Natural Language Processing (2024), 1–24
doi:10.1017/nlp.2024.48

ARTICLE

Reliable uncertainty estimation in emotion recognition in conversation using conformal prediction framework

Samad Roohi[1], Richard Skarbez[1] and Hien Duy Nguyen[2,3]

[1]Computer Science and Information Technology, La Trobe University, Melbourne, Victoria, Australia, [2]Department of Mathematical and Physical Science, La Trobe University, Melbourne, Victoria, Australia, and [3]Institute of Mathematics for Industry, Kyushu University, Fukuoka, Fukuoka Prefecture, Japan
Corresponding author: Samad Roohi; Email: s.roohi@latrobe.edu.au

(Received 2 April 2024; revised 13 August 2024; accepted 13 August 2024)

Abstract

Emotion recognition in conversation (ERC) faces two major challenges: biased predictions and poor calibration. Classifiers often disproportionately favor certain emotion categories, such as *neutral*, due to the structural complexity of classifiers, the subjective nature of emotions, and imbalances in training datasets. This bias results in poorly calibrated predictions where the model's predicted probabilities do not align with the true likelihood of outcomes. To tackle these problems, we introduce the application of conformal prediction (CP) into ERC tasks. CP is a distribution-free method that generates set-valued predictions to ensure marginal coverage in classification, thus improving the calibration of models. However, inherent biases in emotion recognition models prevent baseline CP from achieving a uniform conditional coverage across all classes. We propose a novel CP variant, class spectrum conformation, which significantly reduces coverage bias in CP methods. The methodologies introduced in this study enhance the reliability of prediction calibration and mitigate bias in complex natural language processing tasks.

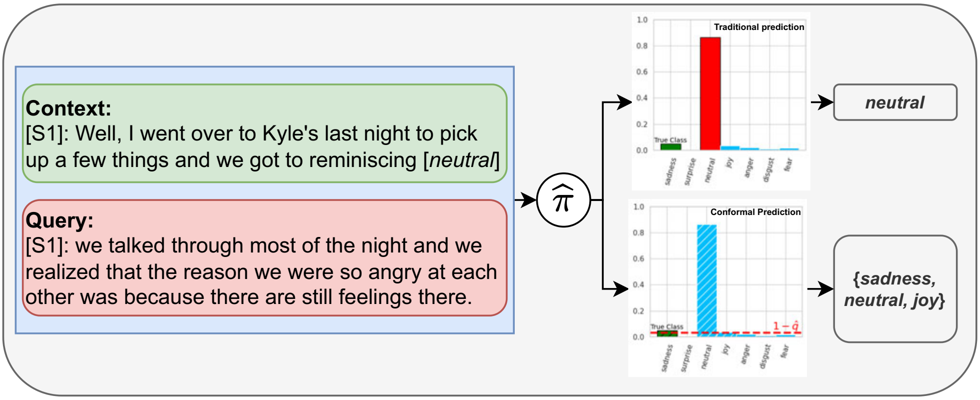
Keywords: Emotion recognition; conversational systems; bias; calibration; uncertainty estimation

1. Introduction

Emotion is a pivotal aspect of natural communication, shaping interpersonal dynamics, driving the flow of conversation, and facilitating the understanding of a speaker's intents, concerns, and desires. A reliable emotion recognition in conversation (ERC) system can enhance user engagement by providing personalized experiences, foster empathetic communication with customers by addressing their emotional needs, and consequently improve their overall satisfaction in customer services (Chen *et al.* 2023). Moreover, it can facilitate the early detection and intervention of emotional distress in mental health applications (Casas *et al.* 2021). An erroneous prediction in high-stake ERC systems can be costly, leading to miscommunication, decreased trust, and potential harm to users (Devillers and Cowie 2023) (Figures 1 and 2).

In recent years, advances in recurrent neural networks (RNNs) and transformer-based pretrained language models (PLMs) have significantly enhanced the performance of ERC systems. However, despite these advancements, such models often fail to achieve fair classification accuracy across different emotion classes. This bias toward overrepresented classes results in skewed performance, thereby hindering the development of reliable models. Furthermore, the cumulative nature of metrics such as the F-score, widely used in many ERC studies (Majumder *et al.* 2019; Ghosal *et al.* 2020; Zhang *et al.* 2023; Feng *et al.* 2023), provides a misleading evaluation

https://doi.org/10.1017/nlp.2024.48 Published online by Cambridge University Press



Context:
[S1]: Well, I went over to Kyle's last night to pick up a few things and we got to reminiscing [*neutral*]

Query:
[S1]: we talked through most of the night and we realized that the reason we were so angry at each other was because there are still feelings there.

Figure 1. Illustration of emotion prediction using traditional and conformal prediction methods. The input into the predictive model $\hat{\pi}$ includes a context and a query text with the true label ‘sadness’. The traditional prediction method outputs a single prediction ‘neutral’, which is incorrect. On the other hand, the conformal prediction method provides a set of plausible emotions ‘sadness’, ‘neutral’, and ‘joy’ considering the specified confidence level, as indicated by the dashed line at $1 - \hat{q}$.

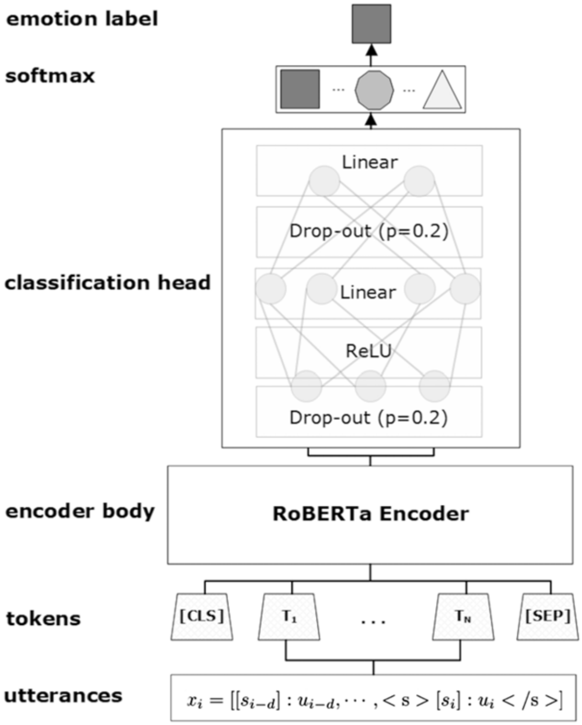


Figure 2. Architecture of a RoBERTa model, fine-tuned for emotion recognition in conversation (ERC). This model can be replaced by any arbitrary classifier.

https://doi.org/10.1017/nlp.2024.48 Published online by Cambridge University Press

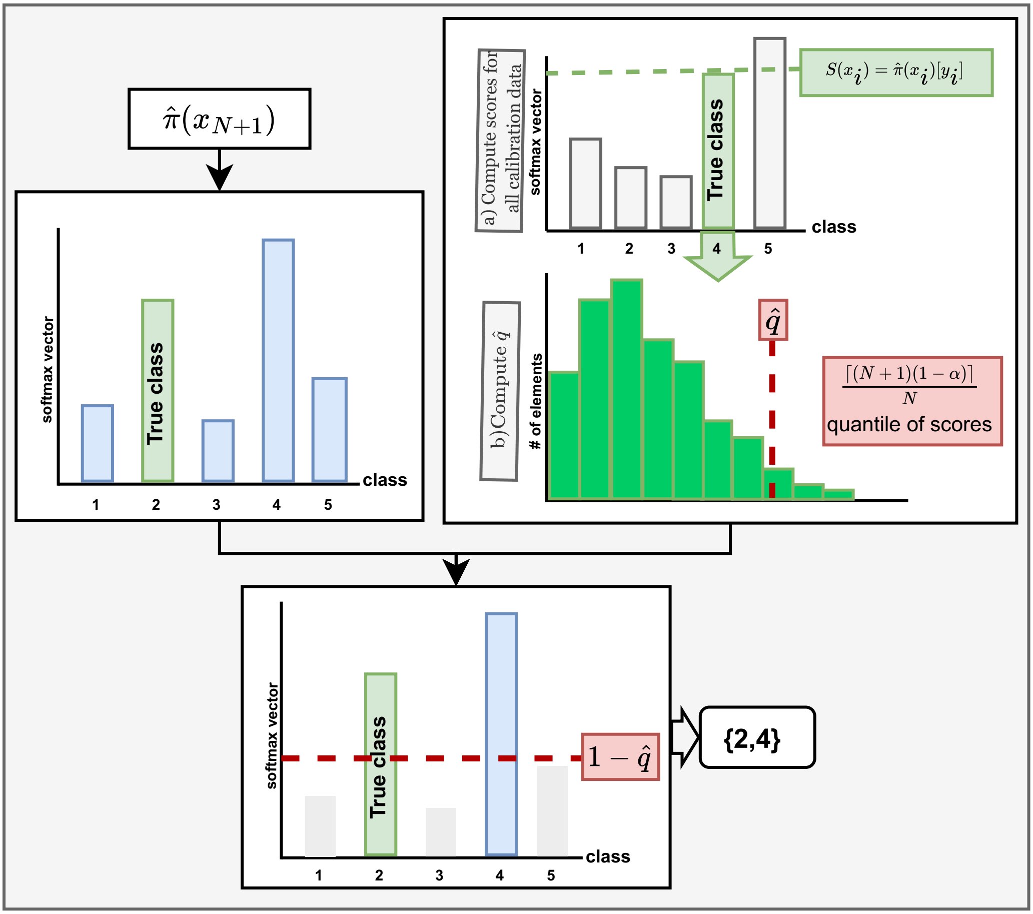


Figure 3. Schematic of the conformal prediction process. The softmax scores produced by the classifier is used to calculate nonconformity measures, which are used to establish a prediction threshold based on a calculated quantile. The final step produces the prediction set {2,4} as potential class labels with the predefined confidence level α.

of model performance by overestimating the model's proficiency in classifying overrepresented classes while neglecting its suboptimal performance in classifying underrepresented classes (Figure 3).

Beyond the issue of bias, empirical evidence across various natural language processing (NLP) tasks indicates that these models frequently demonstrate poor calibration (Sankararaman *et al.* 2022; Guo *et al.* 2017; Jiang *et al.* 2021). This calibration deficiency implies that the predicted probabilities fail to accurately reflect the true likelihood of the corresponding emotions, resulting in either overconfident or underconfident predictions (Figure 4). The problem is particularly pronounced in ERC due to several compounding factors: the intricate nuances of natural language, the multidimensional nature of emotions, and the scarcity of high-quality ERC training datasets (Dragos 2013). Consequently, the probabilities generated by these models often misrepresent the actual likelihood of correct predictions. This poor calibration, coupled with biased predictions, significantly impedes the attainment of reliable quality estimation for individual predictions. Our findings, as illustrated in Figure 5 and Table 1, demonstrate the issues of poor calibration and bias in model predictions, respectively. In selective prediction, predictive uncertainty allows for a nuanced assessment of the risk or reliability associated with individual predictions to achieve a

Table 1. Accuracy of emotion classification on three selected emotional conversation datasets. Labels 0–6 for MELD are sadness, surprise, neutral, joy, anger, disgust, and fear; labels 0–6 for EmoWOZ are neutral, fearful (sad), dissatisfied, apologetic, abusive, excited, and satisfied; and labels 0–3 for Emocx are others, happy, sad, and angry

Dataset	lbl-0	lbl-1	lbl-2	lbl-3	lbl-4	lbl-5	lbl-6
MELD	0.25	0.53	0.83	0.57	0.38	0.10	0.03
EmoWOZ	0.94	0.33	0.66	0.73	0.65	0.47	0.92
Emocx	0.88	0.77	0.71	0.86	NA	NA	NA

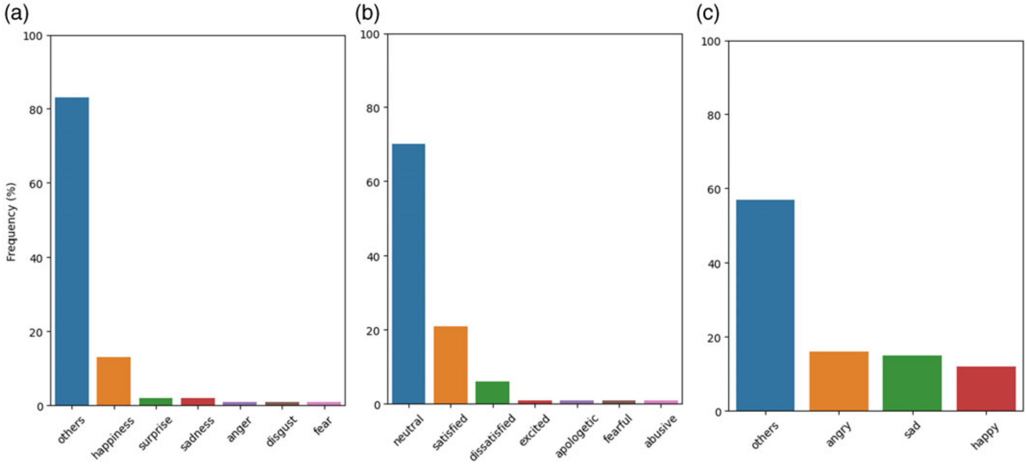


Figure 4. Distribution of (a) MELD, (b) EmoWOZ, and (c) EmoContext datasets utilized in this paper: demonstrating imbalance across emotion categories.

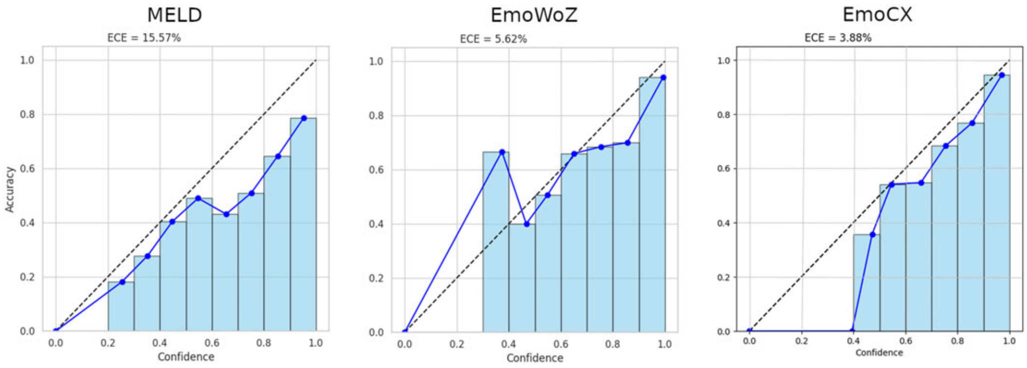


Figure 5. Calibration plots for MELD, EmoWoZ, and EmoCX datasets showing reliability diagrams. ECE and MCE values indicate poor calibration. The bars show accuracy per bin, and the blue line connects the average confidence per bin.

https://doi.org/10.1017/nlp.2024.48 Published online by Cambridge University Press

better decision-making processes (Varshney *et al.* 2022). Uncertainty estimation has a very rich history in learning problems. Various methods such as Bayesian methods (Neal 2012; Harper and Southern 2022), Monte Carlo approximation (Gal and Ghahramani 2016) and ensemble methods (Lakshminarayanan *et al.* 2017) have been widely used in uncertainty-aware deep learning. While promising, integrating these methods with language models are problematic because of proprietary access to the underlying model, demanding high amount of resources and requiring certain assumptions about data distribution (Hu *et al.* 2023). Conformal prediction (CP), introduced by Vovk *et al.* (2005), offers a model-agnostic, distribution-free approach to uncertainty quantification through set-valued classifiers. Unlike traditional methods that output a single label, CP generates a set of potential labels for each input, ensuring the true label is included with a predetermined marginal confidence level. A key strength of CP lies in its calibration efficacy: operating as a post hoc framework, it leverages a hold-out (calibration) dataset to evaluate the 'level of surprise' for each potential label, determining its inclusion in the prediction set (Shafer and Vovk 2008). This calibration process allows CP to maintain its validity across different underlying models and data distributions, providing well-calibrated uncertainty estimates. The size of the prediction set serves as a reliable indicator of predictive uncertainty for individual instances. Larger prediction sets suggest higher uncertainty, indicating that the model is less confident in distinguishing between multiple plausible labels for a given input. Conversely, smaller prediction sets imply lower uncertainty and higher confidence in the model's predictions. Building on this concept, Karimi and Samavi (2023) proposed a formal procedure to quantify the level of uncertainty in predictions using CP output, further enhancing the interpretability and practical utility of CP in uncertainty quantification tasks. Despite the growing interest in uncertainty quantification for large language models (LLMs), the application of CP to NLP tasks remains underexplored. While CP has been extensively studied in unstructured tasks like image classification and synthetic data (Romano *et al.* 2020; Angelopoulos *et al.* 2020; Karimi and Samavi 2023), its use in complex NLP domains is nascent. Kumar *et al.* (2023) demonstrated CP's feasibility for fact-based question answering, but its utility in more nuanced tasks like ERCs is unexplored. Our research addresses this gap, evaluating and improving CP's applicability to uncertainty quantification in conversational emotion recognition. We aim to develop cost-effective strategies that ensure confidence in model predictions, enhancing the reliability of emotion recognition systems in real-world scenarios. Figure 1 illustrates an example output of a distribution-free uncertainty-aware emotion recognition system. To summarize, this paper presents the following significant contributions to the field:

1. **Adaptation of CP framework to ERC task**: We successfully apply two CP techniques—Least Ambiguous Set-Valued Classifiers (LAC) and adaptive prediction sets (APS)—to achieve a set-valued ERC system.
2. **Identification of biased CP**: We examine and highlight the issue of biased CP in ERC, where biased models produce biased (skewed) coverage levels: while the overall marginal promised coverage is maintained, the coverage for overrepresented classes tends to be higher than the promised confidence level, whereas underrepresented classes experience significantly lower coverage rates.
3. **Mitigation of bias in CP**: We propose a novel method, class spectrum calibration, to address and mitigate the coverage bias in conformal ERC models. This approach improves the balance of coverage rates across emotion categories, particularly focusing on enhancing fairness for underrepresented classes. We apply this method to both the LAC and the APS algorithms.

These contributions advance the development of more robust and interpretable emotion recognition systems in conversational contexts. Our experimental results demonstrate the potential of CP methods in this domain and pave the way for future explorations in other NLP tasks.

https://doi.org/10.1017/nlp.2024.48 Published online by Cambridge University Press

2. Background

2.1 Emotion recognition

Emotion fundamentally underpins natural communication. It shapes interpersonal dynamics, drives conversation flow, and facilitates the understanding of a speaker's intents, concerns, and desires. In critical domains, such as clinical mental health, reliable emotion recognition can help in the early detection of emotional distress (Casas *et al.* 2021). Moreover, in customer service support applications, the ability to accurately interpret emotions facilitates the acquisition of realistic feedback from customers, thereby empowering service providers to respond with enhanced effectiveness and empathy (Chen *et al.* 2023).

Broader emotion recognition research includes techniques beyond conversation-specific content, such as facial expression analysis, vocal intonation analysis, and multimodal approaches that integrate visual, auditory, and textual data. For instance, facial expression analysis has been extensively studied for its effectiveness in detecting emotions through subtle changes in facial muscles (Li and Deng 2020). Similarly, vocal intonation analysis leverages acoustic features such as pitch, tone, and rhythm to infer emotional states (Wani *et al.* 2021). Multimodal emotion recognition combines these modalities, providing a comprehensive understanding of emotions by integrating visual, auditory, and textual cues (Sebe *et al.* 2005).

ERC is the process of automatically identifying the emotional state of an utterance considering its context. In recent years, various approaches, including graph learning, sequence learning, and PLMs/LLMs, have been utilized to improve the performance of ERC systems.

2.1.1 Graph learning-based ERC approaches

Graph-based methods have shown promising results in capturing the complex interdependencies in conversational data. Ghosal *et al.* (2019) introduced DialogueGCN, which models both speaker-level and context-level information using graph convolutional networks. Building on this, Shen *et al.* (2021) proposed DAG-ERC, a directed acyclic graph-based approach that better captures the temporal dynamics of conversations. More recently, Gan *et al.* (2024) developed a graph structures that encode temporal order, speaker dependencies, and long-distance context, incorporating a novel context filter and a feature-correction procedure to improve conversational emotion recognition at the utterance level.

2.1.2 Sequence learning-based ERC approaches

As an important development of NLP studies, advanced deep learning techniques, particularly sequence learning models based on recurrent neural networks (RNNs) and transformers, have been widely applied to ERC tasks. Poria *et al.* (2017) proposed an LSTM-based model that considers contextual information from surrounding utterances. Hazarika *et al.* (2018) introduced ICON, an inter-conversational attention network that models both speaker-specific and conversational context. The COSMIC model by Ghosal *et al.* (2020) further enhanced this approach by incorporating commonsense knowledge. Majumder *et al.* (2019) presented DialogueRNN, a recurrent neural network that keeps track of individual speaker states throughout the conversation, showing significant improvements in emotion recognition accuracy.

2.1.3 PLM/LLM-based ERC approaches

The advent of PLMs and LLMs has significantly impacted the field of ERC. Li *et al.* (2020) proposed HiTrans, a hierarchical transformer model that leverages pretrained BERT embeddings. More recently, there has been a shift toward using LLMs for ERC tasks. DialogueLLM Zhang *et al.* (2023) represents a significant advancement in this direction, being the first to apply LLMs specifically to the ERC task. This model demonstrates the potential of large-scale language models in

https://doi.org/10.1017/nlp.2024.48 Published online by Cambridge University Press

understanding and interpreting emotional context in conversations. In a related development, Zhang *et al.* (2024) introduced RGPT for sentiment classification, showcasing the adaptability of LLMs to various text classification tasks, including emotion recognition. Their work on "Pushing The Limit of LLM Capacity for Text Classification" provides valuable insights into optimizing LLMs for specific classification tasks. Additionally, Lei et al., proposed InstructERC which employs a retrieval module to concatenate the historical dialog content statement, label statement and emotional domain demonstrations with high semantic similarity (Lei *et al.* 2024). They also improved the reasoning capabilities of the model by adding emotion alignment through speaker identification and emotion prediction tasks.

2.2 The challenges of current ERC systems

The inherent ambiguities in human language, stemming from linguistic complexities like synonyms and contextual nuances, significantly increase the difficulty of accurately identifying emotions in textual conversation (Dragos 2013; Ott *et al.* 2018; Blodgett *et al.* 2020). This issue is further complicated by the fuzzy and subjective nature of emotional states, where emotions such as 'happiness' and 'joy', for example, often exhibit significant overlap and lack clear boundaries, which complicates distinct classification efforts (Schlosberg 1954). Additionally, the absence of nonverbal cues, such as facial expressions and vocal intonations, in text-based interactions, which are critical in multimodal emotion recognition, presents another layer of complexity in textual ERC (Dessai and Virani 2021; Soundariya and Renuga 2017). These challenges are further exacerbated by disparities in training datasets, which often disproportionately represent certain emotions like 'happiness' or 'neutral'. This imbalance leads to biased and overconfident predictions and diminishes the effectiveness in predicting underrepresented states, such as 'angry' or 'disgust' (Deriu *et al.* 2021).

2.3 Reliable ERC systems through predictive uncertainty

These interconnected challenges significantly impact the entire pipeline of ERC, resulting in biased predictions and limiting the models' capability to generalize across diverse emotional contexts. Traditional evaluation metrics, such as overall accuracy, often fail to accurately reflect the model's performance on individual outputs. This is because overall accuracy is amplified by the model's proficiency in identifying overrepresented labels within the training dataset. A reliable classifier must generate well-calibrated confidence values for its outputs, providing a meaningful indication of the model's uncertainty about its predictions. This form of evaluation is particularly critical in domains like health care or customer service, where inaccuracies in emotion recognition can have substantial consequences (Oh *et al.* 2017; Xu *et al.* 2022). Calibrated confidence measures not only disclose model limitations but also enable informed decision-making.

Uncertainty quantification in machine learning is crucial for understanding and mitigating the inherent uncertainties in model predictions. Methods for uncertainty quantification can be categorized into sampling-based, calibration-based, and distribution-based approaches (Hu *et al.* 2023). Incorporating uncertainty quantification transforms point predictions of a model into robust and reliable confidence values (Sankararaman *et al.* 2022). A naïve technique for estimating uncertainty is using the complement of softmax values (i.e., $1 - \text{softmax}_k(f)$ for each $k = 1, \ldots, C$). However, this method has inherent pitfalls, particularly in deeper architectures, which tend to exhibit overconfidence in their predictions (Guo *et al.* 2021). CP, a calibration-based method, aligns a model's predicted probabilities with the true likelihoods of outcomes, ensuring that the model's confidence accurately reflects its correctness. This method generates predictive uncertainty for each individual prediction, enhancing the reliability and interpretability of model outputs. By leveraging CP, models can provide well-calibrated confidence intervals, offering a robust measure of the quality of each individual prediction (Shafer and Vovk 2008).

https://doi.org/10.1017/nlp.2024.48 Published online by Cambridge University Press

3. Methodology

3.1 Emotion recognition in conversation

The goal of ERC is to categorize the emotional dynamics expressed within dialogues into specific emotional categories, such as 'happiness' or 'sadness'. The supervised methods for ERC employ a discriminative framework which necessitates training on datasets with emotion labels. Formally, consider the dataset as a collection of N pairs $\{(x_i, y_i)\}_{i=1}^{N}$, where each pair (X, Y) consists of a structured dialogue x_i and its corresponding emotion label y_i. The objective of the classification task is to optimize parameters ω of a C-variate function $f^{\omega}(d) = (f_1^{\omega}(d), \ldots, f_C^{\omega}(d))$, where C is the number of emotion classes. This optimization process aims at minimizing the cross-entropy loss:

$$- \sum_{i=1}^{N} \sum_{k=1}^{C} \mathbb{I}[y_i = k] \log \left[\text{softmax}_k \left(f^{\omega}(x_i) \right) \right], \tag{1}$$

where softmax(f) is a C-variate function defined as

$$\text{softmax}_k(f) = \frac{\exp(f_k)}{\sum_{l=1}^{C} \exp(f_l)}, \tag{2}$$

for vectors $f = (f_1, \ldots, f_C)$ and $\mathbb{I}[\text{A}]$ is the indicator function that takes value 1 if statement A is true and 0 otherwise. In the context of multi-class classification tasks, the softmax function transforms a vector of real-valued logits into a set of probabilities, which together sum to 1. Following this, the argmax function is applied to these softmax-derived probabilities to pinpoint the class with the highest probability $\hat{y}_i \in \{1, \ldots, C\}$ for the given input x_i.

In this paper, we used the RoBERTa transformer as the baseline model for ERC. The model architecture consists of two primary components: an encoder module, which processes the input text and extracts contextualized representations of dialogues, and a classification head, designed to predict the corresponding emotion label using latent data. We use the RoBERTa model as the baseline encoder (Liu *et al.* 2019). Throughout the training process, the weights of both the encoder and the classification head are iteratively updated to minimize the loss value between the predicted emotion labels and their ground-truth counterparts. Figure 2 provides a schematic overview of our classifier for the ERC task.

It is noteworthy that the main goal of this paper is to investigate the often overlooked aspect of uncertainty within the ERC task, rather than the development of complex structured methods for the task of emotion recognition. We adapt and introduce distribution-free, uncertainty-aware frameworks to integrate and assess predictive confidence for predicted labels. AsCP is a model-agnostic framework, it can be easily adapted to other applications and models of NLP.

An input to the model is a windowed multi-turn conversation between interlocutors, tagged using special tokens. We set the window size to 3, meaning that each input includes a starting utterance and the subsequent two utterances by the speakers as context and the third utterance as query. Let's consider a conversation represented as a sequence $[(s_1, u_1), \cdots, (s_N, u_N)]$, where each pair (s_i, u_i) includes an utterance u_i and its corresponding speaker s_i for $i = 1, \cdots, N$. The goal of ERC is to predict the emotion label $y \in \mathscr{Y}$ for the query utterance u_i, considering the context provided by a sequence of d preceding utterances $[u_{i-d}, \cdots, u_{i-1}]$, here $d = 2$, and the corresponding speaker information $[s_{i-d}, \cdots, s_{i-1}]$.

Each input text x_i into the prompt generation code model is constructed by appending special tokens to u_i for the query utterance as follows:

$$x_i = [\text{context: } [s_{i-d}]\text{: } u_{i-d}, \cdots, < \text{s} > [s_i]\text{: } u_i < /\text{s} >]. \tag{3}$$

https://doi.org/10.1017/nlp.2024.48 Published online by Cambridge University Press

3.2 Set-valued classifiers

Set-valued classifiers are a category of classifiers that, instead of predicting a singular class label for an input, assign a set of possible class labels. The size of prediction set serves as an indicator of model uncertainty regarding its output. This approach is particularly advantageous in scenarios characterized by ambiguity in the class categories, or when the cost of making an incorrect prediction is high. Formally, let us define $\hat{\pi}$ as a classifier trained on a dataset $\{(x, y)\} \in (\mathscr{X}, \mathscr{Y})$, where $\mathscr{Y} = \{1, \ldots, C\}$, and consider (x_{new}, y_{new}) as a new instance in $(\mathscr{X}, \mathscr{Y})$. In the context of ERC task, C is the number of emotion classes. The objective of a set-valued classifier is to establish a set function Γ to map x_{new} to a subset of $\{1, \ldots, C\}$, ensuring $\Gamma(x_{new})$ encompasses y_{new} with a predefined significance level $\alpha \in (0, 1)$:

$$P[y_{new} \in \Gamma(x_{new})] \geq 1 - \alpha. \tag{4}$$

An initial strategy to achieve this objective might involve sequentially adding labels to the prediction set in descending order of their softmax probability values, as determined by $\hat{\pi}$ and stopping when the aggregative sum of these probability values marginally surpasses $1 - \alpha$ (Guo *et al.* 2017; Platt 1999). As highlighted in Angelopoulos *et al.* (2020), this 'naïve' strategy, albeit intuitive, does not meet the indented coverage criterion in Equation (4) due to reliance on uncalibrated probability estimates and is liable to generate excessively large sets for instances with poorly calibrated tail probabilities, thus failing to precisely reflect the model's intrinsic predictive uncertainty.

3.3 Conformal prediction

CP, as introduced by Vovk *et al.* (2005), is a distribution-free framework for uncertainty quantification in predictions. It guarantees the marginal coverage condition, outlined in Equation (4), across $N + 1$ data points. This post hoc framework assesses the 'level of surprise' associated with observing a specific label for a new input, relative to the hold-out (calibration) data (Shafer and Vovk 2008). CP's procedure begins with an arbitrary pretrained classifier, denoted $\hat{\pi}$, which, in our context. Then using a small amount of data $\{(x_i, y_i)\}_{i=1}^{N} \in (\mathscr{X}, \mathscr{Y})$, which has not been previously seen by the classifier (known as hold-out or calibration set), it understands the predictive performance of $\hat{\pi}$ on future inputs. The baseline CP utilizes the probability vector produced by $\hat{\pi}$ (e.g., softmax values) to calculate the model's error for each instance in the calibration set using a nonconformity (conformity) function. Nonconformity function for classification tasks can be $S(x_i) = \hat{\pi}(x_i)[y_i]$, where $\hat{\pi}(x_i)[y_i]$ refers to the probability value of the correct class y_i, predicted by the classifier $\hat{\pi}$. This step effectively ranks all calibration set examples by their probability. Then these ranked probability values are segregated into certain and uncertain groups based on a predetermined significance level α, a process known as the *calibration* step. This step establishes a decision-making threshold for future predictions. The conformity threshold is calculated using the $\frac{\lceil (N+1)(1-\alpha) \rceil}{N}$-*quantile* of the conformity scores from the calibration set, with $\hat{q}$ representing this quantile.

During the test stage, $\hat{q}$ is employed to generate a prediction set for a new unseen example x_{N+1}. For each potential label $y_c \in \mathscr{Y}$, we accept the hypothesis that y is the correct label for x_{N+1} if $\hat{\pi}(x_{N+1})[y_c]$ does not exceed $\hat{q}$:

$$\Gamma(x_{N+1}) = \{y_c \colon \hat{\pi}(x_{N+1})[y_c] \geq 1 - \hat{q}\}. \tag{5}$$

The overall procedure of CP is illustrated in Figure 3. The length of the prediction set generated by CP is an indicator of predictive uncertainty. This length can be converted into certified boundaries to quantify the model uncertainty on outputs. A formal procedure for converting CP output to uncertainty measures is discussed in Karimi and Samavi (2023).

For any new exchangeable instance x_{N+1}, the conformal coverage theorem, introduced by Vovk *et al.* (1999), guarantees that the prediction set will encompass the correct label y_{N+1} with the probability of at least $1-\alpha$ (see Theorem 1). It is noteworthy that Theorem 1 holds regardless of the underlying data distribution until the exchangeability assumption is met.

Theorem 1 (Conformal coverage guarantee (Vovk *et al.* 1999)). *Let* $(x_1, y_1), \ldots, (x_N, y_N)$ *be a sequence of observations, and let* (x_{N+1}, y_{N+1}) *be a new observation. Assume that all observations are exchangeable. For any significance level* $\alpha \in (0, 1)$, *a conformal classifier provides a prediction set* $\Gamma(x_{N+1})$ *for* y_{N+1} *such that:*

$$P(y_{n+1} \in \Gamma(x_{n+1})) \geq 1-\alpha.$$

However, CP often falls short of meeting the stronger notion of conditional coverage:

$$P[y \in \Gamma(X)|X=x] \geq 1-\alpha. \tag{6}$$

Conditional coverage necessitates valid coverage for any observation in $\mathscr{X}$. While achieving conditional coverage theoretically requires additional modeling assumptions, the goal in conformal methods is to develop classification approaches that maintain marginal coverage, approximate conditional coverage, and ensure robust prediction sets with the possible minimum size as proposed by Barber *et al.* (2019). Various strategies have been developed to balance the trade-offs between approximating conditional coverage and minimizing the size of prediction sets. It is important to note that CP is a flexible framework rather than a fixed methodology, and its specific settings can vary across different implementations. In this paper, we adapted the implementation of LAC as described by Kumar *et al.* (2023) for the baseline CPs.

3.3.1 The issue of bias in ERC

A significant challenge for conventional ERC models is their reliance on datasets characterized by imbalanced emotion classes. Typically, within ERC datasets, certain emotions, for example 'neutral' or positive emotions, are overrepresented. This imbalance results in the development of biased classifiers, which exhibit overconfidence in these overrepresented classes, even in instances with incorrect predictions (Guo *et al.* 2017; Johnson and Khoshgoftaar 2019). The issue of class imbalance is widespread in emotion recognition datasets and significantly affects the reliability of models (see Figure 4). Our findings, detailed in Section 4, demonstrate that this biased classifiers can compromise the efficacy of conformal classifiers during both calibration and prediction stages (Figures 5 and 6).

Although baseline CP methods such as LAC can achieve marginal coverage, as guaranteed by Theorem 1, their application to ERC tasks may not adequately address the desideratum of approximating conditional coverage (Figures 7 and 8). Specifically, the prediction sets generated for underrepresented classes often exhibit coverage levels significantly below the predefined confidence threshold (see Figure 9). This discrepancy severely limits the applicability of such methods in scenarios where accurate estimation of uncertainty for nuanced emotional states is crucial. For instance, in conversational systems used in customer service or mental health support, the reliable detection of complex negative emotions is essential (Tivatansakul *et al.* 2014; Vaudable and Devillers 2012).

To address this issue, this study evaluates the efficacy of two distinct methodologies. Initially, we employed the adaptive prediction set (APS) method introduced by Romano *et al.* (2020), from the domain of image processing, to assess its applicability and performance in the task of ERC. Subsequently, we introduce a novel strategy, termed *class spectrum calibration*. Unlike previous methods that compute a universal single nonconformity metric, our method calculates per-class nonconformity metrics. The core aim of class spectrum calibration is to fulfill several key objectives: maintaining marginal coverage, producing small prediction sets and approximating conditional coverage as stated Section 4. Notably, we integrate the *class spectrum* approach into both the baseline CP and APS.

https://doi.org/10.1017/nlp.2024.48 Published online by Cambridge University Press

3.3.2 Adaptive prediction set (APS)

As discussed before, CP provides a guarantee of marginal coverage. This denotes that, on average, for any new input x_{N+1}, the produced prediction set is expected to encompass the true label y_{N+1} at least with the probability of the predefined confidence level, across the last $N+1$ inputs. This guarantee stands irrespective of the underlying distribution of the data. Nonetheless, in most cases, CP fails to achieve the conditional coverage for a specific class. This shortfall points to a potential discrepancy in the method's efficacy across different class-specific scenarios.

APS introduces a novel nonconformity score to address the coverage discrepancies by calibrating the quantile value. It provides an adaptive score function for the calculation of q-value through computing a generalized inverse quantile conformity score. This process involves summing the sorted predicted probabilities, starting from the highest toward the true class, and using this cumulative value as a nonconformity score for each instance. Then in the prediction step, APS uses a randomization criterion to include or exclude the label that exceeds the cumulative value beyond the threshold. The resulted prediction set is the smallest randomized prediction set with conditional coverage at level $1-\alpha$. This adaptive scoring guarantees finite-sample coverage for future test points. APS provides marginal coverage on synthetic and image data, but its performance on more complex structures such as natural language has not been explored yet. For an in-depth understanding and comprehensive insights, refer to the seminal paper by Romano *et al.* (2020).

3.3.3 Class spectrum conformal prediction (CSCP)

LAC and APS rely on a universal quantile value, overlooking the class distributions in the dataset. Our results shown in Figure 9 indicate that such an approach leads to biased q-values, particularly for the models trained on imbalanced datasets. This bias negatively affects the ability of CP to achieve an approximation of conditional coverage. In this respect, the calibration stage, designed to determine the thresholds for making predictions, often fails to adequately represent minority classes. For overrepresented classes, coverage may exceed the predetermined confidence level, whereas it significantly falls below this threshold for minority classes. Class spectrum conformal prediction (CSCP) is an effective strategy designed to compute calibration values for each class independently. It produces nonconformity scores that are adjusted with the distinct distribution of calibration data and complexities of each class, thereby improving the coverage and fairness of CP.

In CSCP prediction sets for test inputs are generated by leveraging class-specific p-values, derived from nonconformity scores, offering an alternative to quantile-based approaches. These p-values quantify how unusual a specific emotion label demonstrates the emotional state of the new input x_{n+1}. As outlined in Algorithm 1, the process starts with a prediction model $\hat{\pi}$, the calibration data $\{(x_i, y_i)\}_{i=1}^{N} \in (\mathscr{X}, \mathscr{Y})$, and the unseen input x_{N+1} from $\mathscr{X}$. The classifier $\hat{\pi}$ calculates softmax probabilities for each class, which are then used to compute nonconformity scores for the true class across the calibration data. These nonconformity scores measure the alignment between each sample's true label and the predicted probability.

In the testing step, the same classifier $\hat{\pi}$ is utilized to calculate the softmax probability vector for x_{N+1}. Then this vector is used for calculating a p-value for each label as stated in stage 9 of the algorithm. The p-value quantifies the proportion of instances in the calibration set with the same label that exhibit a degree of nonconformity equal to or greater than that of the test instance. These p-values serve as indicators of the probability of each label being included in the prediction set of x_{N+1}, regarding the statistical insights extracted from the calibration data:

$$p\text{-value}[y] = \frac{1}{M+1} \sum_{i=1}^{M} \mathbb{I}[s_i[y] \geq S(x_{N+1}, y)], \forall y \in \mathscr{Y}. \tag{7}$$

https://doi.org/10.1017/nlp.2024.48 Published online by Cambridge University Press

Algorithm 1 Class Spectrum Conformal Prediction (CSCP)

```
1:  Inputs: Prediction model $\hat{\pi}$, calibration data $\{(x_i, y_i)\}_{i=1}^{N}$, new input $x_{N+1}$, and significance level
    $\alpha \in (0, 1)$
2:  for each sample $(x_i, y_i)$ in the calibration data do                    ▷ Calibration stage
3:      Predict probability vector $probs_{cal}[y_i] \leftarrow P(y_i|x_i)$ using $\hat{\pi}$
4:      Compute nonconformity scores using score function $scores[y_i] \leftarrow 1 - probs_{cal}[y_i]$    ▷ can
    can be any arbitrary function
5:  end for
6:  Predict probability vector $probs_{N+1} \leftarrow P(Y|X = x_{i+1})$ using $\hat{\pi}$
7:  $\mathscr{C}_{N+1} = [\,]$                                                  ▷ Prediction stage
8:  for each $y \in \mathscr{Y}$ do                                          ▷ $\mathscr{Y}$ is the label space
9:      $p_y \leftarrow probs_{N+1}[y]$
10:     Compute $p$-value$_{[y]}$ using $p_y$ and $probs_{cal}[y]$
11:     if $p$-value$_{[y]} \geq \alpha$ then add label $y$ into $\mathscr{C}_{N+1}$
12: end for
13: Output: The prediction set $\mathscr{C}_{N+1}$ for the new input $x_{N+1}$
14: end algorithm
```

Here, p-value$[y]$ denotes the p-value for emotion label y, M is the number of calibration instances for each label y, s_i is i'th nonconformity score for the label y in the calibration set, and $\mathbb{I}[\,\cdot\,]$ is the indicator function that returns 1 if the condition inside is true and 0 otherwise. In stage 10, labels are included in the prediction set $\mathscr{C}_{N+1}$ if their associated p-value exceeds or is equal to the predefined significance level α. This criterion ensures that the prediction sets achieve the desired coverage level, reliably containing the true label y_{N+1}, with a confidence level of at least $1 - \alpha$. This approach provides a flexible and robust method for approximating conditional coverage in scenarios with nonuniform class distributions.

3.3.4 Class spectrum adaptive prediction sets (CS-APS)

The last approach we investigated in this paper is the class spectrum adaptive prediction sets (CS-APS), which incorporates the strategy of class spectrum quantile calculation into standard APS algorithm, an adaptation we derived from image processing (Romano *et al.* 2020). In Section 4, we demonstrate that implementing an adaptive set in ERC tasks significantly improves coverage levels across different classes. Despite this improvement, we observed that applying APS to the task of ERC still resulted in imbalanced coverage across emotion classes (see Figure 9). To get a more accurate approximation of conditional coverage, we modified the standard APS by incorporating a class spectrum strategy. This modification involves a tailored calibration process, where for each sample in the calibration set, we append the adaptive nonconformity for the true label to its label-specific list (e.g., $Q[y]$), rather than relying on a universal list for all classes. This process results in distinct arrays of q-values for each class, enabling more precise class-wise coverage adjustment.

To choose the right value from the list of q-values for instance x_{N+1} at test time, one intuitive heuristic is to apply the CP process on each class and then combine their outputs to form the final prediction set. However, this approach tends to generate overly large prediction sets, which can't effectively convey useful information on predictive uncertainty. Instead, we leverage the label predicted by $\hat{\pi}(x_{\mathrm{N+1}})$ to guide the selection of a more appropriate q-value and construct the prediction set accordingly. For instance, if $\operatorname{argmax}(\hat{\pi}(x_{\mathrm{N+1}})) = happy$, we utilize $Q[happy]$ as the target q-value for that instance. This approach is advantageous as it allocates proportional consideration to each class, corresponding to the model's confidence in its prediction.

Algorithm 2 Class Spectrum Adaptive Prediction Set (CS-APS)

```
Inputs: Prediction model $\hat{\pi}$, calibration data $\{(x_i, y_i)\}_{i=1}^{N}$, new input $x_{N+1}$, significance level
    $\alpha \in (0, 1)$
for each example $(x_i, y_i)$ in the calibration data do                ▷ Calibration stage
    Predict probability vector $P_i \leftarrow P(Y|X = x_i)$ using $\hat{\pi}$
    $P_{isorted}$ ←Sort the vector $P_i$ in descending order
    $S_i$ ← calculate the cumulative sum of $P_{isorted}$ starting from index zero to the position of
    true label $Y_i$ in $P_{isorted}$
    Compute generalized inverse quantile conformity score using $S_i$ and add the result into
    $E[y_i]$
end for
for each $y \in \mathcal{Y}$ do                                         ▷ $\mathcal{Y}$ is the label space
    $Q[y] \leftarrow \lceil(1 - \alpha)(1+ |$ calibration data $|)\rceil$ largest values in $E[y]$
end for
$\mathcal{C}_{N+1} = [\,]$                                               ▷ Prediction stage
Predict probability vector $probs_{N+1} \leftarrow P(Y|x_{N+1})$ using $\hat{\pi}$
$prediction \leftarrow \arg\max\,(probs_{N+1})$
$q \leftarrow Q[predition]$
Sort $probs_{N+1}$ in descending order
Add labels into $\mathcal{C}_{N+1}$, starting with the largest probability, until threshold $q$ is reached
Apply randomization criterion
Output: A prediction set $\mathcal{C}_{N+1}$ for the new input $X_{N+1}$
end algorithm
```

It encourages the inclusion of a larger range of classes in the prediction set when the model produces predictions with higher entropy (e.g., when the prediction is an underrepresented class). This approach is particularly beneficial for models prone to bias, as it ensures that the prediction set reflects a more balanced consideration of all possible outcomes. This strategy leads us to enhanced coverage across all classes. We present a modified version of APS, the *class spectrum adaptive prediction set (CS-APS)*, in Algorithm 2.

Beyond ERC, the methods proposed in this paper can be effectively applied to various other NLP tasks that require reliable uncertainty quantification. For instance, CP techniques can be leveraged in sentiment analysis to provide more nuanced and trustworthy sentiment predictions by generating sets of potential sentiments with associated confidence levels. Similarly, in tasks like machine translation and text classification, these methods can enhance the interpretability and reliability of predictions by offering well-calibrated uncertainty estimates. This adaptability underscores the broader applicability of our approach, paving the way for its integration into diverse NLP applications.

4. Experiments

In this paper, we evaluated the performance of LAC, APS, CSCP, and CS-APS focusing on the task of ERC. The inherent complexity of emotion within textual conversational contexts, as discussed in (2), positions ERC as an ideal benchmark for examining the methods under study.

All the experiments and source code in this study are accessible in the project's GitHub repository[a].

4.1 Model and datasets

Pretrained transformers like RoBERTa (Liu *et al.* 2019) have achieved significant attention in recent years across various NLP tasks, primarily for their proficiency in capturing long-range dependencies and synthesizing rich contextual cues, including sentiment analysis (Tan *et al.* 2022), question answering, and text classification (Nassiri and Akhloufi 2023). In this paper, we fine-tuned a RoBERTa model on the task of ERC using three widely used datasets for emotional conversation:

- Multimodal EmotionLines Dataset (MELD): Developed by Poria *et al.* (2019), comprises approximately 1,400 dialogues from the TV series *Friends* and private Facebook Messenger conversations. The utterances are annotated with one of seven emotion categories, extending Ekman's six basic emotions ('joy', 'sadness', 'anger', 'fear', 'surprise', 'disgust') by adding a 'neutral' category.
- EmoWOZ: Developed by Feng *et al.* (2022), features around 11,000 task-oriented human-machine dialogues, annotated with a novel set of seven emotions: 'neutral', 'fearful', 'dissatisfied', 'apologetic', 'abusive', 'excited', and 'satisfied'.
- EmoContext: Published by Chatterjee *et al.* (2019) for the emotion detection challenge of SemEval-2019 Task 3, in which each dialogue includes three turns labeled by using one of the emotion classes: 'happy', 'sad', 'angry', or 'others'. This dataset comprises 30,160 dialogues from a conversational agent for training and two separate sets for evaluation, with 2,755 and 5,509 dialogues, respectively.

The datasets chosen for this research exhibit a pronounced imbalance across different emotion classes, making them a particularly suitable choice for the objectives of this paper. Figure 4 illustrates the distribution of emotion classes within the datasets, organized in order of decreasing frequency.

4.1.1 The problem of biased prediction distribution in transformer-based ERC

The results presented in Table 1 highlight the skewed prediction distribution for RoBERTa model using three emotional conversation datasets: MELD, EmoWOZ, and Emocx. In the MELD dataset, the highest accuracy is achieved for the 'neutral' class (0.83), while the 'fear' class has the lowest accuracy (0.03), indicating significant challenges in correctly classifying fear. For the EmoWOZ dataset, the 'neutral' class also has the highest accuracy (0.94), and the 'fearful (sad)' class has the lowest accuracy (0.33), suggesting that neutral emotions are generally easier to classify. The Emocx dataset shows the highest accuracy for the 'others' class (0.88) and the lowest for the 'sad' class (0.71). These discrepancies point to potential biases in the datasets or the models' difficulty in distinguishing certain emotions, especially those with more subtle or overlapping features.

4.1.2 The problem of poor calibration in transformer-based ERC

Figure 5 represents the calibration performance of the fine-tuned RoBERTa model on three ERC datasets. In all three cases, the model's confidence (x-axis) does not align well with its actual accuracy (y-axis), as evidenced by the deviation of the blue line from the ideal calibration. This poor

[a] https://github.com/samadroohi/cerc

https://doi.org/10.1017/nlp.2024.48 Published online by Cambridge University Press

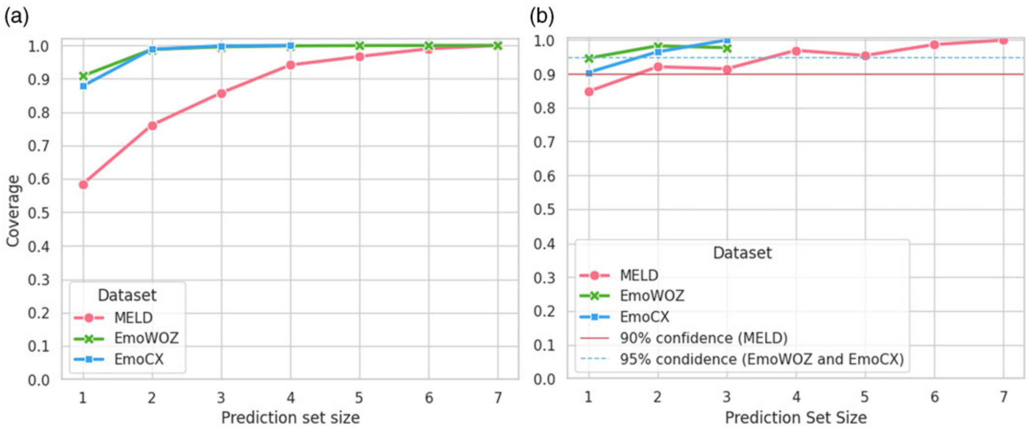


Figure 6. Comparative analysis of prediction set coverage between (a) naive top-*k* prediction sets and (b) size-stratified prediction sets for CP. CP demonstrating consistently higher accuracy and superior performance across various set sizes compared to the naive methods.

calibration is quantified using expected calibration error (ECE). ECE measures the average difference between confidence and accuracy across all predictions. Notably, all three datasets exhibit a tendency toward overconfidence, especially in higher confidence ranges, where the blue line falls below the diagonal. This indicates that the model is often more confident in its predictions than its actual performance.

4.1.3 Comparison of size-stratified CP with naive top-k prediction

In Angelopoulos *et al.* (2020), the size-stratified coverage is used as a metric to assess error rates across prediction sets of varying sizes, highlighting the effectiveness of CP over naive methods. We follow the same paradigm in the first experiment to compare the coverage achieved by the baseline CP algorithm and the conventional naive top-*k* prediction sets, which are simply formed by taking the top-*k* softmax scores across all predictions.

The results illustrated in Figure 6 highlight the superior accuracy of CP in single-size prediction sets when compared to the naive top-1 accuracy. Additionally, while both methods exhibited an increase in coverage with larger set sizes, CP consistently tends to maintain a higher level of accuracy across almost all prediction set sizes. In contrast, the naive top-*k* approach demonstrates sharp improvements but starts from the lower level of accuracy than the desired confidence levels.

This comparative analysis indicates that relying merely on the top-*k* highest probability scores fails to achieve comprehensive coverage effectively. In contrast, CP maintains consistent adherence to the predetermined confidence levels across various prediction set sizes. This observation highlights the superiority of CP methods, which not only provide a more accurate estimation of error rates but also sustain robust performance throughout.

4.1.4 The calibration of conformal prediction coverage

Figure 7 illustrates the calibration of CP, comparing the empirical coverage to the specified confidence level. In contrast to the traditional ERC systems (as presented in Section 4.1.2), CP exhibits superior calibration performance. This suggests that prediction sets provide a true likelihood of the output. Therefore, for each individual prediction, the prediction set size provides a reliable indicator of predictive uncertainty.

https://doi.org/10.1017/nlp.2024.48 Published online by Cambridge University Press

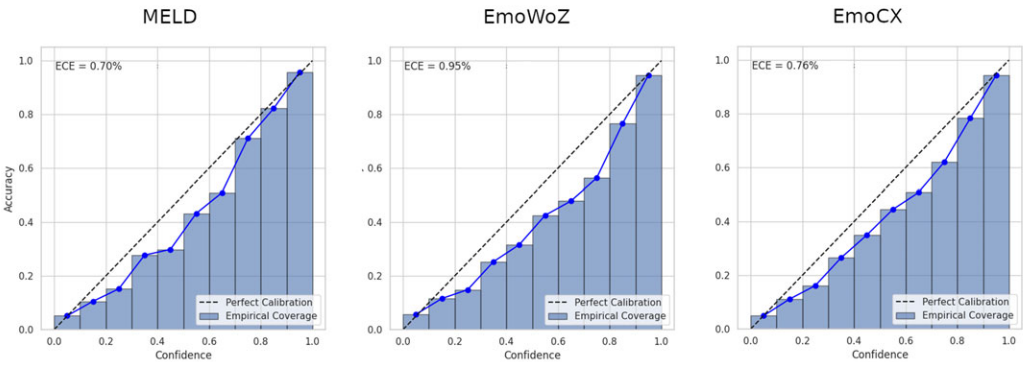


Figure 7. The calibration for conformal prediction considering various confidence levels.

4.1.5 Marginal coverage and average prediction set size

The results of our experiments, which examined four distinct methodologies across designated significance thresholds (90% for MELD and 95% for EmoWOZ and EmoCX), are depicted in Figure 8. This analysis focuses on the marginal coverage and average prediction set sizes achieved by each method, calculated using the outcomes of ten experimental iterations utilizing randomly selected calibration and testing samples. Notably, the coverage level provided by the LAC method did not meet the predetermined confidence criteria for the MELD and EmoWOZ datasets. In comparison, the CSCP approach demonstrated relatively enhanced coverage relative to LAC, albeit at the cost of a noticeable increase in prediction size. Both APS and CS-APS achieved coverage rates surpassing the predetermined confidence levels, consistent with the criterion outlined in Equation (4). Specifically, the CS-APS method outperformed the other approaches in terms of marginal coverage. For EmoWOZ and EmoCX, the CS-APS maintained prediction sizes that were smaller or comparable to prediction sizes generated by APS.

4.1.6 Class spectrum coverage level

The findings from the previous section indicated that, apart from LAC for MELD and EmoWOZ datasets, the other approaches were aligned well with the expected marginal coverage as outlined by CP. However, a notable limitation observed in the results depicted in Figure 8 was the lack of granular insight regarding the class-specific performance of these methods. In an effort to bridge this gap, we conducted a comprehensive class-wise analysis for each method, adhering to the parameters established in the first experiment. The results of this analysis are illustrated in Figure 9, which displays the class spectrum coverage of the methodologies LAC, CSCP, APS, and CS-APS across datasets MELD, EmoWOZ, and EmoCX, with a confidence level of 90% for MELD and 95% for both EmoWOZ and EmoCX. This analysis revealed that, across all examined tasks, LAC consistently underperformed in comparison to its counterparts, while CS-APS demonstrated superior class spectrum coverage, indicating its effectiveness in the respective tasks.

These findings suggest that using a universal quantile value inadequately captures the diverse degrees of uncertainty and the intrinsic distribution characteristic of emotional classes. This issue is particularly pronounced in datasets characterized by uneven representation, leading to a q-value that is affected by the model's bias toward overrepresented classes. For example, in the EmoWOZ dataset, both LAC and APS underperformed in comparison to their class spectrum counterparts, CSCP and CS-APS. Notably, this can be observed in the performance metrics for the underrepresented emotion category 'fearful'. A detailed comparison of coverage values for

https://doi.org/10.1017/nlp.2024.48 Published online by Cambridge University Press

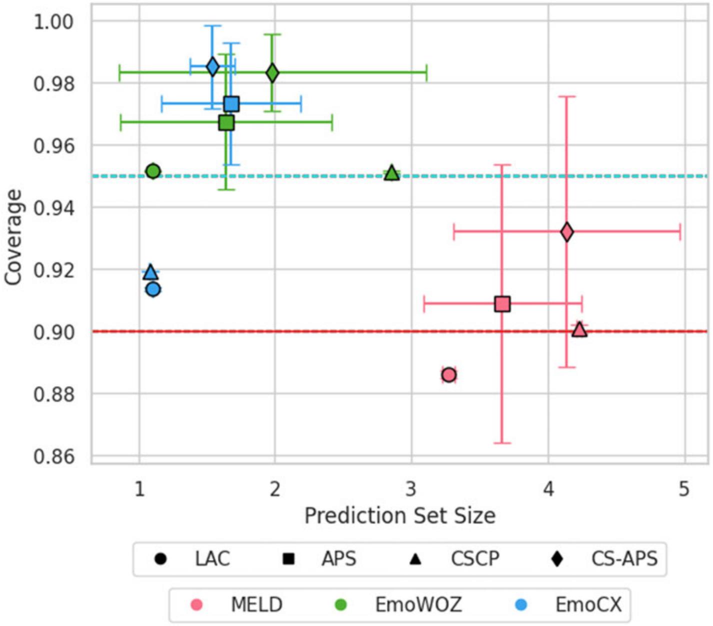


Figure 8. Comparative analysis of marginal coverage and average prediction set sizes for four approaches (LAC, CSCP, APS, CS-APS) across three datasets (MELD, EmoWOZ, and EmoCX). CS-APS outperformed other methods in marginal coverage with an efficient prediction size.

these four methodologies is delineated in Appendix (A), underscoring the significance of this issue.

4.1.7 Class spectrum prediction set size

Figure 10 presents the size of class spectrum prediction sets across methodologies analyzed within this research. The LAC generates relatively small prediction sets. However, its class spectrum coverage, as shown in Figure 9, significantly falls below the predetermined confidence threshold. This suggests that LAC tends to underestimate the uncertainty level for a given input. In contrast, CSCP maintains class spectrum coverage close to the predetermined confidence level, though referring to this as conditional coverage might be somewhat misleading. CSCP tends to produce significantly larger prediction sets, potentially overestimating the actual level of uncertainty, which could limit its utility in practical applications. As highlighted in (3.3), achieving conditional coverage is a challenge, but approximating it is a desired objective in CP methods. Both APS and CS-APS find a good balance between the coverage and the size of prediction sets. This offers an adaptive level of uncertainty that results in prediction sets considerably more compact than CSCPs while achieving a class spectrum coverage level similar to CSCPs.

5. Conclusion

The primary motivation underlying this work was to establish a novel paradigm for leveraging CP in addressing two problems in traditional ERC systems: the biased classification and poor calibration. In our exploration, we focused on three principal objectives of the CP framework: maintaining marginal coverage, minimizing prediction set sizes, and achieving a close approximation of conditional coverage. To this end, we adapted and evaluated two established

https://doi.org/10.1017/nlp.2024.48 Published online by Cambridge University Press

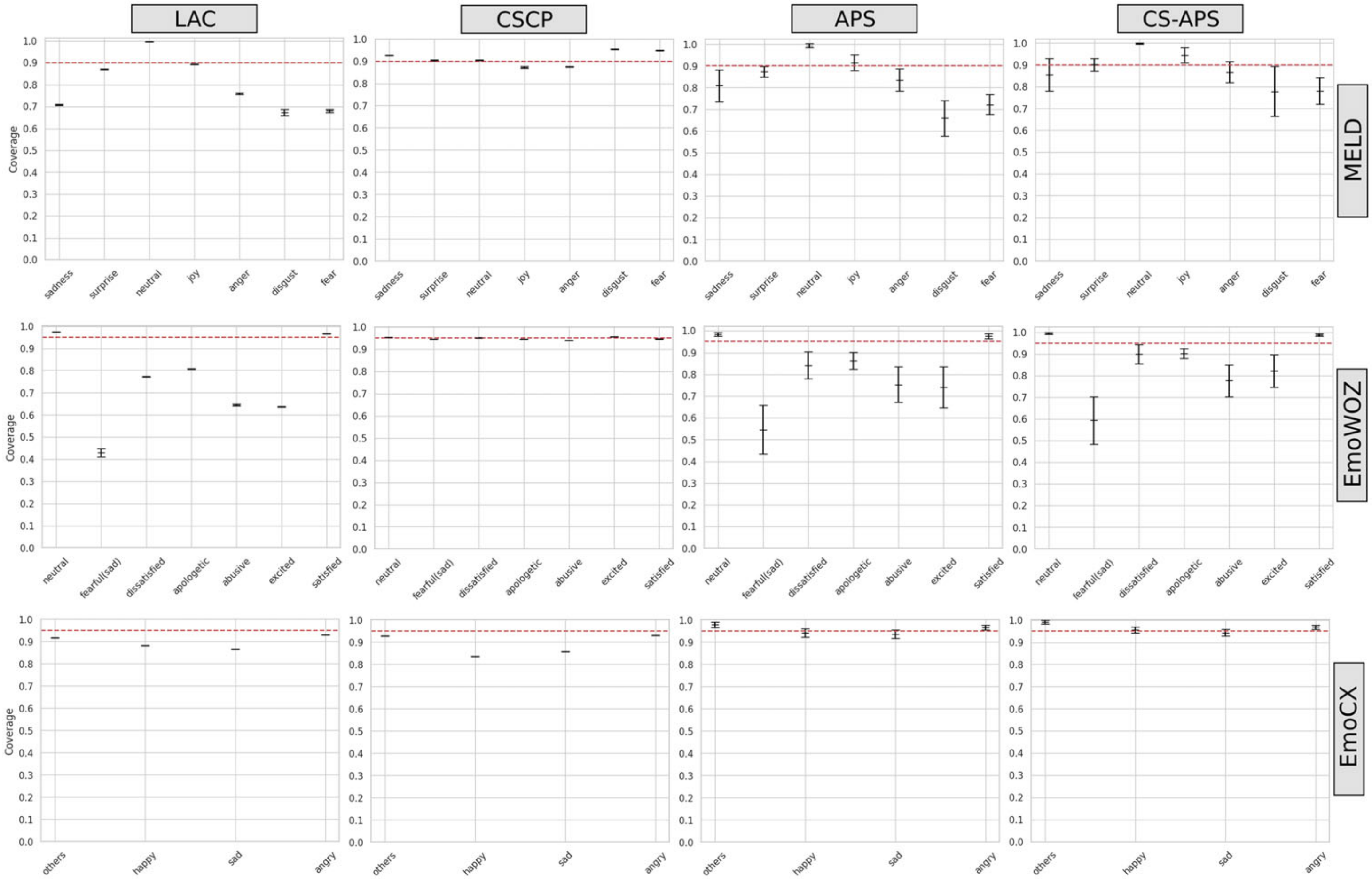


Figure 9. Class spectrum coverage analysis across MELD, EmoWOZ, and EmoCX datasets: comparing the performance of LAC, CSCP, APS, and CS-APS at confidence levels of 90% for MELD and 95% for EmoWOZ and EmoCX, highlighting the inferior performance of LAC and the superiority of CS-APS.

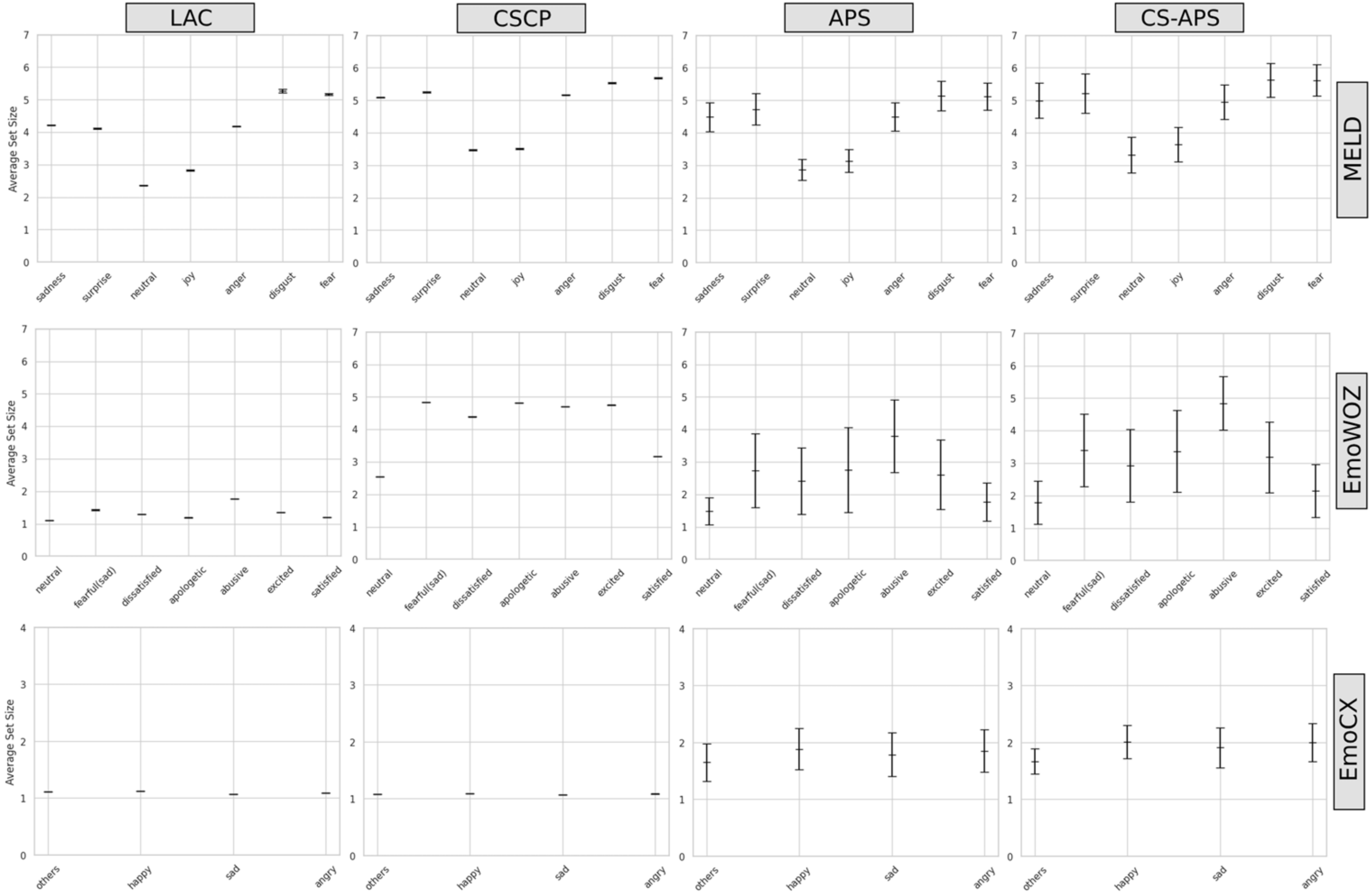


Figure 10. Class spectrum prediction set sizes: comparing LAC, CSCP, APS, and CS-APS with emphasis on LAC's underestimation of uncertainty, CSCP's overestimation, and the balanced adaptivity of APS and CS-APS.

methods—LAC and APS—and introduced two novel approaches: CSCP and CS-APS. Our findings revealed that CP can significantly improve the issue of poor calibration in ERC tasks. Furthermore, we found that while the baseline CP (LAC) generates compact prediction sets, its performance in maintaining marginal coverage fluctuates, likely due to inherent biases in ERC models. Among the evaluated techniques, LAC was also the least capable of approximating conditional coverage. In contrast, CSCP demonstrated superiority in achieving marginal coverage and approximating conditional coverage but at the cost of producing overly large prediction sets, which limits its practical utility. APS and CS-APS maintained consistent marginal coverage and produced smaller prediction sets, comparable to LAC, with CS-APS outperforming in conditional coverage approximation.

These findings highlight the limitations of employing a universal calibration metric, such as a singular q-value, in complex NLP tasks. An effective calibration strategy should dynamically reflect the class distribution in the calibration data. The methodologies devised and applied in this study are easily adaptable to other NLP tasks with minimal modifications to the predictive model.

A critical limitation of our approach lies in handling the class spectrum calibration, wherein we employ calibration values for each class without considering their frequency in the hold-out set, a gap that requires optimization through hyperparameter tuning specific to each class's characteristics. Future research should explore the performance of these algorithms through the lens of uncertainty quantification metrics, such as the area under the receiver operating characteristic curves, thereby offering a more granular understanding of their efficacy and potential applications in predictive uncertainty.

Competing interests. None.

References

Angelopoulos A., Bates S., Malik J. and Jordan M.I. (2020). Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations (ICLR)*. Retrieved from https://iclr.cc/virtual/2021/spotlight/3435

Barber R.F., Candès E.J., Ramdas A. and Tibshirani R.J. (2021). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2), 455–482.

Blodgett S.L., Barocas S., Daumé H. III. and Hal W. (2020). *Language (Technology) is Power: A Critical Survey of "Bias" in NLP*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, pp. 5454–5476. DOI:10.18653/v1/2020.acl-main.485

Casas J., Spring T., Daher K., Mugellini E., Khaled O.A. and Cudré-Mauroux P. (2021). Enhancing Conversational Agents with Empathic Abilities. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents (IVA '21)*, New York, NY, USA: Association for Computing Machinery, pp. 41–47, https://doi.org/10.1145/3472306.3478344

Chen D., Zhengwei H., Yiting T., Jintao M. and Ribesh K. (2023). Emotion and sentiment analysis for intelligent customer service conversation using a multi-task ensemble framework.. *Cluster Computing* **27**, 2099–2115. https://doi.org/10.1007/s10586-023-04073-z.

Chatterjee A., Narahari K.N., Joshi M. and Agrawal P. (2019). *SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text*. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 39–48, https://doi.org/10.18653/v1/S19-2005, https://aclanthology.org/S19-2005

Dessai A. and Virani H. (2021). Emotion detection using physiological signals. *In. 2021 International Conference On Electrical, Computer and Energy Technologies (ICECET)*, vol. 1, pp. 1–4. https://doi.org/10.1109/ICECET52533.2021.9698729.

Devillers L. and Cowie R. (2023). Ethical considerations on affective computing: an overview. In *Proceedings of the IEEE*

Deriu J., Rodrigo A., Otegi A., Echegoyen G., Rosset S., Agirre E. and Cieliebak M. (2021). Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, **54**. Publisher:Springer, pp. 755–810.

Dragos V. (2013). *An ontological analysis of uncertainty in soft data*. In *IEEE Conference Publication — IEEE Xplore*, Available at: https://ieeexplore.ieee.org/document/6641188, Accessed on 21 October 2023.

Feng S., Sun G., Lubis N., Zhang C. and Gašić M. (2023). Affect recognition in conversations using large language models, In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 259–273). Kyoto, Japan: Association for Computational Linguistics.

https://doi.org/10.1017/nlp.2024.48 Published online by Cambridge University Press

Feng S., Lubis N., Geishauser C., Lin H.-C., Heck M. and van Niekerk C. (2022). EmoWOZ: a large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems. In *Proceedings of the 13th Language Resources and Evaluation Conference,* (pp. 4096–4113). Marseille, France: European Language Resources Association. ISBN 979-10-95546-72-6.

Gan C., Zheng J., Zhu Q., Jain D.K. and Štruc V. (2024). A graph neural network with context filtering and feature correction for conversational emotion recognition. *Information Sciences*, **658**, Elsevier, pp. 120017.

Gal Y. and Ghahramani Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, PMLR, pp. 1050–1059.

Ghosal D., Majumder N., Gelbukh A., Mihalcea R. and Poria S. (2020). COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 2470–2481, https://doi.org/10.18653/v1/2020.findings-emnlp.224.

Ghosal D., Majumder N., Poria S., Chhaya N. and Gelbukh A. (2019). *DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics.https://doi.org/10.18653/v1/D19-1015, https://aclanthology.org/D19-1015

Guo H., Pasunuru R. and Bansal M. (2021). *An Overview of Uncertainty Calibration for Text Classification and the Role of Distillation*. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, Association for Computational Linguistics, pages 289–306, DOI:10.18653/v1/2021.repl4nlp-1.29, https://aclanthology.org/2021.repl4nlp-1.29

Guo C., Pleiss G., Sun Y. and Weinberger K.Q. (2017). On calibration of modern neural networks. In *International Conference on Machine Learning*, PMLR, pp. pages.1321–1330.

Harper R. and Southern J. (2022). A bayesian deep learning framework for end-to-end prediction of emotion from heartbeat. *IEEE Transactions On Affective Computing* **13**, 985–991. https://doi.org/10.1109/TAFFC.20201610

Hazarika D., Poria S., Mihalcea R., Cambria E. and Zimmermann R. (2018). ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1280, https://aclanthology.org/D18-1280

Hu M., Zhang Z., Zhao S., Huang M. and Wu B. (2023). Uncertainty in natural language processing: sources, quantification, and applications. CoRR, abs/2306.04459.

Jiang Z., Araki J., Ding H. and Neubig G. (2021). How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics* **9**, 962–977. https://doi.org/10.1162/tacl_a_00407

Johnson J.M. and Khoshgoftaar T.M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data* **6**,27. https://doi.org/10.1186/s40537-019-0192-5

Karimi H. and Samavi R. (2023). Quantifying deep learning model uncertainty in conformal prediction, Publisher: arXiv. http://arxiv.org/abs/2306.00876 (Accessed 2023-07-24.).

Kumar B., Lu C., Gupta G., Palepu Anil, Bellamy D., Raskar R. and Beam A. (2023). Conformal prediction with large language models for multi-choice question answering, In P*roceedings of the 40th International Conference on Machine Learning (ICML 2023)*, PMLR, pp. 493–502.

Lakshminarayanan B., Pritzel A. and Blundell C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles, *Advances in Neural Information Processing Systems (NeurIPS 2017)*, **30**, 6402–6413.

Lei S., Dong G., Wang X., Wang K. and Wang S. (2024). InstructERC: reforming emotion recognition in conversation with a retrieval multi-task LLMs framework, arXiv preprint arXiv: 2309.11911., https://arxiv.org/abs/2309.11911,

Li S. and Deng W. (2020). Deep facial expression recognition: a survey. In *IEEE Transactions On Affective Computing*, **13**, IEEE, pp. 1195–1215.

Li J., Ji D., Li F., Zhang M. and Liu Y. (2020). HiTrans: A Transformer-Based Context- and Speaker-Sensitive Model for Emotion Detection in Conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain: International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.370, https://aclanthology.org/2020.coling-main.370

Liu Y., Ott M., Goyal N., Du Jingfei, Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L. and Stoyanov V. (2019). RoBERTa: a robustly optimized BERT pretraining approach, arXiv: 1907.11692 [cs], Available at: http://arxiv.org/abs/1907.11692.

Majumder N., Poria S., Hazarika D., Mihalcea R., Gelbukh A. and Cambria E. (2019). Dialoguernn: an attentive RNN for emotion detection in conversations. *Proceedings of the AAAI Conference On Artificial Intelligence* **33**, 6818–6825.

Nassiri K. and Akhloufi M. (2023). Transformer models used for text-based question answering systems. *Applied Intelligence* **53**, 10602–10635. https://doi.org/10.1007/s10489-022-04052-8.

Neal R.M. (2012). *Bayesian Learning for Neural Networks*. Springer Science & Business Media. ISBN 978-1-4612-0745-0.

https://doi.org/10.1017/nlp.2024.48 Published online by Cambridge University Press

Oh K.-J., Lee D., Ko B. and Choi H.-J. (2017). A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. In *Proceedings of the 2017 18th IEEE International Conference on Mobile Data Management (MDM)* (pp. 371–375). Daejeon, South Korea. IEEE. https://doi.org/10.1109/MDM.2017.64.

Ott M., Auli M., Grangier D. and Ranzato M. (2018). Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, PMLR 80, 3956–3965. Stockholm, Sweden.

Platt J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, **10**. Cambridge, MA: Publisher, pp. 61–74.

Poria S., Cambria E., Hazarika D., Majumder N., Zadeh A. and Morency L.-P. (2017). Context-Dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 1 Long Papers), Vancouver, Canada: Association for Computational Linguistics, https://doi.org/10.18653/v1/P17-1081, https://aclanthology.org/P17-1081

Poria S., Hazarika D., Majumder N., Naik G., Cambria E. and Mihalcea R. (2019). MELD: a multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 527–536). Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1050

Romano Y., Sesia M. and Candès E.J. (2020). *In Advances in Neural Information Processing Systems (NeurIPS 2020)*, **33**, 1–11.

Sankararaman K.A., Wang S. and Fang H. (2022). BayesFormer: transformer with uncertainty estimation. In *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*, PMLR, pp. 1–15.

Schlosberg H. (1954). Three dimensions of emotion. *Psychological Review* **61**, 81–88. https://doi.org/10.1037/h0054570.

Sebe N., Cohen I. and Huang T.S. (2005). Multimodal emotion recognition. *Handbook of Pattern Recognition and Computer Vision*. World Scientific, pp. 387–409.

Shafer G. and Vovk V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, **9**(3), 371–42.

Shen W., Wu S., Yang Y. and Quan X. (2021). Directed Acyclic Graph Network for Conversational Emotion Recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (vol. 1: Long Papers). Association for Computational Linguistics, pp. 1551–1560.

Soundariya R. S. and Renuga R. (2017). Eye movement based emotion recognition using electrooculography. In *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*, pp. 1–5. https://doi.org/10.1109/IPACT.2017.

Tan K.L., Lee C.P., Anbananthen K.S. and Lim K.M. (2022). RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network. *IEEE Access* **10**, 21517–21525.

Tivatansakul S., Ohkura M., Puangpontip S. and Achalakul T. (2014). Emotional healthcare system: emotion detection by facial expressions using Japanese database. In *Proceedings of the 2014 6th Computer Science and Electronic Engineering Conference (CEEC)*, pp. 41–46. IEEE. https://doi.org/10.1109/CEEC.2014.6958573.

Varshney N.,Mishra S. and Baral C. 2022). Investigating selective prediction approaches across several tasks in IID, OOD, and adversarial settings. In Muresan S., Nakov P. and Villavicencio A., (eds), *Findings of the Association for Computational Linguistics: ACL 2002*. Dublin, Ireland: Association for Computational Linguistics, pp. 1995–2002, https://doi.org/10.18653/v1/2022.findings-acl.158.

Vaudable C. and Devillers L. (2012). Negative emotions detection as an indicator of dialogs quality in call centers. *In. 2012 IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP)* **6**, 5109–5112. https://doi.org/10.1109/ICASSP.2012.

Vovk V., Gammerman A. and Shafer G. (2005). *Algorithmic Learning in a Random World*, volume **29**. New York, NY, USA: Springer.

Vovk V., Gammerman A. and Saunders C. (1999). Machine-Learning Applications of Algorithmic Randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, pp. pages 444–453.

Wani T.M., Gunawan T.S., Qadri S.A.A., Kartiwi M. and Ambikairajah E. (2021). A comprehensive review of speech emotion recognition systems *IEEE Access*, **9**, 47795–47814.

Xu Q., Yan J. and Cao C. (2022). Emotional communication between chatbots and users: an empirical study on online customer service system. In Degen H. and Ntoa S., (eds), *Artificial Intelligence in HCI, Lecture Notes in Computer Science*. Cham: Springer International Publishing, pp. 513–530. https://doi.org/10.1007/978-3-031-05643-7_33.

Zhang Y., Wang M., Wu Y., Tiwari Prayag, Li Q., Wang B. and Qin J. (2023). DialogueLLM: context and emotion knowledge-tuned large language models for emotion recognition in conversations. arXiv.org., https://arxiv.org/abs/2310.11374v4.,

Zhang Y., Wang M., Tiwari P., Li Q., Wang B. and Qin J. (2023). Dialoguellm: context and emotion knowledge-tuned llama models for emotion recognition in conversations, In *Proceedings of the 39th International Conference on Machine Learning (ICML 2023)*, pp. 1–10. PMLR.

Zhang Y., Wang M., Ren C., Li Qiuchi, Tiwari P., Wang B. and Qin J. (2024). Pushing the limit of LLM capacity for text classification, arXiv preprint arXiv: 2402.07470, https://arxiv.org/abs/2402.07470

https://doi.org/10.1017/nlp.2024.48 Published online by Cambridge University Press

Appendix A. Class spectrum coverage values

The coverage values for LAC, CSCP, APS, and CS-APS across datasets MELD, EmoWoZ, and EmoContext are as follows: (Table 2)

Table 2. Results of four approaches on three selected emotional conversation datasets. Labels 0–6 for MELD are sadness, surprise, neutral, joy, anger, disgust, and fear; for EmoWOZ are neutral, fearful (sad), dissatisfied, apologetic, abusive, excited, and satisfied; and labels 0–3 for EmoContext are others, happy, sad, and angry

		Emotion labels						
Dataset	Method	lbl-0	lbl-1	lbl-2	lbl-3	lbl-4	lbl-5	lbl-6
MELD	LAC	70.94	87.10	99.78	89.43	76.03	67.38	68.02
($\alpha = 0.1$)	CSCP	92.67	90.34	90.49	87.28	87.52	95.36	94.88
	APS	80.81	87.28	99.28	91.55	83.56	65.86	72.16
	CS-APS	85.47	90.05	99.78	94.31	86.64	77.87	77.93
EmoWOZ	LAC	97.38	43.13	77.30	80.78	64.48	63.80	96.59
($\alpha = 0.05$)	CSCP	95.23	94.44	95.21	94.49	94.11	95.59	94.56
	APS	98.37	54.64	84.17	84.28	75.29	74.14	97.61
	CS-APS	99.49	59.34	89.89	90.34	77.64	82.16	98.88
EmoContext	LAC	91.69	88.00	86.46	92.94	NA	NA	NA
($\alpha = 0.05$)	CSCP	92.70	83.45	85.60	93.01	NA	NA	NA
	APS	94.01	96.43	97.77	93.47	NA	NA	NA
	CS-APS	99.00	95.58	94.35	96.81	NA	NA	NA

Appendix B. Comparison of prediction size

The detailed analysis of the size of prediction set approaches of LAC, CSCP, APS, and CS-APS across datasets MELD, EmoWoZ, and EmoContext is as follows: (Table 3)

Table 3. Results of four approaches on three selected emotional conversation datasets. Labels 0–6 for MELD are sadness, surprise, neutral, joy, anger, disgust, and fear; for EmoWOZ are neutral, fearful (sad), dissatisfied, apologetic, abusive, excited, and satisfied; and labels 0–3 for EmoContext are others, happy, sad, and angry

		Emotion labels						
Dataset	Method	lbl-0	lbl-1	lbl-2	lbl-3	lbl-4	lbl-5	lbl-6
MELD	LAC	4.21	4.11	2.35	2.82	4.18	5.26	5.16
	CSCP	5.08	5.24	3.46	3.49	5.16	5.54	5.68
	APS	4.48	4.72	2.85	3.13	4.49	5.13	5.12
	CS-APS	4.99	5.21	3.31	3.66	4.95	5.59	5.61

https://doi.org/10.1017/nlp.2024.48 Published online by Cambridge University Press

Table 3. (Continued)

		Emotion labels						
Dataset	Method	lbl-0	lbl-1	lbl-2	lbl-3	lbl-4	lbl-5	lbl-6
EmoWOZ	LAC	1.11	1.43	1.30	1.19	1.76	1.35	1.19
	CSCP	2.54	4.83	4.39	4.82	4.70	4.75	3.16
	APS	1.48	2.72	2.40	2.75	3.79	2.61	1.77
	CS-APS	1.78	3.40	2.92	3.36	4.84	3.18	2.15
EmoContext	LAC	1.10	1.12	1.06	1.09	NA	NA	NA
	CSCP	1.08	1.08	1.06	1.08	NA	NA	NA
	APS	1.65	1.88	1.78	1.84	NA	NA	NA
	CS-APS	1.66	2.00	1.90	1.99	NA	NA	NA

Cite this article: Roohi S, Skarbez R and Nguyen HD. Reliable uncertainty estimation in emotion recognition in conversation using conformal prediction framework. *Natural Language Processing* https://doi.org/10.1017/nlp.2024.48