

# A Self-Attentive Emotion Recognition Network

Harris Partaourides\*, Kostantinos Papadamou\*, Nicolas Kourtellis<sup>+</sup>, Ilias Leontiadis<sup>‡</sup>\*, Sotirios Chatzis\*

\*Cyprus University of Technology, <sup>+</sup>Telefonica Research, <sup>‡</sup>Samsung Research  
c.partaourides@cut.ac.cy, ck.papadamou@edu.cut.ac.cy, nicolas.kourtellis@telefonica.com,  
i.leontiadis@samsung.com, sotirios.chatzis@cut.ac.cy

## Abstract

Modern deep learning approaches have achieved groundbreaking performance in modeling and classifying sequential data. Specifically, attention networks constitute the state-of-the-art paradigm for capturing long temporal dynamics. This paper examines the efficacy of this paradigm in the challenging task of emotion recognition in dyadic conversations. In contrast to existing approaches, our work introduces a novel attention mechanism capable of inferring the immensity of the effect of each past utterance on the current speaker emotional state. The proposed attention mechanism performs this inference procedure without the need of a decoder network; this is achieved by means of innovative self-attention arguments. Our self-attention networks capture the correlation patterns among consecutive encoder network states, thus allowing to robustly and effectively model temporal dynamics over arbitrary long temporal horizons. Thus, we enable capturing strong affective patterns over the course of long discussions. We exhibit the effectiveness of our approach considering the challenging IEMO-CAP benchmark. As we show, our devised methodology outperforms state-of-the-art alternatives and commonly used approaches, giving rise to promising new research directions in the context of Online Social Network (OSN) analysis tasks.

## 1 Introduction

Affective computing is an interdisciplinary research field that aims at bridging the gap between human and machine interactions. To that end, researchers utilize sentiment analysis [17, 22, 29] and emotion recognition [3, 13, 25] algorithms to develop systems that can recognize emotions to properly drive their responses. In this context, the accuracy of emotion recognition is crucial for the success of affective computing solutions. Therefore, it is needed that the machine learning community develops increasing complex models, far and beyond the models used in the related but simpler task of sentiment analysis. Traditionally, to successfully recognize emotions, researchers have to utilize a variety of modalities such as speech, facial expression, body gestures and physiological indices [11, 14, 27]. This combination of distinct modalities ensures algorithm effectiveness.

This work is motivated from the important challenge of online emotion recognition from textual dialog data (online chats).

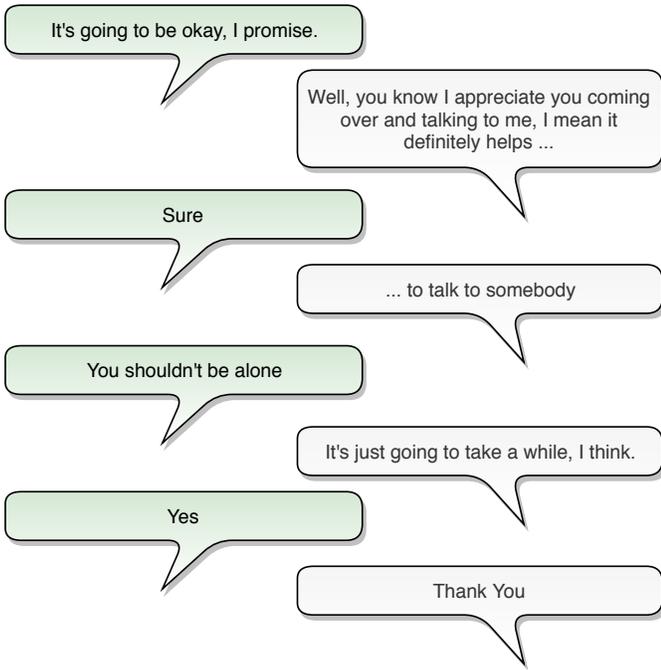
This is a problem of increasing immensity due to the emergence and strong popularity of Online Social Networks (OSNs). Unfortunately, existing algorithms that address this problem suffer from two major shortcomings: 1) they cannot capture temporal affective patterns over long dialogs; this results in missing crucial information that may have appeared many utterances before, but has salient effect on the current emotional state of the speakers, 2) as they have access to only one type of modality, namely text, these algorithms cannot achieve high recognition performance since this typically requires combination of multiple modalities.

In the affective computing literature, we observe a plethora of machine learning algorithms used for emotion recognition, such as linear discriminant classifiers, k-nearest neighbor, decision tree algorithms, support vector machines [4, 32] and artificial neural networks [34]. More recently, the research community has shown that machine learning models with the capacity to capture contextual information are capable of achieving much higher performance, as is well expected due to the nature of dialog data [2, 6, 21, 28]. Indeed, using contextual information to perform emotion recognition is similar to the actual process people use to infer the emotional state of their interlocutor. This becomes apparent in cases where the latest utterance is insufficient in inferring emotions. In such cases, people consider how the conversation evolved to acquire the missing information. Hence, it is indisputable that, more often than not, we need contextual information to accurately predict emotional states from dialog text.

**Example 1.** In the conversation shown in Fig. 1, the last message (“Thank you”) does not imply anything about the speaker’s emotional state. However, by analyzing the conversation up to that point, one can clearly infer the underlying emotional state (sadness).

To address these shortcomings, researchers have relied on models that can capture temporal dynamics, including hidden Markov models (HMMs) and recurrent neural networks (RNNs). These context-aware models have yielded major improvements compared to their context-unaware counterparts [20, 30]. However, both HMMs and RNNs suffer from a major limitation that undermines the effectiveness of emotion recognition in the context of OSNs dialogs: they both are model families that can capture temporal dependencies over short horizons. This implies a clear inability to retain salient information that may affect emotion over a long horizon, spanning the whole

\*<sup>‡</sup>Work done while at Telefonica Research



**Figure 1:** A typical dialog segment (extracted from the IEMOCAP dataset [5])

course of an OSN dialog.

Recently, the machine learning community has attempted to achieve a breakthrough in the performance of emotion recognition systems by relying on neural attention mechanisms [9]. These mechanisms build upon the short-term memory capabilities of RNNs to enable the creation of strong machine learning pipelines that can capture salient temporal dynamics over long horizons. However, their efficacy has been examined only in the context of modeling fused distinct modalities, including speech and facial gestures, in addition to text. Besides, existing works [12, 33] have relied on single-layer Bidirectional RNNs (BiRNNs) [24] for encoding context-level dialog dynamics; a fact that requires a-priori provision of the whole dialog to perform inference. This is clearly limiting, as it renders rather prohibitive the real-time analysis of OSN activity.

This paper offers a coherent solution that addresses the aforementioned limitations of the existing neural attention paradigm in the context of online emotion recognition from OSN dialog text. For the first time in the literature, we introduce a *self-attentive hierarchical encoder* network that is capable of extracting salient information on both the individual utterance level as well as the level of the dialog context, as it has evolved until any given time point. Specifically, our proposed model comprises a hierarchical encoding mechanism that performs representation extraction on two levels: The first employs a Bidirectional Long Short Term Memory (LSTM) [10] that captures word-level contextual information in each individual utterance. The second utilizes a GRU [8] that performs dialog context-level representation to allow for capturing the salient dynamics over the whole dialog span.

The formulated hierarchical encoder is complemented with a

novel self-attention (SA) mechanism. This is carefully designed to generate accurate inferences of how strongly the encoder-obtained representations of the latest utterance at the final layer (*dialog state*) correlate with the corresponding representations pertaining to previous utterances. As these representations constitute RNN states, which inherently encode short-term dynamics, the so-obtained correlation information allows for establishing a notion of attention among the current and the previous *dialog states*. Therefore, this self-attention information can be leveraged to yield meaningful weights for effectively combining the whole history of dialog states into a highly informative dialog context vector; we eventually use the resulting self-attentive context vector to drive an accurate penultimate emotion recognition layer of high accuracy. We emphasize that our use of a simple GRU at the second level of the encoder, as apposed to a bidirectional one, allows for performing emotion inference without requiring a-priori provision of the whole dialog, that is in an online fashion. We dub our proposed approach the Self-attentive Emotion Recognition Network (SERN).

We experimentally evaluate our approach on the IEMOCAP dataset [5] and empirically demonstrate the significance of the introduced self-attention mechanism. Subsequently, we perform an ablation study to demonstrate the robustness of the proposed model. We empirically show an important enhancement of the attainable speaker emotional state inference capabilities. This is of vital importance for OSNs, since they are increasingly associated with distress and negative implications on users' mental health [7].

The remainder of this paper is organized as follows. Section 2 provides a concise review of the methodological background of our approach. In Section 3, we introduce the proposed approach and elaborate on its training and inference algorithms. Section 4 constitutes an in-depth experimental evaluation of the proposed method using a popular benchmark dataset. Finally, in Section 5 we summarize our contribution and conclude this paper by discussing directions for future research.

## 2 Methodological Background

### 2.1 Word Representations

In order for machine learning algorithms to perform analysis of word data, it is needed that the observed words are transformed into a vectorial representation; these are typically referred to as word embeddings in the related literature. Word2Vec [18] is a popular embedding technique based on deep learning principles. It aims at yielding embedding spaces of low dimensionality and high representational power. This is achieved by postulating a one-hidden-layer softmax classifier which is presented with sentence fragments of fixed length, and is trained to predict the next word in the corresponding sentences. In a different vein, GloVe [19] is an unsupervised algorithm for obtaining vector representations of words. Its main principle consists in capturing word-word concurrences based on the frequency that dictionary words co-occur in the available training corpus. In this work, we rely on the Word2Vec scheme; however, we elect to train the representations from

scratch, using our available datasets, as opposed to resorting to the pretrained Word2Vec embeddings.

## 2.2 Recurrent Neural Networks

A recurrent neural network (RNN) is a neural network with the capacity to model the temporal dependencies in sequential data. Given an input sequence  $x = (\mathbf{x}_1, \dots, \mathbf{x}_T)$ , an RNN computes the hidden sequence  $h = (\mathbf{h}_1, \dots, \mathbf{h}_T)$ ; its hidden vector  $\mathbf{h}_t$  constitutes a concise representation of the temporal dynamics in a short-term horizon prior to time  $t$ . At each time step  $t$ , the hidden state  $\mathbf{h}_t$  of the RNN is updated by  $\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{x}_t)$  where  $f$  is a non-linear activation function. Given the state sequence  $h$ , the network eventually computes an output sequence  $y = (\mathbf{y}_1, \dots, \mathbf{y}_T)$ . A significant advantage of RNNs is the fact that they impose no limitations on input sequence length. However, practical application has shown that RNNs have difficulties in modeling long sequences. Specifically, RNNs are notorious for the exploding and vanishing gradients problem, which renders model training completely intractable for applications that entail long sequences.

To resolve these issues, two popular RNN variants are usually employed, namely the GRU [8] and the LSTM [10] network. The hidden state,  $\mathbf{h}_t$ , of a GRU network at time  $t$  is given by:

$$\mathbf{z}_t = \sigma(W_z \mathbf{x}_t + U_z \mathbf{h}_{t-1} + \mathbf{b}_z) \quad (1)$$

$$\mathbf{r}_t = \sigma(W_r \mathbf{x}_t + U_r \mathbf{h}_{t-1} + \mathbf{b}_r) \quad (2)$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \circ \mathbf{h}_{t-1} + \mathbf{z}_t \circ \tanh(W_h \mathbf{x}_t + U_h(\mathbf{r}_t \circ \mathbf{h}_{t-1}) + \mathbf{b}_h) \quad (3)$$

On the other hand the hidden state,  $\mathbf{h}_t$ , of an LSTM network at time  $t$  is given by:

$$\mathbf{f}_t = \sigma(W_f \mathbf{x}_t + U_f \mathbf{h}_{t-1} + \mathbf{b}_f) \quad (4)$$

$$\mathbf{i}_t = \sigma(W_i \mathbf{x}_t + U_i \mathbf{h}_{t-1} + \mathbf{b}_i) \quad (5)$$

$$\mathbf{o}_t = \sigma(W_o \mathbf{x}_t + U_o \mathbf{h}_{t-1} + \mathbf{b}_o) \quad (6)$$

$$\mathbf{c}_t = \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \tanh(W_c \mathbf{x}_t + U_c \mathbf{h}_{t-1} + \mathbf{b}_c) \quad (7)$$

$$\mathbf{h}_t = \mathbf{o}_t \circ \tanh(\mathbf{c}_t) \quad (8)$$

In these equations, the  $W$ ,  $U$ ,  $\mathbf{b}$  are the trainable parameters and  $\sigma$  is the logistic sigmoid function.

Finally, bidirectional formulations of RNNs have great use in natural language processing tasks. Specifically, when dealing with understanding of whole sentences, it is intuitive to jointly model the temporal dynamics that stem from reading the sentence both in a forward and a backward fashion. Indeed, this may facilitate a more complete extraction of syntactic context, which is crucial for language understanding. In essence, bidirectional RNN variants comprise two distinct RNNs, one presented with the observed sequence, and one presented with its reverse. At each time point, the state vectors  $\mathbf{h}_t$  of the two component RNNs are concatenated and presented as such to the penultimate layer of the network.

## 2.3 Neural Attention

Neural attention has been a major breakthrough in Deep Learning for Natural Language Processing, as it enables capturing long temporal dynamics that elude the capacity of RNN variants. Among the large collection of recently devised neural attention mechanisms, the vast majority build upon the concept of soft attention [31]. Given a sequence of hidden states  $\mathbf{h}_t$  ( $t = 1, \dots, T$ ), the attention mechanism computes the context vectors,  $\mathbf{c}_t$ ; these are weighted sums of the available hidden states and are given by:

$$a_s^t = \text{softmax}(\text{score}(\mathbf{h}_t, \mathbf{h}_s)), t \neq s \quad (9)$$

$$\mathbf{c}_t = \sum_s a_s^t \mathbf{h}_s \quad (10)$$

In this expression, typical options for the score function are:

$$\text{score}(\mathbf{h}_t, \mathbf{h}_s) = \begin{cases} \mathbf{h}_t^T \mathbf{h}_s \\ \mathbf{h}_t^T W_a \mathbf{h}_s \\ \mathbf{u}_a^T \tanh(W_a[\mathbf{h}_t; \mathbf{h}_s]) \end{cases} \quad (11)$$

where the  $W_a$  and  $\mathbf{u}_a$  are trainable parameters. In cases where we are dealing with a model generating whole sequences of different length from the input sequence, the  $\mathbf{h}_s$  is the current hidden state of the sequence-generating model component, also known as the decoder. On the other hand, when dealing with frame-level classification tasks where the penultimate network layer is a softmax classifier, as opposed to a decoder, the  $\mathbf{h}_s$  can be the current state of the employed RNN, yielding

$$a_s^t = \text{softmax}(\text{score}(\mathbf{h}_t, \mathbf{h}_s)), t \leq s \quad (12)$$

$$\mathbf{c}_s = \sum_{t \leq s} a_s^t \mathbf{h}_t \quad (13)$$

## 3 Proposed Approach

As already discussed, the ultimate goal of this work is to enable accurate emotion recognition in online text chats. This gives rise to the challenging task of performing natural language understanding at both the utterance level and the dialog context up to the current utterance. This is a problem of immense complexity, since it requires the capacity to perform valid inference at the utterance level and effectively correlate the obtained information over arbitrarily long dialog durations.

To this end, we introduce a novel hierarchical encoder network that is capable of extracting salient information on the individual utterance level, and inferring potent temporal dynamics across the dialog duration. The latter capacity is enabled by appropriately implementing the concept of self-attention as an intrinsic part of our novel architecture.

Let us consider a dialog  $X = \{\mathbf{X}_1, \dots, \mathbf{X}_s\}$ , where the  $\mathbf{X}_s$  represents the  $s$ -th utterance of the dialog, and  $\mathbf{X}_s = \{X_{s1}, \dots, X_{sw_s}\}$ , where  $w_s$  represents the word count in the  $s$ -th utterance. We seek a model capable of correctly recognizing the dominant emotion related to each utterance, summarized into the vector  $\mathbf{e} \in E^S$ , where  $E$  is the considered addressed set

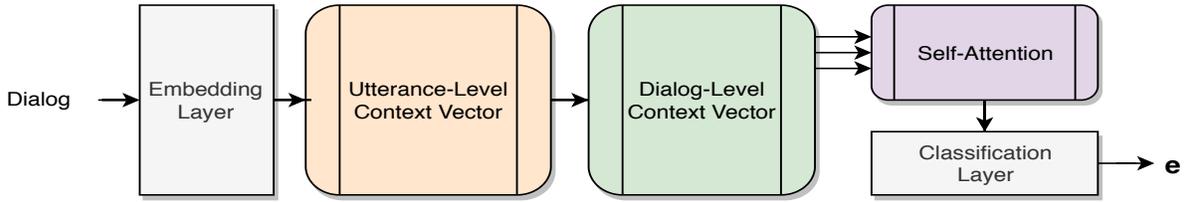


Figure 2: SERN model configuration.

of emotions. In the following, we consider the six main emotion categories, namely angry, happy, sad, neutral, excited and frustrated; these are the emotions with an adequate amount of examples in the IEMOCAP dataset. A descriptive illustration of the proposed model configuration is provided in Fig 2.

Based on this motivation, our proposed approach comprises three consecutive core parts: Initially, a trainable *Word2Vec embeddings* mechanism is presented with the input sequence. Let us denote as  $\mathbf{m}_{st}$  the word embedding pertaining to the  $t$ -th token of the  $s$ -th utterance. Subsequently, a *bidirectional LSTM* (BiLSTM) is used to capture the salient linguistic information contained within each utterance. We use a bidirectional LSTM to ensure optimal inference of syntactic structure at the utterance level, as typical in the literature. Let  $\mathbf{f}_s^{utt}$  be the final state vector of the employed *utterance-level BiLSTM*, presented with the  $s$ -th utterance. This constitutes the latent vector representation fed to the *subsequent dialog-level GRU* network. Specifically, this network uses the BiLSTM-obtained latent vector representations of the preceding utterances to infer salient temporal dynamics at the dialog-context level, useful for driving a penultimate emotion classification layer.

Let us denote a running dialog comprising  $s$  utterances. The postulated GRU network presented with the utterance-level representations  $\{\mathbf{f}_{s'}^{utt}\}_{s'=1}^s$  has generated a set of state vectors  $\{\mathbf{f}_{s'}^{dial}\}_{s'=1}^s$  representing dialog-level semantic information. This could be used to drive a penultimate dialog context-informed emotion classification layer. However, as already discussed, RNN variants are only capable of capturing temporal dependencies over short-term horizons, with exponentially-decreasing temporal effect. As real-world dialogs may be quite long and entail a gradual temporal evolution that spans long time frames, it is imperative that we endow the proposed model with the capacity to capture such long temporal dynamics.

To this end, we deploy a *self-attention layer* on top of the dialog-level GRU network. As discussed in Section 2.3, the postulated self-attention mechanism computes, for the current utterance  $s$ , the affinity of its dialog-level representation,  $\mathbf{f}_s^{dial}$ , with the representations pertaining to the previous utterances. On this basis, it computes an affinity-weighted average of these representations, as described in Eqs. 12-13, which eventually drives emotion classification. We refer to this weighted average as the dialog context vector at step  $s$ ,  $\mathbf{c}_s$ .

We train the devised model in an end-to-end fashion. The employed training objective function is the categorical cross-entropy of the model; this is a natural selection, as we are dealing with a frame-level classification problem. Specifically, we resort to stochastic gradient descent to obtain parameter estima-

Class	Angry	Sad	Happy	Frustrated	Excited	Neutral
Utterances	1,103	1,084	595	1,849	1,041	1,708

Table 1: IEMOCAP dataset: Number of utterances on each emotion.

tors, employing the Adam optimizer [15].

## 4 Experimental Evaluation

In this section, we perform a thorough experimental evaluation of our proposed model. We provide a quantitative assessment of the efficacy and effectiveness of SERN, combined with deep qualitative insights pertaining to the functionality of the self-attention scheme. Furthermore, we perform an ablation study to better illustrate the robustness of our approach. To this end, we utilize a well-known benchmark for emotion recognition, namely the IEMOCAP dataset [5]. We have implemented our model in TensorFlow [1]. The code of our implementation can be found on GitHub<sup>1</sup>.

### 4.1 IEMOCAP Dataset

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database has been collected by emulating conversations in a controlled environment in order to study expressive human behaviors. The conversations have been performed by ten unique speakers over five dyadic sessions in both a scripted but also an improvisation manner, with various audio-visual modalities being recorded. Each utterance in the dataset is labeled by three human annotators using categorical labels; these include angry, sad, happy, frustrated, excited, neutral, as well as other categories which we omit in this study. The available annotation has been performed by three annotators who assess the emotional states of the speakers taking into consideration dialog context. Thus, this dataset requires that we employ a model capable of inferring potent dialog-level contextual dynamics, as is the case with the proposed approach.

In our experiments, we only utilize the textual modality (transcripts) of the dataset and the categorical labels of each utterance. The used label information is derived by performing majority voting on the three available annotations. This dataset contains 151 conversations with a total number of 10,039 utterances. However, only 7,380 utterances contain the six types of emotions we retain in this study; thus, the remaining utterances are omitted. Table 1 provides a breakdown of the resulting dataset.

<sup>1</sup><https://github.com/Partaourides/SERN>

Class	Angry	Sad	Happy	Frustrated	Excited	Neutral	Dialogs
<b>Train+Validation</b>	933	839	452	1,468	742	1,324	120
<b>Test</b>	170	245	143	381	299	384	31

**Table 2:** Model training: Dataset split.

	Accuracy	Precision	Recall	F1 Score
SVM	0.313 ( $\pm 0.00$ )	0.484 ( $\pm 0.00$ )	0.235 ( $\pm 0.00$ )	0.316 ( $\pm 0.00$ )
<i>BiLSTM</i>	0.477 ( $\pm 0.01$ )	0.471 ( $\pm 0.02$ )	0.459 ( $\pm 0.01$ )	0.465 ( $\pm 0.01$ )
<i>BiLSTM<sub>att</sub></i>	0.516 ( $\pm 0.02$ )	0.516 ( $\pm 0.02$ )	0.501 ( $\pm 0.02$ )	0.509 ( $\pm 0.02$ )
SERN	0.522 ( $\pm 0.02$ )	0.544 ( $\pm 0.02$ )	0.517 ( $\pm 0.02$ )	0.530 ( $\pm 0.02$ )

**Table 3:** Performance metrics.

Our data pre-processing regimen consists in word-based sentence segmentation and removing words with low frequency ( $frequency < 5$ ); to this end, we use the NLTK<sup>2</sup> library. Similar with [20], we split the dataset into a training and test set by leaving out the fifth dyadic session. To perform hyperparameter tuning, we hold out a small representative subset of the training set ( $\sim 7\%$ ) to form a validation set. Table 2 summarizes the details of this split.

## 4.2 Quantitative Study

To exhibit the effectiveness of our approach, we compare its performance with the following baselines:

- **Support Vector Machine (SVM):** A simple classifier that does not consider utterance or dialog context information.
- **Bidirectional LSTM (*BiLSTM*):** A neural network with the capacity to capture only the utterance-level contextual information.
- **Bidirectional LSTM with self-attention (*BiLSTM<sub>att</sub>*):** A **single-layer *BiLSTM*** endowed with an additional self-attention mechanism, similar to Eqs. 12-13.

In all cases, we perform stochastic gradient descent by means of the Adam algorithm with an initial learning rate of  $5E^{-3}$ , and epsilon,  $1E^{-8}$ . Hyper-parameter tuning for the SVM is performed under the grid search strategy. We train each model twenty times, with different initializations each time, and calculate the mean and standard deviation of the obtained accuracy, precision, recall and F1 scores. We present our results in Table 3. For exhibition purposes, Table 4 depicts the confusion matrix obtained from a randomly picked experiment repetition, combined with the corresponding precision and recall metrics.

As we observe, SERN yields notable performance improvements over the alternatives in all performance metrics. We emphasize that application of the Student’s t-test corroborates the statistical significance of the observed differences among the alternatives. Specifically, the p-values obtained on all performance metrics is below the 0.05 threshold. The confusion matrix of Table 4 depicts the number of accurate and misclassified predictions; this illustrates the difficulties in predicting the actual emotional state of the speaker. Apparently, the most promi-

	Angry	Excited	Frustrated	Happy	Neutral	Sad	Recall
<b>Angry</b>	110	2	29	0	22	7	0.647
<b>Excited</b>	9	156	8	74	27	25	0.522
<b>Frustrated</b>	71	6	193	1	87	23	0.507
<b>Happy</b>	14	19	0	80	29	1	0.559
<b>Neutral</b>	35	34	83	11	197	24	0.513
<b>Sad</b>	9	12	42	7	11	164	0.669
<b>Precision</b>	0.444	0.681	0.544	0.462	0.528	0.672	

**Table 4:** A randomly-picked confusion matrix and the corresponding precision and recall metrics. The confusion matrix rows and columns depict the ground-truth and the predicted emotions, respectively.

Utterance at $t$	$t-10$	$t-9$	$t-8$	$t-7$	$t-6$	$t-5$	$t-4$	$t-3$	$t-2$	$t-1$	$t$	Emotion
Oh, you infuriate me ...	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	Angry
Yeah, well I ignore ...	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.06	0.20	0.70	Frustrated
And she-she's Larry's ...	0.00	0.00	0.00	0.02	0.00	0.04	0.24	0.01	0.00	0.00	0.69	Frustrated
Well, from your father's ...	0.00	0.00	0.01	0.00	0.00	0.16	0.05	0.02	0.00	0.07	0.68	Frustrated
Cause listen, I'm telling ...	0.01	0.07	0.00	0.03	0.15	0.02	0.00	0.01	0.20	0.28	0.21	Frustrated
What do you want from ...	0.07	0.01	0.00	0.05	0.09	0.13	0.03	0.02	0.22	0.08	0.30	Frustrated
Every time I reach out ...	0.00	0.02	0.12	0.00	0.00	0.00	0.11	0.13	0.07	0.02	0.54	Frustrated
You're a considerate ...	0.02	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.92	Neutral
To hell with that.	0.01	0.10	0.69	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.15	Angry

**Table 5:** Example A. Self-attention weights over windows spanning the latest nine utterances.  $t$  denotes the current timestep in the dialog.

Utterance at $t$	$t-10$	$t-9$	$t-8$	$t-7$	$t-6$	$t-5$	$t-4$	$t-3$	$t-2$	$t-1$	$t$	Emotion
Then she's gone.	0.02	0.45	0.02	0.02	0.03	0.32	0.02	0.08	0.00	0.00	0.04	Sad
It's going to be ...	0.29	0.00	0.00	0.01	0.39	0.00	0.02	0.00	0.00	0.00	0.28	Sad
Well, you know ...	0.04	0.03	0.04	0.32	0.03	0.10	0.01	0.01	0.03	0.33	0.07	Sad
Sure	0.05	0.08	0.21	0.08	0.11	0.07	0.01	0.07	0.11	0.09	0.12	Sad
to talk to somebody ...	0.01	0.51	0.00	0.03	0.00	0.00	0.00	0.30	0.00	0.00	0.14	Sad
you shouldn't be ...	0.32	0.00	0.01	0.00	0.00	0.00	0.10	0.00	0.00	0.02	0.55	Sad
It's just going to ...	0.00	0.01	0.00	0.00	0.00	0.06	0.00	0.00	0.01	0.37	0.54	Sad
Yes.	0.02	0.02	0.00	0.01	0.05	0.02	0.01	0.04	0.22	0.43	0.17	Sad
Ah.	0.00	0.00	0.00	0.13	0.00	0.00	0.04	0.41	0.33	0.00	0.08	Sad
Thank you.	0.00	0.01	0.06	0.03	0.02	0.04	0.16	0.29	0.18	0.09	0.10	Sad

**Table 6:** Example B. Self-attention weights over windows spanning the latest ten utterances.  $t$  denotes the current timestep in the dialog.

nent difficulties arise between the (neutral, frustrated) and (excited, happy) pairs.

## 4.3 Qualitative Study

Here, we exhibit the role of the self-attention mechanism in enhancing the obtained emotion recognition accuracy. To this end, we train a variant of SERN whereby the context vectors are computed over time-windows spanning the latest ten utterances, as opposed to whole course of the dialog. In Tables 5 and 6, we present the so-obtained self-attention weights over two representative dialog segments. It becomes apparent that utterances many steps in the past may play a crucial role in describing the current emotional state; this salient information would have been missed had it not been for the employed self-attention mechanism.

For instance, the last message of the dialog segment shown in Table 6, "Thank you," was uttered by a sad individual; this emotion can only be inferred through the utterances "you shouldn't be alone" and "It's just going to take a while, I think." Alternatively, the last message of the dialog segment in Table 5, "To hell with that," was uttered by an angry individual; this can be

<sup>2</sup><https://www.nltk.org/>

Classes	Accuracy	Precision	Recall	F1 Score
4	0.689 ( $\pm 0.03$ )	0.685 ( $\pm 0.02$ )	0.699 ( $\pm 0.02$ )	0.692 ( $\pm 0.02$ )
5	0.583 ( $\pm 0.02$ )	0.589 ( $\pm 0.02$ )	0.569 ( $\pm 0.02$ )	0.579 ( $\pm 0.02$ )
6	0.522 ( $\pm 0.02$ )	0.544 ( $\pm 0.02$ )	0.517 ( $\pm 0.02$ )	0.530 ( $\pm 0.02$ )

**Table 7:** Performance metrics of *SERN* trained on four, five and six emotions.

Classes	Angry	Excited	Frustrated	Happy	Neutral	Sad	Happy+Excited
4	0.617	-	-	-	0.720	0.667	0.847
5	0.649	0.767	-	0.317	0.685	0.635	-
6	0.444	0.681	0.544	0.462	0.528	0.672	-

**Table 8:** Classification precision breakdown for four, five and six detected emotions.

traced back to his/her emotional state during the first sentence, "Oh, you infuriate me sometimes. You know, it's not just my business if dad throws a fit." Even more crucially, we observe that the last utterance is not often the principal component that drives the emotional state of the speaker. This can be observed at column  $t$  of Tables 5 and 6 which contains self-attention weights which are less than the values at previous columns.

#### 4.4 Ablation Study

To further assess the robustness of our model, we consider a different mixture of recognized emotions. First, we train and test it with five emotions (angry, happiness, sad, excited, neutral), hence ignoring the examples of the "frustrated" class; this is similar to [5]. Then, we also train and test our model with four emotions (angry, happiness, sad, neutral), by merging the "excitement" and "happiness" categories to a single "happiness" category, similar to [16, 23]. In Table 7, we present the obtained performance metrics, while in Table 8 we offer a breakdown for each emotion. To provide deeper insights, Tables 9 and 10 depict the confusion matrices obtained on a randomly-picked experiment repetition, combined with the corresponding precision and recall metrics. We clearly observe that our method retains its robustness in these alternative settings. Interestingly, the angry and neutral emotions become easier to discern when we omit the frustrated class (five classes scenario); a similar improvement is obtained when we combine the happy and excited emotions (four classes scenario). We posit that this improvement stems from the high level of ambiguity between emotional types, even among human annotators.

Finally, we examine the effect of the number of previous utterances used to compute the inferred context vectors. To this end, we repeat our experiments with the *SERN* model using only the five, ten, twenty and forty latest utterances to compute the context vectors, and compare to the outcome of the full-fledged model. Our obtained results are depicted in Table 11; it is obvious that using a window twenty steps long yields the best performance.

## 5 Conclusions

Accurate emotion recognition is a significant challenge for the developers and administrators of modern OSNs. Indeed, the im-

	Angry	Excited	Happy	Neutral	Sad	Recall
<b>Angry</b>	122	0	0	35	13	0.718
<b>Excited</b>	7	102	121	36	33	0.341
<b>Happy</b>	10	18	71	27	17	0.497
<b>Neutral</b>	34	8	21	278	43	0.724
<b>Sad</b>	15	5	11	30	184	0.751
<b>Precision</b>	0.649	0.767	0.317	0.685	0.635	

**Table 9:** Five classes scenario: A randomly-picked confusion matrix and the corresponding precision and recall metrics.

	Angry	Happy+Excited	Neutral	Sad	Recall
<b>Angry</b>	116	1	37	16	0.682
<b>Happy+Excited</b>	19	337	42	44	0.762
<b>Neutral</b>	37	40	277	30	0.721
<b>Sad</b>	16	20	29	180	0.735
<b>Precision</b>	0.617	0.847	0.720	0.667	

**Table 10:** Four classes scenario: A randomly-picked confusion matrix and the corresponding precision and recall metrics.

	Accuracy	Precision	Recall	F1 Score
<i>SERN</i> <sub>5</sub>	0.557	0.563	0.552	0.558
<i>SERN</i> <sub>10</sub>	0.570	0.570	0.591	0.581
<i>SERN</i> <sub>20</sub>	0.584	0.583	0.580	0.582
<i>SERN</i> <sub>40</sub>	0.581	0.595	0.565	0.579
<i>SERN</i>	0.555	0.555	0.570	0.562

**Table 11:** Performance metrics of the proposed model using different window size.

portance of these algorithms lies at facilitating the accurate and timely recognition of the emotional state of the speaker. This renders them the key mechanism that could enable the development of effective mitigation strategies, for instance for dealing with cyberbullying and suicidal ideation in OSNs. However, this necessitates the availability of algorithms with high recognition accuracy.

In response to this need, in this paper we devised a self-attentive emotion recognition network that is composed of novel mixture of hierarchical encoding components and self-attention mechanisms. Our overarching goal was to enable a more potent modeling of the dialog dynamics, with special emphasis on accounting for long-term affective inference. Our formulation is carefully crafted to allow for predicting the emotional state of the speaker via a feed-forward scheme driven from the dialog evolution up to any desired time point. This endows *SERN* with real-time capability, thus permitting its usage directly on OSNs.

We performed a thorough experimental evaluation of our approach using the challenging IEMOCAP benchmark. We provided deep qualitative and quantitative insights to illustrate the efficacy of our modelling selections and the functional characteristics of our approach. In addition, we performed comparisons to a number of state-of-the-art alternatives and showcased the superiority of our approach. These findings vouched for the usefulness of the introduced novel modeling arguments that underlie *SERN*.

The promising findings of this work encourage us to pursue

the further evolution of SERN. We consider methodological extensions, for instance by exploring feedback-driven refinement by means of reinforcement learning techniques [26]. We also actively work on the integration of our model in a real-world OSN enhancement framework. These endeavors constitute our future research directives.

## Acknowledgments

This project has received funding from the European Union’s Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie ENCASE project (Grant Agreement No. 691025). This work reflects only the authors’ views; the Agency and the Commission are not responsible for any use that may be made of the information it contains.

## References

- [1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [2] A. Agrawal and A. An. Unsupervised emotion detection from text using semantic and syntactic relations. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 346–353. IEEE Computer Society, 2012.
- [3] C. O. Alm, D. Roth, and R. Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 579–586. Association for Computational Linguistics, 2005.
- [4] S. Aman and S. Szpakowicz. Using roget’s thesaurus for fine-grained emotion recognition. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, 2008.
- [5] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335, 2008.
- [6] R. T. Cauldwell. Where did the anger go? the role of context in interpreting emotion in speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.
- [7] W. Chen and K.-H. Lee. Sharing, liking, commenting, and distressed? the pathway between facebook interaction and psychological distress. *Cyberpsychology, Behavior, and Social Networking*, 16(10):728–734, 2013.
- [8] K. Cho, B. van Merriënboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, 2014.
- [9] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2122–2132, 2018.
- [10] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [11] L. E. Holzman and W. M. Pottenger. Classification of emotions in internet chat: An application of machine learning using speech phonemes. *Retrieved November, 27(2011):50*, 2003.
- [12] P. Khurana, P. Agarwal, G. Shroff, and L. Vig. Resolving abstract anaphora implicitly in conversational assistants using a hierarchically stacked rnn. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 433–442. ACM, 2018.
- [13] S. M. Kim, A. Valitutti, and R. A. Calvo. Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 62–70. Association for Computational Linguistics, 2010.
- [14] Y. Kim, H. Lee, and E. M. Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 3687–3691. IEEE, 2013.
- [15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [16] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9-10):1162–1171, 2011.
- [17] T. Li, Y. Zhang, and V. Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 244–252. Association for Computational Linguistics, 2009.
- [18] T. Mikolov, K. Chen, G. S. Corrado, and J. A. Dean. Computing numeric representations of words in a high-dimensional space, May 19 2015. US Patent 9,037,464.
- [19] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- [20] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 873–883, 2017.
- [21] Y. Ren, Y. Zhang, M. Zhang, and D. Ji. Context-sensitive twitter sentiment classification using neural network. In *AAAI*, pages 215–221, 2016.
- [22] Y. Ren, Y. Zhang, M. Zhang, and D. Ji. Improving twitter sentiment classification using topic-enriched multi-prototype word embeddings. In *AAAI*, pages 3038–3044, 2016.
- [23] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad. Ensemble of svm trees for multimodal emotion recognition. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, pages

- 1–4. IEEE, 2012.
- [24] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.
- [25] C. Strapparava and R. Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560. ACM, 2008.
- [26] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [27] S. Tripathi, S. Acharya, R. D. Sharma, S. Mittal, and S. Bhattacharya. Using deep and convolutional neural networks for accurate emotion classification on deap dataset. In *AAAI*, pages 4746–4752, 2017.
- [28] A. Vanzo, D. Croce, and R. Basili. A context-based model for sentiment analysis in twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2345–2354, 2014.
- [29] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.
- [30] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *Proc. INTERSPEECH 2010, Makuhari, Japan*, pages 2362–2365, 2010.
- [31] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.
- [32] C. Yang, K. H.-Y. Lin, and H.-H. Chen. Emotion classification using web blog corpora. In *Web Intelligence, IEEE/WIC/ACM International Conference on*, pages 275–278. IEEE, 2007.
- [33] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.
- [34] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, 2017.