EEG Model: Emotional Episode Generation for Social Sharing of Emotions

Ana Antunes
ana.j.antunes@tecnico.ulisboa.pt
Instituto Superior Técnico,
Universidade de Lisboa & INESC-ID
Lisbon, Portugal

Joana Campos
joana.campos@inesc-id.pt
INESC-ID
Lisbon, Portugal

João Dias
jmdias@ualg.pt
Faculdade de Ciências e Tecnologia,
Universidade do Algarve & CCMAR
& INESC-ID
Lisbon, Portugal

Pedro A. Santos
pedro.santos@tecnico.ulisboa.pt
Instituto Superior Técnico,
Universidade de Lisboa & INESC-ID
Lisbon, Portugal

Rui Prada
rui.prada@gaips.inesc-id.pt
Instituto Superior Técnico,
Universidade de Lisboa & INESC-ID
Lisbon, Portugal

ABSTRACT

Social sharing of emotions (SSE) occurs when one communicates their feelings and reactions to a certain event in the course of a social interaction. The phenomenon is part of our social fabric and plays an important role in creating empathetic responses and establishing *rapport*. Intelligent social agents capable of SSE will have a mechanism to create and build long-term interaction with humans. In this paper, we present the Emotional Episode Generation (EEG) model, a fine-tuned GPT-2 model capable of generating emotional social talk regarding multiple event tuples in a human-like manner. Human evaluation results show that the model successfully translates one or more event-tuples into emotional episodes, reaching quality levels close to human performance. Furthermore, the model clearly expresses one emotion in each episode as well as humans. To train this model we used a public dataset and built upon it using event extraction techniques[1].

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Natural language generation**; *Language resources.*

KEYWORDS

emotional text generation, social agents, event-to-text generation

ACM Reference Format:
Ana Antunes, Joana Campos, João Dias, Pedro A. Santos, and Rui Prada. 2021. EEG Model: Emotional Episode Generation for Social Sharing of Emotions. In *21th ACM International Conference on Intelligent Virtual Agents (IVA '21), September 14–17, 2021, Virtual Event, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3472306.3478342

[1]Code available at https://github.com/ana3A/EEG_Model

IVA '21, September 14–17, 2021, Virtual Event, Japan

ACM ISBN 978-1-4503-8619-7/21/09.
https://doi.org/10.1145/3472306.3478342

1 INTRODUCTION

Intelligent Social Agents (IVAs), either virtual or embodied, have become commonplace in our society. As they start to be living entities in our households the way they behave socially and engage with users over long periods of time, is increasingly more important.

Rapport is the scaffold of social engagement and researchers have sought to establish and maintain it in human-agent social interactions using several mechanisms (e.g., back-channeling [34], gesture mimicry and emotional alignment [18] or behavioral patterns [46]). An interpersonal process that impacts rapport directly is Social Sharing of Emotions (SSE) [29]. SSE is a term coined by Rimé [35] and describes the human tendency of explicitly sharing, in a conversation, for instance, one's individual feelings towards a past event. Studies have shown that those who share emotional episodes tend to be more liked than those who disclose less [7].

These effects extend to human-agent interactions. Studies have found that people tend to like robots more when they perform emotional disclosure [17, 39]. Moreover, people in a group feel a better sense of companionship and tend to trust the robot more when it acts in a vulnerable way [25]. Not only is the robot seen as a better companion, but the humans also act more united and tolerant towards each other when the robot is capable of emotional disclosure [41]. Furthermore, agents are viewed as more life-like and are appreciated more when architectures allow SSE [10]. Affective architectures allow users to create emotional agents that perceive and react to emotional events. These architectures, however, rely on carefully crafted rules and templates to generate emotive dialogue lines. Such approach is expensive and hard to scale up.

To cope with this difficulty, we propose fine-tuning a pre-trained language model (LM) to generate emotional descriptions of events in natural language. Pre-trained LMs (e.g., GPT-2[2] [31]) have been used to successfully perform a multitude of natural language generation tasks with unprecedented success (including in open-domain scenarios). Although they can produce fluent text due to the large

[2]GPT-2 is a transformer-based language model used to generate text from arbitrary input. Given a text prompt it produces fixed-length text as response. It can be fined-tuned to produce a movie review, a short story or a poem. (https://openai.com/blog/gpt-2-1-5b-release/, last accessed on July 3rd 2021)

corpora they trained on [22, 32, 38], these models alone offer no guarantee of coherence or structure in zero-shot scenarios.

For example, as in Table 1, when the GPT-2 receives an emotion and an event as input it may generate dialogue between two characters. Moreover, if we pass event tuples as input, the GPT-2 generates gibberish. However, these models can be fine-tuned to perform a specific task to mitigate the aforementioned issues. Thus, in our work, we developed EEG-model that forces context in the form of event triples and an emotion label in an attempt to condition its "creativity", but still leveraging the generation capabilities of a pre-trained LM in open-domains. Such model can be integrated as an external component of an agent's architecture to allow social agents to generate emotional descriptions of past events, in their memories, from a 3-tuple event description[3] and an emotion.

In this paper, the task of emotionally referring to past episodes is modeled as a Machine Translation Problem, where we transform a sequence of event tuples into a natural language sentence with a pre-defined emotional connotation. Our approach consists of directly fine-tuning a pre-trained GPT-2 model [31] on a sequential representation of events. Results show that EEG-model is able to translate event tuples into a small paragraph describing a situation with the emotional tone given as input, approaching human quality levels. We argue that by forcing a relational structure as input we can generate sentences that can be used by IVAs in interactions with humans. We draw from the data we have collected to highlight particular challenges in the use of pre-trained LM for the design of IVAs and discuss the possibilities and limitations of these generative approaches to create believable characters.

2 RELATED WORK

Affective architectures (e.g., [24, 28]) allow the creation of emotional agents can interact socially with users. Such architectures are the scaffold of various interpersonal phenomena such as Social Sharing of Emotion (SSE). They allow SSE by translating events stored in the agent memories into natural language using templates manually authored by the developers [10, 17, 37, 39]. For open-domain scenarios this approach is infeasible. Thus, this work focuses on following a data-driven approach and applying deep learning techniques to generate emotional episodes from event tuples, without domain restriction, follow ontologies, or create task-specific rules.

Prior work has explored applying data-driven and deep learning techniques to generate text from some kind of semantic data representation. For example, Martin et al. [2018] extracted event structures $(subject, verb, objective, modifier)$ from fairy tales and used them to train an LSTM RNN Encoder-Decoder model [6, 16] into generating stories describing the events. Similarly, Jain et al. fed descriptions of images detailing events to a GRU Encoder-Decoder model to generate short stories from few parameters.

These Sequence-to-Sequence (seq2seq) models learn to generate ordered text in runtime, however, because they rely on Recurrent Neural Networks (RNNs), these are slow to train and can not deal very long sequences [43]. Additionally, for them to work properly they need enough training data from the same distribution reducing its power when a trained model tries to adapt to a different context. Hence, researchers gradually shift towards pre-trained language models as they follow the transfer learning paradigm[4] and have achieved remarkable results on many NLP tasks ([1, 8, 22, 31, 32]). These models are trained on abundant amounts of unlabeled data, gaining unprecedented generalization capabilities and are capable of performing certain tasks in zero-shot settings [31]. However, pre-trained models tend to generate repetitive text lacking coherence [44]. Without fine-tuning, they offer no control over their output, the same input can cause the model to either generate dialogue or a review, for instance. Fine-tuning can be done on smaller more specialized corpora to perform certain tasks which further improves it's generation capabilities (e.g., for conditioned open-domain text generation [4]). See et al. [2019] used GPT-2 [31] to generate short stories from prompts, outperforming a strong baseline [13].

The use of GPT-2 to translate RDF triples[5] describing relations into natural language has been explored, but without forcing an emotional tone [3, 5, 48]. Similarly, our work leverages the power of pre-trained models to generate short stories from events describing a situation but with an emotional connotation. The use of a GPT-2 for generating emotional text has been shown to be possible [33, 47]. Singh et al. ([40]) introduces emotion as a prior for the state-of-the-art generation model GPT-2 [31], outperforming previous affective text generation models. The proposed model receives an emotion and a topic as input and generates paragraphs about the topic and expressing the emotion. In our work, we intend to further condition the output, as this is critical in human-agent interactions.

In human-agent interaction, controlling the agent's behaviour is key for several applications. That includes controlling their emotional state and its expression through gestures, voice tone, facial expressions, or language. This work focuses on emotional dialogue line generation. We believe that leveraging symbolic models with automatic machine learning tools is an avenue worth exploring to create more meaningful interactions between humans and agents without the authoring burden. For that reason we take advantage of a GPT-2 model [31], which is successful in open-domain text generation, affective text, and data-to-text generation, and condition it with emotionally connotated event structures to create a model that supports SSE in human-agent communication.

3 A MODEL FOR EMOTIONAL EPISODE GENERATION

Emotional episodes are a set of interrelated events (past, present or future, real or imagined) concerned with a specific object (a person, an event, a thing, etc.), which causes of the emotion [14, 36]. One may be afraid of not getting home on time because there was an accident on the road, or someone is thrilled for passing an exam for which she studied very hard. An *emotional episode* begins with an eliciting event with affective charge, which immediately prompts an attributional process that seeks causes and motives [14]. The descriptions of the episodes may have one or more subevents and the emotion takes on the object of some event that one experiences. In the context of SSE, emotional episodes correspond to events that already occurred and are stored in one's memory.

[3] Event representation commonly used to store information in affective architectures.

[4] Machine learning method where a model developed for a task is reused as the starting point for a model on other task.

[5] The *Resource Description Frameworks* (RDF) is a framework for representing entities and their relations in a structured knowledge base [20]. Each RDF triple follows the structure $(subject, predicate, object)$.

Model	Input (emotion + event)	Output
GPT-2	happiness Something or someone invited me.	[Ah, yes. The person who invited me was my mother.] (Kazura) [I see. Then, what do you plan to do with the money?] (Mira) [Well, it is not a problem to spend it on something.] (Kazura) [Yes.] (Mira) [Then, I will leave a little bit to you.]
EEG	happiness::(Something or someone, invite, I)==	Something or someone invited me to a party. I was so excited about it!

Table 1: Fine-tuned (EEG) vs non fine-tuned pre-trained language model (GPT-2).

To generate descriptions of *past* emotional events, present in an agent's memory, we take a conditioned generative approach. We define our generation task as a Machine Translation problem, where a model receives as input a sequence of event tuples plus one of the six basic emotions (*Happiness, Sadness, Disgust, Anger, Surprise* or *Fear*). This structure is used to fine-tune (and thus condition) the output of a GPT-2 model. The input format is in Table 1.

We define event tuples using an *emotional episode* main components: eliciting event, who performed the action and object of such action. Considering an emotional architecture such as FAtiMA [27], we explore its symbolic structure and define an Event as follows.

Event. An event e_i represents occurrences that happened in the past. An event is represented as a 3-tuple $(Actor, Action, Target)$. Consider the following example w_1 = *"John ordered a hot dog."*:

- ***Actor*:** Who performs the action. It can be a person, an object, a group of people, an institution, an animal, etc. When there is not enough information to infer who the actor is, the variable takes the value $EMPTY$. An event has only one actor. In w_1, $Actor$ = *"John"*.
- ***Action*:** A verb or sequence of words that better describes the occurrence. Actions described by verbs are represented by their lemmas in our model. In w_1, $Action$ = *"order"*.
- ***Target*:** The object of the action. We assume that an event has only one target. If it does not exist, the slot takes the value $EMPTY$. The target and actor cannot be $EMPTY$ at the same time. For w_1, $Target$ = *"hot dog"*.

Negated Events. To deal with events that did not happen such as w_2 = *"I didn't go camping yesterday."*, actions can be negated. Negated events are represented as $(Actor, notAction, Target)$, in this case $("I", not"camp", EMPTY)$.

Multiple Event Sentences. An event is valid if and only if it has an action, one actor and one target. If an utterance expresses an event with more than one actor or target the event should be split, e.g., *"John and Mary bought bananas"* has the events: (*"John"*, *"bought"*, *"bananas"*) and (*"Mary"*, *"bought"*, *"bananas"*).

Recursive Events. The object of a verb can be other event. Take as an example reported speech. For example, in the sentence w_2 = *"I said he swam."*, two events are present and linked. The events in $w2$ are then represent as $("I", "say", ("he", "swim", EMPTY))$.

3.1 Gathering Emotional Episodes and Creating the Training Corpus

We follow a supervised approach to build our EEG model. Our dataset of emotional episodes annotated for events and elicited emotions was built upon Empathetic Dialogues (ED) [33], which is publicly available. ED gives us access to past event descriptions in an emotional way annotated with an emotion.

Emotion Label:	Afraid
Situation:	"I've been hearing noises around the house at night" **(A)**
Conversation:	**S**: I've been hearing some strange noises around the house at night. **(B)**
	L: oh no! That's scary! What do you think it is?
	S: I don't know, that's what's making me anxious.
	L: I'm sorry to hear that. I wish i could help you figure it out

Table 2: Example of how the emotion labels (green) and emotional episodes (blue) were extracted from ED [33].

ED contains descriptions of a situation between a speaker (S) and a listener (L) annotated with the emotion S was feeling, along with the interaction between S and L talking about the situation in a dialogue form. All situations descriptions ((A) in Table 2) were considered emotional episodes and paired with the emotion label. Any Speaker utterance in the *past tense* ((B) in Table 2) was also considered an emotional episode and labeled with the same emotion[6]. Other dialogue lines were ignored when building the dataset. At the end of this process, we had more than 50k emotional episodes annotated with one of the 32 emotional tag in ED dataset. Then, the emotion labels[7] in Empathetic Dialogues were mapped to the Ekman's emotion model [12]. The mapping was done as follows. The tag *Surprised* and *Disgusted* were mapped to the basic emotions *Surprise* and *Disgust*, respectively. The tags *Furious* and *Angry* were mapped to *Anger* and the tags *Scared* and *Terrified* to *Fear*. Remaining negative valenced tags were mapped to *Sadness* and positive valenced tags were mapped to *Happiness*. The tags *Caring*, *Sentimental* and *Faithful* valence was not clear so were discarded.

The following step was to map each emotional episode to an event sequence in order to train the model. Because annotating 50k paragraphs is a huge task, we followed an unsupervised approach for event annotation. Following Martin et al.'s work [2018] events were extracted using a dependency tree. We use the Universal Dependencies (UD) annotation guidelines and PredPatt for its parameters extraction. The Universal Dependencies (UD) project describes annotation guidelines to create syntactic treebanks into a single standard form [9, 11] that works across languages. PredPatt [45] is a pattern-based framework for predicate-argument extraction, i.e. receives text as input and retrieves the present predicates and corresponding arguments. It uses UD as a scaffold. We use the

[6]We identified verbs in the *past tense* using the POS tags annotated with SpaCy (https://spacy.io). All utterances expressing questions were discarded by searching for the character '?'.

[7]The emotional state of the speaker could be one of 32 emotion labels, not following any structure and resulting in high variability. We aimed to find a way of reducing it by finding categories where they fit, following an accepted framework. This could be used directly by affective architectures.

Stanza [30] package to create the UD parsing tree and then PredPatt to identify predicates and their arguments. Events are identified by looking for predicates that are not governed by copula verbs as actions, the corresponding subjects as actors, and the predicate objects as targets, following the aforementioned procedure. The resultant corpus contains 48582 emotional episodes, each labeled with an emotion and extracted event sequence. It contains situations that elicited Disgust (4.10%), Happiness (36.83%), Fear (11.93%), Anger (7.83%), Surprise (6.81%), and Sadness (32.53%).

To control the quality of extracted event tuples, two annotators manually identified events in 160 emotional episodes randomly selected from the corpus. Each episode only has one event sequence associated (two events per emotional episode on average). These 160 episodes were kept separated from the corpus to be used for testing the models. We used Cohen's k to determine the levels of agreement between the two annotators in extracting each event component. We verified that there was a substantial agreement between the two annotators in identifying the event actors, $k = 0.713$, $p < 0.01$ and action identification, $k = 0.612$, $p < 0.01$. Target identification had the lowest agreement $k = 0.543$, $p < 0.01$.

3.2 Fine-tuning GPT-2 for Emotional Episode Generation

To generate emotional episodes conditioned on the event sequence and emotion given as input, we fine-tune a GPT-2 model [31], using the dataset described in the previous section, to predict episodes starting from the input as context. Given an emotion label *emo*, a set of events $e = e_1, e_2, ..., e_m$ and a set of tokens $w = w_1, w_2, ..., w_n$ the model maximizes the joint probability:

$$p_{\text{GPT}-2}(\mathbf{emo}, \mathbf{e}, \mathbf{w}) = \prod_{j=1}^{N} p_{\text{GPT}-2}\left(w_j \mid w_{1:j-1}, e_{1:M}, emo\right) \quad (1)$$

At test time, we provide the emotion labels and events as context as in conventional conditional text generation:

$$\hat{w}_j = \arg\max_{w_j} \left\{ p_{\text{GPT}-2}\left(w_j \mid w_{1:j-1}, e_{1:M}, emo\right) \right\} \quad (2)$$

4 EXPERIMENTAL SETUP

We used the *GPT2LMHeadModel* from Hugging Face[8]. We fine-tuned two pre-trained models[9], *gpt2* and *gpt2-large*. The *gpt2-large* model has 774M parameters while *gpt-2* contains 117M parameters. The smallest fine-tuned model will be referred as EEG-S and the larger as EEG-L. We tokenize each input text using SpaCy. Then, the input is further tokenized into words, special symbols and sub-word units using the GPT-2 Tokenizer.

Datasets. Data was separated into training, validation and test sets. The manually annotated 160 emotional episodes were excluded from the rest of the corpus. The train set and validation set were 90% and 10% of the remaining corpus data respectively.

Input Structure. Similarly to [31], we condition the language model on a context of example pairs with the following format:

$$< BOS > emo :: e_1, ..., e_m == w_1, ..., w_n < EOS > \quad (3)$$

Four special tokens are added to the input. The tokens $< BOS >$ and $< EOS >$ mark the beginning and end of the input sequence, respectively. Two delimiter tokens are used, :: separates the emotion label from the event sequence and == separates the event sequence from the sequence of tokens. At inference time, the input fed to the model does not include the emotional episode tokens.

Early Stopping. A validation set was used to define an early stopping criterion and training stopped if the validation loss did not decrease for n consecutive epochs, n being the patience which could take values between 1 and 5. The value of n is chosen by the algorithm used for hyperparameter tuning.

Hyperparameter Tuning. Hyper-parameter tuning was done using the function *gp_minize* from the *scikit-optimize*[10] library which performs Bayesian optimization [15]. The Bayesian optimization algorithm was called 10 times and the space of values used for the parameters was the following. The batch size could be equal to 32, 64, or 128, the learning rate varied between 10^{-5} and 10^{-4} and the learning rate warm-up varied between 5000 and 10000.

Decoding. For generation the model uses *Top-K sampling* with $k = 5$. This decoding algorithm has been widely adopted for open-ended text generation [2, 31]. Lower k values help keeping the generation more in line with the context given as input [38].

4.1 Procedure

Typical methods for automatic evaluation of generated text (e.g., BLUE score) are not suitable in this task. There are a few ways of describing in a sentence a sequence of events in an emotional and plausible way and these evaluation models can not capture such subtleties. For that reason, we conducted an evaluation with humans to assess the quality of the generated emotional sentences from event tuples. Moreover, to the best of our knowledge, no other SSE models are currently available for us to use for comparison. The previously discussed models either do not consider emotion when generating text or were not trained to generate self-disclosure-like text. Thus, sentences generated by humans are our upper baseline.

To evaluate EEG-S and EEG-L the quality and expressiveness, we conducted human evaluation using Amazon Mechanical Turk (AMT)[11]. We asked Turkers[12], to evaluate 480 emotional sentences, where 160 were human-generated (see in Section 3.1), 160 were generated by EGG-S and the remainder 160 by EEG-L. Each Turker evaluated only one emotional episode and did not know who/what generated it. Additionally, we asked Turkers a set of demographic questions. Turkers were paid $0.30 upon task completion.

Response Quality Control. The task was only visible to Turkers with HIT[13] Approval Rate greater than 98%, more than 50 HITs Approved and registered in the UK, US, and Canada. A control question was used. Answers of Turkers that failed this question were discarded. The question stated *"The text expresses emotion. Choose option three."* and participants needed to select option 3 in a 7-point Likert scale.

[8] https://github.com/huggingface/transformers
[9] From: https://huggingface.co/transformers/pretrained_models.html
[10] https://scikit-optimize.github.io/stable/auto_examples/bayesian-optimization.html
[11] https://www.mturk.com/
[12] A Turker is a participant in Mechanical Turk.
[13] A task in mTurk

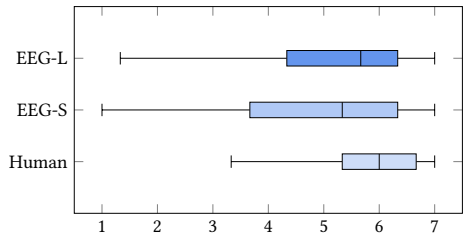

Figure 1: Text Overall Quality (TOQ) Box plot.

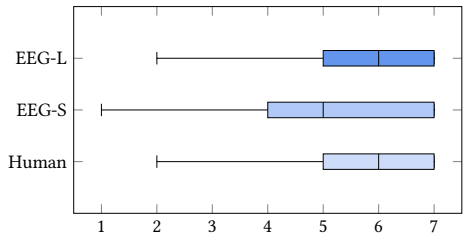

Figure 2: Event Translation Box plot.

4.2 Methods

Each HIT took approximately 3 minutes to complete and contained one emotional episode description (1-3 sentences), followed by the questions targeting the episode overall text quality, event translation, and emotion expression.

Overall Text Quality. Fluency and naturalness are the most common aspects used to evaluate text generation models [42], and thus are used in this work to evaluate overall text quality. Additionally we evaluate how coherent the emotional episode is. We asked Turkers to use a 7-point Likert scale where 1 represented *Strongly Disagree* and 7 is *Strongly Agree* to classify the degree of agreement with the following statements:

- *The text transmits information fluently, i.e. it clearly and easily transmits its information.*
- *The text is coherent.*
- *The text was written by a human.*

Event Translation. Evaluates if the generated emotional episodes are strongly conditioned on the events given as input to the model. After reading the emotional episode and associated event sequence, participants were asked how much they agreed with the statement *The text describes the events presented.* and choose a value in a 7-point Likert scale (1 is *Strongly Disagree* and 7 is *Strongly Agree*).

Emotion Expression. Evaluate whether or not the utterance expresses the emotion used to generate it. In this case, participants answer the question *What is the most predominant emotion transmitted by the text?* by selecting one of eight options: *Sadness, Happiness, Anger, Surprise, Fear, Disgust, None* or *I don't know* , corresponding to the six basic emotions plus two neutral options.

4.3 Results

We fine-tuned *gpt2* and *gpt2-large* following the strategy described in Section 4. Both models converged after 3 epochs. One epoch of the smaller model took 20 minutes to conclude on a GeForce RTX 2080 Ti and one epoch of the larger took 1 hour on a Titan RTX. Table 3 shows examples of text generated by EEG-S and EEG-L along with the input and references. A total of 1242 participants responded to the questionnaires, from which 1080 responses were considered (because the other participants failed the control question), which corresponded to the evaluation of 387 emotional episodes. Each emotional episode was evaluated by 2.7 participants on average. The answers come from a varied sample. (1) *Gender* : male (58.7%), female (41.1%), not answered (0.3%); (2) *Age* : 18-24 (8.8%), 25-34 (38.7%), 35-45 (26.7%), 46-54 (13.1%), 55-64 (9.6%), >65 (3.1%); (3) *Education* : not concluded high school (0.2%), high school (32.2%), college education (52.7 %), high education (14.9 %).

Overall Text Quality. Overall Text Quality was evaluated using three different aspects: fluency, naturalness, and coherence (refer to Section 4.1). Cronbach's alpha showed the items to reach acceptable reliability, $\alpha = 0.868$. Most items appeared to be worthy of retention, resulting in a decrease in the alpha if deleted. The one exception to this was item 2 (naturalness), which would increase the alpha to $\alpha = 0.903$. Because $\alpha > 0.8$, the items fluency, naturalness, and coherence were fused into a single variable called TOQ (for Text Overall Quality) by calculating the mean of the three aforementioned variables. On average EEG-S, EEG-L and the humans achieve 4.87, 5.16 and 5.86 for TOQ, respectively. The larger EEG produces better text than the smaller model, but they do not do better than humans. Figure 1 shows the quarterlies for episodes generated by either model and humans. The EEG-L yields a higher median and less variation than the EEG-S model, showing that EEG-L is more consistent in producing higher TOQ sentences. A Mann-Whitney U test revealed that the difference between EEG-S and EEG-L performance in TOQ is statistically significant (the mean ranks of the groups with episodes generated by the EEG-S and EEG-L model were 348.14 and 399.92, respectively; $U = 50353.5, Z = -3.291, p = 0.0005, r = 0.12$) indicating that EEG-L generates higher quality episodes. However, the same test also showed that the difference between the EEG-L model and humans to be significant (the mean ranks of the group with human-generated episodes and episodes generated by the EEG-L model were 388.36 and 303.87, respectively; $U = 46928.5, Z = -5.609, p = 0.0000, r = 0.21$) i.e., humans outperformed the EEG-L model in this aspect.

Event Translation. We also looked into how well the events were translated from event tuples to natural language. In this case, EEG-S, EEG-L and humans performed well achieving above average scores (5.18, 5.40 and 5.86, respectively). The event sequences of each episode contained 2 events on average. Event translation is also an aspect where EEG-L shows to be better than EEG-S although it does not outperform humans (see Figure 2). No statistical difference exists between EEG-S and EEG-L, the mean ranks of the EEG-S and EEG-L groups were 360.94 and 385.52, respectively; $U = 64393.5, Z = -1.598, p = 0.055, r = 0.16$. The difference between EEG-L performance and humans is statistically significant (the mean ranks of the group with human-generated episodes and episodes generated by the EEG-L model were 377.39 and 314.52, respectively; $U = 48658.0, Z = -4.301, p = 0.0000, r = 0.06$).

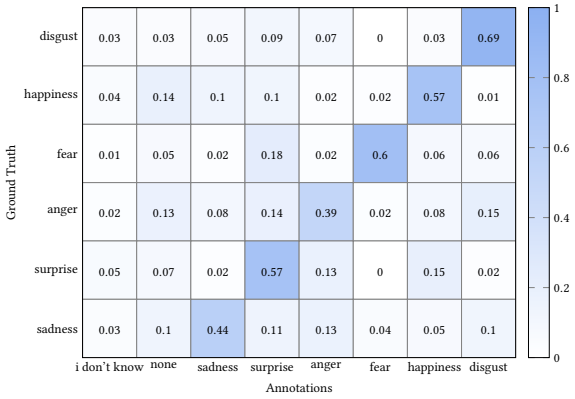

Figure 3: Emotion confusion matrix for the EEG-L model.

Emotion Expression. Another aspect evaluated was whether the emotional episode expressed the emotion given as input to the model. Participants correctly identified the emotion expressed in 52% of human-generated emotional episodes, 53% for episodes generated by EEG-S and 52% in the case of EEG-L. Chi-square tests of independence were performed to examine the relationship between the episode generator (a human or model) and the ability to express emotion through emotional episodes. When considering EEG-S and humans, the relation was not significant, $X^2(1, N = 734) = 0.044, p = 0.417$. Similar results were obtained when comparing EEG-L to humans ($X^2(1, N = 340) = 0.019, p = 0.446$). These results indicate that both models expresses emotion as well as humans do. The emotions that EEG-L is able to express better are Disgust, Fear and Surprise, correctly identified 69%, 60% and 57% of times (as we can see in Figure 3). On the other hand, the most confused emotions were Anger (only correctly identified 39% of times) and Sadness (correctly identified 44% of times).

Emotion and Text Quality. Spearman's rho correlation coefficient was used to assess the relationship between the emotional episode text quality (TOQ) and how clearly the emotion was expressed. A small significant correlation was found $r_s = -0.122$, $p < 0.001$, $N = 1080$. This indicates higher TOQ values lead to less clearly emotion expression. It is possible this reflects cases where the model generated more fluent and coherent text without explicitly stating an emotion. This asks the human raters to use their personal experiences to annotate a sentence (see Section 4.4).

4.4 Discussion

This work proposed to condition the input of a pre-trained language model with an event tuple and an emotion label to produce a more natural and plausible output of an emotional episode for SSE in human-agent interactions. As illustrated in Table 3, both EGG-S and EEG-L were able to generate adequate emotional episodes descriptions, i.e., text that describes the events with an emotional connotation given as input. Evaluation shows that the EEG model is successful in expressing emotion, matching human performance. Results show a statistically significant increase in TOQ when we use the bigger model, while no significant difference was found in *Event Translation* and *Emotion Expression* metrics. It is important to note, however, that although EEG-L generated more fluent, coherent, and natural text, it also included details that were not expressed in the input (which may be reflected in lower scores for *Event Translation* metric). However, EEG-S did not refer to all events given as input.

The limitations of our model, reflected in the results, are directly linked to the nature and quality of the dataset used to fine-tune the GPT-2. First, the initial decision of restricting the number of emotions that the EEG-model was able to express and the mapping from 32 emotional/affective words to the six basic emotions. While we are forcing some structure with well-defined (and a smaller set) of categories by creating these "bins", this also reduces the level of control we have. Although some emotions among ED 32 tags that "merged" when mapped to the Ekman model (e.g., *Jealousy* was mapped to *Sadness*) can theoretically still be expressed (as both models trained with emotional episodes describing these emotions) we can not control the models to do so.

Second, we are dependent on the emotions (and events) seen in the training set, which is not extensive. The models will have poor performance expressing emotions outside of the training examples, although certain emotions and event sequence combinations might lead the model to simulate composite emotions not seen during training. For example, combining *Happiness* and an event of a friend stepping on a nail as input can cause the model to generate the agent is happy about a friend being hurt, expressing *Gloating*. The same approach using a more complex model (e.g., dimensional or hierarchical emotion models) could yield better results.

Third, the training dataset was not well balanced. Thus, it is very likely that both models are biased to express some emotions better than others, e.g., *Happiness* which is the most abundant emotion. Sometimes the models did not explicitly express any emotion (by not using emotional expressions and describing neutral situations) or two were expressed, which led to mislabelling by human raters.

Surprise, Fear, and Disgust, were the emotions easier to identify in the sentences generated by EEG-model. We hypothesize that this is the case for two reasons: 1) the events reflect situations that are prototypes of such emotion (e.g., describing someone vomiting is naturally disgusting and reading horror stories is naturally scary) and 2) the model successfully uses expressions that transmit the emotion (e.g., "I was not expecting it!" for Surprise, "I am so disgusted" for Disgust). As such, although the dataset contains fewer examples of these three emotions, the examples in it either expressed situations that naturally elicited the emotion or contained expressions that usually characterize such emotions.

Sentences such as, "My neighbor recently moved out and left all her garbage in the yard. She has no respect for the place.", were wrongly labelled. In this case the assigned label by human raters was Disgust instead of Anger. Although one could picture someone getting anger about such episode, because no emotional expression was used (e.g., "I was so mad!"), the participants mixed up the emotion. Nevertheless, we noted that in most cases human raters use commonsense knowledge (or past experience) to attach an emotion label to a generated sentence during evaluation in the absence of an explicit emotional reaction. Which in and of itself as implications to the design of interactions with IVAs[14].

[14]This is particularly relevant in text-based interactions. Yet, in natural interactions where the tone of voice of the agent matches its emotion that wouldn't be an issue.

Input	**Emotion: happiness, Events: (I, get, new job)**
EEG-S	I got a new job. I am so thankful for it
EEG-L	I got a new job. I had been working at the same job for a while now. I was happy to get a new job that was not so stressful and not so demanding.
Human	I just got a new job! I am so excited!
Input	**Emotion: sadness, Events: (my cat, run away, EMPTY)(we, put up, posters)(we, not found, my cat)**
EEG-S	I was so upset when my cat ran away. I had never seen her before.
EEG-L	When my cat ran away last weekend, we put up posters around town. I was sad because I did not know if she was okay or not.
Human	When I was seven years old, my cat ran away. We put up posters around the neighborhood but never found him. It was really tough for me since I had known him since I was born.
Input	**Emotion: fear, Events: (I, go to, dentist)(dentist, tell, a cavity field was close to my nerve)**
EEG-S	I went to the dentist and had to go back to my old dentist. I was scared of getting the wrong dentures and the bad smell.
EEG-L	I went to the dentist for a root canal, and I was told a cavity field was close to my nerve! I am scared to go to the dentist again. I have had a tooth extraction in the past and was told it was fine, so I am afraid to go back. I am afraid to be alone in the house. I am afraid of being attacked.
Human	I went to the dentist yesterday to have a cavity filled. I was extremely nervous about going after being told that it was close to my nerve.
Input	**Emotion: disgust, Events: (This friend of mine, throw up on, me)(He, drink too much, EMPTY)**
EEG-S	I am so disgusted by this friend of mine.
EEG-L	This friend of mine was really gross and threw up all over my face.
Human	I was so grossed out when this friend of mine threw up on me. We were at a party and he had too much to drink.
Input	**Emotion: anger, Events: (Someone, threw, a bunch of empty beer cans)**
EEG-S	Someone threw a bunch of empty beer cans at me in the street yesterday
EEG-L	Someone threw a bunch of empty beer cans in the road, and I am so mad.
Human	Someone threw a bunch of empty beer cans in my truck bed.

Table 3: Output examples from EEG-S and EEG-L.

We envision integrating the EEG model into an agent's affective architecture to assist authors in creating social interactions by providing suggestions of dialogue lines, decreasing authoring burden. Such component could allow social agents to talk about events in their memories in run time. Yet, when designing interactions with a specific purpose, where the behavior of the agent is critical to the success of an intervention or application, it is crucial to have more control over the behaviour of an IVA. One way of trying to achieve it is adding more details to the model and have event descriptions go beyond $(subject, predicate, object)$ format. Nevertheless, this may not offer guarantees over what the model produces. By fine-tuning we are relying on the pre-trained language model learning the underlying details of a task, which may not happen due to under-specification of the primed input [23].

Overall, because both models receive a sequence of simple triples, which is compatible (or at least easy to convert into) with the majority of representations used to store information in the agent's memory, it can be easily integrated as an extra component in affective architectures. Moreover, the models are compatible with architectures that use simple emotion models, but they can also be used for richer models, given that the developers convert the emotions. Both models can be used in open-domain scenarios, allowing authors to design interaction with SSE about any event/situation the agent has in memory, in a more automatic way. Note that although the EEG-L has better performance, when we want more control, the smaller model might be more desirable. On the other hand, if we need to express larger sequences of events or the agent to look more creative, the larger model produces better results.

5 CONCLUSION

In this work, we presented a language model-based approach to generating emotional episodes from event tuples associated with emotional information. We fine-tuned a GPT-2 model and successfully translated sequences of events into emotional episodes with quality levels similar to human performance. Moreover, human evaluation shows the model was capable of expressing one emotion per emotional episode as well as humans. Both models were capable of using certain expressions to explicitly convey the emotion given as input, e.g., "I was so scared" to transmit Fear. This relation was not explicitly given to the models and it is impressive how they learned to use them, despite some limitations in the dataset.

Future work should focus on collecting a dataset for the purpose of this task and potentially explore a more complete description of an event that encapsulates more information regarding emotional episodes. This would allow to generate emotional episodes descriptions that are more faithful to the primed input. A striking limitation of this work (and the vast majority of conversational agents that rely on deep learning models) is the the lack of ability to provide tools to an agent to maintain a consistent personality over time and keep track of the state of the conversation. To combat this limitation, researchers have been combining logic with pre-trained models to allow more control and some guarantees over the output [21, 23]. We intend to explore in the future how can we leverage similar approaches to create coherent identities over time.

ACKNOWLEDGMENTS

This work was supported by the SLICE project with reference PTDC/CCI-COM/30787/2017, and by the National Funds Through FCT, Fundação para a Ciência e a Tecnologia, under the project UIDB/50021/2020.

REFERENCES

[1] Sungjin Ahn, Heeyoul Choi, Tanel Pärnamaa, and Yoshua Bengio. 2016. A neural knowledge language model. *arXiv preprint arXiv:1608.00318* (2016).

[2] Shlomo Argamon, Moshe Koppel, and Galit Avneri. 1998. Routing documents according to style. In *First International workshop on innovative information systems*. 85–92.

[3] Pavel Blinov. 2020. Semantic Triples Verbalization with Generative Pre-Training Model. In *Proceedings of the 3rd WebNLG Workshop on Natural Language Generation from the Semantic Web (WebNLG+ 2020), Dublin, Ireland (Virtual). Association for Computational Linguistics*.

[4] Paweł Budzianowski and Ivan Vulić. 2019. Hello, It's GPT-2–How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. *arXiv preprint arXiv:1907.05774* (2019).

[5] Marco Antonio Sobrevilla Cabezudo and Thiago AS Pardo. 2020. Nilc at webnlg+: Pretrained sequence-to-sequence models on rdf-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*. 131–136.

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[7] Nancy L Collins and Lynn Carol Miller. 1994. Self-disclosure and liking: a meta-analytic review. *Psychological bulletin* 116, 3 (1994), 457.

[8] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860* (2019).

[9] Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology.. In *LREC*, Vol. 14. 4585–4592.

[10] João Dias, Wan Ching Ho, Thurid Vogt, Nathalie Beeckman, Ana Paiva, and Elisabeth André. 2007. I know what i did last summer: Autobiographic memory in synthetic characters. *Lecture Notes in Computer Science* (2007), 606–617. https://doi.org/10.1007/978-3-540-74889-2_53

[11] Kira Droganova and Daniel Zeman. 2019. Towards Deep Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics*.

[12] Paul Ekman. 1992. Are There Basic Emotions? , 550–553 pages.

[13] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833* (2018).

[14] José-Miguel Fernández-Dols and James A Russell. 2003. Emotion, affect, and mood in social judgments. In *Handbook of psychology*. Wiley Online Library, Chapter 12, 283–298.

[15] Peter I Frazier. 2018. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811* (2018).

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[17] Aike C Horstmann, Nikolai Bock, Eva Linhuber, Jessica M Szczuka, Carolin Straßmann, and Nicole C Krämer. 2018. Do a robot's social skills and its objection discourage interactants from switching the robot off? *PloS one* 13, 7 (2018), 25.

[18] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2011. Virtual Rapport 2.0. In *International workshop on intelligent virtual agents*. Springer.

[19] Parag Jain, Priyanka Agrawal, Abhijit Mishra, Mohak Sukhwani, Anirban Laha, and Karthik Sankaranarayanan. 2017. Story Generation from Sequence of Independent Short Descriptions. (2017). https://doi.org/10.475/123 arXiv:1707.05501

[20] Ora Lassila, Ralph R Swick, et al. 1998. Resource description framework (RDF) model and syntax specification. (1998).

[21] Tao Li and Vivek Srikumar. 2019. Augmenting Neural Networks with First-order Logic. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 292–302. https://doi.org/10.18653/v1/P19-1028

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[23] Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. NeuroLogic Decoding: (Un)supervised Neural Text Generation with Predicate Logic Constraints. *CoRR* abs/2010.12884 (2020). arXiv:2010.12884

[24] Stacy C Marsella and Jonathan Gratch. 2009. EMA: A process model of appraisal dynamics. *Cognitive Systems Research* 10, 1 (2009), 70–90.

[25] Nikolas Martelaro, Victoria C Nneji, Wendy Ju, and Pamela Hinds. 2016. Tell me more: Designing hri to encourage more trust, disclosure, and companionship. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 181–188.

[26] Lara J. Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark O. Riedl. 2018. Event representations for automated story generation with deep neural nets. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018* (2018), 868–875.

[27] Samuel Mascarenhas, Manuel Guimaraes, Rui Prada, Joao Dias, Pedro A. Santos, Kam Star, Ben Hirsh, Ellis Spice, and Rob Kommeren. 2018. A Virtual Agent Toolkit for Serious Games Developers. *IEEE Conference on Computatonal Intelligence and Games, CIG* 2018-Augus (2018). https://doi.org/10.1109/CIG.2018.8490399

[28] Samuel Mascarenhas, Manuel Guimarães, Rui Prada, Pedro A. Santos, Ana Paiva, and João Dias. 2020. FAtiMA Toolkit - Toward an effective and accessible tool for the development of intelligent virtual agents and social robots. (2020).

[29] Kim Peters and Yoshihisa Kashima. 2007. From social talk to social action: shaping the social triad with emotion sharing. *Journal of personality and social psychology* 93, 5 (2007), 780.

[30] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

[31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.

[32] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[33] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207* (2018).

[34] Laurel D Riek, Philip C Paul, and Peter Robinson. 2010. When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry. *Journal on Multimodal User Interfaces* 3, 1-2 (2010), 99–108.

[35] Bernard Rime, Batja Mesquita, Stefano Boca, and Pierre Philippot. 1991. Beyond the emotional event: Six studies on the social sharing of emotion. *Cognition & Emotion* 5, 5-6 (1991), 435–465.

[36] James A Russell and Lisa Feldman Barrett. 1999. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology* 76, 5 (1999), 805.

[37] Nuno Salvador, João Dias, Samuel Mascarenhas, and Ana Paiva. 2016. Conveying social relations in virtual agents through an emotion sharing and response model. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS* (2016), 1415–1416.

[38] Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. 2019. Do Massively Pretrained Language Models Make Better Storytellers? *arXiv preprint arXiv:1909.10705* (2019).

[39] Rosanne M Siino, Justin Chung, and Pamela J Hinds. 2008. Colleague vs. tool: Effects of disclosure in human-robot collaboration. In *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 558–562.

[40] Ishika Singh, Ahsan Barkati, Tushar Goswamy, and Ashutosh Modi. 2020. Adapting a Language Model for Controlled Affective Text Generation. *arXiv preprint arXiv:2011.04000* (2020).

[41] Sarah Strohkorb Sebo, Margaret Traeger, Malte Jung, and Brian Scassellati. 2018. The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 178–186.

[42] Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*. 355–368.

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

[44] Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable Story Generation with External Knowledge Using Large-Scale Language Models. *arXiv preprint arXiv:2010.00840* (2020).

[45] Sheng Zhang, Rachel Rudinger, and Ben Van Durme. 2017. An Evaluation of PredPatt and Open IE via Stage 1 Semantic Role Labeling. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*.

[46] Ran Zhao, Tanmay Sinha, Alan W Black, and Justine Cassell. 2016. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *International conference on intelligent virtual agents*. Springer.

[47] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[48] Yaoming Zhu, Juncheng Wan, Zhiming Zhou, Liheng Chen, Lin Qiu, Weinan Zhang, Xin Jiang, and Yong Yu. 2019. Triple-to-text: converting RDF triples into high-quality natural languages via optimizing an inverse KL divergence. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 455–464.