Hidden as part of the exercise

1 INTRODUCTION

With the introduction of foundation models [1], [2], a new paradigm to utilise Machine Learning (ML) models was introduced in Natural Language Processing (NLP). The paradigm relies on the emerging capabilities of Large Language Models (LLMs) [3] to perform more complex tasks with scaling. Instead of training specialised models for specific problems, a large general foundation model would be trained once with general knowledge and this would be utilised (via prompting) in many other problems later on. This paradigm has been introduced by language models as few-shot learners [4], and popularised with the launch of the groundbreaking ChatGPT foundation models, namely GPT-3.5 [5] and its superior GPT-4 [6], which utilised techniques like Reinforcement Learning with Human Feedback (RLHF) [5]. The emerging capabilities of ChatGPT are being studied in several domains, mostly either related to general artificial intelligence [7], or traditional NLP problems, namely neural machine translation [8], and named entity recognition (NER) [9].

We examined the emerging capabilities of ChatGPT in affective computing in an early evaluation in a previous work [10], where we evaluated the performance of ChatGPT on three affective computing problems, namely suicide tendency detection, big-five personality assessment, and sentiment analysis. The study has shown interesting initial results confirming the emerging capabilities of ChatGPT, with results comparable to classical NLP methods such as Bag-of-Words (BoW) [11] and Word2Vec [12], whereas worse than fine-tuned language models like RoBERTa [13]. This study was performed on an early stage of ChatGPT, when the API for it was not yet released. Consequently, it was done manually on a small evaluation set. Additionally, given that prompt engineering is an emerging field, the prompting was less developed. Last but not least, the former study of ChatGPT evaluations focused on classification-based affective computing tasks, where regression-based tasks were not evaluated.

Given the trend that started with GPT-3 and is progressing rapidly with foundation models like GPT-3.5 and GPT-4, major changes might occur in the future in the field of ML if this trend continues. Such foundation models could revolutionize the paradigms of ML in the near future, for example, by replacing specialized models trained for specific single tasks. As a result, it is crucial to evaluate the realism of such a claim and the distance between the current state and this major change in ML [14]. The underlying models of ChatGPT are among the most advanced in this sphere, with GPT-4 considered to be the state-of-the-art in this field as the most competent LLM in many tasks, and GPT-3.5 being its most similar predecessor with a significantly smaller number of parameters. The focus of this work is to study the aforementioned trend by analysing two of the currently best foundation models.

In this work, we aim to extend these evaluations beyond the limitations of the previous work [10], namely by exploring better prompting, broader comparison, handling of different learning tasks, and the wider scope of affective computing problems [15], [16]. We

- *Mostafa M. Amin is with the Chair of Health Informatics, Technical University of Munich, 85748 Garching near Munich, Germany; he is with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany, and also with AI the R&D Team, SYNCPILOT GmbH, 86156 Augsburg, Germany. E-mail: mostafa.amin@tum.de.*
- *Rui Mao and Erik Cambria are with the School of Computer Science and Engineering, Nanyang Technological University, 639798 Singapore, Singapore.*
- *Björn W. Schuller is with GLAM – The group on Language, Audio, and Music, Imperial College London, SW7 2BX London, U.K., and also with the Chair of Health Informatics, Technical University of Munich, 85748 Garching near Munich, Germany.*

examine a total of 34 setups for 13 affective computing problems. The problems are aspect extraction, aspect polarity classification, opinion extraction [17], [18], [19], sentiment analysis [20], sentiment intensity ranking [21], emotions intensity ranking [22], suicide tendency detection [23], toxicity detection [24], well-being assessment [25], engagement measurement, personality assessment [26], sarcasm detection [27], [28], and subjectivity detection [29]. The evaluation of these tasks is conducted due to their recognised prominence in the field of affective computing. These tasks collectively encompass various dimensions of understanding affective language and are widely acknowledged as pivotal in this domain.

Our evaluation yields the following findings: ChatGPT models excel in sentiment-related tasks, outperforming supervised baselines in opinion extraction, emotions, and sentiment intensity ranking. They particularly excel in identifying extremely negative emotions, notably in well-being assessment and toxicity detection, likely due to safety-focused training. However, they struggle with implicit signal tasks such as engagement measurement, personality assessment, sarcasm detection, and subjectivity detection.

The contributions of this paper are as follows:

1) Executing a wide evaluation of the performance of the ChatGPT models, namely GPT-3.5 and GPT-4, on a wide range of affective computing problems.
2) Introducing a prompting framework that can be applied to LLMs for a wide range of affective computing problems.
3) Introducing a framework for the conversion of a regression task into a pairwise ranking classification task for the purpose of evaluating ChatGPT models.

The paper is organised as follows: we discuss related work in the next section; afterwards, we introduce our method; then, we present and discuss the experimental results; finally, we provide concluding remarks.

2 RELATED WORK

We introduce related works with a primary focus on the evaluation of foundation models in affective computing or NLP tasks. Affective computing is an important research domain in NLP, including diverse tasks, such as sentiment analysis [17], [30], emotion detection [31], [32], sarcasm detection [27], [33], personality analysis [34], mental health analysis [25], [35], figurative language processing [36], [37], and more. [38] evaluates ChatGPT on various sentiment analysis-related tasks, including aspect extraction. [39] evaluates ChatGPT on sentiment analysis, while [40] evaluates ChatGPT's ability to predict personality. The works of [41], [42], [43] explore the capabilities of ChatGPT on a wide range of NLP problems including sentiment analysis and emotion recognition, and other NLP tasks like NER and text summarising. Furthermore, we evaluated the fusion capabilities of ChatGPT with traditional NLP methods in [44].

The aforementioned ChatGPT models assessments concentrated on a confined subset of tasks within the domain of affective computing, notably sentiment analysis and emotion detection. Nonetheless, considering the expansive scope of the affective computing field, there exists merit in conducting a comprehensive evaluation of the proficiency exhibited by the ChatGPT models within this domain. This is motivated by the recognition that distinct affective computing tasks encapsulate intricate emotional nuances, demanding ChatGPT models to adeptly discern them across diverse application contexts. A comprehensive evaluation encompassing a wider array of tasks also serves as a means to uncover any inherent biases that might manifest in the performance of the ChatGPT models.

On the other hand, the latest survey on the assessments of the ChatGPT models has unveiled a notable discrepancy in findings, stemming from the employment of diverse prompting strategies [45]. Moreover, the evaluation process has scarcely encompassed regression tasks, due to the difficulty of prompting objective scores from LLMs in different task setups. This highlights the imperative of adopting a unified prompting framework to counteract the biases introduced by distinct affective computing evaluation tasks. Furthermore, the establishment of an efficacious evaluation methodology for assessing LLMs on regression tasks holds extra utility for forthcoming research endeavours.

3 METHODOLOGY

3.1 Utilising ChatGPT

We utilise the API provided by OpenAI to query ChatGPT[1], using 'gpt-3.5-turbo-0301' for GPT-3.5 and 'gpt-4-0314' for GPT-4. For each problem and target label combination, we construct a prompt for it, which is used as the system prompt. For each given example in the evaluation data, we send two messages to ChatGPT, namely, the system prompt of the problem and a user message that includes only the input text of the example. The assistant response acquired by ChatGPT is then the answer for the example. We exclude the examples that did not have simple parsing criteria. Unlike the previous work [10], the extra specificity in the system prompt made the vast majority of examples easy to parse.

3.2 Prompting

The design of the prompts is of a crucial importance in determining the performance of an underlying LLM. The prompting structure uses and extends the prompting structure from [10], [44]; the extension mainly aims to add specificity on the output formats. The prompting structure below aims to reach the best affective performance, while producing stable easy-to-parse results that facilitate clear evaluation [46].

The prompt design follows a general pattern:

1) Define the role of the assistant as an expert in the given problem, and add a problem description.
2) Define the prediction task, namely, by specifying the type of labels and stating that the user will give an input and the assistant should reply to that by predicting the label for that example.
3) Emphasise the answer format in bullet points of the exact requested format, and also the *not* allowed formats. We do not allow it to explain the reasoning behind the decision, or mentioning statements like *My guess is ⋯*.

3.2.1 *Word-level labels prompts*

1) **Aspect extraction with polarity classification prompt**

You are an aspect-based sentiment analysis expert, you will be given a sentence by the user and you will list all the aspect words target objects. List the words in bullet points. The aspect targets are objects that are classified by a corresponding one of four sentiment targets: positive, negative, neutral, and conflict. It is possible that a word has

---

1. https://platform.openai.com/docs/guides/gpt/chat-completions-api

no target, which is defined as a background target. Use the following format:
* You will output a list of words in bullet points.
* Each bullet point will be on the form: "word" is target.
* The target is one of the four targets, do not report background targets.
* You will not mention any other text like "My guess is ..." or "I think ...".
* If all words have background target, then you return the word "BACKGROUND" without any bullet points.

2) **Opinion extraction prompt**

You are an aspect-based sentiment analysis expert, you will be given a sentence by the user that contains aspect words objects. Your task is to list all the sentiment opinionated words / expressions, that are corresponding to the aspect in the text (if any). You just need to list the words/expression in bullet points without classifying them. There will be many words without sentiment, these should not be listed. Use the following format:
* You will output a list of words in bullet points.
* Each bullet point will be on the form (without quotations): "* expression"
* You should mention words that are explicitly in the text.
* You will not mention implied sentiment.
* You should mention the words exactly how they are written in the input, even if they have typos.
* You will not mention any other text like "My guess is ..." or "I think ...".
* If all words have no sentiment, then you respond with the word "BACKGROUND" without any bullet points.

3.2.2 *Sentence-level labels prompts*

The following text is appended at the end of each of the following prompts.

Use the following format:
* You are only allowed to answer "{label$_A$}" or "{label$_B$}".
* Don't write an explanation of the answer.
* Don't write things like "My guess is...", or "I think ...". Just write {label$_A$} or {label$_B$}, but nothing else.

1) **Sentiment analysis prompt**

You are an expert at sentiment analysis. Given a text by the user, analyze the sentiment of the text if it is 'positive' or 'negative'. You are not allowed to answer 'neutral', try to narrow it down to 'positive' and 'negative'.

2) **Sentiment intensity ranking prompt**

You are an expert at sentiment analysis. Given a pair of texts A and B from the user, you will output which text expresses more positive sentiment.

3) **Emotion intensity ranking prompt**

You are an expert at emotion analysis. Given a pair of texts A and B from the user, you will output which text expresses higher intensity of the {emotion} emotion.

4) **Suicide detection prompt**

You are an expert at psyche analysis. Given a text by the user, solve the binary classification of analysing if the text expresses a tendency for suicide.

5) **Toxicity detection prompt**

You are an expert at toxicity analysis. Assume that we have the capability of analysing 6 toxicity traits. "toxic", "severe

toxic", "obscene", "threat", "insult", "identity hate". Your task is to make binary classification for the trait {trait}, and not the remaining traits. Given a text by the user, estimate if the given text displays the trait {trait} or not.

6) **Well-being assessment prompt**

You are an expert at psyche analysis. Given a text by the user, estimate if the given text talks about a stress-related topic, or expresses emotional stress be it implicit or explicit.

7) **Engagement measurement prompt**

You are an expert at social media analysis. Given a pair of texts A and B representing tweets, estimate which text is more engaging. You will achieve this by estimating which text is more viral, by estimating which one has a higher number of retweets.

8) **Personality assessment prompt**

You are an expert at the big-five personality traits assessment. Given a pair of texts A and B from the user, you will output which text expresses higher intensity of the {trait} trait, from the big-five OCEAN personality traits.

9) **Sarcasm detection prompt**

You are an expert at sarcasm analysis. Given a text by the user, estimate if the given text is sarcastic or not.

10) **Subjectivity detection prompt**

You are an expert at language and sentiment analysis. The user will give you a text, your task is to make a binary classification on the text, if the given text is opinionated / subjective / biased, or if it is non-opinionated / objective / descriptive / factual. Please note that this is about "how" the text is described and not "what" it describes, so the text can still "objectively" describe a fictional story with some emotional terms.

3.3 Pairwise comparison to solve regression tasks

Querying and prompting ChatGPT models for classification tasks can be readily accomplished by instructing them to select a suitable label from the predefined label set of classification tasks. Nonetheless, the challenge arises when transitioning to regression tasks, as querying an objective score from the ChatGPT models becomes intricate due to the variations in scales and the inherent subjectivity associated with dataset annotation criteria. For example, given a regression label that is normally distributed within $[0, 1]$, the ChatGPT model might be more inclined to predict uniform values like $0.0, 0.1, 0.2, \cdots, 0.9, 1.0$ similar to rating problems on $1 - 10$ likert scale, or vice versa. Evaluating ChatGPT on this basis compared to traditional ML models would be unfair for ChatGPT, since it did not train on the specific distribution label with its distribution and scaling, unlike the traditional ML models which learn this jointly in the training process on the specific training dataset. Thus, we evaluate ChatGPT on regression problems by modelling regression problems as a pairwise ranking classification problem. In other words, given an evaluation dataset on a regression label that consists of N examples, we can remodel the problem into a classification problem by sampling M pairs (a, b), and solve the binary classification problem '*is $y_a > y_b$?*'. A crucial aspect of this framework is how to sample the M pairs to be as few as possible whilst being representative. We employ the small-world graph generation algorithm [47], which is a densely connected graph without having a very high number of connections. In our setup we sample $M = 4N$ pairs.

This modelling can be further applied for predictions not just evaluations. A prediction procedure would utilise successive halving of a range of answers by comparisons with the median item within the remaining range, until the range has a small width, similar to the binary search algorithm. However, we do not use this prediction mechanism in this work.

3.4 Compared Model

To evaluate the performance of the ChatGPT models, we compare them with a Recurrent Neural Network (RNN)-based framework that consists of a single bidirectional LSTM layer (with L units in each direction) followed by N fully-connected layers with U units (with ReLU activations), then a final prediction layer. The final layer uses sigmoid activation in most setups, except for aspect-based problems (softmax for multi-class classification), sentiment ranking (tanh), and engagement measurement (ReLU). This framework leverages different features that will be introduced in the following section.

Adam [48] is employed as an optimisation algorithm, with a learning rate α. The loss function is crossentropy for classification tasks [11], and Mean Absolute Error (MAE) for regression tasks [11]. For problems with imbalanced datasets, namely toxicity detection and aspect extraction, we make use of a weighting parameter λ that discounts the weight of the '0/negative' class (typically the over-represented class). In order to tune the hyperparameters λ, α, N, U, L, we opt to utilise the hyperparameter optimisation toolkit SMAC [49] to select the best hyperparameters for each problem. The hyperparameter space is $N \in [0, 1]$, $U \in [64, 512]$ (log-sampled), $L \in [16, 64]$ (log-sampled), $\alpha \in [10^{-6}, 10]$ (log-sampled), and $\lambda \in [10^{-3}, 1]$ (log-sampled). We sample 25 combinations for each problem. In each training run, we train for a total of 300 epochs with early stopping (30 epochs). The model with the best validation score is the one used for testing.

3.5 Text Features

We utilise two different textual representations of the input texts, and then we train an LSTM on top of that. The representations are the raw text as sequence of word IDs (we call that the end-to-end (E2E) approach), and RoBERTa features [13] In the E2E approach, the IDs are limited to the most common 2,000 words in the training set of a given problem, and we train embeddings of dimension 128 for that, jointly with the rest of the model.

RoBERTa Language Model We employ the RoBERTa language model to extract features for a given sentence, by basically running the RoBERTa-base model[2], on a given sentence to acquire a sequence of features, corresponding to features of the subwords. Each subword is represented by a feature vector of size 768. The RoBERTa-base model [13] is the smaller variant of the RoBERTa architecture, which is based on the BERT transformer architecture [50]. We do not adopt the RoBERTa-large variant since its improvements above the RoBERTa-base for affective tasks are not major [51], [52], in addition to the fact that the RoBERTa-base was shown on few affective computing tasks to already surpass the ChatGPT model (GPT-3.5) by a wide margin [10].

In this work, we do not make use of the pooling of the RoBERTa model, but instead we use an LSTM model on the features sequence extracted. Using an LSTM on top of RoBERTa has been shown to

yield some improvements [53], [54]. As a result we opt to use the LSTM variant to have a unified architecture that is also capable of making sequence predictions for the aspect-based problems.

Furthermore, we do not fine-tune RoBERTa, instead we only use it to extract features. Even though fine-tuning can lead to some improvements, the feature-based approach has been shown to be quite effective in many setups [50], [55], especially when combined with LSTMs [54] to bridge the gap between feature-based and fine-tuning use of RoBERTa. We use simpler expert models to contrast the differences between LLMs and these expert models across various affective computing tasks. The complexity of fine-tuning RoBERTa is significantly higher than that of the model we are applying. For instance, RoBERTa-base consists of 12 layers of Transformer, whereas our basic expert model comprises just a single layer of bidirectional LSTM. Our experimental design does not aim to emphasise their comparable accuracy through increasing the complexity of expert models. Instead, the comparison between a simple classifier and an LLM can provide a more objective reflection of the limitations of the inspected LLM. Otherwise, it would not be surprising if LLMs fall behind complex expert models and even state-of-the-art expert models. Moreover, this setup provides the advantage of the possibility to execute a wide range of experiments, since there are 19 training tasks (as seen in Table 1), with 25 different hyperparameters combinations per task as mentioned in Section 3.4. Fine-tuning RoBERTa for all of these 475 combinations would be tedious to execute.

For word-based labelling in the aspect-related problems, we give all subwords the same label as the corresponding word at training, while using the prediction of the first subword of a word as the corresponding word prediction at evaluation similar to [50].

3.6 Datasets

We employ 13 datasets for the 13 evaluation tasks in the affective computing domain. The summary of the statistics of our employed datasets is given in Table 1. For training and validation, we make use of the original splits provided with the datasets, or we split them otherwise[3]. For benchmarking, we always downsample the testing set if the original testing set is not small enough, due to the very tight restrictions on scaled usage of GPT-4.

The Aspect-based Sentiment Analysis Datasets are from the SemEval 2014 [18] and SemEval 2015 [19] shared tasks. The shared tasks sourced data from laptop and restaurant reviews, termed lap14, res14, and res15 in the later result table. We employ the split of training, validation, and testing sets from the work of [56]. The dataset contains three sets of word annotations labels for aspect extraction, aspect polarity prediction, and opinion extraction tasks.

The Sentiment Analysis Dataset is the Twitter140 dataset [20] which consists of tweets and their corresponding sentiment negative or positive labels. Given the extremely small size of the provided testing set, we downsample and split the large training set into the Train/Dev/Test sets we use.

The Sentiment Intensity Ranking Dataset is from the SemEval-2017 Shared Task 5 [21]. The dataset sourced data from two domains, namely, microblog messages and news headlines. The sentiment intensity scores are within $[-1, 1]$. Our experiments are conducted on the microblog data.

---

2. Acquired on 28.07.2023 from https://huggingface.co/docs/transformers/model_doc/roberta

3. Splits procedures will be provided on: https://github.com/mostafa-mahmoud/chatgpt-wide-evaluation.

	Problem	Train	Dev	Test
ABSA	res14	2,436	608	800
ABSA	lap14	2,439	609	800
ABSA	res15	1,052	263	685
	Sentiment Analysis	100,000	10,000	2,500
	Sentiment Ranking	1,000	300	365
Emotion	Sadness	786	74	673
Emotion	Joy	823	79	714
Emotion	fear	1,147	110	995
Emotion	Anger	857	84	760
	Suicide	23,398	5,611	2,345
	Toxicity	30,000	6,864	959
Well-be.	Reddit bodies	1,511	458	935
Well-be.	Reddit titles	3,538	996	998
Well-be.	Twitter denoised	851	400	800
Well-be.	Twitter full	5,900	1,500	1,500
	Engagement	30,037	5,000	4,000
	Personality	5,992	2,000	1,996
	Sarcasm	18,709	4,000	4,000
	Subjectivity	6,000	2,000	2,000

TABLE 1: Datasets sizes statistics; these are shown on our selection of the data, after we downsample (and possibly split) some of the datasets. ABSA denotes aspect-based sentiment analysis that includes aspect extraction, opinion extraction and aspect polarity classification subtasks.

The Emotion Intensity Dataset is from the WASSA-2017 Shared Task on Emotion Intensity (EmoInt) [22]. The dataset includes four emotions, namely joy, sadness, fear, and anger and emotion intensity scores, ranging within $[0, 1]$.

The Suicide Tendency Detection Dataset is from the work of [23], which is a dataset collected from Reddit under depression and teenagers Subreddits, for positive and negative labels, respectively.

The Toxicity Detection Dataset is from the Toxic Comment Classification Challenge[4]. The dataset encompasses an extensive collection of Wikipedia comments that have been meticulously annotated by humans. These instances of toxicity encompass various categories, namely: toxicity, severe toxicity, obscenity, threats, insults, and identity-based hate. We split the training part into Train/Dev sets, and downsample the negative class because it is extremely over-represented in comparison.

The Well-being Assessment Dataset is from the work of [25], where the data were from Reddit and Twitter, with two benchmarks on each. For the (Combi) Reddit benchmark, we make use of the bodies of the posts, instead of the titles. Binary labels are used in this dataset, indicating stress-negative and stress-positive text.

The Engagement Measurement Dataset is from Kaggle TEDTalks Tweets[5]. The dataset was obtained from Twitter and pertains to discussions associated with TEDTalks. The dataset contains tweet content, and the number of likes. This allows us to conduct a comparative analysis, discerning the relative favourability of different content. We use the $\log_{10}(\text{number of retweets} + 1)$ as the target label, since it represents the labels distribution effectively.

The Personality Assessment Dataset is the First Impressions dataset [57], [26], from which we use the transcripts of personality annotated videos. The labels are expressed by the big-five personality model (Openness, Conscientiousness, extraversion, agreeableness, and neuroticism), with labels within $[0, 1]$.

The Sarcasm Detection Dataset is from the work of [28] (Version 1). The data were sourced from *TheOnion* and *HuffPost* news

4. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge
5. https://www.kaggle.com/datasets/thedevastator/social-media-interactions-on-tedtalks-dataset

headlines, associated with binary labels, indicating if a news headline is sarcastic or not.

The Subjectivity Detection Dataset is from the work of [29]. The data were sourced from movie reviews from *Rotten Tomatoes*, associated with binary labels, indicating if a sentence or a piece of text is subjective or objective.

4 RESULTS

The main results of the experiments are shown in Table 2, where we evaluate classification accuracy and Unweighted Average Recall (UAR), which is the unweighted average of the accuracy of classifying each class separately [58]. For the time-series labels, we use micro-averaging for the measuring the performance. For all results, we utilise a two-tailed randomised permutation test to check for the statistical significance of the differences in performance compared to GPT-3.5 [59].

The results show that the LSTM trained with RoBERTa features has the best performance in the majority of problems, for both metrics; in most of these, it is even significantly better than GPT-3.5. Furthermore, GPT-4 comes at second best overall, with the best performance on some of the setups. In the instances where GPT-3.5 is better than RoBERTa, it is with a difference that is not regarded as statistically significant, and it is often due to the fact that the RoBERTa-based approach prioritised improving UAR instead of accuracy. On the other hand, comparing GPT-3.5 to the simpler baseline E2E presents a different picture; E2E is sometimes significantly worse and sometimes significantly better than GPT-3.5.

For aspect extraction and aspect target predictions, RoBERTa has the best performance overall especially when considering UAR, followed by GPT-3.5. However, E2E is better than GPT-3.5 (considering UAR only) to identify the aspect but not its polarity. Surprisingly, GPT-4 has the worst performance only on this problem; upon inspection of few of the GPT-4 results, we found that despite its right identification of the aspect expressions, still it tends to include more surroundings words, which is probably a major reason behind this deterioration in performance. For the opinion extraction problem, GPT-4 has the best performance, especially considering UAR, followed by RoBERTa, then GPT-3.5.

In sentiment related problems, RoBERTa and GPT-4 are far better than both GPT-3.5 and E2E. GPT-3.5 and E2E have relatively close performance on sentiment-based problems. The results of GPT-3.5 on sentiment analysis are similar to previous work [10], [44].

In emotions intensity ranking, GPT-4 is the best model (significantly better than GPT-3.5 and E2E), followed by RoBERTa, then GPT-3.5, then E2E (significantly worse than GPT-3.5). An interesting observation is the strong performance of GPT-3.5 in identifying the emotion sadness; since ChatGPT models seem to be very competent in problems related to identifying negative emotions as we will elaborate.

On psychology-related problems with extreme negative emotions, namely the detection of suicide tendency, well-being, and toxicity, the ChatGPT models have a strong performance in many cases that significantly outperforms E2E. The results of the suicide-detection are consistent with previous work [10], [44], where RoBERTa is significantly better than GPT-3.5. Moreover, ChatGPT models seem to thrive with longer texts, as seen in the results of the well-being problem part Reddit bodies, which is the only part consisting of long texts of the full body of Reddit posts, where both

Dataset		Accuracy [%]				UAR [%]			
		E2E	RoBERTa	GPT-3.5	GPT-4	E2E	RoBERTa	GPT-3.5	GPT-4
Aspect Extraction	res14	81.73**	92.00**	86.95	71.50**	81.18**	91.54**	76.18	73.23**
	lap14	78.22**	87.19**	84.60	70.32**	82.21**	90.64**	77.62	75.94
	res15	81.28**	73.02**	84.57	70.05**	72.28**	85.13**	78.56	73.81**
Aspect Polarity	res14	86.10*	71.85**	85.13	69.30**	49.72	58.09*	49.96	48.26
	lap14	72.57**	90.22**	82.23	67.63**	45.10	58.23**	47.37	44.54**
	res15	79.08**	84.31**	82.38	67.51**	36.84**	48.96	47.79	44.27
Opinion Extraction	res14	81.61**	93.26**	91.04	80.93**	80.87	80.06	77.28	83.02*
	lap14	74.33**	73.81**	89.43	76.90**	66.43*	76.73	73.51	77.42
	res15	79.42**	89.16	89.32	78.10**	68.59	77.31	76.90	77.19
Sentiment Analysis		78.87	88.74**	80.54	84.09**	78.84	88.75**	79.93	83.69**
Sentiment Ranking		70.88	72.37	69.30	73.21**	70.83	72.41*	68.69	73.08**
Emotion Ranking	Joy	66.49**	75.41	74.07	78.46**	66.51**	75.40	74.29	78.63**
	Fear	68.65**	76.83**	72.76	73.96	68.65**	76.83**	72.86	74.09
	Anger	67.63**	73.47	72.12	75.58**	67.60**	73.46	72.09	75.49**
	Sadness	72.41**	76.06	78.19	78.55	72.40**	76.05	78.22	78.58
Suicide Detection		84.75**	98.43**	89.46	93.46**	85.32**	98.46**	89.44	93.32**
Toxicity	Toxic	81.85**	85.23	87.37	89.29	82.75**	86.01	87.19	89.65*
	Severe toxic	87.65**	80.07**	66.55	75.52**	82.48	84.78*	80.65	85.29**
	Obscene	85.40	84.83	83.45	88.16**	86.62*	86.60*	83.48	86.78**
	Threat	94.05**	95.54**	70.59	91.99**	73.99	87.43*	80.12	91.51**
	Insult	84.65**	87.25**	80.14	80.70	84.64	89.28**	83.21	84.89
	Identity hate	90.52**	90.98**	66.82	82.66**	81.61	86.16**	78.61	87.88**
Well-being	Reddit bodies	84.50**	89.88	91.93	93.33	68.82**	86.16	84.41	78.63**
	Reddit titles	86.60**	96.75**	80.61	89.54**	85.62**	96.65**	80.05	89.77**
	Twitter denoised	43.36**	93.23**	60.53	72.31**	45.45**	93.14**	65.05	73.45**
	Twitter full	80.39**	84.39**	66.24	75.25**	80.39**	84.39**	66.26	75.25**
Engagement		71.02**	79.18**	51.92	54.15**	71.02**	79.19**	51.85	53.80**
Personality	Openness	58.36**	60.54**	50.11	54.75**	58.36**	60.56**	50.60	54.59**
	Conscient.	56.79	61.59**	55.54	57.44*	56.78	61.59**	55.84	57.33
	Extraversion	56.51**	59.03**	53.55	55.90**	56.51**	59.02**	53.38	55.90**
	Agreeable.	57.81**	58.12**	51.67	54.04**	57.80**	58.14**	52.10	54.05*
	Neuroticism	58.60**	59.86**	48.94	49.68	58.60**	59.86**	49.04	49.73
Sarcasm		63.14**	90.66**	59.13	66.66**	66.29**	90.70**	56.82	69.45**
Subjectivity		87.28**	95.56**	59.56	88.38**	87.19**	95.51**	58.59	88.19**

TABLE 2: Classification accuracy and UAR scores for all the problems. The bold scores correspond to the best method on the given metric and problem. *,** correspond to scores with statistically significant results (with p-values $< 5\%$ and $< 1\%$, respectively) as compared to GPT-3.5.

GPT-4 and GPT-3.5 are achieving the best results; on the shorter texts, RoBERTa is the best model by a wide margin. Furthermore, the ChatGPT models are showing the best results for the toxicity problem. This might come as non-surprising given the fact that the ChatGPT models from OpenAI generally are tuned to identify toxicity as part of applying safety policies[6]. However, they are not as competent at specifying further the reason behind toxicity, given their inferiority compared to RoBERTa about the more specific toxicity labels.

For the tasks with more implicit or latent social signals, namely engagement measurement, personality assessment, sarcasm detection, and subjectivity detection, GPT-3.5 has very poor performance that is significantly worse than E2E in most cases and significantly worse than RoBERTa in all cases. GPT-4 shows minor improvement over GPT-3.5 on these problems, where it slightly surpasses E2E only in the sarcasm and subjectivity detection problems. Similar to previous work [10], [44], the results of the personality from GPT-3.5 are the worst, even compared to the simple baseline BoW. The results of the ChatGPT models on the engagement measurement problem are extremely poor, close to a random predictor (UAR 50 %). In earlier experiments, we attempted to make GPT-3.5 solve different formulations of the engagement measurement problem; the first was binary classifying if the number of retweets is < 10, the second was classifying if the number of retweets is < 100, and the third is 3-class classifying

6. https://openai.com/policies/usage-policies

the number of retweets $[0, 10)$, $[10, 100)$, and $[100, \infty)$. It still yielded the same poor results in all of them.

The GPT models seem to excel in tasks where the signal is explicitly present in the text, e. g., in sentiment analysis, sentiment ranking, emotion ranking, and opinion extraction. This is an expected behaviour to emerge, since training language models require the model to predict next tokens, which would make it capable to understand affective contexts and directly predict words displaying emotional content or other similar signals. GPT models excel even more in problems displaying negative emotions, where they are better than specialised models. In addition to the aforementioned reason, the two additional reasons can be attributed to the fact that GPT models are fine-tuned for toxicity detection, and potentially that humans are more likely to speak about negative emotions than positive emotions [60]. On the other hand, GPT models can have much weaker performance in tasks where the labels are implicit in the text or subjective, meaning that a group of human annotators would not easily agree on a given label. By definition, tasks with implicit signals would have weaker predictive abilities which would make such abilities harder to emerge for a language model compared to tasks with explicit signals.

The effectiveness of the introduced regression evaluation technique in section 3.3 is demonstrated by the results of the personality assessment problem. In the previous work [10], [44], personality assessment was evaluated as a binary classification problem, while the results of GPT-3.5 here are evaluated using

the newly introduced pairwise ranking evaluation framework. The consistency of the results and their relative order across the five independent labels indicate the effectiveness of the technique. The same applies to the engagement measurement problem as mentioned earlier.

The issues of parsing results mentioned in the early evaluation [10] are mostly resolved within this study. This is due to the system prompt (which enforces instructions) introduced in the API version of ChatGPT models (which was not available at the early evaluation [10]), and our more precise formulation of the prompts (see section 3.1). A crucial aspect is highlighting the possible answering formats while disallowing improper formats. The redundancy is also helpful in this regard, since we specify the formatting in a general form once on the problem description (like [10]), then, we specify it a second time in a precise manner in the notes bullet points (unlike [10]). This led to the behaviour that very few examples are not following the described format, unlike [10]. However, the issue still stands for the compound predictions, in particular, we faced the same parsing issues for the aspect-related problems because the predictions contained multiple labels, where ChatGPT models were improvising in the response format or modifying some of the mentioned input words. As a result, the responses of ChatGPT models are not fully reliable for properly formatting complex predictions, only straight-forward predictions with a short answer.

In summary, the results indicate that foundation models display their biggest advantages in affective signals that are explicitly present in the data, where they get better with larger models, even exceeding specialised models trained for such cases. Even though, only GPT-4, the state-of-the-art foundation model, is able to reach or exceed the performance of specialised models in some of these tasks, future foundation models may likely replace specialised models in many of the affective tasks, but it is not generally the case today. Therefore, it is still advised to use specialised models to gain best performance, and to avoid issues related to parsing of outputs or high costs of running foundation models; however, they offer easy-to-use benchmarks that are at least as good as LSTM models only based on their emergent abilities.

5 CONCLUSION

In this paper, we evaluated ChatGPT models, namely GPT-3.5 and GPT-4, on 13 affective computing problems. We prompted the ChatGPT models methodically for each problem, then, we used the API to retrieve predictions. We compared the ChatGPT models against two traditional natural language processing models, namely Recurrent Neural Networks (RNNs) trained end-to-end (E2E) using LSTMs, and RNNs trained using the input features from the RoBERTa model. The results have shown that the general performance order of the models is championed by RoBERTa, followed by GPT-4, then, E2E and GPT-3.5 are the following rivals with no clear winner for all problems. GPT-4 is mostly better than GPT-3.5, except for the aspect extraction problem, with a statistically significant difference in most cases.

The ChatGPT models showed strong performance in sentiment-related problems, where GPT-3.5 is usually better than E2E, and GPT-4 often outperforms RoBERTa, e. g., in the opinion extraction, and emotions and sentiment intensity ranking problems. Both ChatGPT models have shown their strongest performance in problems identifying extremely negative emotions, especially for well-being assessment over long texts and toxicity detection; this

is probably due to the extra training by OpenAI in an attempt to enforce safety policies. The problems that the ChatGPT models fell very short of were problems with implicit signals, like engagement measurement, personality assessment, sarcasm detection, and subjectivity detection. GPT-3.5 has shown significantly worse performance than E2E for most of these, and GPT-4 was at best slightly better than E2E if not significantly worse for two of these problems.

General reasons for the over-performance of ChatGPT models compared to RNNs is the fact that they are pretrained on a very large corpus of text, namely the whole internet, which includes huge amounts of text with affective data, be it implicit like sarcasm, or explicit like sentiment or emotions. ChatGPT models seem to have affective capabilities that are superior on signals present explicitly in text. Additionally, the underlying transformers architecture of the GPT models is far more capable of generalisations due to the dense attention mechanism.

Future efforts can focus on reinforced prompt design, evaluating and fine-tuning open-source foundation models like Llama2 [61] and Mistral [62].

REFERENCES

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson *et al.*, "On the Opportunities and Risks of Foundation Models," *arXiv:2108.07258*, 2021.

[2] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, H. Peng, J. Li, J. Wu, Z. Liu, P. Xie, C. Xiong, J. Pei, P. S. Yu, and L. Sun, "A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT," *arXiv:2302.09419*, 2023.

[3] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent Abilities of Large Language Models," *arXiv:2206.07682*, 2022.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child *et al.*, "Language Models are Few-Shot Learners," in *NeurIPS*, 2020, pp. 1877–1901.

[5] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *NeurIPS*, vol. 35, 2022, pp. 27 730–27 744.

[6] OpenAI, "GPT-4 Technical Report," *arXiv:2303.08774*, 2023.

[7] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang, "Sparks of Artificial General Intelligence: Early experiments with GPT-4," *arXiv:2303.12712*, 2023.

[8] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, and H. H. Awadalla, "How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation," *arXiv:2302.09210*, 2023.

[9] J. Li, H. Li, Z. Pan, and G. Pan, "Prompt ChatGPT In MNER: Improved multimodal named entity recognition method based on auxiliary refining knowledge from ChatGPT," *arXiv:2305.12212*, 2023.

[10] M. M. Amin, E. Cambria, and B. W. Schuller, "Will Affective Computing Emerge from Foundation Models and General AI? A First Evaluation on ChatGPT," *IEEE Intelligent Systems*, vol. 38, no. 2, pp. 15–23, 2023.

[11] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York City, NY, USA: Springer, 2006.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *NeurIPS*, Lake Tahoe, NV, USA, 2013, pp. 1–11.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv:1907.11692*, 2019.

[14] E. Cambria, R. Mao, M. Chen, Z. Wang, and S.-B. Ho, "Seven pillars for the future of artificial intelligence," *IEEE Intelligent Systems*, vol. 38, no. 6, pp. 62–69, 2023.

[15] E. Cambria, "Affective Computing and Sentiment Analysis," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 102–107, 2016.

[16] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. New York City, NY, USA: John Wiley & Sons, 2013.

[17] R. Mao and X. Li, "Bridging Towers of Multi-task Learning with a Gating Mechanism for Aspect-based Sentiment Analysis and Sequential Metaphor Identification," *Proceedings of AAAI*, vol. 35, pp. 13 534–13 542, 2021.

[18] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 Task 4: Aspect Based Sentiment Analysis," in *Proceedings of SemEval 2014*, 2014, pp. 27–35.

[19] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "SemEval-2015 Task 12: Aspect Based Sentiment Analysis," in *Proceedings of SemEval 2015*, 2015, pp. 486–495.

[20] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," *CS224N project report, Stanford*, vol. 1, no. 12, p. 2009, 2009.

[21] K. Cortis, A. Freitas, T. Daudert, M. Huerlimann, M. Zarrouk, S. Handschuh, and B. Davis, "SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News," in *Proceedings of SemEval-2017*, 2017, pp. 519–535.

[22] S. Mohammad and F. Bravo-Marquez, "WASSA-2017 Shared Task on Emotion Intensity," in *Proceedings of WASSA 2017*, 2017, pp. 34–49.

[23] V. Desu, N. Komati, S. Lingamaneni, and F. Shaik, "Suicide and Depression Detection in Social Media Forums," in *Smart Intelligent Computing and Applications, Volume 2*. Singapore, Singapore: Springer Nature Singapore, 2022, pp. 263–270.

[24] cjadams, J. Sorensen, J. Elliott, L. Dixon, M. McDonald, and C. W. nithum and, "Toxic Comment Classification Challenge," 2017. [Online]. Available: https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge

[25] A. Rastogi, Q. Liu, and E. Cambria, "Stress Detection from Social Media Articles: New Dataset Benchmark and Analytical Study," in *IJCNN*, 2022, pp. 1–8.

[26] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "Chalearn lap 2016: First Round Challenge on First Impressions - Dataset and Results," in *ECCV*, Cham, Switzerland, 2016, pp. 400–418.

[27] A. K. Jayaraman, T. E. Trueman, G. Ananthakrishnan, S. Mitra, Q. Liu, and E. Cambria, "Sarcasm Detection in News Headlines using Supervised Learning," in *International Conference on Artificial Intelligence and Data Engineering (AIDE)*. Karkala, India: IEEE, 2022, pp. 288–294.

[28] R. Misra and P. Arora, "Sarcasm Detection using News Headlines Dataset," *AI Open*, vol. 4, pp. 13–18, 2023.

[29] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," in *Proceedings of ACL*, 2004, pp. 271–278.

[30] E. Cambria, X. Zhang, R. Mao, M. Chen, and K. Kwok, "SenticNet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing," in *International Conference on Human-Computer Interaction (HCII)*, 2024.

[31] W. Li, L. Zhu, R. Mao, and E. Cambria, "SKIER: A Symbolic Knowledge Integrated Model for Conversational Emotion Recognition," *Proceedings of AAAI*, vol. 37, pp. 13 121–13 129, 2023.

[32] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1743–1753, 2023.

[33] T. Yue, R. Mao, H. Wang, Z. Hu, and E. Cambria, "KnowleNet: Knowledge fusion network for multimodal sarcasm detection," *Information Fusion*, vol. 100, p. 101921, 2023.

[34] L. Zhu, W. Li, R. Mao, V. Pandelea, and E. Cambria, "PAED: Zero-Shot Persona Attribute Extraction in Dialogues," in *Proceedings of ACL*, 2023, pp. 9771–9787.

[35] S. Han, R. Mao, and E. Cambria, "Hierarchical Attention Network for Explainable Depression Detection on Twitter Aided by Metaphor Concept Mappings," in *Proceedings of COLING*, 2022, pp. 94–104.

[36] R. Mao, C. Lin, and F. Guerin, "Word Embedding and WordNet Based Metaphor Identification and Interpretation," in *Proceedings of ACL*, 2018, pp. 1222–1231.

[37] R. Mao, X. Li, K. He, M. Ge, and E. Cambria, "MetaPro Online: A Computational Metaphor Processing Online System," in *Proceedings of ACL (System Demonstrations)*, 2023, pp. 127–135.

[38] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing, "Sentiment Analysis in the Era of Large Language Models: A Reality Check," *arXiv:2305.15005*, 2023.

[39] Z. Wang, Q. Xie, Z. Ding, Y. Feng, and R. Xia, "Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study," *arXiv:2304.04339*, 2023.

[40] Y. Ji, W. Wu, H. Zheng, Y. Hu, X. Chen, and L. He, "Is ChatGPT a Good Personality Recognizer? A Preliminary Study," *arXiv:2307.03952*, 2023.

[41] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz *et al.*, "ChatGPT: Jack of all trades, master of none," *Information Fusion*, p. 101861, 2023.

[42] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, "Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT," *arXiv:2302.10198*, 2023.

[43] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, "Is ChatGPT a General-Purpose Natural Language Processing Task Solver?" *arXiv:2302.06476*, 2023.

[44] M. M. Amin, E. Cambria, and B. W. Schuller, "Can ChatGPT's Responses Boost Traditional Natural Language Processing?" *IEEE Intelligent Systems*, vol. 38, no. 5, pp. 5–11, 2023.

[45] R. Mao, G. Chen, X. Zhang, F. Guerin, and E. Cambria, "GPTEval: A survey on assessments of ChatGPT and GPT-4," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024.

[46] M. M. Amin and B. W. Schuller, "On Prompt Sensitivity of ChatGPT in Affective Computing," *arXiv:2403.14006*, 2024.

[47] B. Chen, S. Escalera, I. Guyon, V. Ponce-López, N. Shah, and M. Oliu Simón, "Overcoming Calibration Problems in Pattern Labeling with Pairwise Ratings: Application to Personality Traits," in *Computer Vision – ECCV 2016 Workshops*, 2016, pp. 419–432.

[48] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *ICLR*, San Diego, CA, USA, 2015, pp. 1–13.

[49] M. Lindauer, K. Eggensperger, M. Feurer, A. Biedenkapp, D. Deng, C. Benjamins, T. Ruhkopf, R. Sass, and F. Hutter, "SMAC3: A Versatile Bayesian Optimization Package for Hyperparameter Optimization," *Journal of Machine Learning Research*, vol. 23, no. 54, pp. 1–9, 2022.

[50] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL*, 2019, pp. 4171–4186.

[51] D. Faraj and M. Abdullah, "SarcasmDet at SemEval-2021 Task 7: Detect Humor and Offensive based on Demographic Factors using RoBERTa Pre-trained Model," in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Online: Association for Computational Linguistics, 2021, pp. 527–533.

[52] J. Barnes, "Sentiment and Emotion Classification in Low-resource Settings," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 290–304.

[53] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network," *IEEE Access*, vol. 10, pp. 21 517–21 525, 2022.

[54] N. A. Semary, W. Ahmed, K. Amin, P. Pławiak, and M. Hammad, "Improving sentiment classification using a RoBERTa-based hybrid model," *Frontiers in Human Neuroscience*, vol. 17, 2023.

[55] H. Huo and M. Iwaihara, "Utilizing BERT pretrained models with various fine-tune methods for subjectivity detection," in *Web and Big Data: 4th International Joint Conference, APWeb-WAIM 2020, Part II 4*. Tianjin, China: Springer, 2020, pp. 270–284.

[56] Z. Chen and T. Qian, "Relation-Aware Collaborative Learning for Unified Aspect-Based Sentiment Analysis," in *Proceedings of ACL*, Online, 2020, pp. 3685–3694.

[57] H. J. Escalante, H. Kaya, A. A. Salah, S. Escalera, Y. Güçlütürk, U. Güçlü, X. Baró, I. Guyon, J. C. J. Junior, M. Madadi *et al.*, "Modeling, Recognizing, and Explaining Apparent Personality From Videos," *Transactions on Affective Computing*, vol. 13, no. 2, pp. 894–911, 2020.

[58] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proceedings INTERSPEECH*, 2013, pp. 148–152.

[59] P. Good, *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York City, NY, USA: Springer, 1994.

[60] P. Rozin and E. B. Royzman, "Negativity Bias, Negativity Dominance, and Contagion," *Personality and Social Psychology Review*, vol. 5, no. 4, pp. 296–320, 2001.

[61] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," *arXiv:2307.09288*, 2023.

[62] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7B," *arXiv:2310.06825*, 2023.

Mostafa M. Amin is currently working toward the Ph.D. degree with the Chair of Health Informatics at the Technical University of Munich (formerly the Chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg), while working as Senior Research Data Scientist at SYNCPILOT GmbH in Augsburg, Germany. His research interests include Affective Computing, Audio and Text Analytics. He received a M.Sc. degree in Computer Science from the University of Freiburg, Germany. Contact him at mostafa.mohamed@uni-a.de

Rui Mao is a research fellow and lead investigator at Nanyang Technological University. He obtained his Ph.D. degree in Computing Science from the University of Aberdeen. His research interest lies at NLP, affective computing, and their applications in finance and cognitive science. He and his funded company (Ruimao Tech) have developed an end-to-end system (MetaPro) for computational metaphor processing and a neural search engine (wensousou.com) for searching Chinese ancient poems with modern language. Contact him at rui.mao@ntu.edu.sg.

Erik Cambria is a professor of Computer Science and Engineering at Nanyang Technological University, Singapore. His research focuses on neurosymbolic AI for explainable natural language processing in domains like sentiment analysis, dialogue systems, and financial forecasting. He is an IEEE Fellow and a recipient of several awards, e. g., IEEE Outstanding Career Award, was listed among the AI's 10 to Watch, and was featured in Forbes as one of the 5 People Building Our AI Future. Contact him at cambria@ntu.edu.sg.

Björn W. Schuller is currently a professor of Artificial Intelligence with the Department of Computing, Imperial College London, UK, where he heads the Group on Language, Audio, & Music (GLAM). He is a full professor and the head of the Chair of Health Informatics at the Technical University of Munich, Germany. He is also with the Munich Data Science Institute (MDSI), the Munich Center for Machine Learning (MCML), and the Konrad Zuse School of Excellence in Reliable AI (relAI), all in Munich, Germany. Additionally, he is a guest professor and a former head of the Chair of Embedded Intelligence for Health Care and Wellbeing with the University of Augsburg, Germany. He is the Founding CEO and current Chief Scientific Officer of audEERING. He is an IEEE Fellow alongside other Fellowships. Contact him at schuller@tum.de.