowledge, this is the first attempt to explore the feasibility and effectiveness of the LA theory for modeling the vocabulary production of users suffering from a mental disorder. We conduct our experiments using the speech transcripts from the DAIC-WOZ [20] and E-DAIC [11] datasets, examples of clinical interviews performed by a virtual agent. Our experimental results obtain a macro $F1 = 0.64$ in each corpora, and a $F1 \approx 0.83$ when outside data is used for evaluation, outperforming very competitive baselines based on contextual word embeddings, e.g., BERT [54].

Our key contributions are: *i)* a method grounded on psycholinguistics theory, representing the very first time the LA theory is applied in a data-driven approach for identifying relevant features and use them in a classification process; *ii)* a competitive and less computationally expensive approach; and, *iii)* an interpretable model that allows understanding how the classifier is making its decisions.

2 PRELIMINARIES AND MOTIVATION

The study of *Lexical Availability* (LA) dates to the late 1950s and emerges as a result of issues found on second language courses [25, 46]. Generally speaking, LA studies propose an approach to add vocabulary to educational material as an alternative for compensating the shortcomings of frequency-based data, as they serve to foreground relevant more accessible vocabulary that does not appear in frequency counts [35]. In recent years, the LA theory is being applied in the socio-linguistics field, mainly as a method for studying variation and lexical norms from specific communities [15, 51]; and in psycho-linguistics and cognition, where the emphasis is the study of semantic relations between the *available lexicon* and the *mental lexicon*, that is, the way an individual stores and retrieves lexical units given a specific communicative intent [16, 23, 30].

In summary, the LA is measured through a test designed to reflect participant's spontaneous lexical production in relation to a particular center of interest (i.e., topic). The test can be in an oral or written form [21] and usually has a time limit (2 or 5 minutes). As a result of this test, a set of N lists of terms are collected, one for each participant.[1] During the test, the center of interest acts as a trigger for accessing the *mental lexicon*, thence a participant has to mention, orally or written; terms (items) associated with that center of interest. Traditionally, employed centers of interest are topics related to daily life, such as 'food and drink', 'health and medicine', etc. However, there are studies that proposed other topics that allow them to evaluate the lexical production in very particular situations; for instance, smells that can be perceived with the nose [47], greetings and farewells [44], intelligence [22], or to study *available lexicon* in insults from speech production [39]. In the later, Pérez Durán uses the transcripts of informal conversations from young students, and through manual analysis, it is found that not always the most frequent insults are the most accessible.

According to cognitive linguistics, each individual expresses their thoughts in a specific way that is pertinent only to them. However, since we all experience situations in a shared social context, it may be possible for the LA to capture the *available lexicon* relevant for a particular community [38]. Moreover, previous research [22, 33], stated that the *mental lexicon* of a community reveals the type, size, and richness of their vocabulary as well as provides evidence of the community member's understanding of a particular culture, or the structure of their context and the existing regularities present, i.e., it reflects complex cognitive processes. Thus, the final step when collecting the *available lexicon* from a community is computing the availability *score* that reflects the importance of a term in the language production of the analyzed population. The complete list of terms with their respective availability scores is what we refer to as the *available lexicon* of that population.

The LA test is considered a category fluency test and has some relationship with the free word association task [5], which taps directly into the semantic information of the *mental lexicon* [10]. Although the original LA methodology can be seen as artificial (due to the prior assumptions regarding the relationship among topics and produced items), our method aims at tackling this artificiality using the LA theory in a fully communicative intent generated in the context of a clinical interview mediated by an automatic agent.

3 PROPOSED METHOD

Previous research has shown that the way people speak/write (i.e., their language production processes) is different for participants suffering from a mental disorder [48]. Thus, we propose a novel method for modeling the vocabulary production of participants suffering from a mental disorder. Our main hypothesis establishes that it could be possible to approximate the *available lexicon* for a group of persons suffering from depression (the shared social context). Contrary to the traditional LA elicitation test, we aim to demonstrate it is possible to approximate such *available lexicon* by analyzing participants' responses, produced during a similar semi-structured communication process i.e., a clinical interview. To the best of our knowledge, this is the first attempt to adapt the Lexical Availability theory to: *i)* approximate the *available lexicon* from depressive and non-depressive groups, *ii)* use the found lexicon to build a non-sparse representation in order to train an automatic classifier.

[1] Obtained lists represent the relevant (more accessible) vocabulary for each person, and are used to obtain the *available lexicon* of the analyzed community.

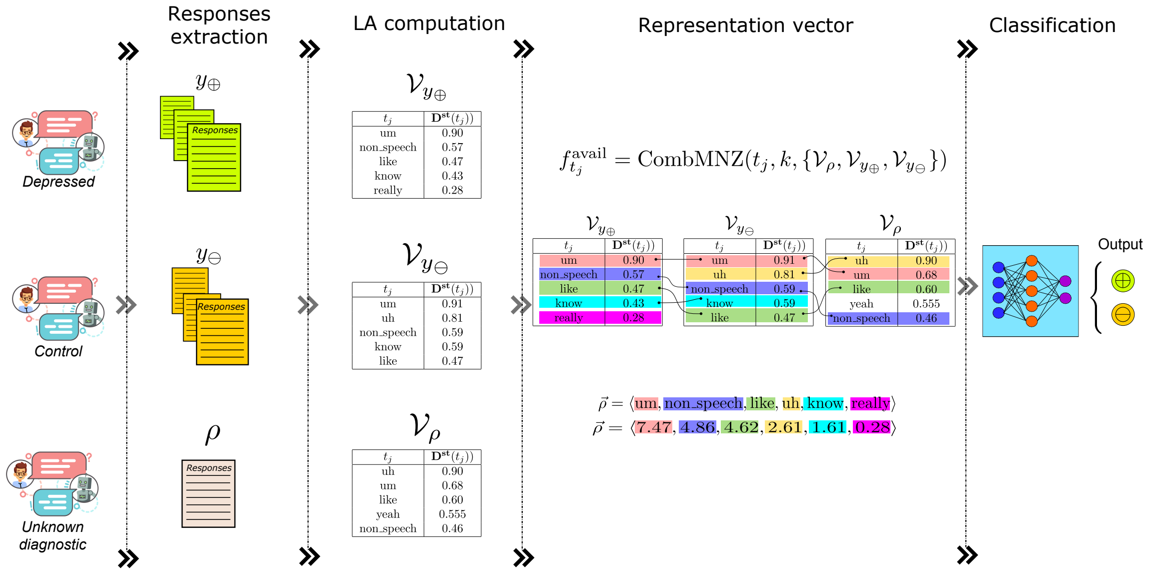

Responses extraction
LA computation
Representation vector
Classification
Depressed
Control
Unknown diagnostic
$y_\oplus$
$y_\ominus$
ρ
Responses
$\mathcal{V}_{y_\oplus}$
t_j | $\mathbf{D}^{st}(t_j))$
um 0.90
non_speech 0.57
like 0.47
know 0.43
really 0.28
$\mathcal{V}_{y_\ominus}$
t_j | $\mathbf{D}^{st}(t_j))$
um 0.91
uh 0.81
non_speech 0.59
know 0.59
like 0.47
$\mathcal{V}_\rho$
t_j | $\mathbf{D}^{st}(t_j))$
uh 0.90
um 0.68
like 0.60
yeah 0.555
non_speech 0.46
$f_{t_j}^{\text{avail}} = \text{CombMNZ}(t_j, k, \{\mathcal{V}_\rho, \mathcal{V}_{y_\oplus}, \mathcal{V}_{y_\ominus}\})$
$\vec{\rho} = \langle$um, non_speech, like, uh, know, really$\rangle$
$\vec{\rho} = \langle 7.47, 4.86, 4.62, 2.61, 1.61, 0.28 \rangle$
Output
⊕
⊖

Figure 1: The general overview of the proposed architecture based on the Lexical Availability theory.

Figure 1 shows the main components of our proposed method. Generally speaking, our approach relies on identifying the *available lexicon* of each population (i.e., participants with a mental disorder and participants without a mental disorder) and then use it to generate a non-sparse text representation to train an automatic classification model to distinguish between *depressed* and *control* participants. More formally, let $\mathcal{D} = \{(d_1, y_1), \ldots, (d_h, y_h)\}$ be a training set of h-pairs of documents d_i and class labels y_i, where $y_i \in \mathcal{Y} = \{y_\oplus, y_\ominus\}$.[2] The first step of our method consists of obtaining the available lexicon ($\mathcal{V}$) for each category, i.e., $\mathcal{V}_{y_\oplus}$ and $\mathcal{V}_{y_\ominus}$ for the documents belonging to *depressed* and *control* categories respectively. Thus, the resultant available lexicon for each category y_i is a list of n-pairs of the form $\mathcal{V}_{y_i} = \{(t_1, \mathbf{D}^{\mathbf{st}}(t_1)), \ldots, (t_n, \mathbf{D}^{\mathbf{st}}(t_n))\}$, where each term t_j is accompanied by its lexical availability score $\mathbf{D}^{\mathbf{st}}(t_j)$. All the details on how to compute the available lexicon are in §3.1.

Then, for generating the representation of the undiagnosed participant ρ (new interviewee), first we obtain its own available lexicon ($\mathcal{V}_\rho$). Next, we calculate its *availability* features (f^{avail}) by means of a fusion strategy among the top k terms from $\mathcal{V}_{y_\oplus} \cup \mathcal{V}_{y_\ominus}$ and $\mathcal{V}_\rho$ (see §3.2), resulting in a representation vector like:

$$\vec{\rho} = \langle f_{t_1}^{\text{avail}}, \ldots, f_{t_j}^{\text{avail}}, \ldots, f_{t_k}^{\text{avail}} \rangle \tag{1}$$

Once we have this representation, we can follow the traditional machine learning pipeline for training an automatic classifier.

3.1 Lexical availability computation

As mentioned in §2, the LA test produces a single word list, which is referred to as the *available lexicon* (with its corresponding availability scores), for each community. To compute the availability scores of this available lexicon, we have to analyze the responses of each individual in that population (see Figure 1 columns 1 to 3); to that end, we use the formulation described in Callealta Barroso and Gallego Gallego, defined as follows:

$$\mathbf{D^{st}}_{w,k,m}(t_j) = \sum_{i=1}^{n} w^{\left(\frac{i-1}{k-1}\right)^m} \times \frac{f_{ji}}{I} \tag{2}$$

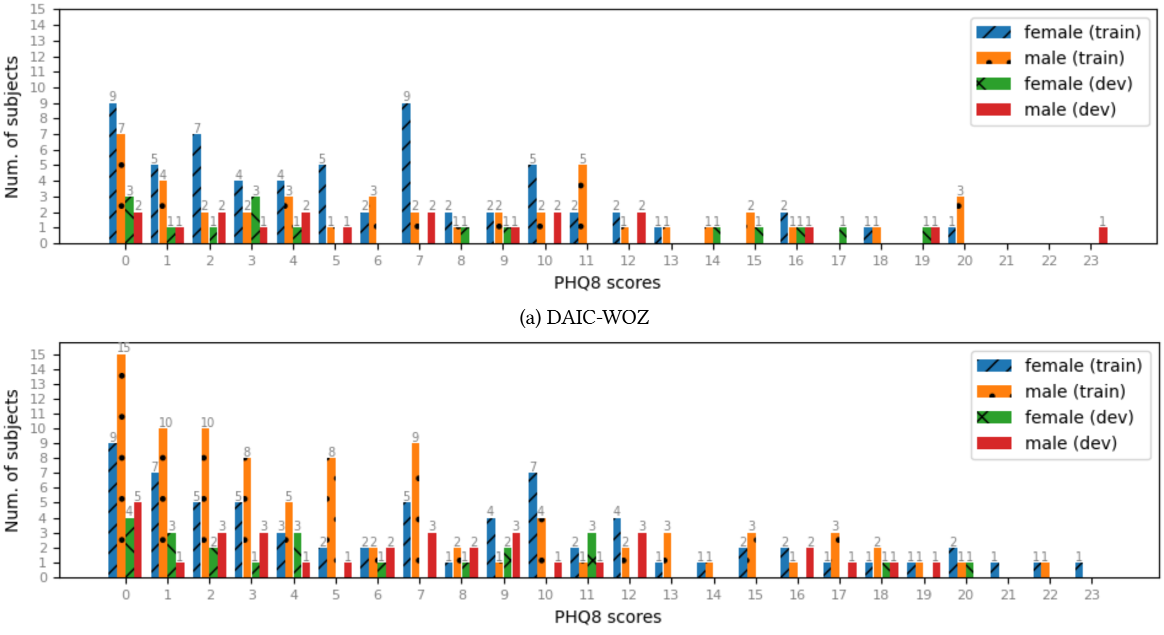
where t_j represents the lexical term for which we want to know its availability score; i is the position indicator where t_j is mentioned in the considered individual responses; n is the maximum position reached by term t_j in all the considered responses; I serves as a normalization factor and is defined as $I = most\_freq\_term$, which depicts the highest frequency found in the vocabulary of the population being analyzed[3]; f_{ji} is the number of participants who produced term t_j at position i in their respective responses; k indicates the position value where the score will be equal to w; w is the desired weight (normally close to 0) for position k, and m modulates the weight decay across terms in the final *available lexicon*.

The $\mathbf{D^{st}}$ equation will assign higher scores (close to 1) to the most available words produced by the analyzed participants. Conversely, it assigns progressively lower scores to less accessible words until reaching value w in position k, at a weight decay intensity defined by the parameter m. Intuitively, the smaller the value of m, the faster the weight decay across words in consecutive positions. For all our experiments, we defined $w = 0.001$ and $m = 1.0$. Thus, as established in [8], the $\mathbf{D^{st}}$ (Eq. 2) represents a standardized LA metric that allows direct comparisons among studies independently from the size of the produced vocabulary lists.[4]

[2] We'll refer as documents to the transcribed text obtained from the participants' communication processes.

[3] In the original formulation proposed by Callealta Barroso and Gallego Gallego, I is assigned the total number of participants who participated in the test.

[4] Our Python implementation to compute the *available lexicon* (Eq. 2) can be found at: https://github.com/gabyrr/lexical_availability_score



(a) DAIC-WOZ

(b) E-DAIC

Figure 2: Distribution of the PHQ-8 scores from participants in DAIC-WOZ and E-DAIC datasets. We make a distinction between *female/male* and *train/dev* participants. All participants with a PHQ-8 ≥ 10 were labelled as depressed [11, 20].

3.2 Availability features

The goal of this step is to generate the $\vec{\rho}$ vector showed in Eq. 1, i.e., to compute the *availability* features (as we illustrate in Figure 1 under the Representation column). We defined the *availability* features (f^{avail}) as the single (most representative) LA score for each term $t_j \in (\mathcal{V}_{y_\oplus} \cup \mathcal{V}_{y_\ominus})$. Thus, to obtain the $f^{\text{avail}}_{t_j}$ score of term t_j we apply the CombMNZ [18] data-fusion strategy. Data-fusion strategies aim at integrating many possible answers (scores) for an object into a single best representative score. Therefore, to compute the representative score of t_j we first obtain the available lexicon $\mathcal{V}_\rho$ from instance ρ applying Eq. 2. Then, for obtaining the $f^{\text{avail}}_{t_j}$ we fuse the scores of word t_j from the list $\mathcal{V}_\rho$ with the previously computed available lexicons from the populations of interest, in this case $\mathcal{V}_{y_\oplus}$ and $\mathcal{V}_{y_\ominus}$. For this process, we do as follows:

$$f^{\text{avail}}_{t_j} = \text{CombMNZ}(t_j, k, \{\mathcal{V}_\rho, \mathcal{V}_{y_\oplus}, \mathcal{V}_{y_\ominus}\}) \quad (3)$$

where t_j is the word for which we want a fused score, k indicates the maximum position where t_j will be searched in the input lists, and $\mathcal{V}$'s are the set of lists that will be considered during the fusion process. Notice that k has the same interpretation of that in Eq. 2; intuitively, it indicates the number of words (features) to be considered for building the representation vector.

Thus, assuming N as the number of ranked lists to be fused, D^c as the normalized score of term t_j in list c, and $|D^c > 0|$ as the number of non-zero scores given to t_j by any list c, the final score for each unique term t_j is computed as follows:

$$\text{CombMNZ}_{t_j} = \sum_{c}^{N} D^c \times |D^c > 0| \quad (4)$$

To understand what is happening after applying Eq. 3, let's assume the term t_j, is used by both populations, i.e., by *depressed* and *control* participants. If this term has a low LA score in $\mathcal{V}_{y_\ominus}$, and a high score in $\mathcal{V}_{y_\oplus}$, it indicates (to some extent) that t_j is more relevant (more accessible) for depressed participants. Now, assuming we want to generate the representation of participant ρ. Following the previous example, if the same t_j has a low score in $\mathcal{V}_\rho$, the result of the fusion strategy will produce a $f^{\text{avail}}_{t_j}$ with a low value, i.e., ρ uses t_j as the control population does. Thus, in the end, $\vec{\rho}$ represents a set of features where its weights indicate to what category (population) they adjust the best.

4 DATASET

For the experiments, we use the Distress Analysis Interview Corpus - wizard of Oz (DAIC-WOZ) dataset [20] and the Extended Distress Analysis Interview Corpus (E-DAIC) [11], which is an extended version of the DAIC-WOZ. Both datasets contain semi-structured clinical interviews, performed by an animated virtual interviewer,[5] designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic disorder. Datasets are multimodal corpora, i.e., they include audio and video recordings, the transcribed text from the interviews, and the questionnaire responses. Nevertheless, for performing our experiments we use only the speech transcripts from the participant's responses during the interview. All the recordings contain spontaneous North American English, and every speaker has a Patient

[5] For DAIC-WOZ the virtual interviewer is human controlled, while for the E-DAIC the virtual interviewer is fully automatic.

Table 1: Composition of the DAIC-WOZ and E-DAIC datasets for depressed (*D*) and control (*C*) participants. Column '#S's' depicts the number of participants, 'Vocabulary' represents the average number of unique terms in the transcripts, and 'Responses' is the average size of the interview (i.e., the number of utterances produced by the interviewed participant)

Dataset	Category	***train***			***dev***		
		#S's	Vocabulary (σ)	Responses (σ)	#S's	Vocabulary (σ)	Responses (σ)
DAIC-WOZ	*D*	30 (*28%*)	266.96 (±90.64)	153.26 (±70.37)	12 (*34%*)	302.83 (±96.66)	232.08 (±84.21)
	C	77 (*72%*)	264.90 (±108.27)	159.84 (±68.53)	23 (*66%*)	266.73 (±87.69)	169.30 (±61.97)
E-DAIC	*D*	37 (*23%*)	402.48 (±133.19)	104.37 (±33.22)	12 (*21%*)	402.08 (±119.09)	107.5 (±38.04)
	C	126 (*77%*)	430.25 (±145.77)	89.47 (±30.05)	44 (*79%*)	383.36 (±123.12)	94.31 (±27.01)

Health Questionnaire (PHQ-8) [26] score, which indicates the severity level of clinical depression. Figure 2 shows the distribution of the PHQ-8 scores across the interviewed participants (*female/male* and *train/dev* differentiated) in the DAIC-WOZ (Fig. 2a) and E-DAIC (Fig. 2b) respectively. According to DeVault et al., Gratch et al., participants with a PHQ-8 ≥ 10 are labelled as *depressed* participants.

The DAIC-WOZ was used during the AVEC 2016 challenge [53], and contains audio-visual interviews of 189 participants: 107 for training, 35 for development, and 47 for test.[6] The E-DAIC was used during the AVEC 2019 challenge [42], and contains audio-visual interviews of 275 participants: 163 for training, 56 for development, and 56 for test.[7] Table 1 depicts the composition of the datasets and some statistics.

5 EXPERIMENTAL SETUP

As a normalization step, we applied the following; all the common contractions, e.g., *we'll*, *can't*, etc., are converted to its formal writing, i.e., *we will*, *can not*, etc. Disfluencies are preserved, i.e., all repetitions and hesitations produced in the speech process (e.g., *i i i think. uh*, *um*, etc.). All non-speech phenomena, e.g., `<cough>`, `<laughter>`, etc., are labeled as `<NON_SPEECH>`. All punctuation marks are removed, and number occurrences are labeled as `<NUMBER>`. Finally, we lower case all the text.

5.1 Baselines

As first baseline we compare against a traditional Bag-of-Words (BoW) using the top 1000 most frequent words under a Term Frequency Inverse Document Frequency *tf-idf* weighting scheme. As second baseline we use the Linguistic Inquiry and Word Count (LIWC) [36] categories for representing the documents. LIWC psychological categories provide a way to capture the semantic content of the language produced [49], e.g., it is possible to detect positive or negative emotion words, words referencing family, friends or society, pronouns which can capture inclusive language (e.g., us, we), exclusive language (e.g., you, they, them), and words referencing how the person is feeling (e.g., sad, anxious, sleep).

As third baseline, we evaluate the impact of recent transformer-based models [54] as a language representation strategy. For our experiments we test an English pre-trained BERT model.[8] As known, the [CLS] token acts an "aggregate representation" of the input tokens, and is considered as a sentence representation for many classification tasks [12]. Accordingly, for generating the representation of each document, we split the document into smaller chunks (with maximum length of 512 tokens), obtain the [CLS] encoding of each chunk, and we apply a *mean* pooling to obtain the final document representation. Finally, we also show the performance of a *naive* method that assigns the majority class.

5.2 Evaluation

We evaluate our method using three machine learning algorithms, Table 2 shows the considered algorithms and its parameters.

Table 2: Learning algorithms and its parameters

Algorithm	Parameters
Support Vector Classifier (SVC)	`kernel='linear', random_state=0`
Multi-layer Perceptron (MLP)	`activation='relu', alpha=1e-5, solver='lbfgs', random_state=0, max_iter=300`
Logistic Regression (LR)	`solver='lbfgs', random_state=0`

For evaluating the performance, we use the F score ($F1$) for both the *depressed (D)* and *control (C)* classes, and the Macro-F for the problem *(All)*. This decision was made in agreement with previous work that reports these metrics [13, 41, 53]. We acknowledge the limitations regarding the small size of the corpora, however, this is a common shortcoming of studies that use clinical datasets. Thus, in order to achieve stable and robust results, we decide to report performance results, for each dataset, under three validation strategies: i) we report the average performance over a stratified 10 cross-fold-validation technique using the *train* partition (10-CFV), ii) we report the performance over the *dev* partition only, and, iii) we report the performance of our method when trained in one dataset and evaluated in the other, i.e., when outside data is used as test. We do not report results for the *test* partitions, since these are not publicly available.

6 EXPERIMENTAL RESULTS

Table 3 summarizes our results for the experiments using a 10-CFV strategy over the *train* partition; Table 4 shows the performance of the experiments performed on the *dev* partition of each dataset; and Table 5 shows the results obtained when our model is evaluated

[6] A portion of the DAIC-WOZ transcriptions were generated using the ELAN tool from the Max Planck Institute for Psycholinguistics [6]

[7] All E-DAIC transcriptions were generated using Google Cloud's ASR service.

[8] https://huggingface.co/transformers/pretrained_models.html

Table 3: Average results in the *train* partitions over 10-CFV

Dataset	Method	Classifier	F1 score		
			All	*D*	*C*
DAIC-WOZ	*naive*	-	*0.41*	*0.00*	*0.83*
	BoW	MLP	0.61	0.40	0.82
	LIWC	MLP	0.53	0.34	0.72
	BERT	SVC	**0.69**	0.51	**0.86**
	LA_{100}	SVC	0.57	0.41	0.73
	LA_{500}	MLP	0.68	**0.53**	0.83
	LA_{1000}	MLP	0.66	0.45	**0.86**
E-DAIC	*naive*	-	*0.43*	*0.00*	*0.87*
	BoW	MLP	0.55	0.25	0.84
	LIWC	MLP	0.58	0.36	0.80
	BERT	MLP	0.60	0.38	0.82
	LA_{100}	LR	**0.68**	**0.49**	0.87
	LA_{500}	LR	0.65	0.41	**0.88**
	LA_{1000}	MLP	0.59	0.33	0.85

Table 4: Experimental results on *dev* partitions

Dataset	Method	Classifier	F1 score		
			All	*D*	*C*
DAIC-WOZ	*naive*	-	*0.39*	*0.00*	*0.79*
	BoW	MLP	0.51	0.30	0.72
	LIWC	MLP	0.49	0.29	0.69
	BERT	MLP	0.56	0.38	0.73
	LA_{100}	MLP	**0.64**	**0.52**	0.77
	LA_{500}	LR	0.53	0.27	**0.80**
	LA_{1000}	MLP	0.42	0.12	0.72
E-DAIC	*naive*	-	*0.44*	*0.00*	*0.88*
	BoW	MLP	0.58	0.27	0.89
	LIWC	MLP	0.56	0.32	0.81
	BERT	LR	0.59	0.29	0.90
	LA_{100}	LR	0.56	0.25	0.88
	LA_{500}	LR	0.62	0.35	0.88
	LA_{1000}	LR	**0.64**	**0.38**	**0.90**

with outside data. Given our space restrictions, we only report the obtained results from the best learning algorithm (Classifier column). For the experiments using the LA method, the number in the sub-index indicates the value of the k from Eq. 2, i.e., the number of terms considered for obtaining the available lexicon lists.

For the 10-CFV experiments (Table 3), observe that the worst performance is obtained by the LIWC configuration, $F1 = 0.53$ and $F1 = 0.58$ for DAIC-WOZ and E-DAIC datasets respectively. This result indicates that using the psychological categories from LIWC for representing participants' interactions do not provide good discriminant features. On the contrary, notice that transformer-based methods such as BERT are able to obtain the best performance among the considered baselines; $F1 = 0.69$ in the DAIC-WOZ dataset, and a $F1 = 0.60$ in the E-DAIC. Notice that our LA method achieves a competitive performance in the DAIC-WOZ dataset ($F1 = 0.68$) considering $k = 500$. Although the overall F1 from BERT method is better, the LA method is better at identifying participants from the *depressed* category ($F1_D = 0.53$), this means that the LA approach has better recall values ($R_D = 0.50$) in comparisson to the BERT based-approach ($R_D = 0.40$). The impact of the LA method is more evident on the E-DAIC dataset, where the LA_{100} obtains an overall $F1 = 0.68$ and $F1_D = 0.49$, outperforming BERT.

A similar behavior can be observed on the experiments performed on the *dev* partition from each of the considered datasets (Table 4). For the DAIC-WOZ dataset, notice that the LA_{100} obtains an overall $F1 = 0.64$, outperforming BERT with $F1 = 0.56$. In the same way, for the E-DAIC dataset, the LA_{1000} obtains an overall $F1 = 0.64$, outperforming BERT ($F1 = 0.59$). One important observation in this experiment is regarding the number of required terms for the LA method. While for the DAIC-WOZ the top 100 terms are sufficient, in the case of the E-DAIC it becomes necessary to consider up to 1000 terms. We argue that this variation in the value of k is related to the size of the respective datasets. Observe in Table 1 (*dev* column); the vocabulary size in the DAIC-WOZ dataset is smaller than the vocabulary in the E-DAIC; suggesting a lesser variation of terminology in the provided answers. At the same time, the number of responses provided by participants is bigger in the DAIC-WOZ dataset than in the E-DAIC, i.e., interviews' duration are longer on the DAIC-WOZ dataset. All this means that we have more samples of the communicative process for both *depressed* and *control* users in the DAIC-WOZ, with smaller variability of lexical units, allowing our method a good performance with a low k.

Finally, in Table 5, we report a series of experiments aiming at validating the generalization process of our method and its applicability to external data. Thus, we use as training data one complete dataset, and evaluate the performance of the learned model on different (outside) data. First, notice that for the experiments using the DAIC-WOZ dataset as training and the dev partition from E-DAIC as test, the performance of the BoW obtains the best overall F1 score ($F1 = 0.81$) while the LA_{1000} obtains the second best result ($F1 = 0.80$). Even though these results are similar in terms of the F1 score, when analyzing the details of each method, we found that our LA method obtains a better recall performance on the depress class ($R_D = 0.92$) in comparison with the BoW method ($R_D = 0.75$). In other words, our method is able to detect 11 out of 12 true depressed participants, while the BoW approach detects only 9 out of 12. Given the nature of the task, we consider this an important factor when choosing over the two configurations. On the other hand, for the experiments using as training the E-DAIC dataset, and as test the *dev* parition from DAIC-WOZ, our LA method is able to outperform all the considered baselines obtaining an overall $F1 = 0.83$. Generally speaking, results reported in Table 5 show the applicability of our method in outside data, and, represent to the best of our knowledge, the first time such experimental setup is evaluated on these two datasets. Hence, we consider these experiments as an additional contribution of this paper.

6.1 Comparative evaluations

Most of the related work that has done experiments in the considered datasets have either tackled the problem using multimodal

Table 5: Experimental results when training is done using one complete dataset, and evaluation is made on external data. The following configurations were tested: *(1)* training: DAIC-WOZ, test: E-DAIC; *(2)* training: E-DAIC, test: DAIC-WOZ

Training set	Evaluation set	Method	Classifier	F1 score			Precision			Recall		
				All	*D*	*C*	*All*	*D*	*C*	*All*	*D*	*C*
DAIC-WOZ	**E-DAIC**	BoW	SVC	**0.81**	**0.71**	**0.91**	**0.81**	0.69	0.93	0.83	0.75	0.90
		LIWC	SVC	0.63	0.37	0.90	0.79	**0.75**	0.83	0.61	0.25	**0.98**
(*train+dev*)	(*dev*)	BERT	MLP	0.59	0.35	0.83	0.59	0.36	0.82	0.59	0.33	0.84
D:42 (*30%*)	***D***:12 (*21%*)	LA$_{100}$	MLP	0.67	0.48	0.85	0.66	0.46	0.86	0.67	0.50	0.84
C:100 (*70%*)	***C***:44 (*79%*)	LA$_{500}$	LR	0.78	0.67	**0.89**	0.76	0.60	0.93	0.81	0.75	0.86
		LA$_{1000}$	SVC	**0.80**	**0.71**	**0.89**	0.78	0.58	**0.97**	**0.87**	**0.92**	0.82
E-DAIC	**DAIC-WOZ**	BoW	MLP	0.73	0.59	0.87	0.88	1.00	0.77	0.71	0.42	1.00
		LIWC	SVC	0.40	0.00	0.79	0.33	0.00	0.66	0.50	0.00	1.00
(*train+dev*)	(*dev*)	BERT	MLP	0.48	0.15	0.81	0.84	1.00	0.68	0.54	0.08	1.00
D:49 (*22%*)	***D***:12 (*34%*)	LA$_{100}$	MLP	0.70	0.64	0.76	0.70	0.56	0.84	0.72	**0.75**	0.70
C:170 (*78%*)	***C***:23 (*66%*)	LA$_{500}$	MLP	**0.83**	**0.76**	**0.90**	0.87	0.89	**0.85**	**0.81**	0.67	0.96
		LA$_{1000}$	LR	0.82	0.74	**0.90**	**0.91**	**1.00**	0.82	0.79	0.58	**1.00**

approaches, or have faced the problem as a regression task (i.e., prediction of the PHQ-8 score). Thus, it is very difficult to directly compare our results.

During the AVEC-2016 challenge, the proposed baseline, an SVM trained on audio and video features, obtains an $F1 = 0.50$ over the *dev* partition [53]. Similarly, Al Hanai et al. evaluate the performance of an LSTM recurrent network, using only the transcripts, reporting $F1_D = 0.67$. Rinaldi et al. proposed a deep learning method, named Joint Latent Prompt Categorization (JLPC), and report a performance of $F1_D = 0.44$. More recently, in [55] a late fusion approach between the LA method and a acoustic-based modality obtains a $F1_D = 0.80$.

For the E-DAIC dataset, most of the related work report results in terms of the Root Mean Squared Error (RMSE). For instance, the proposed baseline during AVEC-2019 [42], based on audio and visual features, obtains a $RMSE = 5.03$. In [40], authors trained an attention based BLSTM network, and report a $RMSE = 4.37$. Zhang et al., employs doc2vec embeddings as features to train a multitask-DNN; reports a Micro-$F1 = 0.90$ and $RMSE = 4.66$.

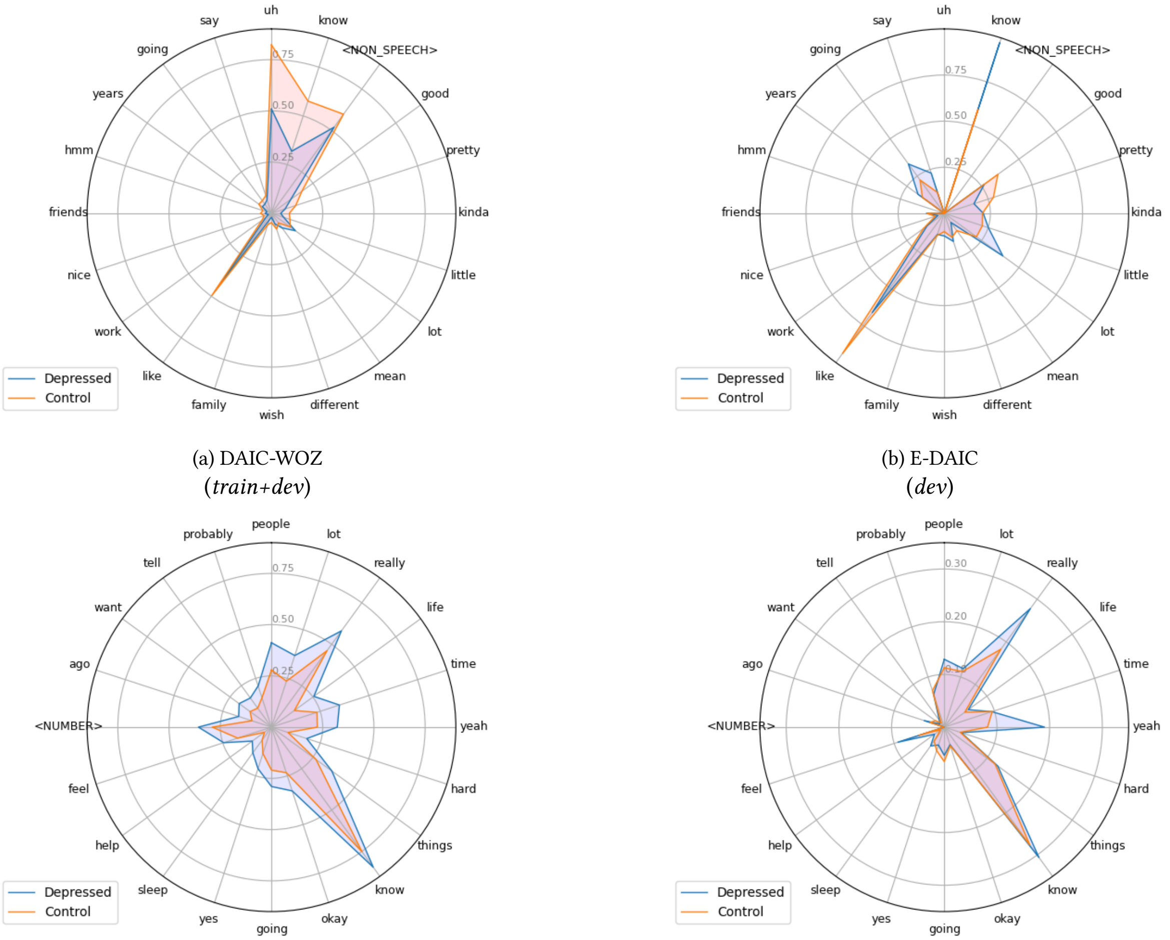
6.2 Category-based analysis of the available lexicon

For evaluating our model's interpretability, we explore how *depressed* and *control* communities employ the identified *available lexicon* in each of the considered datasets. For this analysis, we compute the *available lexicon* from each category, i.e., *D* and *C*, as explained in §3. Then, to select the most representative terms we compute the absolute difference among all the availability scores from $\{\mathcal{V}_{y_\oplus} \cap \mathcal{V}_{y_\ominus}\}$, we rank the terms according to the obtained value, and select the top *n* terms. Selected terms will represent, to some extent, the subset of lexical units more relevant (i.e., more accessible), among the analyzed communities.

For the following examples, we select the best configuration from the experiments reported in Table 5, i.e., (*i*) training with the DAIC-WOZ (*train+dev*) with LA$_{1000}$, and (*ii*) training using the E-DAIC (*train+dev*) with LA$_{500}$. For each case, we followed the procedure described above to obtain the top 20 most representative terms. Figure 3 illustrates how the identified *available lexicon* is employed by the analyzed participants, i.e., how users, from the respective partitions, recur to these subset of words.

Accordingly, Fig. 3a and 3b correspond to case *(i)*, while Fig. 3c and 3d represent the case *(ii).* The first observation we can make is regarding the available lexicon obtained for the training data, i.e., Fig. 3a and Fig. 3c. Observe that, the selected terms have a clear distinct available score among *D* and *C* categories, meaning that such terms, although relevant, are not equally accessible for the two populations. While this is true for both datasets, it is more evident in Fig. 3c, where the shadow generated by the available lexicon of the *control* participants is subsumed under the availability scores of *depressed* participants. On the other hand, when we analyze the availability of these terms on the *dev* sets (Fig. 3b and Fig. 3d), notice that for the case *(i)* there are some differences in comparison to the respective train partition, e.g., term 'know' is more accessible for depressed participants in the *dev* set, while in the *train* is more accessible for control participants. Conversely, for case *(ii)* the availability scores are equivalent to those observed in the training partition. This reason explains why the results under this configuration (case *ii*) are better that those obtained in case *(i)* (see Table 5).

During these analysis we identified a few examples that are partially aligned with previously reported work from the psychological theory. For instance, the term: 'know', commonly used in the expression 'you know', is employed as a way to ensure that the interlocutor is clear about what is being said. According to previous studies, using this particular expression opposes to overconfidence, which is a trait of *cognitive rigidity*, a characteristic of depression [1]. Similarly, word 'kinda' denotes opposition to *dichotomous thinking* [50], also a trait of language on depressed people. Disfluencies ('uh', 'hmm', <non_speech>) are considered as a way of *ruminate response style* [14], and also notice the *absolutist term* 'lot' [3]. As future work we plan to perform a deeper analysis to evaluate how the available lexicon correlates with aspects such as social skill



(a) DAIC-WOZ (*train+dev*)

(b) E-DAIC (*dev*)

(c) E-DAIC (*train+dev*)

(d) DAIC-WOZ (*dev*)

Figure 3: Illustration of the most relevant (*available*) terms identified in the depressed (*D*) and control (*C*) communities.

deficits [45], the use of vague language [4], and self-focused and detached expressions [43].

7 CONCLUSIONS

This paper addressed the problem of detecting depression from clinical interview transcripts. Inspired by the Lexical Availability theory, we propose a method that approximates the *mental lexicon* through the identification of the *available lexicon* for depressed and non-depressed participants, and use it in a classification process to detect depression. An in-depth evaluation of our method, in two well known clinical datasets (DAIC-WOZ and E-DAIC), shows that the LA method is able to outperform very recent transformer-based models (BERT), with a relative improvement, in terms of the F1 metric, of 12.5%, and 7.8% in the DAIC-WOZ and the E-DAIC respectively. Additionally, our method showed consistent performance when outside data is used as test, reaching a macro $F1 = 0.83$ when the *dev* partition of the DAIC-WOZ is employed, representing state-of-the-art results on this particular dataset.

Our work represents the first attempt to approximate the *available lexicon*, from participants' responses produced in a semi-structured communication process, and successfully use it in a classification process. An exploration of the information provided by the LA method showed the interpretation capabilities of our approach, its relation with previous psycholinguistcs findings, helping to explain the model's decisions. As future work directions we plan to evaluate the impact of our proposed approach in a multimodal scenario (audio and visual), and the influence of the parameters from Eq. 2. Automated detection methods may help to identify depressed individuals, however, several ethical issues arise. We strongly believe that further research is needed in order to minimize potential adverse effects of fully automatic systems.

ACKNOWLEDGMENTS

Esaú Villatoro-Tello, was supported partially by Idiap, SNI CONACyT, and UAM-Cuajimalpa Mexico. Gabriela Ramírez-de-la-Rosa would like to thank UAM-Cuajimalpa for their support.

REFERENCES

[1] Mari Aguilera, Clara Paz, Victoria Compañ, Juan Carlos Medina, and Guillem Feixas. 2019. Cognitive rigidity in patients with depression and fibromyalgia. *International Journal of Clinical and Health Psychology* 19, 2 (2019), 160–164.

[2] Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. 2018. Detecting Depression with Audio/Text Sequence Modeling of Interviews.. In *Interspeech*. 1716–1720.

[3] Mohammed Al-Mosaiwi and Tom Johnstone. 2018. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science* 6, 4 (2018), 529–542.

[4] Nancy JC Andreasen and Bruce Pfohl. 1976. Linguistic analysis of speech in affective disorders. *Archives of General Psychiatry* 33, 11 (1976), 1361–1367.

[5] Antonio Manuel Ávila Muñoz and José María Sánchez Sáez. 2011. La posición de los vocablos en el cálculo del índice de disponibilidad léxica: procesos de reentrada en las listas del léxico disponible de la ciudad de Málaga. (2011).

[6] Hennie Brugman and Albert Russel. 2004. Annotating Multi-media/Multi-modal Resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA), Lisbon, Portugal.

[7] Sergio G Burdisso, Marcelo Errecalde, and Manuel Montes-y Gómez. 2019. A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications* 133 (2019), 182–197.

[8] Francisco J. Callealta Barroso and Diego J. Gallego Gallego. 2016. Medidas de disponibilidad léxica: comparabilidad y normalización (Measures of lexical availability: comparability and standardization). *Boletín de filología* 51, 1 (2016), 39–92.

[9] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social Media As a Measurement Tool of Depression in Populations. In *Proceedings of the 5th Annual ACM Web Science Conference* (Paris, France) *(WebSci '13)*. ACM, New York, NY, USA, 47–56. https://doi.org/doi.org/10.1145/2464464.2464480

[10] Simon De Deyne, Steven Verheyen, and Gert Storms. 2016. Structure and organization of the mental lexicon: A network approach derived from syntactic dependency relations and word associations. In *Towards a theoretical framework for analyzing complex linguistic networks*. Springer, 47–79.

[11] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 1061–1068.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[13] S Pavankumar Dubagunta, Bogdan Vlasenko, and Mathew Magimai Doss. 2019. Learning Voice Source Related Information for Depression Detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6525–6529.

[14] Thomas E Ellis and Billy Rutherford. 2008. Cognition and suicide: Two decades of progress. *International Journal of Cognitive Therapy* 1, 1 (2008), 47–68.

[15] Maitena Etxebarria. 2014. Lexical variation and bilingual education in the Basque country. *Spanish in Context* 11, 1 (2014), 50–75.

[16] Roberto Ferreira and Max S. Echeverría. 2010. Redes semánticas en el léxico disponible de inglés L1 e inglés LE. *Onomázein* (2010). https://www.redalyc.org/articulo.oa?id=134513546005

[17] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health* 4, 2 (2017), e19.

[18] Edward A Fox and Joseph A Shaw. 1994. Combination of multiple searches. *NIST special publication SP* 243 (1994).

[19] Georges Gougenheim, R. Michéa, P. Rivenc, and A. Sauvageot. 1964. *L'élaboration du français fondamental (1er degré): étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Vol. 1. Chilton Books.

[20] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The Distress Analysis Interview Corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, 3123–3128.

[21] Natividad Hernández Muñoz. 2010. Social aspects of oral and lexical written production in Spanish. *SKY Journal of Linguistics* 23 (2010), 101–123.

[22] Natividad Hernández-Muñoz, Cristina Izura, and Carmela Tomé. 2014. Cognitive factors of lexical availability in a second language. In *Lexical availability in English and Spanish as a second language*. Springer, 169–186.

[23] Natividad Hernández Muñoz and Carmela Tomé Cornejo. 2017. Léxico disponible en primera y segunda lengua: bases cognitivas. In *Palabras vocabulario léxico: la lexicología aplicada a la didáctica ya la diacronía*. Edizioni Ca Foscari, 99–122.

[24] Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth* 6, 11 (2018), e12106.

[25] Rosa María Jiménez Catalán. 2013. *Lexical availability in English and Spanish as a second language*. Vol. 17. Springer.

[26] Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. 2009. The PHQ-8 as a measure of current depression in the general population. *Journal of affective disorders* 114, 1-3 (2009), 163–173.

[27] Adrian Pastor López-Monroy, Fabio A. González, Manuel Montes, Hugo Jair Escalante, and Thamar Solorio. 2018. Early Text Classification Using Multi-Resolution Concept Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1216–1225. https://doi.org/10.18653/v1/N18-1110

[28] David E Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of eRisk: Early Risk Prediction on the Internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 343–361.

[29] David E. Losada, Fabio Crestani, and Javier Parapar. 2020. eRisk 2020: Self-harm and Depression Challenges. In *Advances in Information Retrieval*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer International Publishing, Cham, 557–563.

[30] Geral Eduardo Mateus-Ferro, Laura Marcela Castiblanco, and Pedro Augusto Álvarez Bermúdez. 2018. Mecanismos lógicos y analógicos en la producción del léxico disponible. *Revista Folios* (2018).

[31] Colin D Mathers and Dejan Loncar. 2006. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS medicine* 3, 11 (2006), e442.

[32] Michelle Morales, Stefan Scherer, and Rivka Levitan. 2018. A Linguistically-Informed Fusion Approach for Multimodal Depression Detection. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. ACL, New Orleans, LA, 13–24.

[33] Heili Orav. 2006. Lexical Knowledge of Personality Traits. In *Proceedings of the GWC 2006*. 239–343.

[34] Michael J Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *Fifth International AAAI Conference on Weblogs and Social Media*.

[35] Chris Payne. 2016. Lexical availability. *English teaching professional* 102 (2016), 18–20.

[36] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[37] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54, 1 (2003), 547–577.

[38] Marco Antonio Perez Durán. 2019. Un estudio sobre los macrocentros de interés (McI). Una aportación teórica para los estudios actuales de disponibilidad léxica. *Cuadernos de Lingüística de El Colegio de México* 6 (2019).

[39] Marco Pérez Durán. 2020. Análisis del léxico disponible del centro de interés del insulto en estudiantes de secundaria de San Luis Potosí, México. *Revista de Filología y Lingüística de la Universidad de Costa Rica* 46, 1 (mar. 2020), 261–278.

[40] Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg. 2019. Multi-Level Attention Network Using Text, Audio and Video for Depression Prediction. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop* (Nice, France) *(AVEC '19)*. Association for Computing Machinery, New York, NY, USA, 81–88.

[41] Alex Rinaldi, Jean Fox Tree, and Snigdha Chaturvedi. 2020. Predicting Depression in Screening Interviews from Latent Categorization of Interview Prompts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, Online, 7–18.

[42] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. 2019. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 3–12.

[43] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion* 18, 8 (2004), 1121–1133.

[44] Gabriela Ríos González. 2008. Diferencias léxicas entre el hombre y la mujer en tres centros de interés: Saludos, Temas de conversación y Despedidas. *Revista de Filología y Lingüística de la Universidad de Costa Rica* 33, 1 (mar. 2008), 151–166.

[45] Chris Segrin. 1990. A meta-analytic review of social skill deficits in depression. *Communications Monographs* 57, 4 (1990), 292–308.

[46] Marjana Šifrar Kalan. 2015. Lexical Availability and L2 Vocabulary Acquisition. *Journal of Foreign Language Teaching and Applied Linguistics* 2, 2 (2015).

[47] Urmas Sutrop. 2001. List Task and a Cognitive Salience Index. *Field Methods* 13, 3 (2001), 263–276. arXiv:https://doi.org/10.1177/1525822X0101300303

[48] Allison M Tackman, David A Sbarra, Angela L Carey, M Brent Donnellan, Andrea B Horn, Nicholas S Holtzman, To'Meisha S Edwards, James W Pennebaker, and Matthias R Mehl. 2019. Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis. *Journal of personality and social psychology* 116, 5 (2019), 817.

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

[50] John D Teasdale, Jan Scott, Richard G Moore, Hazel Hayhurst, Marie Pope, and Eugene S Paykel. 2001. How does cognitive therapy prevent relapse in residual depression? Evidence from a controlled trial. *Journal of consulting and clinical psychology* 69, 3 (2001), 347.

[51] Ester Trigo, Manuel-Francisco Romero, and Inmaculada-Clotilde Santos-Díaz. 2019. Empirical approach from lexical availability to the influence of sociolinguistic factors on mastery of spelling / Aproximación empírica desde la disponibilidad léxica a la influencia de los factores sociolingüísticos en el dominio ortográfico. *Culture and Education* 31, 4 (2019), 814–844. https://doi.org/doi.org/10.1080/11356405.2019.1659007

[52] Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. 2018. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering* (2018).

[53] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*. ACM, 3–10.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[55] Esaú Villatoro-Tello, Pavankumar Dubagunta, Julian Fritsch, Gabriela Ramírez-de-la Rosa, Petr Motlicek, and Mathew Magimai.-Doss. 2021. Late Fusion of the Available Lexicon and Raw Waveform-based Acoustic Modeling for Depression and Dementia Recognition. In *Proceedings of the INTERSPEECH 2021. ISCA-International Speech Communication Association 2021.*

[56] Esaú Villatoro-Tello, Gabriela Ramírez-de-la Rosa, and Héctor Jiménez-Salazar. 2017. UAM's Participation at CLEF eRisk 2017 task: Towards Modelling Depressed Blogers.. In *CLEF (Working Notes)*.

[57] Charles Welch, Allison Lahnala, Veronica Perez-Rosas, Siqi Shen, Sarah Seraj, Larry An, Kenneth Resnicow, James Pennebaker, and Rada Mihalcea. 2020. Expressive Interviewing: A Conversational System for Coping with COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020.*

[58] World Health Organization. 2017. *Depression and Other Common Mental Disorders Global Health Estimates*. Retrieved May 11, 2021 from ht+tps://www.who.int/mental_health/management/depression/prevalence_global_health_estimates/en/

[59] Til Wykes, Jessica Lipshitz, and Stephen M Schueller. 2019. Towards the design of ethical standards related to digital mental health and all its applications. *Current Treatment Options in Psychiatry* 6, 3 (2019), 232–242.

[60] Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. 2020. Identifying Depressive Symptoms from Tweets: Figurative Language Enabled Multitask Learning Framework. In *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online), 696–709.

[61] Ziheng Zhang, Weizhe Lin, Mingyu Liu, and Marwa Mahmoud. 2020. Multimodal deep learning framework for mental disorder recognition. In *2020 15th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2020). IEEE.*