Speech and Text Processing for Major Depressive Disorder Detection

Edward L. Campbell¹, Laura Docío-Fernández¹, Nicholas Cummins², Carmen García-Mateo¹

¹GTM research group, AtlantTIc Research Center, University of Vigo, Spain
²Institute of Psychiatry, Psychology and Neuroscience, King’s College London, London, UK

Abstract

Major Depressive Disorder (MDD) is a common mental health issue these days. Its early diagnosis is vital to avoid bigger consequences and provide an appropriate treatment. Speech and utterance’s transcription of patients’ interviews contain useful information sources for the automatic screening of MDD. In this sense, speech- and text-based systems are proposed in this paper, using the DAIC-WOZ dataset as experimental framework. The speech-based one is a Sequence-to-Sequence (S2S) model with a local attention mechanism. The text-based one is based on GloVe features and a Convolutional Neural Network as classifier. A description of some of the more relevant results achieved by other research publications on DAIC-WOZ are described as well. The goal is to provide a better understanding of the context of our systems results. In general, the S2S architecture provides mostly better results than previous speech-based systems. The GloVe-CNN system shows even a better performance, leading to the idea that text is a more suitable information source for the detection of MDD when it is manually developed. However, to automatically obtain high quality transcriptions is not a straightforward task, which makes necessary the development of effective speech-based systems as the presented in this research work.

Index Terms: Major Depressive Disorder, Sequence-to-Sequence modelling, GloVe features, DAIC-WOZ corpus

1. Introduction

Mental health includes our emotional, psychological, and social well-being [1]. Stress, depression, anxiety are some of its negative manifestation in our every-day-life. Institutions such as the World Health Organization have developed campaigns to increase social awareness [2, 3, 4]. The illustration guide “Doing What Matters in Times of Stress” [5] is a clear example. Major Depressive disorder (MDD) is one of the most common mental health diseases and is frequently linked to increases in suicidal behaviours [6].

MDD can cause a persistent feeling of sadness and loss of interest. Although depression may occur only once during your life, people typically have multiple episodes. During these episodes, symptoms occur most of the day, nearly every day and may include [7]:

- Feelings of sadness, tearfulness, emptiness or hopelessness
- Loss of interest or pleasure in most or all normal activities such as hobbies or sports
- Slowed thinking, speaking or body movements
- Frequent or recurrent thoughts of death, suicidal thoughts, suicide attempts or suicide
- “Patients speak in a low voice, slowly, hesitatingly, monotonously, sometimes stuttering, whispering, try several times before they bring out a word, become mute in the middle of a sentence” [8, 9].

Speech affected by depression is often subjectively characterised in clinical settings by decreases in verbal activity, decreases in utterance length, a reduction in speech rate and an increase in long silent pauses [9, 10, 11, 12]. Speech has many properties which make it an attractive candidate for use in an automated screening tool; it can be measured cheaply, remotely, non-invasively and non-intrusively [13]. On the other side, previous studies have already shown that depression also has an effect on the language used by affected individuals [14]. For example, a more frequent use of first person singular pronouns in spoken language was first observed in 1981 [15, 16].

Because of the complexity of speech production, speech is a sensitive signal; slight physiological and cognitive changes potentially can produce noticeable acoustic changes (Scherer, 1986). Voice has five main information levels that compound a pyramidal structure [17, 18]. Acoustical and prosodical levels are the basis. From these, we can extract information related to how a patient speak (e.g., energy levels, speech-rate and fluency). Mel Frequency Cepstral Coefficient (MFCC), for example, are a well-know acoustical feature extraction method with a widespread use in speech tasks [19, 20, 21]. Mel-Spectrum is a variant of MFCC, where the Discrete Cosine Transform is omitted [22]. A system based on these features and a classifier should be able to detect speech patterns such as depression symptoms. On the other side, transcriptions of patients’ interviews carry semantic information that supports the MDD screening process [23, 24]. There are cues (e.g., increase in first person pronouns and negative words) in the linguistic style that show symptoms related to MDD [25, 26]. [27] was one of the first research to analyze this concept.

In this paper, a speech- and text-based algorithms is presented. The pipeline of the speech-based system is compound by Mel-Spectrum as feature extraction and a Sequence-to-Sequence (S2S) model with local attention mechanism [28, 29] as classifier. The S2S was a pre-trained model on the RADAR-MDD unscripted speech corpus with the Hard-training regularization methodology [30]. Therefore, a fine-tune process on the DAIC-WOZ dataset was needed. A Convolutional Neural Network (CNN) and GloVe vectors were the main modules of the text-based architecture proposed.

The paper structures is as follows: the experimental framework is described in Section 2. Section 4 is focused on the proposed text and speech-based systems. Then a brief survey about last research results on DAIC-WOZ dataset is made in Section 3. Section 5 and discusses our results.
2. DAIC-WOZ corpus

The Distress Analysis Interview Corpus (DAIC) was designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post traumatic stress disorder [31]. It contains clinical interviews conducted by humans, human controlled agents and autonomous agents [31]. The corpus refereed as DAIC-WOZ (DAIC- Wizard-of-Oz), and used in this paper, only contains interviews conducted by an animated virtual interviewer called Ellie . This virtual interviewer was controlled by a human in another room. Participants were recruited through on-line ads posted on Craigslist.org. All DAIC-WOZ’s interviews were conducted at the USC Institute for Creative Technologies (ICT) in Los Angeles, California. The interview duration per participant is between 5-20 minutes

To label the presence of MDD in participants, the PHQ-8 [32] assessment tool was provided in each interview. It is established as a valid diagnostic and severity measure for depressive disorders in large clinical studies [32, 33]. PHQ-8 is a questionnaire where an score between 0 and 24 is assigned as an equivalent to the depression level of participants. The higher the score, the greater the likelihood of MDD presence. As general rule, the decision cut-point is picked at 10, meaning that participants with a score over or equal to 10 are more likely to have depression symptoms [32].

DAIC-WOZ corpus is split into training, development and test sets. Figure 1 illustrates the score distribution of the training samples. There are more participants without MDD symptoms (PHQ-8 < 10 ) than with them (PHQ-8 >= 10). This data imbalance condition would cause a strong decision bias during the training of a machine learning model [34]. Therefore, techniques such as resampling [34], class weight [35] are encourage to process this database.

![Histogram of PHQ-8 scores of the training set from DAIC-WOZ](image.png)

Figure 1: Histogram of PHQ-8 scores of the training set from DAIC-WOZ

Table 1 shows the number of participants per classes in every set. The class imbalance is still present in the development and test sets.

Table 1: Number of participants with/without depression according to the PHQ-8 score.

<table>
<thead>
<tr>
<th>Set</th>
<th>Positives</th>
<th>Negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>31</td>
<td>76</td>
</tr>
<tr>
<td>Development</td>
<td>12</td>
<td>23</td>
</tr>
<tr>
<td>Test</td>
<td>14</td>
<td>33</td>
</tr>
</tbody>
</table>

3. Previous work on DAIC-WOZ corpus

Before getting in our system description and results, we would like to describe some of the systems and last results obtained by other researches. This can help the reader to understand the context of our work.

Through the years, several approaches have been investigated from speech and/or text-based systems to multimodal ensembles. A Linear Support Vector Machine classifier with eGeMAPS as feature extraction algorithm was proposed as baseline in AVEC 2016 [36]. Hierarchical attention mechanism were also investigated in addition to the use of GloVe features [37]. The experience obtained from the Speaker Recognition field, with the i-vector methodology, was another interesting approach [38]. The advantage of multimodal approaches was also considered. For example, the fusion of audio and video modalities was accomplished in [39]. Table 2 depicts our selection of more relevant speech-based results achieved on the DAIC-WOZ corpus. Because of the unbalance between classes, F-1 and Unweighted Average Recall (UAR) are given as performance measures. To the best of the authors knowledge, no works have fused these GloVe and S2S models with local attention mechanism together.

Table 2: Published performance on DAIC-WOZ corpus

<table>
<thead>
<tr>
<th>System</th>
<th>set</th>
<th>UAR</th>
<th>F-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>GloVe-HCAN</td>
<td>dev</td>
<td>.54</td>
<td>.51</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>.66</td>
<td>.63</td>
</tr>
<tr>
<td>AVEC-2016 [36]</td>
<td>dev</td>
<td>.70</td>
<td>.57</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>.66</td>
<td>.50</td>
</tr>
<tr>
<td>DCT features [40]</td>
<td>dev</td>
<td>-</td>
<td>.73</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>-</td>
<td>.67</td>
</tr>
<tr>
<td>i-vector &amp; G-PLDA [38]</td>
<td>dev</td>
<td>.73</td>
<td>.73</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DepAudioNet [41]</td>
<td>dev</td>
<td>.77</td>
<td>.61</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

4. Proposal systems. Speech and linguistic domain

Speech and text are two valuable information sources to work on during the detection of MDD. The first one is easier to collect but the detection efficacy used to be lower. In the case of transcriptions, they carry more detailed semantic information from participant’s utterances. Since their benefits differ, we decide to focus on both of them. The next subsections depict our speech- and text-based system proposals.

4.1. Speech processing by a Sequence-to-Sequence model

The detection process has to main steps: feature extraction and classification. Regarding the speech-based architecture, Mel-Spectrum was selected as feature extraction algorithm. As this features are compound by a long-term sequential vectors, a Sequence-to-Sequence (S2S) model (Figure 2) with local attention mechanism [42, 28] was used as classifier. It was based on our previous work at [30]. S2S is built on Recurrent Neural Network (RNN) architectures which have been shown to be effective in the processing of sequential data [43]. RNNs ability to track and store dependencies throughout a sequence has been
key in tasks such as Stock Price Pattern Recognition [44, 45] and health care [46]. The use of RNN’s is further justified given that depression has been shown to alter temporal properties of speech [47, 48].

The manual transcription of training samples is likely to be the repel.

Space where similar words cluster together and different words probabilities and their vector representations are located in a case of test. This result is achieved by our proposed ensemble version trained from scratch on the development set.

To validate the benefits of the pre-trained model, we run an experiment where ten S2S models were trained from scratch, and their results were compared with other ten fine-tune S2S models previously trained on the RADAR corpus. Table 3 illustrates average results, showing a clear improvement in the fine-tuned model.

Regarding the text-based model, we extracted 50-dimensional GloVe features using a model trained on Wikipedia 2014 and Gigaword 5. The CNN is structured by 2 pairs of convolutional layers with a max pooling layer. Then, another Convolutional layer is stacked with a global max-pooling layer. The last layer is a feedforward one with a drop-out of 0.5. This model had a total of 1 083 534 parameters of which 786 434 were trainable. This is a widespread CNN configuration where 1-dimensional convolving filters are used to extract high level features, meanwhile the max-pooling layer gets a more abstract version of them [51, 52]. The function of the feedforward layer is to make the final decision [52].

Figures 3a and 3b show the results of our proposed systems during 30 runs on the development set. The S2S model had a more steady performance meanwhile the GloVe-CNN model had a higher pick. Then, an ensembling strategy is followed in test for each modality (speech/text). The test set is processed by 15 S2S model on trained on the training and development sets, obtaining 15 final scores per interview. Then, these scores are averaged to get an ensembling score. The same methodology is followed with 15 instances of the GloVe-CNN architecture.

Table 4 shows our final results in test and development. The results on the development is an average of the 30 runs. In the case of test, this result is achieved by our proposed ensemble speech- and test-based systems.

We can see that the text-based system outperform the S2S. The manual transcription of training samples is likely to be the

Figure 2: Sequence to sequence model architecture with local attention mechanism

S2S is a modelling paradigm that uses two sets of RNNs to convert one sequence of items in one domain into a sequence in another domain [49]. The first RNN network is known as the encoder and the second one as decoder. The encoder learns to process each item of an input sequence and converts this information into a fixed (static) representation vector known as the context vector. The decoder then learns to convert this static representation into a new sequence.

The core component of the encoder and decoder set-up is the RNN blocks. In our work, they are realised by a Gated Recurrent Unit (GRU) layers (Fig. 2). In the input, a Batch Normalization layer is applied for decreasing the training time of the model. Linear transformations is also applied for guarantee matrix compatibility between some consecutive blocks. In the decoder, we added a local attention mechanism in order to consider information relevance when processing the output.

To sum up, the methodology of this system is the following: An interview is divided into a sequence of Mel-Spectrum features which are classified by a Sequence-to-Sequence model with a local attention mechanism. Each sequence gets a score assigned by the classifier. Finally, those sequence scores are averaged to get a final score that represents the participant depression state.

4.2. Text processing through GloVe features

The pipeline of the proposed text-based system is compound by the Global Vectors for Word Representation (GloVe) [50] algorithm and a Convolutional Neural Network (CNN). Glove is an unsupervised method for learning word representation by its statistic occurrence in a corpus [50]. Its main advantage is that it does not rely just on local statistics but incorporate global ones [50]. Words meaning is extracted directly from co-occurrence probabilities and their vector representations are located in a space where similar words cluster together and different words repel.

During a participant’s interview, each utterance is divided into a sequence of embeddings (GloVe features) to be processed by a CNN classifier. Then, we have a score per sequence associated to the depression level of its respective utterance. As final step, the interview scores are averaged, getting a final score as representation of the depression likelihood of participant.

5. Discussion and Results

Because of the unbalance condition between classes in the DAIC-WOZ corpus, we decided to fine-tune our previously trained model on the English and unscripted set of the RADAR-MDD corpus [30]. The RADAR-MDD corpus is a collection of individuals’ speeches with different levels of depression symptoms. The English-unscripted set is about 17 hours long and presents an appropriate class distribution. This methodology helps to mitigate the model’s decision bias caused by a higher number of negative classes (non-depressed) in the DAIC-WOZ corpus. The dimensionality of the Mel-Spectrum feature vector of this system was 40.

To validate the benefits of the pre-trained model, we run an experiment where ten S2S models were trained from scratch, and their results were compared with other ten fine-tune S2S models previously trained on the RADAR corpus. Table 3 illustrates average results, showing a clear improvement in the fine-tuned model.

Table 3: Average performance of the fine-tuned model and the version trained from scratch on the development set.

<table>
<thead>
<tr>
<th>system</th>
<th>set</th>
<th>F-1</th>
<th>UAR</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2S (fine-tuned)</td>
<td>dev</td>
<td>.65</td>
<td>.66</td>
<td>.71</td>
</tr>
<tr>
<td>S2S (from scratch)</td>
<td>dev</td>
<td>.60</td>
<td>.61</td>
<td>.68</td>
</tr>
</tbody>
</table>

Figure 2 illustrates average results, showing a clear improvement in the fine-tuned model.
Figure 3: Performance of proposed speech- and text-based systems on the development set of DAIC-WOZ through 30 runs.

key factor that support this behaviour. We do not expect the same results in a real scenario with automatic transcriptions.

Table 4: Ensembling system results on development and test sets

<table>
<thead>
<tr>
<th>system</th>
<th>set</th>
<th>F-1</th>
<th>UAR</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2S (Score-Ensembling)</td>
<td>dev</td>
<td>.65</td>
<td>.64</td>
<td>.74</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>.63</td>
<td>.68</td>
<td>.64</td>
</tr>
<tr>
<td>GloVe-CNN</td>
<td>dev</td>
<td>.73</td>
<td>.70</td>
<td>.86</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>.68</td>
<td>.67</td>
<td>.74</td>
</tr>
<tr>
<td>Fusion</td>
<td>test</td>
<td>.63</td>
<td>.62</td>
<td>.72</td>
</tr>
</tbody>
</table>

6. Conclusions

Two systems, based on different information sources, were described in this paper. The first one was a Sequence-to-Sequence model with local attention mechanism (speech-based). The second one was a based on GloVe features and a CNN as classifier (text-based). Both systems outperform in their respective modality to those published in last year on the DAIC-WOZ corpus. In our case, the GloVe-CNN system, as a text-based system, had even better results than the S2S. We attribute this behaviour to the manual elaboration of interview transcriptions collected in the experimental corpus of DAIC-WOZ, which guarantees text-based features of great quality. On the other side, the S2S model shows a more steady performance through different runs. Speech and text modalities are key for the detection of MDD. Its difference is what make them so valuable. The one we choose for a real application just depends on our need.

7. Acknowledgements

This work has received financial support from Axudas propias para a mobilidade de Personal Investigador da Universidade de Vigo 2021, the Xunta de Galicia (Centro singular de investigación de Galicia accreditation 2019-2022), Consellería de Cultura (Educación e ordenación Universitaria; axudas for a consolidación e estruturación de unidades de investigación competitivas do Sistema Universitario de Galicia -ED431B 2021/24), the European Union (European Regional Development Fund -ERDF) and also from Axudas de apoio á etapa predoutoral da Consellería de Cultura, Educación e Ordenación Universitaria da Xunta de Galicia.

8. References


