2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII) | 979-8-3503-2743-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ACII59096.2023.10388123

Robustness Analysis uncovers Language Proficiency Bias in Emotion Recognition Systems

Quynh Tran
PricewaterhouseCoopers GmbH
10117 Berlin, Germany
quynh.t.tran@pwc.com

Krystsina Shpileuskaya
PricewaterhouseCoopers GmbH
10117 Berlin, Germany
krystsina.shpileuskaya@pwc.com

Elaine Zaunseder
Heidelberg University
69120 Heidelberg, Germany
elaine.zaunseder@uni-heidelberg.de

Josef Salg
PricewaterhouseCoopers GmbH
60327 Frankfurt am Main, Germany
josef.salg@pwc.com

Larissa Putzar
Hamburg University of Applied Sciences
22081 Hamburg, Germany
larissa.putzar@haw-hamburg.de

Sven Blankenburg
PricewaterhouseCoopers GmbH
10117 Berlin, Germany
sven.blankenburg@pwc.com

***Abstract*—Emotion recognition in conversations (ERC) has rapidly emerged as a vital instrument in enhancing human-computer interactions. However, concerns about the fairness and biases of these ERC systems persist and remain to be addressed by assessing their robustness. This study presents a methodology to analyze the robustness and bias of an ERC system by including complexities of user input with varying English language proficiency. We develop a novel, hybrid approach to create text perturbations by combining natural language generation techniques with rule-based constraints to simulate language proficiency levels. Specifically, we utilize the capabilities of GPT-3 to generate text modifications based on language proficiency characteristics introduced by the internationally recognized Common European Framework of Reference (CEFR). Based on the application of the widely-used COSMIC model, our robustness analysis discloses that the ERC system's performance decreased as language proficiency diminished. Hence, this study demonstrates the presence and implications of language proficiency bias in ERC systems, resulting in discriminatory consequences for non-native English speakers. Overall, our perturbation exhibits versatility for diverse analysis objectives. For instance, it allows investigating gender bias and examining unilateral linguistic bias involving native and non-native speakers. By making our implementation publicly accessible, we aim to foster the advancement of fair ERC systems.**

***Index Terms*—Affective Computing, Emotion Recognition, Bias, Robustness, Artificial Intelligence, Generative AI, GPT-3**

I. Introduction

Emotions play a pivotal role in personal communication [1], which has become even more apparent in our increasingly digital world [2]. The COVID-19 pandemic accelerated this shift, transforming business meetings, educational classes, and medical consultations into virtual experiences without face-to-face communication [3]. These new means of communication pose new challenges to people in terms of emotional stress and could be detected by artificial intelligence (AI) systems to identify the emotional condition of individuals. AI affects various aspects of human life, such as emotion recognition in conversations (ERC), presented in speech and facial expressions [4], written text [5], gestures [6] and biometric information [7]. ERC systems can leverage interactions by utilizing machines to interpret, process, and respond to human emotions. For example, emotionally aware AI can facilitate mental health support [8], customer service [9], and education [10] by tailoring responses to individual emotional states. In virtual conversations, cameras are often absent due to privacy concerns or technical limitations, making ERC more challenging and increasing the reliance on textual input. This input can include grammatical inaccuracies and typographical variabilities, for example from non-native speakers or dialects like African American English, which have experienced racial bias in hate speech detection systems [11], [12]. Thus, it is crucial to examine the impact of an individual's language proficiency on ERC and determine whether linguistic proficiency acts as a discriminatory element. Consequently, ensuring the fairness of ERC requires assessing their robustness to inputs from varying language proficiencies. Levels for linguistic proficiency can be categorized according to the internationally recognized CEFR (Common European Framework of Reference for Languages). To model individuals' language proficiency, we present a hybrid-based perturbation that applies the CEFR level descriptions and the *GPT-3* (Generative Pre-trained Transformer 3) model [13] by OpenAI. GPT-3, a prominent large language model, is renowned for its high capabilities in various natural language generation (NLG) tasks, such as text summarization [14] and question-answering [15]. In this study, we leverage GPT-3 to mimic different CEFR language proficiency levels using dialogues from the IEMOCAP dataset [16] for emotion detection to investigate the robustness of the COSMIC system [5] towards potential bias related to language proficiencies.

II. Related Work

Investigations on bias and fairness of machine learning (ML) models are an ongoing research topic. This is of utmost importance as ML systems in the context of ERC often process sensitive private data as in the case of empathetic embodied agents [17] or therapeutic dialogue systems [8]. It has been shown that bias often exists within the generation process of datasets for training affective computing systems [12].

979-8-3503-2743-4/23/$31.00 ©2023 IEEE

Hence, choosing an appropriate, unbiased dataset for ERC systems is of high importance. Dialogue datasets that are frequently used are, for instance, MELD [18], DailyDialog [19] and IEMOCAP [20] [8]. Whereas, the latter one has shown to be the most common and established [20]. Prior studies examined ML models' ability to detect language proficiency of written texts [21], [22]. It has been shown that such systems are prone to bias that can originate from the model or the training data [23]. Bhardwaj et al. [24] demonstrated the presence of gender bias in BERT, and Ballier et al. [21] revealed lexical bias in the EFCAMDAT dataset [25] commonly used for language proficiency investigation. Modifying the test data is an effective tool for investigating a model's generalization ability. Tran et al. [26] demonstrated realistic natural language perturbations, mimicking scanning, typing, and speech recognition errors occurring in text inputs to evaluate ML models' robustness. Further, Ohashi et al. [27] implemented an NLG module for generative adaptive utterances simulation of speakers with low vocabulary levels and modelled speech recognition errors caused by background noises aiming to increase the system's robustness. Additionally, capabilities of large NLG models like GPT-3 have been used, e.g. to study the ability to detect hate speech [28].

Our Contribution In this research, we develop a novel method to investigate language proficiency bias in ERC systems. We perform realistic language proficiency transformations utilizing GPT-3 and based on the international CEFR standards. Finally, we apply our method on a state-of-the-art (SOTA) benchmark ERC system to investigate the impact of language proficiency on the model's robustness.

III. Material and Methods

In this section, we introduce our novel perturbation technique (Sec. III-C), underlying our robustness analysis. We present details on our methodologies, including the utilized data (Sec. III-A), the investigated ERC system (Sec. III-B) and its specifications as well as used metrics (Sec. III-D).

A. Datasets

We conduct our experiments on a benchmark dataset for ERC in English: The Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [20]. The IEMOCAP text corpus was created by recording ten actors engaging in spoken acted and spontaneous communication sessions, which were subsequently transcribed. These conversations involve paired interactions between a female and male participant about everyday life subjects. In total, 151 dialogues were recorded, amounting to 7,433 utterances. Each utterance has been labelled by multiple annotators with an emotion, i.e. *anger*, *happiness*, *sadness*, *frustration*, *excitement* or *neutral state*. These labels aim to capture the emotional expression conveyed through the content and context of each utterance. Throughout our research, we employ 70-20-10 train-test-validation data splittings, using the test dataset to investigate potential English language proficiency bias effects of the trained model. In addition, we generate ten independent random realizations of train-test-validation samples to ensure a balanced distribution of all emotions. A comprehensive analysis of the dataset splitting and the corresponding results can be found in Sec. IV-A.

B. Models

1) Emotion Recognition System: Among the wide variety of ERC systems in dialogues, we employ the well-known ML-based framework COSMIC (COmmonSense knowledge for eMotion Identification in Conversations) [5] as a representative model for our study[1]. COSMIC achieves high results for emotion detection of utterances on a range of benchmark conversational datasets and is, alongside models like RoBERTA [29] and ROBERTA+DialogueRNN [30], the best-performing ERC system on the IEMOCAP dataset with an accuracy of 65.28% [5]. Notably, COSMIC addresses the task of identifying emotions at the utterance level in text-based conversations by leveraging a large commonsense knowledge base to discern a speaker's and listener's reactions, effects, and intentions. COSMIC comprises three submodules: *a) Context Independent Feature Extraction* employing a fine-tuned RoBERTA model [29] with 24 layers, 16 self-attention heads per block, a hidden dimension of 1024, resulting in overall 355 million parameters, *b) Commonsense Feature Extraction* utilizing the commonsense transformer model COMET [31], based on the pre-trained language model GPT [32] and *c) Commonsense Conversational Model* combining an internal, external and intent state to model different mental states, actions and events of each speaker as well as a context and emotion state to determine the respective context and mood of the speaker and utterance. For further details on the implementation of COSMIC, we refer to the original GitHub repository, available at https://github.com/declare-lab/conv-emotion.

2) GPT-3: GPT-3 is a SOTA language model developed by OpenAI. It is based on the transformer architecture and consists of 175 billion parameters [13]. It uses unsupervised learning through self-attention to learn the statistical structure of language and generate contextually relevant responses [13]. Trained on more than 45TB of text, GPT-3 generates human-like text. The training process involves masked language modelling, where parts of the input text are hidden, and the model predicts the subsequent text from the input, i.e. the missing tokens [33]. This method enables the model to handle varying input lengths and adapt predictions based on the given context. GPT-3 has showcased remarkable performance in various NLG tasks due to its few-shot learning capabilities [13]. Few-shot learning allows the model to learn and perform tasks with minimal task-specific training, relying on its extensive pre-training to generate appropriate responses.

[1]Experiments were implemented in Pytorch, run on two NVIDIA RTX 3090 GPUs, and took approx. 120 hours to complete the training of 10 models with early stopping on the validation dataset.

C. Language Proficiency Perturbation

In this work, we present a novel technique to model the language proficiency of text realistically. We apply rule-based approaches by using explicit linguistic characteristics based on language proficiency levels introduced by the CEFR. Additionally, we integrate generative AI in our perturbation, specifically the GPT-3 language model introduced by OpenAI to create natural text. Particularly, we apply the effective text generative model *text-davinci-003* and utilize the OpenAI API to submit textual requests (prompts) to the model. To attain reduced linguistic proficiency, the input text is simplified, with the optional inclusion of mistakes for the specific proficiency level. Consequently, eight proficiency levels are represented, ranging from A1 (basic) to B2 (independent), along with the corresponding proficiency levels with typical mistakes: from A1$^-$ (A1 with typical mistakes) to B2$^-$ (B2 with typical mistakes). Fig. 1 displays the output of the language proficiency perturbation applied on an example utterance for CEFR level A1 (low proficiency) and B1 (intermediate proficiency), without and with mistakes (in orange).

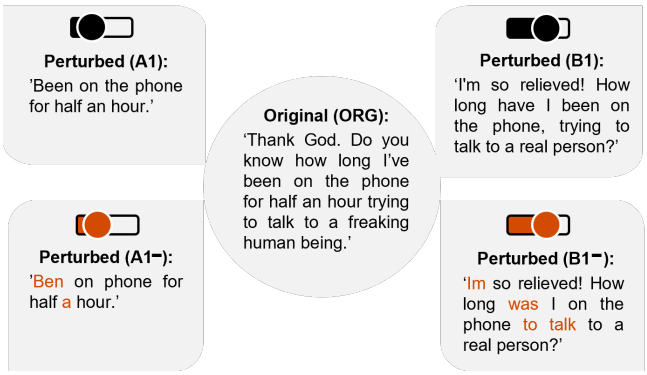

Fig. 1. Language Proficiency Perturbation applied on original utterance. Output for CEFR level A1 and B1 along with A1$^-$ and B$^-$ showing the corresponding language levels including typical mistakes (in orange).

Table I displays the prompts submitted to GPT-3 for the corresponding CEFR levels. The used language proficiency attributes are based on the CEFR descriptions [34] and related studies [35], [36] that analyze different qualitative and quantitative features of language proficiency. For instance, the MERLIN annotation scheme [36] identifies language proficiency across multiple characteristics, including sentence length, structure, tense and vocabulary size. Based on this, Alexopoulou et al. [35] examined the typical sentence length and used verb tenses for CEFR levels on the EFCAMDAT dataset [25]. We apply these insights and findings as the foundation for our language proficiency perturbation. Additionally, we use the information given by the CEFR [34] on learners' sentence structure and vocabulary size to describe their language ability. The output text becomes more prolonged, intricate, and sophisticated as the language proficiency level increases. This results in lengthier sentences, more complex sentence structure, varying usage of verb tenses and a higher range of vocabulary. In addition to the typical CEFR level attributes such as sentence length and structure, grammar, and vocabulary selection, we optionally incorporate typical mistakes into the transformed text. We include errors based on the results of Alexopoulou et al. [35] that identified different types of errors, such as orthographic and grammatical mistakes and improper word selection for specific CEFR levels.

TABLE I
CEFR ATTRIBUTES THAT TARGET VARIOUS LANGUAGE PROFICIENCY ASPECTS, INCLUDING SENTENCE LENGTH (SL), SENTENCE STRUCTURE (SS), TENSES (T) AND VOCABULARY (V).

CEFR PROMPT DESCRIPTION		
CEFR	WITHOUT MISTAKES (BASELINE) [34]	WITH MISTAKES (BASELINE+ERRORS) [35]
A1	* SL: 5-7 words * SS: Very simple subject-verb-object structure * T: Simple present, present continuous, past simple, future simple * V: Greetings, basic questions and answers	* Confuse singular and plural nouns * Confuse adjectives and adverbs * Misuse subject-verb agreement * Neglect articles
A2	* SL: 7-10 words * SS: Simple subject-verb-object structure * T: Simple present, present continuous, present perfect, past simple, future simple * V: Everyday expressions, simple topics	* Confuse adjectives and adverbs * Misuse subject-verb agreement
B1	* SL: 10-15 words * SS: Complex sentences with more varied structure * T: All tenses	* Misplace modifiers * Confuse conjunctions * Incorrectly construct passive voice sentences * Make incorrect verb tense choices
B2	* SL: 15-20 words * SS: Complex sentences with more varied structure * T: All tenses	* Misplace modifiers * Confuse conjunctions * Incorrectly construct passive voice sentences

In order to understand the concept of our perturbation, we present a pipeline in Fig. 2 that visualises our procedure from data preprocessing of utterances in dialogues to the modification of language proficiency and incorporation of typical mistakes. The text perturbation is executed incrementally. Step a) to d) in Fig. 2 describe the data preprocessing part wherein the content is formatted to provide compatible and comprehensible input for the GPT-3 model. The conversations in the test dataset (Fig. 2a) serve as a basis for examining language proficiency bias in a trained emotion recognition model. Based on the indications in the IEMOCAP dataset, whether an utterance involves a female or male speaker, the names 'Alice' and 'Bob' are assigned to each utterance (Fig. 2b) to differentiate between the speakers accordingly. After distinguishing Alice and Bob's speech components, which allow separate perturbation of each speaker, every utterance is sequentially numbered (Fig. 2c) to record the sequence of utterances for each speaker. To overcome current challenges posed by using large text parts as input in the GPT-3 model,

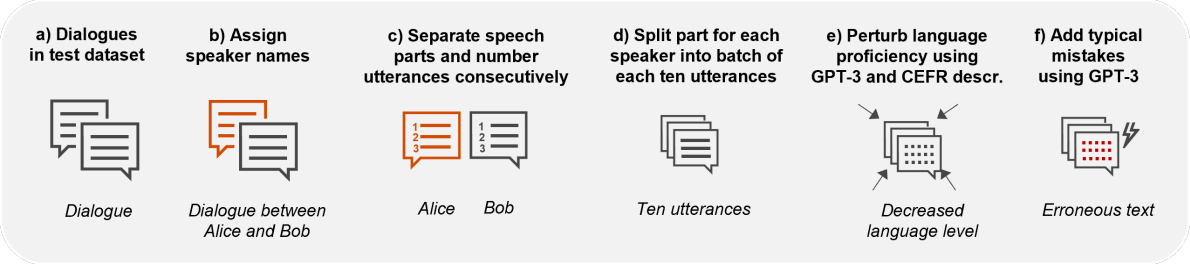

Fig. 2. Pipeline to alter the Language Proficiency of utterances in dialogues using GPT-3. An example of the output in step e) and f) is given in Fig. 1.

which exceed the 4097-token limit shared between prompt and completion, we divide each speaking part into batches of ten utterances for every speaker (Fig. 2d). This batch serves as input to perturb the language proficiency (Fig. 2e) through the following prompt request:

> **Original numbered utterances**
> Transform every utterance above from **Start number** to **End number** with the following characteristics: **CEFR description**

The initial prompt of the request **Original numbered utterances** serves as a placeholder for the input text comprising ten utterances. The second prompt includes three placeholders **Start number**, **End number** and **CEFR description**. Both **Start number** and **End number** indicate the beginning and end, facilitating the perturbation of each utterance individually. The **CEFR description** corresponds to the CEFR characteristics given in Tab. I for the language proficiency level. After applying the perturbation[2], the output text is simplified according to the characteristics of the proficiency level. Despite modifications in sentence length, structure, grammar, tense, and vocabulary, the semantic content of each utterance remains consistent (Fig. 1). To further explore the impact of linguistic errors, characteristic mistakes for the proficiency level are optionally incorporated into the previously modified text (Fig. 2f). To assess the effectiveness of our perturbation technique, we employ a pre-trained RoBERTA model [29] that utilizes supervised learning to estimate the CEFR levels of essays written by students within the EFCAMDAT dataset [25] achieving a 96% test accuracy. Predominately, the trained model classifies the modified texts into the intended CEFR levels, thus verifying the effectiveness of our perturbation. Furthermore, a manual qualitative assessment of a data sample was carried out, confirming the desired perturbation outcomes.

D. Evaluation Metrics

1) Model's Robustness Measure: The F1-score is a harmonic mean of precision and recall, two widely used metrics for evaluating the performance of classification systems. The F1-score is defined as [37]:

$$\text{F1-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \tag{1}$$

[2]Experiments were implemented in Python and took ten hours to create 59,464 $(7,443 \times 8)$ utterances depending on the availability of GPT-3.

The F1-score takes values between 0 and 1, with one being the highest possible score. In this study, the F1-score is chosen as a preferred evaluation metric for several reasons:

- Imbalanced datasets concerning emotion distribution,
- Importance of false positives and false negatives,
- Multi-class classification.

For the reasons mentioned, the F1-score is preferable to other measures such as accuracy, precision, or ROC-curve [37]. We use the F1-score with respect to variations of the test data to measure the model's robustness regarding those modifications.

2) Text Similarity Measures: For our analysis, we apply the BLEU [38] (Bilingual Evaluation Understudy) and ROUGE [39] (Recall-Oriented Understudy for Gisting Evaluation) metrics to investigate the effects of our developed language proficiency 'translation'. In practice, both BLEU and ROUGE are widely used and often complement each other in evaluating text summarization [40] and machine translation [41] performance to measure the quality of machine-generated text against human-generated references. BLEU and ROUGE range from 0 to 1, where one perfectly matches the reference text.

IV. Results

In this section, we present the numerical results regarding the impact of varying language proficiency levels on the COSMIC system's robustness. First, we describe the results of generating independent realizations of train-test-validation splittings of the IEMOCAP dialogue dataset (Sec. IV-A). Subsequently, in Sec. IV-B, the ERC model's bias regarding language proficiency is studied by means of so-called robustness curves. Finally, the correlation between the model's robustness and text similarity measures is shown in Sec. IV-C.

A. Realizations of independent train-test-validation splittings

Generating unbiased and statistically independent samples from a specific dataset is of utmost importance to investigate the impact of a model's robustness. A random dataset splitting could result in overrepresenting attributes such as gender, emotion, and tokens in the train, test or validation set. Consequently, we employ Monte Carlo methods [42] to split the dataset into train, test, and validation by minimizing the cross-entropy of the distributions while maintaining balanced emotions between male and female speakers. We aim for a close resemblance between the train, test, and validation

dataset distribution according to the mentioned features (emotion, speakers' gender, etc.). Throughout our modifications, we preserved the dataset's original dialogue structure.

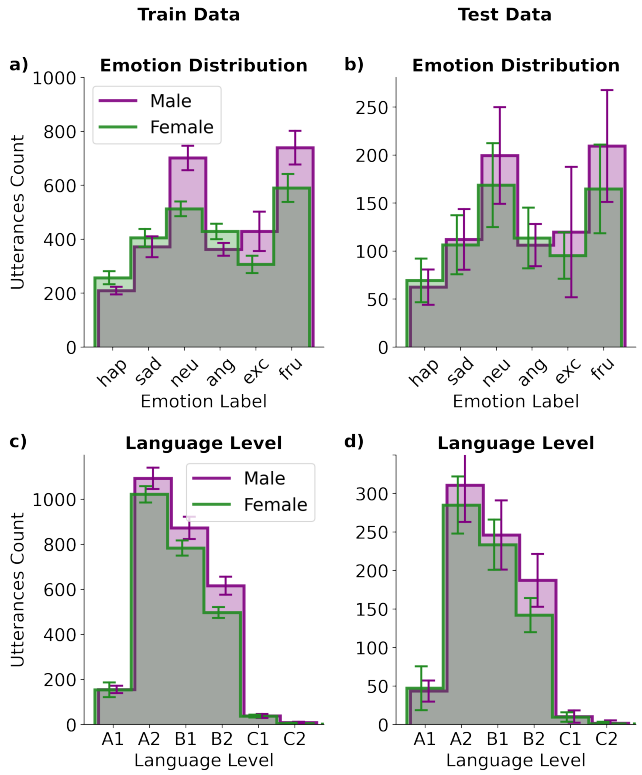

Fig. 3. Statistics of 10 independent realizations of the data splitting regarding Emotion Label and Language Level Distribution for training data (a,c) and test data (b,d). Error bars indicate 2-sigma-confidence interval.

In Fig. 3a-d, the distributions of the train and test dataset derived from ten statistically independent realizations are shown, distinguishing emotion and language level distributions for male (purple) and female (green) speakers. The error bars indicate the 2-sigma confidence interval, i.e. 4.55% error probability. First, it is observable that the distributions across the datasets exhibit a high degree of similarity, implying that the training and test datasets accurately represent one another (cf. Fig. 3a, c and Fig. 3b, d). Second, the emotion distribution among utterances of male and female speakers remains consistent for each dataset. To conclude, we generated unbiased and balanced realizations of the IEMOCAP dataset for a fair robustness analysis of ERC systems.

B. Impact of the Language Proficiency based on entire dialogues on the models' Robustness

We altered all utterances within the test dataset to study the impact of the dialogue partner's language proficiency level (Sec. III-C). We assessed ten independent random realizations of the train-test-validation data as shown in Fig. 3.

We conducted training and testing of the COSMIC model for each data realization, resulting in ten independent training sessions and 80 test sessions (8 language proficiency

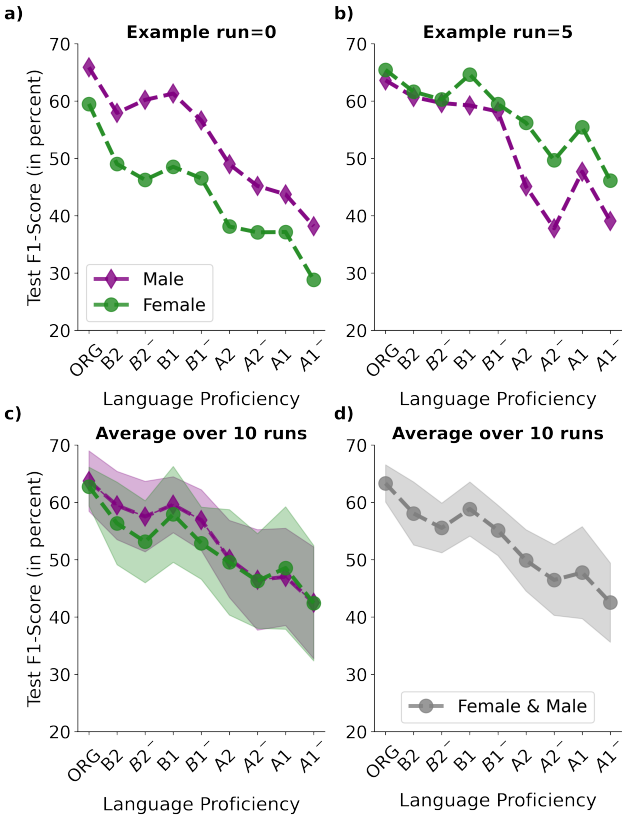

Fig. 4. Robustness Curves influenced by Language Proficiency for male (purple) and female (green) speakers separately for train-test-realization 0 (a) and 5 (b). Averaging over all realizations for male and female speakers separately (c) and for all speakers combined (d).

levels $\times$ 10 trained models). In Fig. 4, the weighted F1-score (Sec. III-D1) for test dialogues across different language proficiency levels for utterances spoken by male (purple) and female (green) speakers is shown for two realizations (runs), i.e. run=0 (Fig. 4a) and run=5 (Fig. 4b). Both realizations (Fig. 4a and b) show the impact of decreasing language proficiency from B2 to A1$^-$ on the system's robustness for male and female speakers. However, in Fig. 4a, the robustness curve for female speakers (green) is systematically lower than for male speakers (purple). This indicates a stronger language proficiency bias for female speakers. Comparing this finding with Fig. 4b, it is observable that the systematic difference between robustness curves for male and female speakers has switched. Thus, the impact of language proficiency level on the ERC system's robustness depends on the specific realization. We averaged across ten independent realizations of the robustness curves to gain statistically significant insights, displayed in Fig. 4c. Here, it can be deduced that the robustness curves for both speaker types are indistinguishable within statistical uncertainties (indicated by the 2-sigma confidence interval, i.e. opaque areas). The robustness curve for male (purple) and female speaker (green) without altering the dialogues (at language proficiency level ORG) is almost identical. For a detailed numerical comparison of the robustness values for

male and female speakers displayed in Fig. 4c, we refer to Tab. II. The robustness curve decreases from 62.8% to 42.4% for female and from 63.7% to 42.5% for male spoken utterances as the language proficiency diminishes (Tab. II).

TABLE II
F1-SCORE AS DESCRIBED IN SEC. IIID. MEAN VALUE AND CONFIDENCE INTERVAL (2-SIGMA) ARE BASED ON 10 INDEPENDENT SIMULATIONS.

LANGUAGE PROF. LEVEL	FEMALE SPEAKER (F1-SCORE)	MALE SPEAKER (F1-SCORE)
ORG	62.8 ± 3.4	63.7 ± 5.3
B2	56.3 ± 7.2	59.5 ± 5.9
B2$^-$	53.1 ± 7.2	57.5 ± 6.1
B1	57.9 ± 8.4	59.6 ± 4.9
B1$^-$	52.9 ± 6.3	56.9 ± 5.3
A2	49.5 ± 9.2	50.1 ± 6.7
A2$^-$	46.3 ± 8.2	46.5 ± 8.8
A1	48.5 ± 10.7	47.0 ± 8.5
A1$^-$	42.4 ± 10.1	42.5 ± 9.7

The confidence interval and the overlap between the two areas in Fig. 4c show no significant difference in the robustness curves between male and female speakers. The robustness curves for both speaker types show weak resonance effects at language levels B1 and A1 (Fig. 4c). One might think that these resonances are due to a close similarity between the train and test datasets at the specific language level. However, the resonance effects cannot be explained by a dominance of specific language proficiencies, particularly B1 and A1, within the training dataset (cf. Fig. 3c with Fig. 4c). Therefore, we investigate those resonances in more detail in Sec. IV-C.

Averaging the robustness curves over male and female speakers and ten independent realizations is shown in Fig. 4d. Here, the negative impact of decreasing language proficiency on the robustness curve is shown, as well as the observed resonance effects at language proficiency level B1 and A1. Thus, the result of diminishing language proficiency on the entire dialogue corresponds to a decrease in the model's robustness as measured by the F1-score. Although a bias related to language proficiency is observed, no statistically significant gender bias is detected (Fig. 4c). However, the ERC system is trained with the speaker's gender as a feature. To understand the impact of language proficiency on specific emotions, we examine the robustness curves for each emotion label separately.

In Fig. 5a-f, the robustness curves are displayed for female (green) and male (purple) spoken utterances for each emotion label. Apart from sadness, the robustness differences between male and female speakers are statistically insignificant. Interestingly, the robustness of the ERC system is emotion-specific. For happiness and neutral (Fig. 5a, f), the effect of language proficiency remains nearly constant. In contrast, for emotions such as excitement and anger, frustration and sadness (Fig. 5b-e), the effect of decreasing language proficiency level is accompanied by a decline in robustness indicating the model's bias for those emotions.

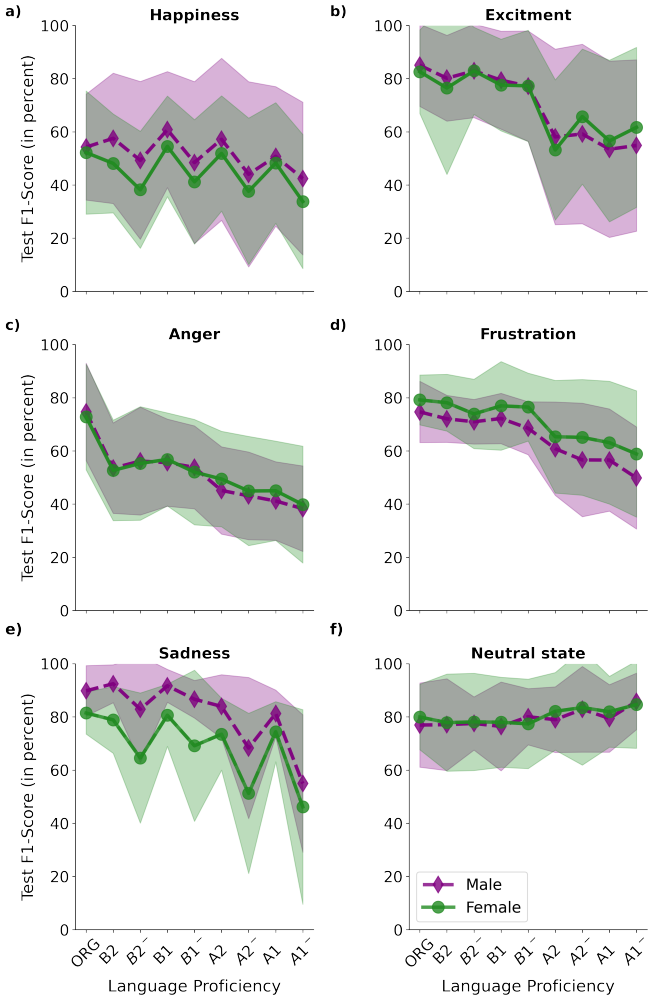

Fig. 5. Language Proficiency Robustness Curves for different emotion labels. Shown are F1-scores with respect to the Language Proficiency for utterances spoken by males (purple) and females (green) for emotion labels: Happiness (a), Excited (b), Anger (c), Frustration (d), Sadness (e), and Neutral (f).

C. *Correlation between Robustness and Text Similarity Scores.*

Having demonstrated that the robustness of ERC systems depends on the language proficiency level, we investigate a possible underlying mechanism in this section. First, we compare the model's robustness with well-known text-similarity measures (as defined in Sec. III-D2). For example, the effect of decreasing robustness might be correlated with a decreased text similarity, i.e. data-specific, between the original and the language proficiency-altered dialogues. In Fig. 6, the scaled BLEU (blue) and ROUGE scores (red) between the original and altered dialogues are shown alongside with the models' overall robustness (grey) as measured by the test F1-score, i.e. model-specific (Fig. 6a) and separately (Fig. 6b).

As observed in Fig. 6a for intermediate language proficiency (B-level), the robustness curve shows the same qualitative behaviour as the text similarity measures (Sec. III-D2). Thus, we infer that the model's robustness decline is due to a decrease in the similarity compared with the training data.

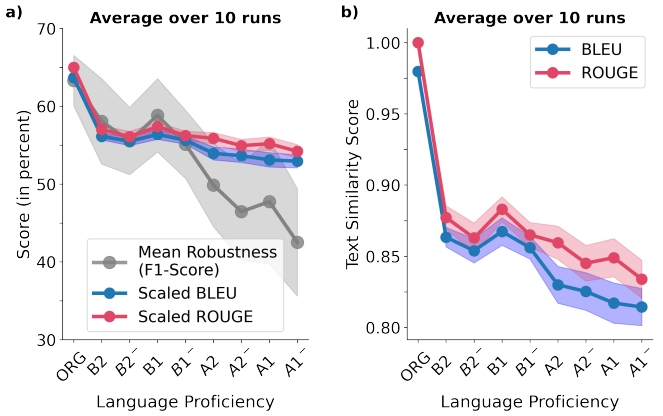

Fig. 6. Comparison between Language Proficiency Robustness Curves and text similarity measures: scaled BLEU (blue) and ROUGE score (red) multiplied with factor 65 with respect to language proficiency level of utterances (a) and unscaled text similarity scores (b).

However, for intermediate to low language proficiencies (A-level), the model's performance decreases, even though the text similarity measures remain almost constant (Fig. 6a). Consequently, based on the text similarity, the model's robustness decreases substantially more than expected for low language proficiency. This observation indicates that the substantial decrease is more model-specific (measured by F1-score) than data-driven (measured by text similarity).

Additionally, in Fig. 6b, it is evident that the text similarity curve of the BLEU and ROUGE score shows a resonance at level B1 similar to the F1-score (Fig. 6a, grey curve) of the ERC model. However, only the ROUGE score displays the second resonance at level A1. Thus, we can conclude that the performance dependency on language proficiency for the ERC system is due to discrepancies between training and perturbed test data at higher language levels (B-level). In contrast, for intermediate and low language proficiencies (A-level), the model's performance decreases much more substantially than anticipated by the text similarity measures.

V. Summary and Discussion

In this study, we developed a novel, hybrid approach to create text perturbations by combining GPT-3 with rule-based constraints to assess the robustness of emotion recognition systems regarding language proficiency. Specifically, we developed a machine learning pipeline which transforms a given text into different language proficiency levels ranging from intermediate to low without changing the semantic content. This modification is based on CEFR proficiency descriptions and GPT-3. Our perturbation is applied in this study to investigate performance bias in emotion recognition systems for conversations regarding the language proficiency of the speakers. Furthermore, the presented approach allows us to study model agnostic effects of dialogue speakers' language proficiency. We applied our approach to the COSMIC model and the IEMOCAP dialogue dataset for emotion recognition. We showed that such systems exhibit a performance bias regarding the language proficiency of dialogue speakers. Although the training dataset entails dialogues with varying language proficiencies, the ERC system's robustness is highly susceptible to diminished language proficiency. Furthermore, when applying the perturbation separately for each speaker, we observed gender-specific robustness differences for female and male speakers solely on single model instances. The presented perturbation mimics differences in language proficiency levels based on grammar, sentence structure, tenses and vocabulary. In future work, our approach could be generalized by modelling the effects of different word pronunciations and considering speakers' social and cultural backgrounds. Additionally, the current perturbation pipeline utilizes GPT-3, but future iterations could use more advanced models, such as GPT-4, to further improve the quality of the perturbation. Future work should examine language proficiency bias in multi-modal emotion recognition where audio and visual input are used. Moreover, exploring bias mitigation techniques to strengthen models against varying language proficiencies should be addressed.

In summary, we have demonstrated significant and discriminatory consequences of language proficiency bias on non-native English speakers using systems for emotion recognition. Consequently, it is crucial to address this issue during the development and evaluation of these models to ensure fair and unbiased systems. This research contributes to fairer output in applications where emotion recognition plays an essential role. Hence, we provide an implementation and propose a pipeline to investigate performance bias induced by language proficiency. Finally, this strengthens the trustworthiness of artificial intelligence in business and society.

Code Availability

A detailed open-source version of the data and Python code implementing our approach and reproducing our numerical benchmark results has been made publicly available at https://github.com/blankenburg-science/ai_language_bias.

Ethical Impact Statement

Investigating language proficiency bias in emotion recognition systems is crucial as these systems significantly impact human-computer interactions. Given the diverse textual input with potential susceptibility to errors of users, influenced by their language proficiency and background, emotion recognition systems could have considerable discrimination consequences for non-native English speakers. By providing a publicly accessible implementation, the study offers a versatile perturbation for replicating text with varying language proficiency to assess potential biases. Ultimately, this study contributes to developing fairer emotion recognition systems and enhancing human-computer interplay.

Acknowledgment

We want to thank the Assurance Digital Team of PwC Germany, particularly Christine Flath, Hans-Peter Dittmar, Rüdiger Loitz and Jan Grabowsky, for supporting our work.

REFERENCES

[1] H. Zhang, A. Jolfaei, and M. Alazab, "A Face Emotion Recognition Method Using Convolutional Neural Network and Image Edge Computing," *IEEE Access*, vol. 7, pp. 159 081–159 089, 2019.

[2] J. Petrova and O. Vasichkina, "Creativity and Emotions in the Digital World," in *Technology, Innovation and Creativity in Digital Society*. Springer, 2022, pp. 512–521.

[3] C. de Las Heras-Pedrosa, P. Sánchez-Núñez, and J. I. Peláez, "Sentiment Analysis and Emotion Understanding during the COVID-19 Pandemic in Spain and Its Impact on Digital Ecosystems," *International Journal of Environmental Research and Public Health*, vol. 17, no. 15, 2020.

[4] J. Zhao, S. Chen, S. Wang, and Q. Jin, "Emotion Recognition using Multimodal Features," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction*. IEEE, 2018, pp. 1–6.

[5] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "COSMIC: COmmonSense knowledge for eMotion identification in conversations," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2470–2481.

[6] J. Guo, "Deep learning approach to text analysis for human emotion detection from big data," *Journal of Intelligent Systems*, vol. 31, no. 1, pp. 113–126, 2022.

[7] N. Novielli, D. Grassi, F. Lanubile, and A. Serebrenik, "Sensor-Based Emotion Recognition in Software Development: Facial Expressions as Gold Standard," in *2022 10th International Conference on Affective Computing and Intelligent Interaction*. IEEE, 2022, pp. 1–8.

[8] A. Zygadlo, "A Therapeutic Dialogue Agent for Polish Language," in *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos*. IEEE, 2021, pp. 1–5.

[9] V. Petrushin, "Emotion in Speech: Recognition and Application to Call Centers," in *Proceedings of Artificial Neural Networks in Engineering*, vol. 710, 1999, p. 22.

[10] W. Wang, K. Xu, H. Niu, and X. Miao, "Emotion Recognition of Students Based on Facial Expressions in Online Education Based on the Perspective of Computer Simulation," *Complexity*, pp. 1–9, 2020.

[11] T. Davidson, D. Bhattacharya, and I. Weber, "Racial Bias in Hate Speech and Abusive Language Detection Datasets," *CoRR*, pp. 25–35, 2019.

[12] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The Risk of Racial Bias in Hate Speech Detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1668–1678.

[13] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language Models are Few-Shot Learners," *ADV NEUR IN*, vol. 33, pp. 1877–1901, 2020.

[14] B. Chintagunta, N. Katariya, X. Amatriain, and A. Kannan, "Medically Aware GPT-3 as a Data Generator for Medical Dialogue Summarization," in *Machine Learning for Healthcare Conference*, 2021, pp. 354–372.

[15] P. Bongini, F. Becattini, and A. Del Bimbo, "Is GPT-3 All You Need for Visual Question Answering in Cultural Heritage?" in *Computer Vision - ECCV 2022 Workshops*. Springer, 2023, pp. 268–281.

[16] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, "M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation," in *Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2022, pp. 4651–4660.

[17] S. DiPaola and O. N. Yalcin, "A multi-layer artificial intelligence and sensing based affective conversational embodied agent," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos*. IEEE, 2019, pp. 91–92.

[18] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations," pp. 527–536, 2019.

[19] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset," pp. 986–995, 2017.

[20] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[21] N. Ballier, S. Canu, C. Petitjean, G. Gasso, C. Balhana, T. Alexopoulou, and T. Gaillat, "Machine learning for learner English: A plea for creating learner data challenges," *International Journal of Rearner Corpus Research*, vol. 6, no. 1, pp. 72–103, 2020.

[22] T. Gaillat, A. Simpkin, N. Ballier, B. Stearns, A. Sousa, M. Bouyé, and M. Zarrouk, "Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach," vol. 34, no. 2. Cambridge University Press, 2022, p. 130–146.

[23] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 108–132, 2023.

[24] R. Bhardwaj, N. Majumder, and S. Poria, "Investigating Gender Bias in BERT," *Cognitive Computation*, vol. 13, no. 4, pp. 1008–1018, 2020.

[25] U. of Cambridge, "English for academic purposes cambridge learner corpus (EFCAMDAT)," Accessed: 2023-03-28, university of Cambridge, Language Technology Lab. [Online]. Available: https://corpus.mml.cam.ac.uk/

[26] Q. Tran, K. Shpileuskaya, E. Zaunseder, L. Putzar, and S. Blankenburg, "Comparing the Robustness of Classical and Deep Learning Techniques for Text Classification," in *2022 International Joint Conference on Neural Networks*, 2022, pp. 1–10.

[27] A. Ohashi and R. Higashinaka, "Adaptive Natural Language Generation for Task-oriented Dialogue via Reinforcement Learning," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 242–252.

[28] K.-L. Chiu, A. Collins, and R. Alexander, "Detecting Hate Speech with GPT-3," *arXiv preprint arXiv:2103.12407*, 2021.

[29] 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 preprint arXiv:1907.11692*, 2019.

[30] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6818–6825.

[31] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "COMET: Commonsense Transformers for Knowledge Graph Construction," in *Association for Computational Linguistics*, 2019.

[32] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving Language Understanding by Generative Pre-Training." OpenAI, 2018.

[33] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Language Models are Unsupervised Multitask Learners." *OpenAI Blog*, vol. 1, no. 8, 2019.

[34] C. of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, 2001.

[35] T. Alexopoulou, M. Michel, A. Murakami, and D. Meurers, "Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing Natural Language Processing techniques," *Language Learning*, vol. 67, pp. 180–208, 2017.

[36] A. Boyd, J. Hana, L. Nicolas, D. Meurers, K. Wisniewski, A. Abel, K. Schöne, B. Stindlová, and C. Vettori, "The MERLIN corpus: Learner language and the CEFR," in *Proceedings of the 9th International Conference on Language Resources and Evaluation*, 2014, pp. 1281–1288.

[37] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008, vol. 39.

[38] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation." in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[39] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. Association for Computational Linguistics, 2004, pp. 74–81.

[40] Y. Graham, "Re-evaluating Automatic Summarization with BLEU and 192 Shades of ROUGE," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 128–137.

[41] S. Stoll, N. C. Camgöz, S. Hadfield, and R. Bowden, "Sign Language Production using Neural Machine Translation and Generative Adversarial Networks," in *Proceedings of the 29th British Machine Vision Conference*, 2018.

[42] S. Mordechai, *Applications of Monte Carlo Method in Science and Engineering*. InTech, 2011.