Late Fusion of the Available Lexicon and Raw Waveform-based Acoustic Modeling for Depression and Dementia Recognition

Esau Villatoro-Tello1,2, S. Pavankumar Dubagunta1,3, Julian Fritsch2,3, Gabriela Ramírez-de-la-Rosa1, Petr Motlicek2, Mathew Magimai-Doss2

1 Universidad Autónoma Metropolitana Unidad Cuajimalpa, Mexico City, Mexico
2 Idiap Research Institute, Martigny, Switzerland
3 École polytechnique fédérale de Lausanne (EPFL), Switzerland

evillatoro@cua.uam.mx, pavankumar.dubagunta@idiap.ch, julian.fritsch@idiap.ch, gramirez@cua.uam.mx, petr.motlicek@idiap.ch, mathew.magimaidoss@idiap.ch

Abstract

Mental disorders, e.g. depression and dementia, are categorized as priority conditions according to the World Health Organization (WHO). When diagnosing, psychologists employ structured questionnaires/interviews, and different cognitive tests. Although accurate, there is an increasing necessity of developing digital mental health support technologies to alleviate the burden faced by professionals. In this paper, we propose a multi-modal approach for modeling the communication process employed by patients being part of a clinical interview or a cognitive test. The language-based modality, inspired by the Lexical Availability (LA) theory from psycho-linguistics, identifies the most accessible vocabulary of the interviewed subject and use it as features in a classification process. The acoustic-based modality is processed by a Convolutional Neural Network (CNN) trained on signals of speech that predominantly contained voice source characteristics. In the end, a late fusion technique, based on majority voting, assigns the final classification. Results show the complementarity of both modalities, reaching an overall Macro-F1 of 84% and 90% for Depression and Alzheimer’s dementia respectively.

Index Terms: Depression Detection, Alzheimer’s Disease, Mental Lexicon, Raw Speech, Multi-modal Approach.

1. Introduction

Mental disorders represent a major public health concern, with considerable associated socio-economic costs, and are recognized as a major cause of disability affecting a great number of people. According to the World Health Organization (WHO), depression and dementia are among the main types of mental disorders and are categorized as priority conditions [1][2]. 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 health 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 [3]. Such situation has become more evident with the current COVID-19 pandemic. Interested readers are referred to [4][5][6] to know efforts towards this direction.

Accordingly, the research community has been interested in making first steps towards computer-supported detection of mental disorders during face-to-face interviews/tests [7][8][9]. The underlying hypothesis of most of previous work relies on the notion of the language as a powerful indicator about our personality, social, or emotional status, and mental health [10][11].

In dementia, for instance, previous research indicates that assessing the language production represents a useful strategy in detecting early markers of dementia [12]. Thus, designed tests for evaluating the language production in elderly patients such as word association tasks, description of objects in pictures, elicitation exercises, etc., aim at measuring the expository speech, oral expression, as well as comprehension. Similarly, for depression, previous research suggest that using excessive self-focused language, and negative emotions represent important markers for screening depressed users [10][13][14] and, recent studies have documented how depressed users suffer some kind of impairment in their speech motor control [15], such as prosodic abnormalities, articulatory and phonetic errors.

Although multi-modal approaches have been explored before [16][17][18], the key novelty of our work is to leverage the psycholinguistics theory for approximating the mental lexicon of analyzed subjects for processing the language-based modality. The acoustic-based modality aims at modeling the patients’ speech in an end-to-end fashion from raw waveform-based CNNs. In conjunction, both modalities allow modeling the language production process employed by subjects with a mental disease during a clinical interview/test. We performed experiments in two well-known clinical datasets, using individual modalities, and in a multi-modal fashion, where a voter makes the final decision through a majority voting mechanism.

2. Methodology

The proposed language-based modality aims at modeling the vocabulary production of subjects suffering from a mental disorder through the Lexical Availability (LA) theory [19]. The LA test is associated with the category fluency tests and the free word association tasks, which taps directly into the semantic information of the mental lexicon [20]. Hence, our main hypothesis establishes that it could be possible to approximate the available lexicon for a group of people suffering from a mental disease. Contrary to the traditional LA elicitation test, we aim to demonstrate that it is possible to approximate the available lexicon by analyzing subjects’ responses in a semi-structured communication process (e.g. a clinical interview/test). To the best of our knowledge, this is the first time the LA theory is adapted to: i) obtain the available lexicon from utterances produced during a clinical test and use the extracted features in a traditional

1 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 [19].
classification pipeline; ii) fuse its predictions in a multi-modal fashion with a raw waveform-based acoustic approach.

Figure 1 shows the main components of our LA method. First, we identify the available lexicon from each population (i.e., subjects with a mental disorder and control subjects) and then use it to generate a non-sparse text representation (i.e., subjects with a mental disorder and control subjects) and then use it to generate a non-sparse text representation (i.e., subjects with a mental disorder and control subjects). Then, we train a classification model to distinguish between these two classes (i.e., control and subjects). Finally, we fuse its predictions in a multi-modal classification pipeline; ii) fuse its predictions in a multi-modal fashion with a raw waveform-based acoustic approach.

2.1. Lexical availability computation

Traditionally, the LA test produces a single word list, i.e., the available lexicon (with its corresponding availability scores), for each analyzed community. To compute the availability scores of this available lexicon, we have to analyze the responses of each individual in that population (see Fig. 1 columns 1-3); to that end, we use the formulation proposed by [24, 25], defined as follows:

$$D^{st}_{w,k,m}(t_j) = \sum_{i=1}^{n} w_i^{(w+1)^{m}} \times f_{ji}$$

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 $I = \max_{freq}$, which depicts the highest frequency found in the vocabulary of the population being analyzed; $f_{ji}$ 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 a value close to 0) for position $k$, and $m$ is a parameter that modulates the weight decay across terms in the final mental lexicon.

Eq. 2 represents a standardized LA metric that allows direct comparisons among studies independently from the size of the produced vocabulary lists of different communities [21]. Accordingly, the $D^{st}$ equation will assign higher scores (close to 1) to the most available words produced by the analyzed subjects. 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 $w = 0.0001$ and $m = 0.8$.

2.2. Availability features

We defined the availability features ($f_{avail}$) as the single (most representative) LA score for each term $t_j \in (V_{\rho} \cup V_{\varnothing})$. Thus, to obtain the $f_{avail}$ score of term $t_j$, we apply the CombMNZ [22] 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 \( V_{y_j} \) of the instance \( \rho \) applying Eq. 2. Then, for obtaining the \( f^{\text{avail}}_{j} \) we fuse the scores of word \( t_j \) from the list \( V_y \) with the available lexicons \( V_{y_{q_0}} \) and \( V_{y_{q_0}'} \). For this process, we do as follows:

\[
\text{CombMINZ}(t_j, k; \{V_y, V_{y_{q_0}}, V_{y_{q_0}'}\}) = \sum_{c} D^c \times |D^c > 0| 
\]

(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 the \( V \)'s are the set of lists to be considered for 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 \( N = \text{len}(\{V_y, V_{y_{q_0}}, V_{y_{q_0}'}\}) \), \( D^c \) as the score of \( t_j \) 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:

\[
\text{CombMINZ}(t_j, k; \{V_y, V_{y_{q_0}}, V_{y_{q_0}'}\}) = \sum_{c} D^c \times |D^c > 0| 
\]

(4)

Broadly speaking, the \( f^{\text{avail}}_{j} \) of term \( t_j \) represent a weight value indicating to what category it adjust the best.

2.3. Correlation degree features

The correlation degree features aim at measuring the relationship between the two sets of paired words, particularly we compute \( \text{cor}(V_{y_{q_0}}, V_y) \), and \( \text{cor}(V_{y_{q_0}'} , V_y) \). The correlation (cor) value will be an indicator of the association between the available lexicon form subject \( \rho \) and the corresponding \( V_{y_{q_0}} \) and \( V_{y_{q_0}'} \) categories. For the experiments performed in this paper, every \( f^{\text{avail}}_{j} \) feature is formed by two values, the Spearman’s correlation coefficient and its corresponding p-value.

2.4. Acoustic based method

The acoustic based method directly models raw waveforms to predict the class-conditional probabilities using a CNN-based architecture. As described in [23], the architecture consists of four 1-D convolutional layers, followed a hidden layer and an output-layer. In order to guide the learning procedure, depending on the task, different approaches were previously proposed: We distinguish between sub-segmental and segmental filtering (see [23] Table 1); raw waveforms can be filtered to extract voice-source related characteristics to guide the learning procedure. Specifically, for the depression detection task, the primary method (denoted as 1stAcoustic) uses zero frequency filtering to get a signal that characterizes the glottal excitation. The secondary method (denoted as 2ndAcoustic) consists of modeling speech at a frame level using linear prediction and subtracting it from the original speech to get the linear prediction residual, which contains voice source related characteristics, while both use an input length of 250ms. However, for Alzheimer’s detection, both systems use 4 second length inputs of zero frequency filtered signals, where the primary method (denoted as 1stAcoustic) applies a sub-segmental filtering stage, the secondary method (denoted as 2ndAcoustic) a segmental filtering stage.

2.5. Late fusion

Once both the language-based and acoustic-based modalities are trained independently, the late fusion approach consists of a voter that takes as inputs the predictions made by the language-based and acoustic-based approaches. The final decision is made by means of a majority voting mechanism, where if tied, the output will be always labeled as C (i.e., control).

3. Experimental Setup

For the experiments, we use the Distress Analysis Interview Corpus - wizard of Oz (DAIC-WOZ) dataset [24] and the Alzheimer’s Dementia Recognition through Spontaneous Speech (ADReSS) dataset [9]. The DAIC-WOZ dataset contain semi-structured clinical interviews, performed by a (human controlled) animated virtual interviewer, designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic disorder. This dataset was used during the AVEC 2016 challenge [17], and contains audio-visual interviews of 189 participants: 107 for training, 35 for development, and 47 for test. The ADReSS data, introduced for the Interspeech 2020 ADReSS challenge [9], consists of speech recordings and transcripts of spoken picture descriptions elicited from participants through the Cookie Theft picture from the Boston Diagnostic Aphasia Exam [25]. It contains speech and transcripts information from 156 participants: 108 for training, and 48 for test. In the DAIC-WOZ dataset approximately \( 30\% \) of the subjects are labeled as depressed (D), while the ADReSS dataset is perfectly balanced. It is worth mentioning that the labeling of each dataset was done by expert mental healthcare providers, interested reader is referred to [24,9].

We evaluate the performance of three well-known text-based methods. First, a traditional Bag-of-Words (BoW) using the top 1000 most frequent words under a Term Frequency Inverse Document Frequency if-idf weighting scheme. Secondly, we use the Linguistic Inquiry and Word Count (LIWC) [26] categories for representing the documents. LIWC psychological categories capture the semantic content of the language produced [27], e.g., allow to detect positive vs. negative emotions, words referencing family/friends/society, pronouns which can capture inclusive language vs. exclusive language, and words referencing how the person is feeling.

As third baseline, we evaluate the impact of recent transformer-based models [28] as a language representation strategy. For our experiments we test an English pre-trained BERT model. As known, the [CLS] token acts an “aggregate representation” of the input tokens, and is considered as a sentence representation for many classification tasks [29]. Accordingly, for generating the representation of each document, we split the document into smaller chunks (maximum length of 512 tokens), obtain the [CLS] encoding of each chunk, and we apply
4. Results and discussion

As previous research [7,8,23,50,11], performance is reported in terms of the F score (F1) for both control (C) and depression/dementia (D) classes, and the Macro-F for the overall problem (O). We acknowledge the limitations regarding the small size of the corpora, however, this is a common shortcoming of all studies that use clinical datasets. Thus, to achieve stable and robust results, we applied two validation strategies: i) the average performance over a stratified 10 cross-fold-validation using train partition (10-CFV), and ii) the performance over the dev partition for the DAIC-WOZ dataset and on the test partition for the ADReSS dataset.

For the proposed Lexical Availability method, we performed a series of experiments using: i) only the availability degree features (LA-A), and ii) the combination of availability and correlation (LA-AC) as in Eq. 4. Table 4 summarizes our results for the experiments using a 10-CFV strategy. Table 2 shows the performance of the experiments performed on the dev and test partitions, and Table 3 shows the results of the fused predictions. Given our space restrictions, we only report results from the best learning algorithm (Class. column)4 for the experiments using the LA-A/LA-AC methods, the number in the sub-index indicates the value of the k parameter.

Clearly, from Tables 1 and 2 we conclude that our LA method outperforms all the proposed textual-based baselines, including very recent transformer-based models (i.e., BERT). Moreover, the LA-A/AC configuration for both tasks (see Table 2) with \( k = 500 \) for ADReSS. This variation in the value of \( k \) is related to the size of the respective datasets. For instance, the DAIC-WOZ corpus, contrary to the ADReSS dataset, contains more samples of the communicative process (i.e., several utterances from interviewed subject) with a smaller variability of lexical units (i.e., small vocabulary), hence paying attention to a reduced set terms is enough for the LA method.

For the multi-modal experiments (Table 3) we took the best configurations based on the performance on the dev/test sets (Table 2). We compare our results against two recent multi-modal approaches. For depression, we considered the work of [10], which evaluates the performance of a multi-modal LSTM recurrent network. For dementia, [17] combines the outputs of CNN-LSTM model and a Speech-GRU cell for making the predictions. As can be observed, our LA fusion strategy, outperforms very recent approaches by an important margin5.

5. Conclusions

We addressed the problem of detecting mental disorders from clinical tests. Inspired by the LA theory, our method approximates the mental lexicon through the identification of the available lexicon for mentally ill and control subjects, and use it in a classification process to detect depression/dementia. Additionally, based on previous studies that demonstrated the suitability of raw waveform CNNs, we designed a multi-modal approach, where a voter makes the final decision using a majority vote mechanism. A thorough evaluation in two well known clinical datasets (DAIC-WOZ and ADReSS), shows that the LA method fused with the raw waveform-based CNN is able to outperform, by a large margin, very recent deep NN techniques.

6. Acknowledgements

Esaú Villatoro-Tello, was supported partially by Idiap, SNI-CONACyT, and UAM-Cuajimalpa Mexico. S Pavankumar Dubagunta was supported by Innosuisse under the project Conversation Member Match (CMM) Julian Fritsch was funded through the EU’s Horizon H2020 MSCA-ITN-ETN project TAPAS grant agreement No. 766287. Gabriela Ramirez-de-la-Rosa would like to thank UAM-Cuajimalpa for their support.

Table 2: Obtained performance over the dev and test partitions for DAIC-WOZ and ADReSS datasets respectively.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Fused approaches</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>O</td>
<td>D</td>
</tr>
<tr>
<td>DAIC-WOZ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LA-AC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[LA-AC, LA-A]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADReSS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[LA-AC, LA-A]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[LA-AC, LA-A]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Obtained performance of the late fusion approach. The reported performance in (7) for depression was F1=0.58, while for ADReSS in (9) the best reached score was F1=0.75.
7. References


