IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 November 2023, accepted 16 January 2024, date of publication 2 February 2024, date of current version 12 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3361750

RESEARCH ARTICLE

Coupling Sentiment and Arousal Analysis Towards an Affective Dialogue Manager

ADRIA MALLOL-RAGOLTA[1,2,4] **AND BJÖRN SCHULLER**[1,2,3,4,5]**, (Fellow, IEEE)**
[1]EIHW—Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, 86159 Augsburg, Germany
[2]CHI—Chair of Health Informatics, MRI, Technical University of Munich, 81675 Munich, Germany
[3]MDSI—Munich Data Science Institute, 85748 Garching, Germany
[4]MCML—Munich Center for Machine Learning, 80333 Munich, Germany
[5]GLAM—Group on Language, Audio, & Music, Imperial College London, SW7 2AZ London, U.K.

Corresponding author: Adria Mallol-Ragolta (adria.mallol-ragolta@informatik.uni-augsburg.de)

This work was supported in part by the European Union's Horizon 2020 Research and Innovation Programme [Smart environments for person-centered sustainable work and well-being (sustAGE)] under Grant 826506, in part by the German Research Foundation's Reinhart Koselleck Project (AUDI0NOMOUS) under Grant 442218748, and in part by the Bavarian Ministry of Science and Arts through the ForDigitHealth Project funded by the Bavarian Research Association on Healthy Use of Digital Technologies and Media.

ABSTRACT We present the technologies and host components developed to power a speech-based dialogue manager with affective capabilities. The overall goal is that the system adapts its response to the sentiment and arousal level of the user inferred by analysing the linguistic and paralinguistic information embedded in his or her interaction. A linguistic-based, dedicated sentiment analysis component determines the body of the system response. A paralinguistic-based, dedicated arousal recognition component adjusts the energy level to convey in the affective system response. The sentiment analysis model is trained using the CMU-MOSEI dataset and implements a hierarchical contextual attention fusion network, which scores an Unweighted Average Recall (UAR) of 79.04 % on the test set when tackling the task as a binary classification problem. The arousal recognition model is trained using the MSP-Podcast corpus. This model extracts the Mel-spectrogram representations of the speech signals, which are exploited with a Convolutional Neural Network (CNN) trained from scratch, and scores a UAR of 61.11 % on the test set when tackling the task as a three-class classification problem. Furthermore, we highlight two sample dialogues implemented at the system back-end to detail how the sentiment and arousal inferences are coupled to determine the affective system response. These are also showcased in a proof of concept demonstrator. We publicly release the trained models to provide the research community with off-the-shelf sentiment analysis and arousal recognition tools.

INDEX TERMS Affective dialogue manager, sentiment analysis, arousal recognition, emotional artificial intelligence, human–computer interaction.

I. INTRODUCTION

The market penetration of *smart* devices is increasing every year and is changing the way how users interact with the technology. For instance, the launch of voice-based *Virtual Assistants* (VAs) – such as Siri™ (Apple), Alexa™ (Amazon), Cortana™ (Microsoft), Bixby™ (Samsung), Celia™ (Huawei), or Google Assistant™ – has advanced the *Human-Computer Interaction* (HCI) field, as these deploy hardware and software components that allow users to interact verbally with these assistants towards a natural interaction. Current VAs focus on the analysis of the linguistic information to provide this sort of natural interaction. Nevertheless, human-human communication is more complex, as the nonverbal communication is a fundamental and decisive aspect of the interaction. Hence, to boost the user experience when interacting with VAs towards a more natural and realistic interaction, there is a need to power these assistants with affective capabilities by means of affective computing technologies [1], [2].

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

Research works on VAs with affective capabilities can be found in the literature. Among the most recent examples, we highlight the EMPATHIC Virtual Coach [3] and the Ryan agent [4]. The former [3] modifies the agent's voice based on the user's emotional state, which is inferred from the user's face and the paralinguistic information embedded in the user's voice recorded during the interaction. The latter [4] includes an affective dialogue manager able to generate responses based on the inferred emotions of the users. Despite considering multimodal information – as the system features sentiment analysis and face emotion recognition –, the information inferred from a single modality is sufficient to determine the affective response.

We present the technologies developed for sentiment and arousal analysis, so that a speech-based dialogue manager can adapt the system response to the sentiment and arousal level conveyed by the user during the interaction. We utilise a customised smartphone app as the gateway for users to communicate and interact with the system. The dialogue manager features a dedicated sentiment analysis component, which exploits the linguistic information embedded in the user's voice, and a dedicated arousal recognition component, which analyses the paralinguistic information. While the output of the former determines the body of the system answer, the output of the latter conditions the level of energy to convey in the response. We detail two of the sample dialogues deployed at the back-end of the system – to exemplify the system logic in the specific use case of an agent that engages its users with short affective dialogues at different points throughout their working day [5] – and provide a proof of concept demonstrator to showcase the implemented affective dialogue manager. An additional contribution of this work is the public release of the *Application Programming Interfaces* (API) developed to interact with the models trained in an attempt to provide the research community with off-the-shelf sentiment analysis and arousal recognition tools.

The scientific contribution of this work focuses on determining the optimal sentiment analysis and arousal recognition models to deploy in the system, which are trained using the CMU-MOSEI dataset [6], and the MSP-Podcast corpus [7], respectively. The CMU-MOSEI dataset is annotated in terms of both sentiment and emotion. Although the sentiment annotations are in the continuous space, the emotional annotations are in the categorical space. Hence, for a fine-grained arousal recognition, we opt for the MSP-Podcast corpus, as it provides affective annotations in the continuous space. In the sentiment analysis literature, a range of conventional [8], [9] and deep learning [10] approaches have been explored. *Recurrent Neural Networks* (RNN) are a specific deep learning technique suitable for sentiment analysis, as it is a sequence modelling task with variable length inputs. The goal of an RNN is to learn an embedded representation of the input sequence, which is then coupled with a classification block responsible for the actual inference. This embedded representation usually corresponds to the hidden state of the RNN produced at the last time step of the input sequence, which encodes information from the whole sequence, but excludes the previous hidden states from the preceding computations, losing potential information. The experiments we conduct target the assessment of this aspect, as we hypothesise that an effective fusion of the hidden states learnt at each time step could help improve the performance of the sentiment analysis models. Following the current trends in the *Artificial Intelligence* (AI) domain, researchers have recently started investigating the utilisation of Transformers [11], [12] and *Large Language Models* (LLM) [13] for sentiment analysis. In the paralinguistic-based affective computing literature, a wide range of feature representations [14] and network architectures [15], [16] have been studied, highlighting the dependency of the models performance on the available data and the targeted application. Thus, we compare the performance of arousal recognition models trained with different neural network architectures exploiting hand-crafted and deep-learnt representations extracted from the speech signals.

TABLE 1. Statistics of the resulting CMU-MOSEI dataset per partition after aligning the linguistic and word embedding representations of the original data, and segmenting the original videos into the corresponding sentences.

Partition	Sentences	Averaged words per sentence	Words of the longest sentence
Train	16 327	21.16	310
Devel	1 871	21.84	306
Test	4 662	21.29	271

The rest of the paper is organised as follows. Section II introduces the datasets explored to train the sentiment analysis and the arousal recognition models. Section III describes the methodology followed. Specifically, this section provides an overview of the composite system and reports on the research conducted in both research areas. Section IV summarises and analyses the results obtained from the experiments conducted. Section V provides a proof of concept demonstrator, showcasing the overall affective dialogue manager implemented, and Section VI concludes the paper.

II. DATASETS

This section introduces the two datasets exploited in this work. Section II-A presents the CMU-MOSEI dataset [6] – used to train the sentiment analysis models –, while Section II-B describes the MSP-Podcast corpus [7] – employed to train the arousal recognition models.

A. CMU-MOSEI DATASET

The data used for training the sentiment analysis model belongs to the CMU-MOSEI dataset [6]. This is one of the largest gender-balanced multimodal datasets for sentiment analysis and emotion recognition in English,

TABLE 2. Number of negative and positive samples belonging to the train, development, and test partitions of the resulting CMU-MOSEI dataset when considering the task as a binary classification problem.

Partition	Sentiment		$\sum$
	Negative	Positive	
Train	4 739	8 050	12 789
Devel	506	932	1 438
Test	1 350	2 287	3 637

TABLE 3. Number of low, mid, and high arousal samples belonging to the train, development, and test partitions of the resulting MSP-Podcast corpus when considering the task as a three-class classification problem.

Partition	Arousal			$\sum$
	Low	Mid	High	
Train	3 248	25 767	9 084	38 099
Devel	375	5 033	2 106	7 514
Test	994	8 419	3 469	12 882

containing more than 3 000 video clips with language, vision, and acoustic features extracted from over 65 hours of video. To download and process the data, we use the CMU Multimodal Data SDK[1] [17]. For the purpose of our study, we exploit the original sequences available, and their corresponding 300-dimensional word embedding representations extracted using *Global Vectors* (GloVe) [18]. We opt for the exploitation of the GloVe word embeddings for consistency with previous works in the literature exploiting the CMU-MOSEI dataset [6], [17]. Early processing of the corpus using the available SDK includes the alignment of both linguistic representations and the segmentation of the original videos into the corresponding sentences, and their splitting into the pre-defined train, development, and test partitions. The compiled vocabulary contains 16 824 tokens. Table 1 synthesises the statistics of the resulting data.

Each sentence in the dataset is annotated with a sentiment score in the range $[-3, 3]$, determined by 3 crowdsourced annotators. These scores correspond to highly negative (-3), negative (-2), weakly negative (-1), neutral (0), weakly positive (1), positive (2), and highly positive (3) sentiments. In this work, we aim to tackle the sentiment analysis task as a binary classification problem to properly support the envisioned use cases of the presented dialogue manager (cf. Section V). Consequently, we map the scores $\in [-3, -1]$ to the negative class, and the scores $\in [1, 3]$ to the positive class. Although related works in the literature based on this corpus cluster the sentences corresponding to the neutral sentiment into the negative class [19], we exclude these sentences to minimise biasing our models towards the negative class. Table 2 summarises the number of positive and negative sentences belonging to the resulting train, development, and test partitions.

B. MSP-PODCAST CORPUS

The data explored for training the arousal recognition model belongs to the MSP-Podcast corpus [7], which was gathered from freely available English podcasts. The selected podcasts were converted into the audio format 16 kHz/16 bit single-channel PCM. The resulting recordings were segmented, so that information from a single speaker was contained in each audio segment. The corpus was annotated via crowdsourcing in terms of emotional attributes (arousal, valence, and dominance), and categorical emotions. For the purpose of our study, we focus only on the arousal-related annotations, as arousal seems to be more prominent in the paralinguistic information embedded in the user's voice [20].

To annotate the audio segments in terms of arousal, the annotators rated the perceived level of arousal of the speaker using a seven-point Likert scale; i. e., the annotators were asked to rate whether the speaker was perceived to be very calm (1), calm (2), somewhat calm (3), neutral (4), somewhat active (5), active (6), or very active (7). Each segment was evaluated by several annotators, and the gold standard was determined as the average value among the annotations provided by the individual annotators.

As the envisioned affective dialogue manager does not need to infer arousal information with this level of granularity, we simplify the problem by clustering the arousal annotations in three different levels [21]: the annotations $\in [1, 3]$ are assigned to the low arousal class, the annotation in $\in (3, 5]$, to the mid arousal class, and the annotations $\in (5, 7]$, to the high arousal class. Table 3 summarises the number of audio samples assigned to the low, mid, and high arousal classes belonging to the resulting train, development, and test partitions.

III. METHODOLOGY

The architecture and the information workflow of the overall system is depicted in Figure 1. A smartphone app acts as a gateway, so the users can record their own voice to communicate and interact with the system. The resulting media file is then transferred via the Internet to the system back-end. Upon reception, the speech file is transcribed using the off-the-shelf *Automatic Speech Recognition* (ASR) service provided by Google Cloud. The benefit of this approach is that the back-end can exploit the recorded speech file, and the resulting transcription, separately.

The proposed back-end architecture contains three main blocks: i) the sentiment analysis component, ii) the arousal recognition component, and iii) the dialogue manager component. Sections III-A and III-B describe the methodology followed to determine the best sentiment analysis, and arousal recognition models, respectively, to deploy in the respective components. The implementation of the dialogue manager component is detailed in Section V. This engineering-based section emphasises how the sentiment and the arousal information inferred is coupled to affectively adapt the system response to the current affective state of the user interacting with the system.

[1] https://github.com/A2Zadeh/CMU-MultimodalDataSDK

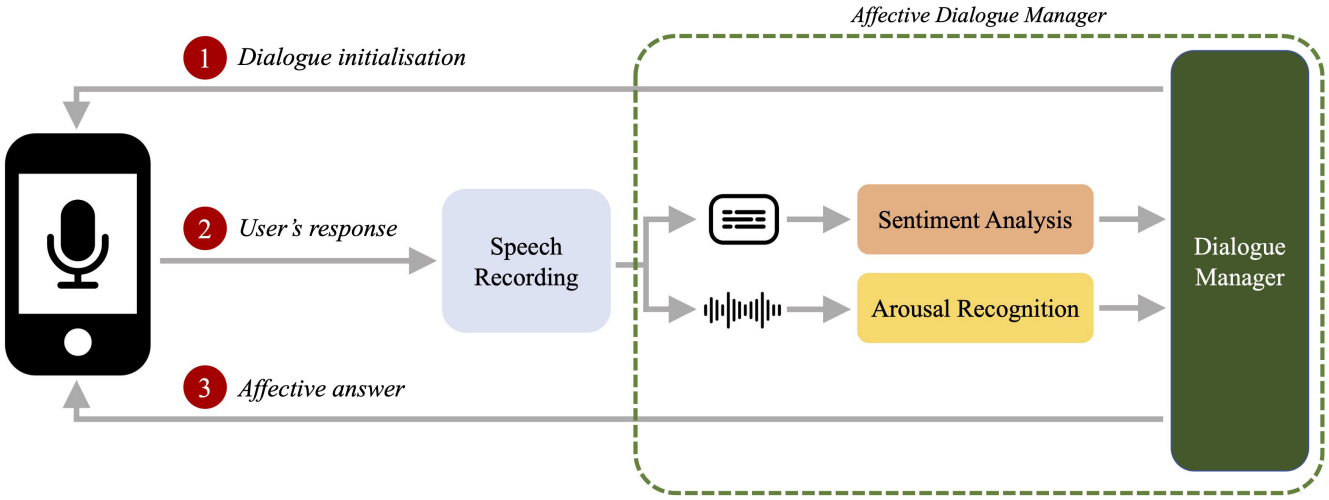

FIGURE 1. Block diagram illustrating the architecture of the affective dialogue manager system implemented and presented in this work, including the interaction workflow. The system answer is determined by considering the sentiment and the arousal information inferred from the linguistic and the paralinguistic analysis, respectively, of the voice-based user's response.

A. SENTIMENT ANALYSIS COMPONENT

This section describes the methodology followed to determine the best sentiment analysis model. Section III-A1 details the pre-processing applied to the sentences belonging to the CMU-MOSEI dataset (cf. Section II-A), Section III-A2 introduces the models implemented, and Section III-A3 summarises their training details.

1) DATA PREPARATION

Each sentence in the CMU-MOSEI dataset is composed of a different number of tokens. The first step is, therefore, the homogenisation of the sequence lengths, so these can be used to train our neural networks. According to the results obtained from our data analysis (cf. Table 1), the longest sentence belongs to the training partition and has a total of 310 tokens. Thus, we fix the length of the sequences to train our networks to 310 time steps. This parameter determines the maximum length of the sentences that can be analysed with our models at inference time. So that the training sentences have a length of 310 time steps, we opt for repeating the sequence of tokens for the shorter sequences until reaching the desired sequence length, avoiding zero-padding. Nonetheless, the sentences with their original lengths are used when evaluating the performance of the models. To overcome the imbalanced data in terms of the positive and the negative sentiments (cf. Table 2), we upsample the under-represented classes via replication, so that the same number of samples for each sentiment is used for training the networks at each epoch.

2) MODELS DESCRIPTION

The sentiment analysis networks implemented in this work are composed of two main blocks: the first block is responsible for learning the embedded representations of the input sequences, while the second block, for the actual classification. The first block features a single-layer, bidirectional *Gated Recurrent Unit – Recurrent Neural Network* (GRU-RNN) with 128 hidden units. We select the use of a GRU-RNN to overcome the vanishing gradient problem suffered by other RNNs, such as the *Long Short-Term Memory – Recurrent Neural Network* (LSTM-RNN). The embedded representation learnt at the output of this block can be mathematically represented as

$$\boldsymbol{h}_i = \left[\overrightarrow{GRU}\left(\boldsymbol{w}_i\right), \overleftarrow{GRU}\left(\boldsymbol{w}_i\right)\right], \tag{1}$$

where $\boldsymbol{w}_i$ corresponds to the word embedding representations extracted from the sequence of words $[w_1 \cdots w_s]$ in the sentence. The second block is composed of two-stacked fully connected layers, preceded by two dropout layers with probability 0.3. The first layer contains 32 neurons and uses the *Rectified Linear Unit* (ReLU) as the activation function. The second layer has as many neurons as classes we need to classify our samples and uses Softmax as the activation function, so that the outputs of the network can be interpreted as probability scores.

The embedded representations learnt at the output of the first block, $\boldsymbol{h}_i$, encapsulate the salient information from the input sequences. Hence, the way how this information is exploited determines the performance of the overall model. In this work, we exploit the embedded representations $\boldsymbol{h}_i$ using the following network architectures.

i) **Baseline Network (Baseline RNN)**. The baseline network uses the last hidden state of the GRU-RNN as a standalone representation of the input sequence, $\tilde{\boldsymbol{h}}$. This embedded representation is then fed to the second block of the network for the actual classification.

ii) **Hierarchical Naïve Fusion Network (H-N)**. This network fuses the sequence of embedded representations, $\boldsymbol{h}_i$, by averaging the representations over all the

sequence. This can be mathematically formulated as:

$$\tilde{\boldsymbol{h}} = \frac{\sum_{i=1}^{S} \boldsymbol{h}_i}{S}. \tag{2}$$

We refer to this approach as a naïve fusion method, since no parameters need to be trained by the network.

iii) **Hierarchical Contextual Attention Fusion Network (H-CA)**. Based on the methodology presented in [22] and adapted from [23], this approach fuses the information by computing contextual attention scores as follows:

$$\boldsymbol{u}_i = \tanh\left(\mathbf{W}\boldsymbol{h}_i + \mathbf{b}\right), \tag{3}$$

$$\alpha_i = \frac{\exp\left(\boldsymbol{u}_i^T \mathbf{u}\right)}{\sum_{i=1}^{S} \exp\left(\boldsymbol{u}_i^T \mathbf{u}\right)}, \tag{4}$$

$$\tilde{\boldsymbol{h}} = \sum_{i=1}^{S} \alpha_i \boldsymbol{h}_i. \tag{5}$$

In this approach, $\mathbf{W}$, $\mathbf{b}$, and $\mathbf{u}$ are defined as trainable parameters. The parameter $\mathbf{u}$ can be interpreted as a contextual tensor, which contributes to the identification of the relevant words in the sentences.

iv) **Convolutional Fusion Network (CNN)**. The fusion of the sequence of embedded representations is performed using a 1-dimensional convolutional layer with 256 and 128 input and output channels, respectively, a kernel size of 3, and a stride of 1. The parameters selected guarantee a smooth integration of this convolutional block into the baseline network for a fair and effective comparison between the models. Batch normalisation is applied to the output of the convolution, and the resulting representation is transformed using a ReLU function. Finally, a 1-dimensional adaptive average pooling is applied to obtain 2 values as a result of the fusion. The final representation is reshaped into a 1-dimensional tensor $\tilde{\boldsymbol{h}}$, ready to be fed into the classification block of the network.

v) **Convolutional Contextual Attention Fusion Network (CNN-CA)**. This final network combines the approaches described for the H-CA and the CNN networks. First, the contextual attention scores from the sequence of embedded representations are computed as defined in Equations (3) and (4). Then, $\boldsymbol{h}_i$ is transformed into an intermediate representation mathematically defined as

$$\boldsymbol{h'}_i = \alpha_i \boldsymbol{h}_i. \tag{6}$$

This new representation $\boldsymbol{h'}_i$ is then exploited using a 1-dimensional convolutional layer, as described for the CNN network.

3) NETWORKS TRAINING

At the initialisation of each network, the pseudo-random number generator is manually seeded for a fair comparison, and reproducibility of the results. The models described in Section III-A2 are trained using the Categorical Cross-Entropy as the loss to optimise. As the optimiser, we use Adam with a fixed learning rate of 10^{-4}. The network parameters are updated in batches of 256 samples, and their gradients are clipped at 1. The networks are trained during a maximum of 100 epochs, and we implement an early stopping mechanism to stop training when the validation loss does not improve for 20 consecutive epochs. Using this early stopping mechanism, we determine the number of epochs needed for training the networks, while minimising the chances of overfitting.

B. AROUSAL RECOGNITION COMPONENT

This section describes the methodology followed to determine the best arousal recognition model. Section III-B1 details the pre-processing applied to the speech samples belonging to the MSP-Podcast corpus (cf. Section II-B), Section III-B2 introduces the models implemented, and Section III-B3 summarises their training details.

1) DATA PREPARATION

The feature representations to extract from the original audio files play a vital role in the paralinguistic analysis. Hence, we aim to compare the performance of the arousal models when exploiting the functionals of the *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) [24] extracted using openSMILE [25], and the Mel-spectrogram representations of the audio signals. The former extracts an 88-dimensional feature vector representation of each audio signal as a whole. The Mel-spectrograms are computed using 128 Mels and a hope size of 128 samples. The audio signals in the MSP-Podcast corpus have different durations. To homogenise their duration for training the models, we window the Mel-spectrogram representations so that they contain the information equivalent to 5 seconds of the original audio signals using an overlap of 50 %. Each windowed representation is stored as an image of 224×224 pixels for further processing. As the speech samples are imbalanced with respect to the arousal classes (cf. Table 3), we use a weighted random sampler to select the samples to use for training the models at each epoch. With this strategy, the samples corresponding to the less represented classes are used more often for training the models than the samples corresponding to the most represented classes.

2) MODELS DESCRIPTION

To model the different features extracted from the speech files (cf. Section III-B1), we explore different network architectures, which we proceed to describe.

i) **MLP**. The eGeMAPS features are modelled using a *Multi-Layer Perceptron* (MLP) composed of two main blocks. The first block acts as a feature adapter, as it uses a linear layer to convert the original features into a 512-dimensional representation. This 512-dimensional representation is then fed to the classification block, which implements two linear layers with 32 and 3 output

neurons, respectively, preceded by 2 dropout layers with probability 0.3. The outputs of the first linear layer are transformed using a ReLU function, and the outputs of the second linear layer use a Softmax activation function, so the network outputs can be interpreted as probability scores.

ii) **Scratch CNN**. This network exploits the Mel-spectrogram representations of the audio signals using a CNN trained from scratch. This network is composed of two main blocks. The first block extracts deep learnt representations from the input Mel-spectrograms. For this, we implement 3 convolutional layers with 32, 64, and 128 filters each, a kernel size of 3×3 and a stride of 1. After each convolutional layer, we use batch normalisation, and the network outputs are transformed using a ReLU function. The first two layers use a 2-dimensional max-pooling layer with a kernel size of 2×2, while the third layer uses a 2-dimensional adaptive average pooling layer, so the outputs of this feature extraction block produce a 512-dimensional representation of the input data. The second block of the network is responsible for the actual classification and implements the same architecture as the classification block of the MLP network described above.

iii) **Pre-trained CNN**. This network also exploits the Mel-spectrogram representations of the audio signals, but using a pre-trained CNN. This network is also composed of a feature extraction and a classification block. We choose the same architecture for the classification block as in the MLP and the Scratch CNN architectures. The difference, however, lies in the feature extraction block. In this case, we opt for applying a pre-trained Resnet-18 [26] network without the last layer to extract deep learnt representations from the input Mel-spectrograms. We fine-tune the network during the training process. This network produces a 512-dimensional representation of the input data. For this reason, we engineered the previous network architectures to produce a 512-dimensional representation at the output of the feature extraction block. This way, we can fairly compare the performance of the three different architectures proposed.

3) NETWORKS TRAINING

At the initialisation of each network, the pseudo-random number generator is manually seeded for a fair comparison, and reproducibility of the results. We train the models described in Section III-B2 to minimise the Categorical Cross-Entropy loss, using Adam as the optimiser with a learning rate of 10^{-3}. The networks are trained in batches of 128 samples and during a maximum of 200 epochs. We implement an early stopping mechanism to stop training when the validation error does not improve for 20 consecutive epochs. With this early stopping mechanism, we determine the number of epochs needed for training the networks, while minimising the chances of overfitting. We decide for the Unweighted Average Recall (UAR) as the metric to compare the ground truth and the inferred arousal annotations, and, therefore, we define $(1 - \text{UAR})$ as the validation error to monitor the training process.

IV. EXPERIMENTAL RESULTS

This section reports the results obtained from the experiments conducted. Section IV-A compares the performance of the sentiment analysis models that implement the different network architectures described in Section III-A2. Section IV-B analyses how the performance of the arousal recognition models is impacted by choosing different feature representations of the speech signals and different network architectures to analyse the information extracted (cf. Section III-B2).

A. SENTIMENT ANALYSIS MODELS

To assess the performance of our sentiment analysis models, we compute the UAR between the inferred and the ground truth annotations. We consider the UAR as the most suitable metric to use in this case, as it is not impacted by the imbalanced data. Hence, the chance level in terms of UAR for the binary classification problem is 50.00 %.

The performance of the binary sentiment analysis models trained is summarised in Table 4. To contextualise the performance of our models, we apply a state-of-the-art transformer-based binary sentiment analysis model to infer the sentiment corresponding to the sentences belonging to both the development and the test partitions. Specifically, we select the pre-trained, off-the-shelf binary sentiment model available from the *pipeline* API of the Transformers library[2] [27]. This model was trained based on the DistilBERT architecture [28] and fine-tuned on the SST2 dataset [29]. The results obtained with this pre-trained model are included in Table 4.

Comparing the results obtained, we observe that all our models achieve a higher performance than the state-of-the-art transformer-based binary sentiment model on the test set. The highest performance on the development partition is obtained with the baseline RNN, scoring a UAR of 75.84 %. Nevertheless, the H-CA network scores the best performance on the test set, with a UAR of 79.04 %, surpassing the baseline network.

The sentiment analysis component in the affective dialogue manager architecture (cf. Section III) deploys the H-CA network-based sentiment analysis model trained. The sentiment analysis component is implemented through a simple API, which is publicly available with the aim to provide an off-the-shelf sentiment analysis tool to the community[3].

B. AROUSAL RECOGNITION MODELS

The results obtained from the network architectures described in Section III-B2 are reported in Table 5. As it can be observed, the best performance is obtained using the

[2]https://huggingface.co/docs/transformers/main_classes/pipelines
[3]https://github.com/EIHW/sustAGE_SentimentAnalysis

TABLE 4. Summary of the results obtained in terms of UAR (%) when tackling the sentiment analysis task as a binary classification problem. The performance of the models is assessed in both development and test partitions.

Model	Devel. Set	Test Set
Wolf et al. [27]	75.35	74.02
Baseline RNN	75.84	75.28
H-N	74.31	74.99
H-CA	75.33	79.04
CNN	75.24	78.39
CNN-CA	75.49	77.61

TABLE 5. Summary of the results obtained in terms of UAR (%) when tackling the arousal recognition task as a 3-class classification problem. The performance of the models is assessed in both development and test partitions.

Model	Devel. Set	Test Set
MLP	56.62	57.81
Scratch CNN	61.61	61.11
Pre-trained CNN	54.34	57.98

scratch CNN architecture exploiting the Mel-spectrogram representations of the audio files with a UAR of 61.61 % and 61.11 % on the development and the test partitions, respectively. The lowest UAR of 57.81 % on the test set is obtained with the MLP architecture exploiting the functionals of the eGeMAPS feature set, similar to the performance obtained with the pre-trained CNN on the Mel-spectrogram representations, which scores a UAR of 57.98 % on the test set. These results suggest the suitability of exploiting the Mel-spectrogram representations of the audio signals with CNNs trained from scratch for arousal recognition.

The arousal recognition component in the affective dialogue manager architecture (cf. Section III) extracts the Mel-spectrogram representations of the input speech signals, and deploys the scratch CNN-based model trained. The arousal recognition component is implemented through a simple API, which is publicly available with the aim to provide an off-the-shelf arousal recognition tool to the community[4].

V. PROOF OF CONCEPT DEMONSTRATOR

Coupling the sentiment analysis and arousal recognition technologies developed and hosted in their corresponding components (cf. Figure 1), we can power a dialogue manager with affective capabilities. As depicted in Figure 1, the proposed system initialises the dialogue, and, then, the users record their voice via the smartphone app to answer. The content of the answer is open to the user. Upon reception of the recorded file, the back-end of the system runs the generated transcription and the received speech file through the sentiment analysis and the arousal recognition components, respectively. Open dialogue systems are a challenging research topic, out of the scope of this work. Hence, we opt for a rule-based dialogue manager. We employ predefined sentences – containing the body of the system answer – and interjections – to convey different levels of energy in the system response –, which the dialogue manager selects according to the sentiment and arousal information inferred by the corresponding components to determine an affective system response to the users' input.

Dialogue 1 Affective dialogue designed at wake-up time considering the outcomes of the sentiment analysis and the arousal recognition models when analysing the open user's response

← **[SYSTEM]** Good morning, [NICKNAME]! How did you sleep tonight?
→ **[USER'S RESPONSE]**
if *(sentiment == positive)* **then**
  if *(arousal == high)* **then**
    ← **[SYSTEM]** I like the spirit! It is great to start a new day full of energy.
  else if *(arousal == mid) or (arousal == low)* **then**
    ← **[SYSTEM]** Glad to hear this. It is great to start a new day full of energy.
else if *(sentiment == negative)* **then**
  ← **[SYSTEM]** I am sorry to hear this. Try to sleep a bit more if you can, or perform 2-3 times the exercise 4-7-8 (4 sec breath in – 8 sec hold – 7 sec slowly breath out).
else
  ← **[SYSTEM]** I think it is still too early in the morning for you.
end

As a proof of concept demonstrator, we integrate the proposed affective dialogue manager at the back-end of a larger companion system which interacts with its users at specific, relevant points in time during the day and gathers users' and context-related information to determine timely and personalised recommendations that can support wellbeing, wellness, and productivity [5]. For this use case, the affective dialogue manager aims at improving the user experience when interacting with the companion system. For a more natural interaction, the dialogue manager addresses the users using a nickname of their choice, which is adapted to each user on the fly. Considering the deployment scenario, we define two of the short dialogues in which the users are engaged at wake-up time (cf. Dialogue 1), and at the end of the working time (cf. Dialogue 2). We also showcase these dialogue scenarios in a video demonstration[5].

It is worth mentioning that the smartphone app showcased in the video demonstration utters the dialogue manager responses with the *Text-to-Speech* (TTS) functionality provided by Android. The authors agree that emotional TTS is an emerging research field [30], which could potentially be

[4]https://github.com/EIHW/sustAGE_ArousalRecognition

[5]https://youtu.be/7Hq1q146P7Q

Dialogue 2 Affective dialogue designed at the end of the working day considering the outcomes of the sentiment analysis and the arousal recognition models when analysing the open user's response

← **[SYSTEM]** Hi [NICKNAME], your working day is finally over. Do you have some plans for this afternoon?
→ **[USER'S RESPONSE]**
if *(sentiment == positive)* **then**
  if *(arousal == high)* **then**
    ← **[SYSTEM]** It sounds exciting! Safe trip back home and enjoy your afternoon.
  else if *(arousal == mid) or (arousal == low)* **then**
    ← **[SYSTEM]** It is good to hear this. Safe trip back home and enjoy your afternoon.
else if *(sentiment == negative)* **then**
  ← **[SYSTEM]** You deserve doing some activity that makes you happy.
else
  ← **[SYSTEM]** You do not seem to be very talkative. I am sure you are tired after working all day long.
end

applied in the proposed affective dialogue manager for a more natural system response. Nevertheless, as herein we focus on the affective capabilities of a dialogue manager from a user analysis perspective, we consider the synthesis aspect of the affective dialogue manager as future work.

Ethical concerns are inherent to voice-based HCI applications; especially those related to privacy [31]. In our case, users actively press a button on the smartphone interface to start and stop the audio recording. We opted for this approach to gain users' trust, avoiding them having the impression their were continuously recorded. When sending sensible data – such as voice – throughout the Internet, the connection between the smartphone and the system back-end needs to be secured and encrypted; for instance, using the HTTPS protocol. Finally, the raw recordings should be deleted after processing and providing the answer to the users in order not to store personal data and minimise the damage of potential data leaks associated to exposing the back-end system to the public Internet.

VI. CONCLUSION AND FUTURE WORK

In this work, we presented a speech-based affective dialogue manager system powered by sentiment analysis and arousal recognition capabilities to create an instantaneous affective profile of the user, so it can be used to condition and adjust the system response. The research conducted on the sentiment analysis problem focused on analysing the information loss experienced by using the hidden state of a recurrent neural network produced at the last time step as the embedded representation encoding the whole input sentence. The best model implemented a hierarchical contextual attention fusion network, which exploited the hidden states produced during all the time steps of the input sentence as the embedded representations. The research conducted on the arousal recognition problem focused on assessing the suitability of using different feature representations of the speech signals and using different network architectures to exploit the information extracted. The best model extracted the Mel-spectrogram representations of the speech signals and used a CNN trained from scratch to generate deep learnt representations. Overall, the deployed sentiment analysis model was able to infer whether the input sentence conveyed a negative or a positive sentiment, while the deployed arousal recognition model was able to infer whether the speaker conveyed a low, mid, or high level of arousal. Furthermore, we provided a proof of concept demonstrator of the implemented affective dialogue manager and presented two of the dialogues supported at the back-end of the system, exemplifying how the inferred affective information determined the system response.

Future work includes the assessment of the proposed affective dialogue manager with real users. To evaluate the effectiveness of the proposed solution, it would be relevant to also compare it with existing dialogue managers. Further investigations could consider exploring more advanced neural network architectures to improve the performance of the models trained. Motivated by the recent trends in the field of AI, a throughout analysis of Transformer-based architectures for sentiment analysis could be conducted with the aim to understand and determine which architecture works best and why. Furthermore, the study of LLMs in this problem is also a promising research direction, based on the excellent performance of such models in a wide range of problems and applications. Additionally, powering the dialogue manager system with natural language understanding capabilities and emotional TTS to utter the system responses would be encouraging directions to support open dialogues between the users and the system and to increase the affective perception of the system by the users, respectively.

REFERENCES

[1] I. Leite, A. Pereira, S. Mascarenhas, C. Martinho, R. Prada, and A. Paiva, "The influence of empathy in human–robot relations," *Int. J. Hum.-Comput. Stud.*, vol. 71, no. 3, pp. 250–260, Mar. 2013.

[2] B. De Carolis, S. Ferilli, and G. Palestra, "Simulating empathic behavior in a social assistive robot," *Multimedia Tools Appl.*, vol. 76, no. 4, pp. 5073–5094, Feb. 2017.

[3] J. M. Olaso et al., "The EMPATHIC virtual coach: A demo," in *Proc. 23rd Int. Conf. Multimodal Interact.*, Montréal, QC, Canada, 2021, pp. 848–851.

[4] H. Abdollahi, M. Mahoor, R. Zandie, J. Sewierski, and S. Qualls, "Artificial emotional intelligence in socially assistive robots for older adults: A pilot study," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2020–2032, Jul./Sep. 2023.

[5] A. Mallol, I. Varlamis, M. Pateraki, M. Lourakis, G. Athanassiou, M. Maniadakis, K. Papoutsakis, T. Papadopoulos, A. Semertzidou, N. Cummins, B. Schuller, I. Karolos, C. Pikridas, P. Patias, S. Vantolas, L. Kallipolitis, F. Werner, A. Ascolese, and V. Nitti, "sustAGE 1.0—First prototype, use cases, and usability evaluation," in *Proc. Human Interact. Emerg. Technol. (IHIET-AI), Artif. Intell. Future Appl.* Lausanne, Switzerland: Springer, 2022, doi: 10.54941/ahfe100895.

[6] A. Zadeh, P. P. Liang, J. Vanbriesen, S. Poria, E. Tong, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.
[7] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, Oct. 2019.
[8] K. Lavanya and C. Deisy, "Twitter sentiment analysis using multi-class SVM," in *Proc. Int. Conf. Intell. Comput. Control (IC)*, Coimbatore, India, Jun. 2017, pp. 1–6.
[9] K. Lu and J. Wu, "Sentiment analysis of film review texts based on sentiment dictionary and SVM," in *Proc. 3rd Int. Conf. Innov. Artif. Intell.*, Suzhou, China, 2019, pp. 73–77.
[10] Z. Jianqiang, G. Xiaolin, and Z. Xuejun, "Deep convolution neural networks for Twitter sentiment analysis," *IEEE Access*, vol. 6, pp. 23253–23260, 2018.
[11] T. Zhang, X. Gong, and C. L. P. Chen, "BMT-Net: Broad multitask transformer network for sentiment analysis," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 6232–6243, Jul. 2022.
[12] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022.
[13] X. Deng, V. Bashlovkina, F. Han, S. Baumgartner, and M. Bendersky, "LLMs to the moon? Reddit market sentiment analysis with large language models," in *Proc. Companion ACM Web Conf.*, Austin, TX, USA, Apr. 2023, pp. 1014–1019.
[14] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech*, Lyon, France, Aug. 2013, pp. 148–152.
[15] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Hong Kong, 2003, doi: 10.1109/ICASSP.2003.1202279.
[16] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5200–5204.
[17] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proc. 32nd Conf. Artif. Intell.*, New Orleans, LA, USA, 2018, pp. 1–8.
[18] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, Doha, Qatar, 2014, pp. 1532–1543.
[19] J. Mingyu, Z. Jiawei, and W. Ning, "AFR-BERT: Attention-based mechanism feature relevance fusion multimodal sentiment analysis model," *PLoS ONE*, vol. 17, no. 9, Sep. 2022, Art. no. e0273936.
[20] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Synth. Emotions*, vol. 1, no. 1, pp. 68–99, Jan. 2010.
[21] A. Triantafyllopoulos and B. W. Schuller, "The role of task and acoustic similarity in audio transfer learning: Insights from the speech emotion recognition case," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, ON, ON, Canada, Jun. 2021, pp. 7268–7272.
[22] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2016, pp. 1480–1489.
[23] A. Mallol-Ragolta, Z. Zhao, L. Stappen, N. Cummins, and B. Schuller, "A hierarchical attention network-based approach for depression detection from transcribed clinical interviews," in *Proc. Interspeech*, 2019, pp. 221–225.
[24] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr. 2016.
[25] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, Firenze, Italy, Oct. 2010, pp. 1459–1462.
[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, vol. 16, Las Vegas, NV, USA, 2016, pp. 770–778.
[27] J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proc. EMNLP*, 2020, pp. 38–45.
[28] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," in *Proc. 5th Workshop Energy Efficient Mach. Learn. Cognit. Comput., Co-Located, 33rd Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2019, p. 5. [Online]. Available: https://www.emc2-ai.org/assets/docs/neurips-19/emc2-neurips19-paper-33.pdf
[29] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng, "Parsing with compositional vector grammars," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*. Sofia, Bulgaria, 2013, pp. 455–465.
[30] C.-B. Im, S.-H. Lee, S.-B. Kim, and S.-W. Lee, "EMOQ-TTS: Emotion intensity quantization for fine-grained controllable emotional text-to-speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 6317–6321.
[31] W. Seymour, X. Zhan, M. Cote, and J. Such, "A systematic review of ethical concerns with voice assistants," in *Proc. 6th Conf. AI, Ethics, Soc.*, Montreal, QC, Canada, 2023, pp. 131–145.

ADRIA MALLOL-RAGOLTA received the B.Sc. degree in audiovisual telecommunication systems engineering from Universitat Pompeu Fabra, Barcelona, Spain, in 2016, and the M.Sc. degree in electrical engineering from the University of Colorado, Colorado Springs, USA, in 2018. He is currently pursuing the Ph.D. degree with the Chair of Health Informatics, Technical University of Munich, Munich, Germany. He has coauthored more than 30 publications in peer-reviewed books, journals, and conference proceedings leading to more than 600 citations (H-index = 12).

BJÖRN SCHULLER (Fellow, IEEE) received the Ph.D. and Habilitation degrees in electrical engineering and information technology from the Technical University of Munich (TUM), Munich, Germany, in 2006 and 2012, respectively. He is currently a Full Professor and the Head of the Chair of Health Informatics, TUM, and a Professor in artificial intelligence with the Department of Computing, Imperial College London, U.K. He has coauthored seven books and more than 1 000 publications in peer-reviewed books, journals, and conference proceedings leading to more than 61 200 citations (H-index = 109).