Contents lists available at *Dergipark*

Journal of Scientific Reports-B

journal homepage: *https://dergipark.org.tr/en/pub/jsrb*

E-ISSN: 2717-8625 ***Number 12, April 2025***

RESEARCH ARTICLE

Receive Date: 20.05.2025 *Accepted Date: 23.06.2025*

A systematic testbed for evaluating emotion classification in large language models

Seda Nur Altun[a,*], Murat Dörterler[b]

[a]*Gazi University, Graduate School of Natural and Applied Sciences, Department of Computer Engineering, 06560, Ankara, Türkiye*
ORCID:0000-0001-9717-0759
[b]*Gazi University, Faculty of Technology, Department of Computer Engineering, 06560, Ankara, Türkiye*
ORCID: 0000-0003-1127-515X

Abstract

The advent of large language models (LLMs) in the domain of natural language processing (NLP) has engendered novel opportunities for the resolution of intricate tasks, such as emotion classification. However, achieving effective emotion analysis with LLMs requires more than simply choosing a ready-made model. In addition, the implementation of specially designed prompt structures, the alignment of the model with tokenisers, the meticulous formatting of both input and output data, and the regulated management of the generation process are imperative. The present paper sets out a technically detailed, reproducible framework for zero-shot and few-shot emotion classification using generative LLMs. The objective of this study is not to assess the efficacy of a given model, but rather to furnish researchers with a comprehensive manual outlining the essential components necessary to construct an LLM-based emotion recognition system from its fundamental principles. Utilising the Meta-LLaMA3 8B Instruct model and the DailyDialog dataset, the study demonstrates that prompt engineering tailored to the purpose, vocabulary-compatible tokenisation strategies, logit-level output constraint mechanisms and structured output normalisation can enable accurate and interpretable emotion classification, even in environments with limited or no labels. The objective of this paper is to furnish a practical and adaptive resource on the construction of LLM infrastructures that are context-sensitive, resilient to class imbalances and suitable for flexible task-oriented applications.

Keywords: emotion classification; zero-shot learning; few-shot learning; large language models

[*]* Corresponding author.
E-mail address: seda.nur.altun@gazi.edu.tr

1. Introduction

The advent of cutting-edge technological innovations has precipitated a paradigm shift in the utilization of computer systems across a myriad of sectors, including but not limited to banking, healthcare, transportation, and communication. The increasing importance of computer systems has led to a gradual expansion in the scope of human-computer interaction. As interaction has increased, the need for systems to communicate more effectively with the user has emerged. To increase this effectiveness, context-sensitive and emotionally coherent communication has become necessary [1]. To address this need, research in the domain of emotion analysis has witnessed a significant surge in recent years. The objective of emotion analysis studies is to facilitate the development of systems capable of recognizing an individual's emotional state and responding accordingly to ensure seamless, natural interaction. The findings obtained through the analysis of big data sources, such as social media content, customer feedback, product reviews, and call center records, further reinforce the importance of emotion analysis in terms of both increasing user satisfaction and improving the effectiveness of communication processes.

In NLP, emotion analysis is a fundamental task related to determining emotional states in data [2]. In studies on determining emotional states, we come across two concepts: emotion analysis and sentiment analysis. Although these two concepts are sometimes used interchangeably, they are different from each other. Sentiment analysis labels the text as positive, negative or neutral, while emotion analysis determines emotions such as anger, fear, joy, excitement, sadness in the text. At the beginning of the studies in this field, the aim was to determine the emotional tone in the text and the classification was handled in a binary structure [3], [4]. In the following studies, it was developed a little more and started to work with multi-class models and more detailed classifications were made as "very positive", "positive", "neutral", "negative" and "very negative". Emotion analysis aims to analyze the emotional states in the text in a more detailed way than sentiment analysis by classifying the text as joy, anger, sadness and surprise, fear, disgust, joy [5]. In the studies, more complex emotional structures can be analyzed better with the use of deep learning and LLMs in emotion analysis tasks.

Approaches applied in emotion analysis studies can be examined under two main headings as traditional and modern structures. In the traditional approach, dictionary-based methods and various supervised machine learning algorithms are used to detect emotional expressions. Dictionary-based methods aim to analyze the emotional intensity of words in the text using predefined emotion dictionaries. In these methods, the weighted average of the emotion score assigned to each word is taken to determine the general emotional tone of the text. However, limitations such as not taking into account contextual information, ignoring syntactic structure and misinterpreting ambiguous words reduce the accuracy of these methods [6]. These difficulties become more evident in speech data where emotional tone depends not only on words but also on the context, discourse history and speaker's intention. Since traditional methods have difficulty in situations where emotion can change depending on the context, the need for models that determine emotion by taking context into account has increased.

Machine learning-based approaches require labeled datasets to train classification models with various algorithms. While machine learning-based approaches are better at capturing contextual and syntactic features than dictionary-based methods, the success of these models varies depending on sufficient, balanced and correctly labeled data. Data tagging requires domain expertise and can be a bottleneck in model development. While this approach outperforms the lexicon-based approach, it is limited by the structural and semantic complexity of the language, and this highlights the need for deeper and more contextual approaches to emotion analysis.

LLMs have been used to overcome these challenges. Unlike traditional models, which require task-specific training, LLMs can perform tasks such as emotion or sentiment classification by interpreting instructions and producing the correct output. This eliminates the need for extensive retraining and makes them more useful in areas with limited resources. The ability of LLMs to understand context from a small amount of data, or even none at all, provides a valuable alternative when labelled data is difficult to obtain. Existing literature particularly emphasises studies focusing on fine-tuned scenarios, where models are trained with domain-specific labelled data. While these approaches have achieved remarkable results, zero-shot and few-shot learning methods have not received sufficient attention.

In response to this gap, the present study employs the Meta-LLaMA3 8B Instruct model to perform emotion classification in a zero-shot and few-shot setting, using the DailyDialog dataset as a testbed. Rather than focusing on the accuracy of the model itself, the objective is to construct and evaluate a complete infrastructure that includes prompt design, token control, output regulation, and label alignment. This ensures a reproducible and context-aware emotion analysis pipeline. The primary contribution of this work is the design of a technically grounded, task-agnostic pipeline for emotion classification using LLMs without task-specific training. The present study proposes a reusable framework that achieves a balance between model flexibility and task specificity through the implementation of structured prompt engineering and output normalization. The proposed configuration functions as a pragmatic exemplar for researchers seeking to execute cost-effective and scalable emotion classification in real-world, label-scarce environments.

2. Literature review

Text-based emotion detection is a significant research domain within the field of NLP. Initially developed through keyword and rule-based methods, this area of research has since evolved into machine learning, deep learning and, most recently, transformer-based models.

Keyword-based approaches are predicated on the analysis of the frequency of words associated with particular emotions in the text; however, they are inadequate for complex emotional expressions due to their insufficient understanding of the context. Machine learning-based methods have been developed to overcome these limitations and have achieved higher success rates, especially with the use of models trained with supervised learning techniques. These models provide a better understanding of the complex contextual framework inherent in text, leading to more significant improvements in research results. For example, Li et al. [7] developed a CNN-based model with the EACWT dataset generated on the Chinese social media platform Weibo and achieved an F1 score of 23.6%. Toçoğlu et al. [8] used artificial neural networks (ANN), convolutional neural networks (CNN) and long short-term memory (LSTM) models using Turkish tweet data and obtained accuracy rates of 71.42%, 74.00% and 72.28% respectively. Batbaatar et al. [9] evaluated the performance of CNN and BiLSTM models using DailyDialog, CrowdFlower, TEC and ISEAR datasets and achieved 84.8% accuracy on DailyDialog dataset. Using a GRU-based model, Jiao et al. [10] obtained F1 scores of 74.4%, 77.1% and 82.1% with Bi-GRU on Friends, EmotionPush and IEMOCAP datasets, respectively. The findings from all these studies show the effectiveness of deep learning on traditional classification tasks.

Despite the success of deep learning-based approaches, sentence-based classification alone has proven to be insufficient to comprehensively account for emotion transitions in speech. The Emotion Recognition in Speech (ERC) approach has emerged as a solution to this problem. The ERC approach facilitates the modeling of context and speaker relationships between texts, enabling context-sensitive emotion analysis. In recent years, significant progress has been made in the field of ERC and many different models have been proposed. These can be categorized into three different groups: recall-based methods, graph neural network (GNN)-based methods and transducer-based methods. Recall-based methods use recurrent neural networks (RNNs), long short-term memory (LSTM) and closed recursive unit (GRU) to analyze the emotional state of speech. Recall-based methods are designed to capture emotional states through the attention mechanism [11]. In this context,

Hu et al.[12] , inspired by human cognitive processes, tested the DialogueCRN model, which combines LSTM structures and attentional mechanisms to analyze speech context, on IEMOCAP and MELD datasets, achieving weighted F1 scores of 66.20% and 60.73% respectively. Later, the same research group developed the SACL-LSTM model [13] and obtained weighted F1 scores of 69.22%, 66.45% and 39.65% on IEMOCAP, MELD and EmoryNLP datasets, respectively. The EmotinIC model proposed by Liu et al. [14] provides a more comprehensive contextual analysis by considering emotional continuity and cognitive features in the emotion recognition process. With the EmotinIC model, weighted F1 scores of 69.61%, 66.32% and 40.25% were obtained in IEMOCAP, MELD and EmoryNLP datasets, respectively, and a macro F1 score of 54.19% was obtained in DailyDialog dataset. In GNN-based methods, interpersonal relationships in speech are represented by graph structures. This structure allows for more effective modeling of emotional interactions. The DialogueGCN model developed by Ghosal et al. [15] achieved a weighted F1 score of 64.18% on the IEMOCAP database. The model aims to capture the relationships between speakers with graphs. Shen et al. [16] used directed acyclic graphs (DAG-ERC) to structure speech content and obtained weighted F1 scores of 68.03%, 63.65%, 59.33% and 39.02% on IEMOCAP, MELD, DailyDialog and EmoryNLP datasets.

Transformer-based approaches are designed to monitor emotional relations in speech over extended periods. The SPCL+CL model developed by Song et al. [17] aims to enhance emotion recognition performance with a BERT-based structure, achieving 69.74%, 67.25%, and 40%, respectively. The KET model, as proposed by Zhong et al. [18], enhanced the transformer architecture by incorporating external information sources, thereby attaining 58.18%, 34.39%, and 59.56% weighted-F1 on the MELD, EmoryNLP, and IEMOCAP datasets, respectively. Additionally, the model achieved 73.48% and 53.37% micro-f1 on the EC and DailyDialog datasets. Furthermore, the CoG-BART model developed by Li et al. [19] and the TODKAT model developed by Zhu et al. [20] utilise generative approaches to ensure emotion continuity. In the context of the DailyDialog, MELD, IEMOCAP, EmoryNLP and CoG-BART models, the following weighted-F1 scores were obtained: 56.09%, 64.81%, 66.18% and 39.04%, respectively. Similarly, the TODKAT model obtained the following weighted-F1 scores: 52.56%, 68.23%, 61.33% and 43.12%, respectively.

However, the majority of the aforementioned models are reliant on task-specific training and substantial annotated datasets, which limits their generalisability and scalability. In view of this, the present study aims to develop a flexible and reproducible infrastructure for zero-shot and few-shot learning emotion classification using LLMs without requiring fine-tuning.

3. Material and method

3.1. Dataset and preprocessing

In this study, the English DailyDialog [21] dataset was utilised, a resource that is frequently employed in emotion classification tasks. The dataset was labelled in accordance with Ekman's six basic categories of emotions [22]. According to Ekman's model, emotions are categorised into six fundamental types: sadness, joy, fear, anger, disgust, and surprise [23], in addition to a neutral category designated as "no emotion".

Fig. 1. Ekman's emotion model [2]

The implemented data preprocessing process was designed in accordance with the structure of the Meta-LLaMA3 8B Instruct model and adopted a prompt-based approach as opposed to traditional encoder-based models. Initially, Unicode character distortions were eliminated, and all text was standardised using ASCII characters. Each dialog sample was processed in its original form according to the number of utterances it contained; no padding was performed to generate a fixed-length utterance count. It is important to note that token lengths were not filtered directly; however, the lengths of the input and output texts were controlled to a certain limit for the purpose of efficient use of the model's memory. Fig. 2 presents the average token length in the dataset as tokenized by the LLaMA 3 tokenizer. Each example was initialised with a guiding system message so that the model could comprehend the task correctly, and then context information, including previous lines of speech and the target utterance, was combined into a custom structure and presented to the model. The detailed structure of the prompt used to deliver each example to the model, including the system message, dialog context, and utterance, will be elaborated in Section 3.3.

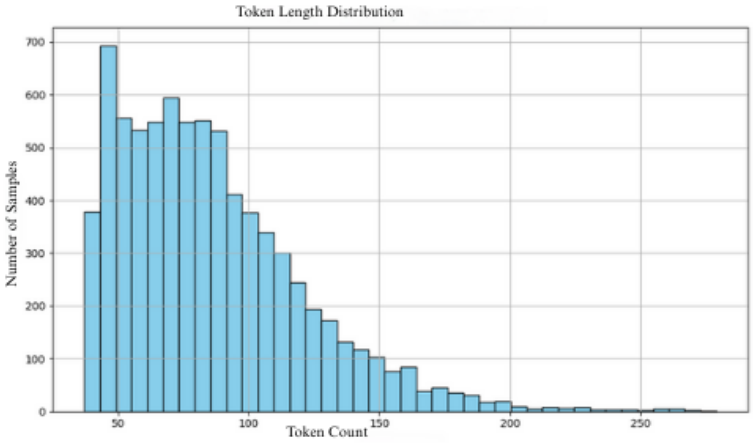

Fig. 2. token length distribution of input prompts using the LLaMA 3 tokenizer

The emotion labels were stored directly as textual expressions, without being converted into numeric format. As

the model outputs were also produced in the same format, these expressions were analysed by matching them to predefined classes. The process of dividing the texts into their constituent parts and presenting them to the model was carried out through the word parser of the model used; the pad process was structured in accordance with the expectations of the model. In this study, no training process was carried out; only the inference phase was performed, and all evaluations were made on the test data. The test data was used in accordance with the specified official ratio. The dataset under scrutiny contains 7,740 instances and exhibits a pronounced class imbalance. The examples are contextualised and each of them contains the fields of speech history, utterance and emotion tag.

Table 1. Key statistics of the DailyDialog dataset

Statistic	DailyDialog
#Dialogues	13118
#Test Dialogues	1000
Utterances	102979
Test Utterances	7740
Classes	7

Table 2. Emotion distribution in the DailyDialog test set

Emotion	Count	Percentage (%)
no emotion	6321	81.67
surprise	116	1.5
fear	17	0.22
happiness	1019	13.17
sadness	102	1.32
anger	118	1.52
disgust	47	0.61

This process is indicative of a contemporary preprocessing approach that has been optimised for zero-shot and few-shot emotion analysis tasks for LLMs.

3.2. Model

The LLaMA 3 8B Instruct model, which was developed by Meta AI and is based on transformer architecture, was utilised as the model in question. The architectural design of the model is illustrated in Fig. 3.

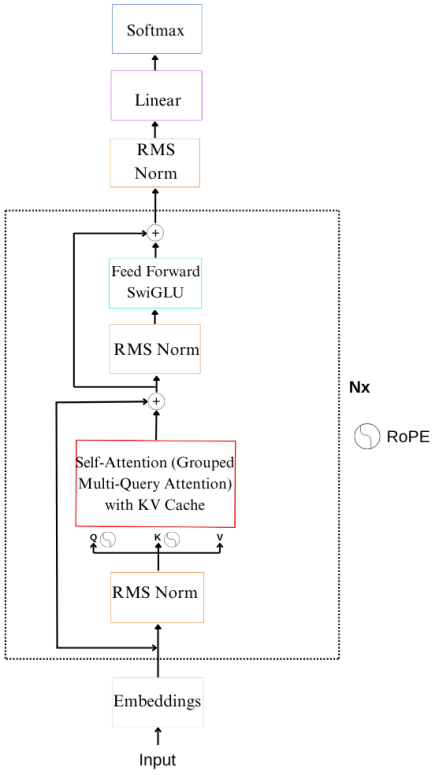

Fig. 3. LLaMA architecture

According to the official report published by Meta AI [24], the model consists of 32 layers in total, with each layer comprising a hidden representation space of size 4096. In the feed-forward network structure, the SwiGLU activation function is employed to enhance the nonlinear learning capacity. In the attention mechanism, there are 32 multiple headers in each layer, and the number of key-value attention headers is limited to 8. The structure has been designed in accordance with the Grouped Query Attention (GQA) principle, with the objective of reducing the computational load in the inference process and providing memory efficiency in the use of the KV cache.

In the normalization layers of the model, RMSE normalisation is applied in lieu of the classical LayerNorm, which provides enhanced stability during model training. The Rotary Positional Embeddings (RoPE) technique is employed for positional encoding; the model's capacity to manage extensive context windows is augmented by increasing the base frequency value in RoPE to 500,000. Tokenization is achieved through the utilisation of a bespoke tokenizer, which possesses a substantial vocabulary of 128,000 units. This facilitates the model's enhanced representation of the linguistic diversity. The architectural characteristics of LLaMA 3 render it a context-sensitive, consistent and powerful option in zero-shot and few-shot learning scenarios. The open-source nature of the model provides significant advantages in terms of reproducibility and research contribution.

3.3. Prompt design with zero shot approach

The initial stage of the zero-shot emotion analysis method employed in the present study is the creation of a prompt to guide the model correctly. The test samples employed encompass the preceding lines of speech and the final

utterance to be classified, with the objective of reflecting the context in which the utterance occurs. The presence of prior speech lines is imperative for the model to accurately interpret the meaning and predict the emotion with a higher degree of precision. The system message is incorporated into the prompts with the objective of enhancing the comprehensibility of the task. The objective of the present study is to restrict the output of the model to seven predefined emotions. In order to achieve this objective, the output field was subject to structural constraints due to the incorporation of the term "Emotion:" at the conclusion of the prompt. The model's efficacy in classifying emotions was demonstrated through its ability to do so based exclusively on task descriptions, with no requirement for the presentation of exemplars. This outcome suggests that the model has been successful in implementing a zero-shot approach. As demonstrated in Fig. 4, the zero-shot setting employs a prompt structure that integrates the instruction, dialog history, and target utterance to guide the model. The present study has demonstrated that such prompt designs offer a dual benefit in that they not only support the semantic consistency of the model, but also increase the reliability of the evaluation processes by facilitating the normalisation of the outputs.

```
System: You are an expert emotion classifier. Respond with one of the following labels: no
emotion, anger, disgust, fear, happiness, sadness, surprise
Previous conversation:
A: I can’t believe this is happening!
B: It’s been so stressful lately.

Now the utterance is:
I feel like I’m losing control.

Emotion:
```

Fig. 4. example prompt structure for zero-shot emotion classification

3.4. Prompt design with few shot approach

In the few-shot learning emotion analysis method, the prompt structure was restructured to be example-based in order to take advantage of the model's ability to learn from context. The prompt for each test case was extended to include not only the contextual information about the utterance to be classified and the previous lines of speech, but also a few representative examples of each emotion class. The purpose of these exemplifications is to facilitate the comprehension of the model with regard to the task description, thereby enabling the model to make appropriate classifications for analogous situations. In addition to the capacity to adhere to instructions, as exhibited by the LLaMA 3 8B Instruct model, the model's capacity to draw inferences from a limited set of examples was enabled. Furthermore, the model's proficiency in emotion classification was evaluated through a process of example-based assessment. At the conclusion of the prompt, the term "Emotion:" was employed once more to prompt the model to produce a particular response format, thereby sustaining the commitment to the classification task. The few-shot approach constitutes an effective strategy, especially for LLMs that are able to generalise from contextual examples. It was anticipated that the model would produce more accurate and targeted predictions in comparison to the zero-shot scenario. As demonstrated in Fig. 5, a representative example of the prompt structure employed in the few-shot setup is presented. This structure integrates contextual examples alongside the target utterance, thereby enhancing the model's comprehension.

```
System: You are an expert emotion classifier. Respond with one of the following labels: no emotion,
anger, disgust, fear, happiness, sadness, surprise

Example 1:
Previous conversation:
A: I just got promoted today.
B: That's amazing!

Now the utterance is:
I feel so proud and excited.
Emotion: joy

[Test instance]
Previous conversation:
A: I can't believe this is happening!
B: It's been so stressful lately.

Now the utterance is:
I feel like I'm losing control.
Emotion:
```

Fig. 5. example prompt structure for few-shot emotion classification

3.5. Tokenization

Language models cannot directly process explicit natural language input; therefore, text must be broken down into smaller meaningful units known as tokens before being converted into numerical representations. This process is called "tokenization". Tokenization is a fundamental step in the interpretation of text by the model. In this work, the text input is processed by a tokenizer that is trained to be fully compatible with the LLMs used. The functionality of this tokenizer is designed to facilitate the identification and understanding of new expressions by coding at the subword level. A special padding token is used to equalize the input dimensions of the model and to ensure alignment of strings during batch processing. The addition of this token standardizes input strings of different lengths. Here, the model's own terminator token is used as a padding token to meet the model's technical requirements. The construction aims to preserve the model's holistic data structure. As a result, a methodically structured tokenization process facilitates the model to process contextual information effectively, thereby improving classification performance.

3.6. Output generation

The tokenised prompts are presented as input to the LLaMA 3 8B Instruct model. The model's structural design enables it to generate the emotion label in accordance with the specified explicit task description. However, it should be noted that, by their very nature, LLMs may exhibit random deviations during the generation process. This can result in irrelevant, stylistically incorrect or overly descriptive output despite the instructions presented to the model. In order to circumvent the aforementioned issue, the study incorporated a logits processor component that intervened directly in the generation process.

Logits processor is a mechanism that intervenes in the probability distribution generated by the language model at each step. The logit values that are computed for the words that the model is capable of producing at each step are $\boldsymbol{z} = [z_1, z_2, ..., z_v] \in \mathbb{R}^V$. In this context, V denotes the total size of the vocabulary. The logits processor is a specialised

algorithm that functions to ensure the preservation of only those positions on the vector that correspond to predefined valid tags. In the event of positions not corresponding to valid tags, the logits processor replaces the logit values of these positions with negative infinity.

$$z_i = \begin{cases} z_i, if\ i \in V_{allowed} \\ -\infty, otherwise \end{cases} \quad (1)$$

Consequently, the probability of invalid words is set to zero prior to the application of the softmax function, thereby compelling the model to produce a single label. This results in a significant reduction of the model's output space, leading to the development of a production process that is directly aligned with the classification task. The implementation of limited generation in conjunction with logit-level intervention has been demonstrated to be a strategy that optimises both accuracy and stability in zero-shot tasks.

3.7. Output normalization

Despite the fact that solely tokens corresponding to specific emotion labels are permitted during the production process, the ordering or completion of these tokens during production may be erroneous. Consequently, it is imperative that the model outputs are subjected to a process of normalisation prior to the evaluation stage. The normalization process applied in this study is based on converting the model-generated response to lowercase, removing peripheral spaces and punctuation, and then comparing the output to the predefined set of tags. In the event that the output does not correspond to a valid label, it is automatically assigned the label "unknown". Consequently, the model's bias can be assessed independently of the evaluation metrics. This normalization step facilitates a more precise analysis of the model's output behaviour in zero-shot tasks.

4. Results

4.1. General performance evaluation

In this study, the Meta-LLaMA3 8B Instruct model was utilised to assess zero-shot and few-shot emotion analysis scenarios. The model demonstrated a 49.8% accuracy rate, a 27.9% macro-F1 score, and a 57.6% weighted-F1 score when evaluated with a task description-only prompt, representing a zero-shot scenario. In the few-shot scenario, where only one example representing each emotion class was presented to the model as prior knowledge, the success metrics were 60.4%, 30.6% and 67.1%, respectively. These findings indicate a substantial enhancement in the contextual generalization capability of the model with a restricted number of examples.

Table 3. Comparison of overall performance metrics for zero-shot and few-shot learning

Metric	Zero-Shot	Few-Shot
Accuracy	0.4979	0.6044
Macro-F1	0.2787	0.3060
Weighted-F1	0.5759	0.6716

The confusion matrix analysis revealed that the "no emotion" and "happiness" classes exhibited high prediction accuracies in both scenarios. However, the "fear", "disgust" and "surprise" classes demonstrated a substantial enhancement in classification accuracy with the few-shot approach.

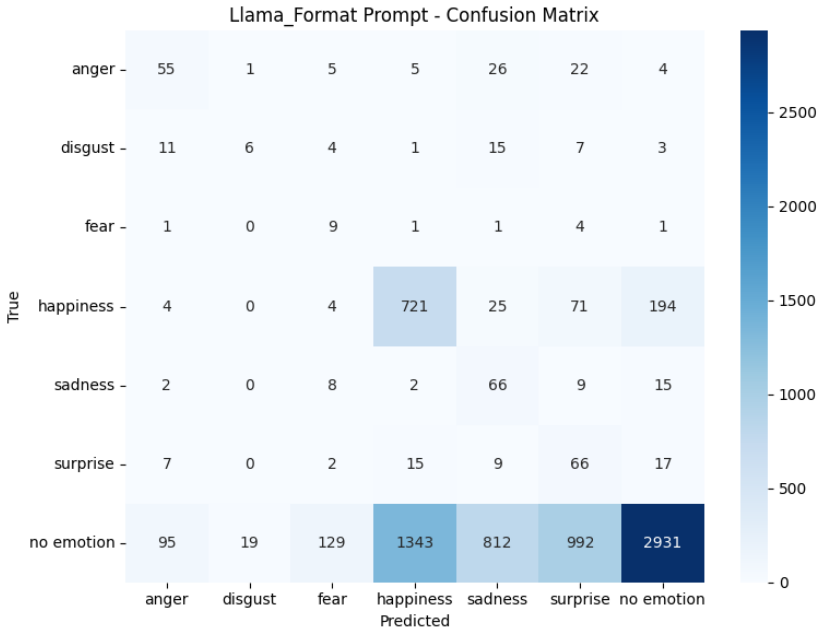

Fig. 6. zero-shot confusion matrix

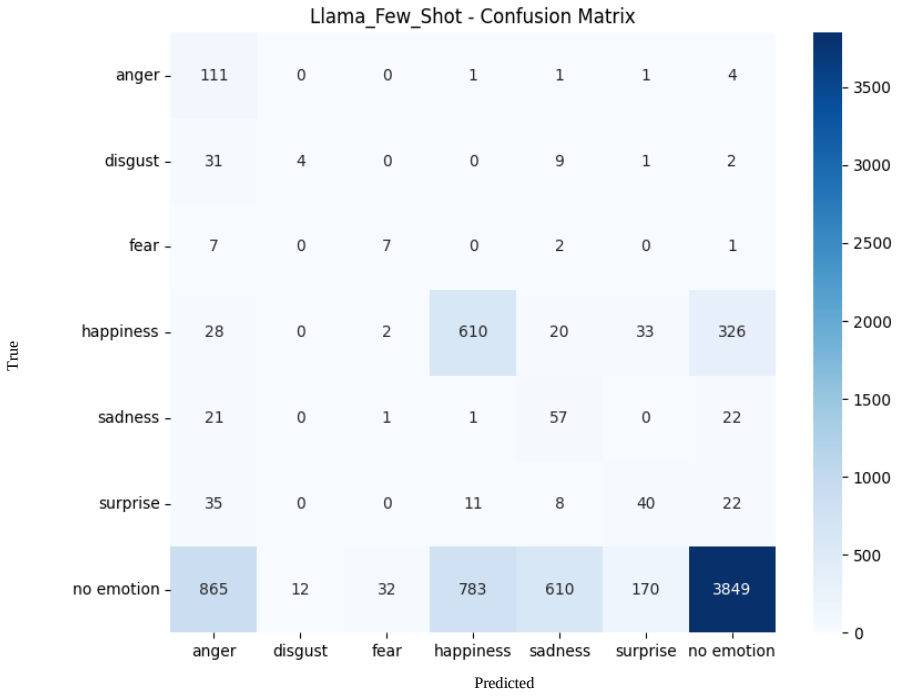

Fig. 7. few-shot confusion matrix

Specifically, the F1 score demonstrated a notable increase from 10% to 24% in the "fear" class and from 10% to 22% in the "surprise" class. The most successful class was "no emotion", where the F1 score was 62% in zero-shot and 73% in few-shot.

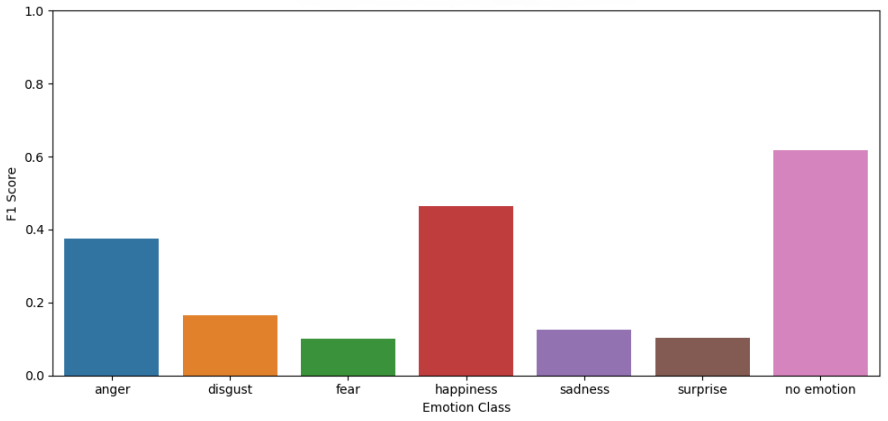

Fig. 8. zero-shot f1 scores per emotion class

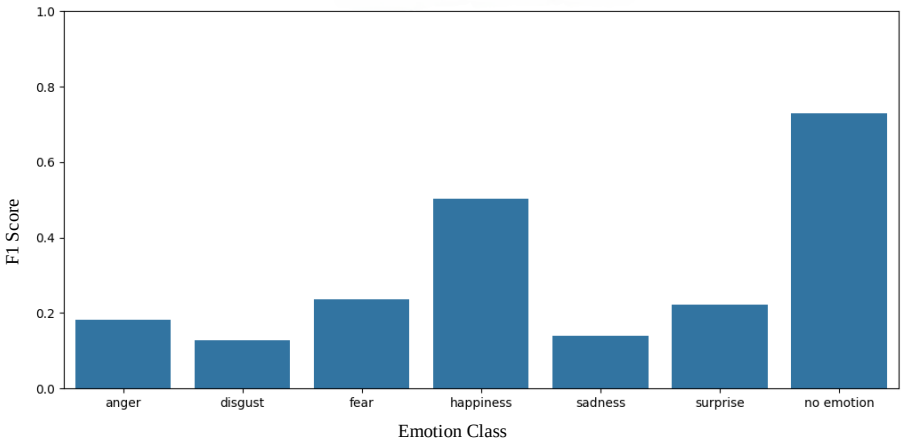

Fig. 9. few-shot f1 scores per emotion class

These disparities indicate that the model's output is influenced by the presence of guiding examples, suggesting its capacity to draw meaningful inferences from a limited number of examples, despite the discrepancy in representation between classes. A subsequent analysis of the distribution of labels reveals that the "no emotion" class accounts for over 81% of the dataset, while classes such as "fear" and "disgust" account for less than 1%.

This imbalance is a primary factor contributing to the suboptimal performance of the model, particularly in the zero-shot setting. However, it is concluded that the few-shot structure compensates for this imbalance to a certain extent, thereby improving the prediction accuracy by providing contextual information for low-frequency classes.

Table 4. Detailed class-wise result between zero-shot (ZS) and few-shot learning (FS)

Emotion	Precision (ZS)	Recall (ZS)	F1-Score (ZS)	Precision (FS)	Recall (FS)	F1-Score (FS)	Δ F1
anger	0.31	0.47	0.38	0.10	0.94	0.18	-0.20
disgust	0.23	0.13	0.16	0.25	0.09	0.13	-0.03
fear	0.06	0.53	0.10	0.17	0.41	0.24	+0.14
happiness	0.35	0.71	0.46	0.43	0.60	0.50	+0.04
sadness	0.07	0.65	0.12	0.08	0.56	0.14	+0.02
surprise	0.06	0.57	0.10	0.16	0.34	0.22	+0.12
no emotion	0.93	0.46	0.62	0.91	0.61	0.73	+0.11

A comparative analysis of the increase in weighted-F1 scores indicates that the LLaMA 3 model is capable of effectively operating with guided samples in unlabeled or under-labeled data environments. However, a noteworthy discovery emerged from the analysis: a decline in F1 scores was observed in the "anger" and "disgust" categories in the few-shot scenario. The aetiology of this phenomenon may be multifactorial. Firstly, it is important to note that the example sentences employed for these classes may not have adequately expressed the intended emotion. Secondly, given the limited number of instances in each class, the model may have lacked the capacity to accurately delineate the boundaries between the classes. This finding indicates that the selection of exemplars has a substantial influence on the emotion classification task.

4.2. Impact of output constraints and prompt design

In order to evaluate the individual contributions of output constraint mechanisms and instruction anchoring components, targeted ablation experiments were conducted by disabling the logits processor module and removing the emotion tag from the prompt structure. Detailed results when logits processor and emotion tag are disabled are given in Table 5.

Table 5. Ablation results: logits processor and emotion tag effect

Metric	Effect of removing the logits processor	Effect of removing the emotion tag
Accuracy	0.0084	0.6776
Macro-F1	0.0692	0.2627
Weighted-F1	0.0088	0.7143
Unknown Count	7575	3

When logits processor was disabled, which only allowed valid sentiment labels to be generated, the model's performance was observed to deteriorate significantly. The weighted F1 score dropped from 0.67 to 0.0088, and more than 97% of the predictions were invalid or irrelevant to the label set. This demonstrates the critical role of constrained decoding in maintaining the consistency of the outputs. Without this mechanism, the model produced long and meaningless texts that did not match valid labels, resulting in almost all instances being classified as "unknown". As shown in Fig. 10, when the logits processor component was disabled, the model's predictions were severely degraded. In particular, more than 97% of its outputs failed to match any valid sentiment labels, as evident in the sharp increase in "unknown" classifications.

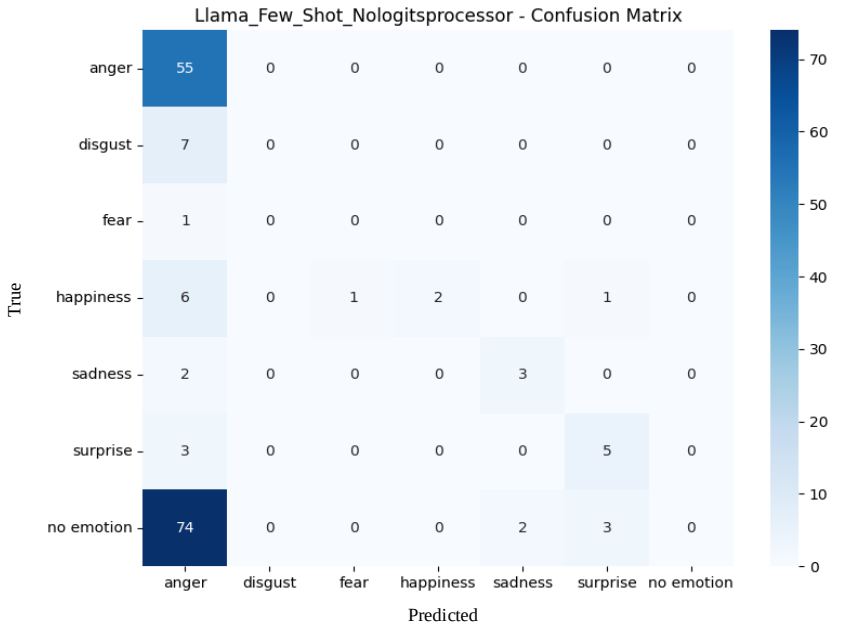

Fig. 10. confusion matrix for few-shot inference without logits processor

When the Emotion instruction tag was removed from the prompt structure, a significant change was observed in the general behavior of the model. Although the increase in the weighted F1 score may seem like a positive development at first glance, this increase does not reflect the real success of the model. Because this increase is due to accurate predictions especially on the dominant and easy classes; a significant uncertainty is observed on the rare or emotionally more complex classes. Removing the Emotion label caused the model to lose its instructional signal; this disrupted the formal consistency of the responses and led to structural deviations in the predictions. Qualitative analyses showed that the model sometimes described the emotion labels with indirect expressions or explanatory sentences instead of directly stating them. For example, instead of "anger", subjective expressions such as "This seems upsetting" were produced that did not fit the label system. In addition, although the number of predictions marked as "unknown" seemed to decrease, this decrease did not actually mean more consistent classification. On the contrary, the model produced responses that were semantically inappropriate but formally valid, and incorrect but recognizable predictions increased. This made the assessment appear better on the surface but open to contextual misinterpretation. As a result, it appears that target format anchoring elements such as "Emotion:" are not only guiding in the context of few-shot learning, but also a critical component that stabilizes model behavior. The absence of these signals weakens the model's predictions and reduces reliability on short or contextually ambiguous utterances. Therefore, despite some rising metrics, overall task performance is observed to decline. As can be seen in Fig. 11, removing the instruction label resulted in a wider dispersion of predictions, especially for the ambiguous classes.

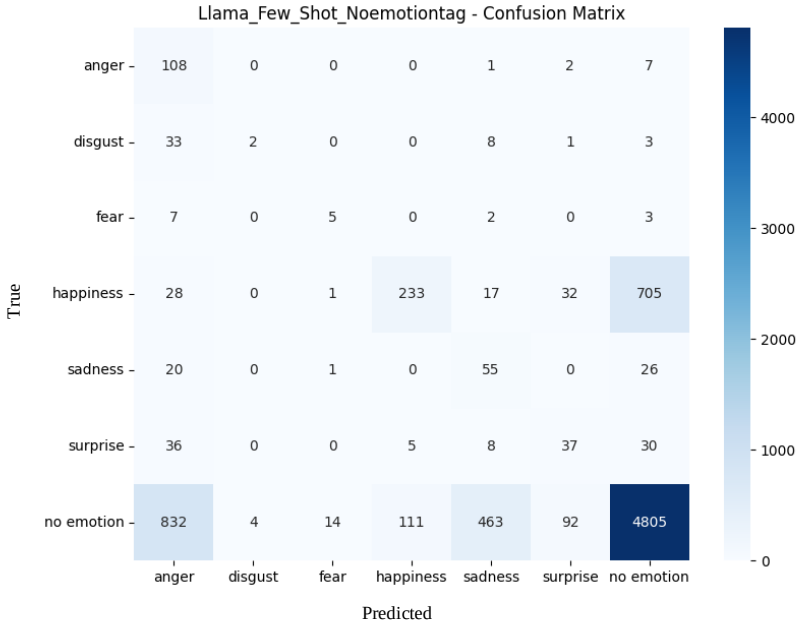

Fig. 11. confusion matrix for few-shot inference without emotion tag

5. Discussion

The present study investigates the efficacy of context-sensitive prompt generation processes, output control mechanisms, and structured input formats in zero-shot and few-shot learning scenarios. The Meta-LLaMA3 8B Instruct model was selected as the primary subject of analysis for this study. The findings of the present study demonstrate that LLMs possess the capability to make significant classifications in unlabelled environments. While the model demonstrates a higher level of success in high representation classes, it has been observed that the model is unable to generalise in low representation classes due to the difficulty of the contextual inference process. This underscores the model's sensitivity to the data distribution and underscores the significance of output routing components, such as the "Emotion:" phrase appended to the end of the prompt and the LogitsProcessor. In the Few-shot scenario, the exclusion of multiple examples per class from the prompt resulted in a substantial enhancement in performance metrics for low-frequency classes. This demonstrates the efficacy of contextual sampling-based prompt structures in enhancing class awareness. Conversely, the performance decrease observed in the “anger” and “disgust” categories in the few-shot scenario necessitates a more comprehensive evaluation of the representativeness of the examples belonging to these classes. In the study, although improvements were achieved in many emotion classes under the few-shot configuration, the F1 score of the “anger” class decreased from 0.38 to 0.18, and the F1 score of the “disgust” class decreased from 0.16 to 0.13. This unexpected situation indicates a structural limitation resulting from the fact that the provided examples were not emotionally specific and distinctive enough. As a result of the analysis conducted on the examples used for the few-shot scenario, it was seen that some sentences presented with the “anger” tag, such as “This is not true.” or “I could have actually bought the bag for less.”, contained polysemous structures that could be interpreted as regret, disappointment, or sadness depending on the context rather than a strong expression of anger. Similarly, from the examples belonging to the “disgust” class, “I don’t like this at all.” or “Maybe that's why it's cheap here.” do not clearly reflect the physical revulsion specific to the emotion of disgust; they rather display a general distaste or sarcasm. Although such examples seem structurally correct, they are weak in terms of

emotional clarity and semantic specificity. Moreover, considering that the "anger" and "disgust" classes in the dataset only constitute less than 2% of the total examples, these ambiguous examples further increase the uncertainty in the model's decision-making process, causing it to gravitate towards dominant and semantically overlapping classes such as "sadness" or "no emotion." This situation reveals that model performance in underrepresented emotion classes is directly related not only to the number of examples, but also to the emotional clarity and contextual richness of the examples. The fact that these classes obtain higher F1 scores in the zero-shot scenario indicates that the model may be negatively affected by inputs with insufficient emotional clarity when generalizing based on examples. In other words, the zero-shot structure driven only by the task definition was able to produce more stable results compared to the negative impact of low-quality few-shot examples.

The findings indicate that the efficacy of emotion classification with LLMs is contingent not solely on the model's capacity, but also on the quality of the infrastructure that governs the input-output flow. It is imperative to acknowledge the pivotal role that aspects such as token length constraints, encoding of speech context, separation of prompt fields, and output filtering at the logits level play in ensuring the alignment of the model output with the task goal. In this manner, the present study makes a contribution to the extant literature by means of proposing a low-cost, reusable and context-sensitive classification framework that can be implemented without the necessity for fine-tuning. In future work, the optimization of sample selection or the development of dynamic prompt templates depending on the data distribution has the potential to further advance this approach.

6. Conclusion and future work

The primary contribution of the study lies in the conceptualisation of a reusable, prompt-based infrastructure that can be employed in settings characterised by limited or absence of annotation. This infrastructure is further substantiated through experimental validation of its efficacy. The combination of system message, context coding and output limiting structures resulted in satisfactory classifications even for low frequency classes. This clearly demonstrates the impact of technical components such as prompt consistency, output control and token alignment on classification reliability.

In future studies, the performance of the LLaMA3 model can be improved with parameter-efficient fine-tuning techniques such as LoRA. With this approach, it is possible to update only a small subset of parameters without retraining all its parameters, and make a specific adaptation to the emotion recognition task. In addition, the language independence and transferability of the model can be evaluated by testing the zero-shot transfer performance using multilingual datasets such as MELD, which include examples from different language families. In addition, prompt structures can be enriched with structural metadata such as speaker role, topic labels, or speech purpose. Such information can provide a clearer separation of contextual emotion changes. This metadata can be added to the prompt as declarative fields in the prompt, and the model performance can be measured with controlled ablation experiments. In order to provide better adaptability of the model, retrieval-augmented prompt structures based on semantic similarity can be used. Here, the top-k semantically similar few-shot exemplars are selected and included in the prompt. This dynamic selection method can provide higher accuracy compared to static prompts. Finally, multi-step or chain-of-thought prompting strategies that promote reasoning before prediction could be investigated. For example, it could first generate an explanation by asking, "What emotion could this expression reflect and why?", and then classify by asking, "Now choose the most appropriate label." Such chain-of-thought prompts could help the model make more accurate decisions on ambiguous expressions and more effectively extract implicit emotional cues.

Acknowledgements

This study was conducted without any financial support from institutions or organizations. The author bears full responsibility for its content.

References

[1] A. Uçan, "TÜRKÇE HİS ANALİZİNDE OPTİMİZASYON VE ÖN-EĞİTİMLİ MODELLERİN KULLANIMI," Hacettepe University, Ankara, Turkey, 2020.
[2] E. Akçapınar Sezer *et al.*, "Türkçe bilgisayarlı dil bilimi çalışmalarında his analizi," *tday*, no. 70, pp. 193–210, Dec. 2020, doi: 10.32925/tday.2020.48.
[3] B. Pang *et al.*, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, Not Known: Association for Computational Linguistics, 2002, pp. 79–86. doi: 10.3115/1118693.1118704.
[4] J. Wiebe *et al.*, "Annotating Expressions of Opinions and Emotions in Language," *Language Res Eval*, vol. 39, no. 2–3, pp. 165–210, May 2005, doi: 10.1007/s10579-005-7880-9.
[5] S. Aman *et al.*, "Identifying Expressions of Emotion in Text," in *Text, Speech and Dialogue*, vol. 4629, V. Matoušek and P. Mautner, Eds., in Lecture Notes in Computer Science, vol. 4629. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 196–205. doi: 10.1007/978-3-540-74628-7_27.
[6] W. Medhat *et al.*, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.
[7] R. Li *et al.*, "EmoMix: Building an Emotion Lexicon for Compound Emotion Analysis," in *Computational Science – ICCS 2019*, vol. 11536, J. M. F. Rodrigues, P. J. S. Cardoso, J. Monteiro, R. Lam, V. V. Krzhizhanovskaya, M. H. Lees, J. J. Dongarra, and P. M. A. Sloot, Eds., in Lecture Notes in Computer Science, vol. 11536. , Cham: Springer International Publishing, 2019, pp. 353–368. doi: 10.1007/978-3-030-22734-0_26.
[8] M. A. Tocoglu *et al.*, "Emotion Analysis From Turkish Tweets Using Deep Neural Networks," *IEEE Access*, vol. 7, pp. 183061–183069, 2019, doi: 10.1109/ACCESS.2019.2960113.
[9] E. Batbaatar *et al.*, "Semantic-Emotion Neural Network for Emotion Recognition From Text," *IEEE Access*, vol. 7, pp. 111866–111878, 2019, doi: 10.1109/ACCESS.2019.2934529.
[10] W. Jiao *et al.*, "HiGRU: Hierarchical Gated Recurrent Units for Utterance-level Emotion Recognition," presented at the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019), Minneapolis, Minnesota: Association for Computational Linguistics, Apr. 2019, pp. 397–406. doi: 10.18653/v1/N19-1037.
[11] Z. Gou *et al.*, "TG-ERC: Utilizing three generation models to handle emotion recognition in conversation tasks," *Expert Systems with Applications*, vol. 268, p. 126269, Apr. 2025, doi: 10.1016/j.eswa.2024.126269.
[12] D. Hu *et al.*, "DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, 2021, pp. 7042–7052. doi: 10.18653/v1/2021.acl-long.547.
[13] D. Hu *et al.*, "Supervised Adversarial Contrastive Learning for Emotion Recognition in Conversations," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada: Association for Computational Linguistics, 2023, pp. 10835–10852. doi: 10.18653/v1/2023.acl-long.606.
[14] Y. Liu *et al.*, "EmotionIC: emotional inertia and contagion-driven dependency modeling for emotion recognition in conversation," *Sci. China Inf. Sci.*, vol. 67, no. 8, p. 182103, Aug. 2024, doi: 10.1007/s11432-023-3908-6.
[15] D. Ghosal *et al.*, "DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation," presented at the 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, Aug. 2019, pp. 154–164. doi: 10.18653/v1/D19-1015.
[16] W. Shen *et al.*, "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 (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, 2021, pp. 1551–1560. doi: 10.18653/v1/2021.acl-long.123.
[17] X. Song *et al.*, "Supervised Prototypical Contrastive Learning for Emotion Recognition in Conversation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 5197–5206. doi: 10.18653/v1/2022.emnlp-main.347.
[18] P. Zhong *et al.*, "Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations," presented at 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, Oct. 2019, pp. 165–176. doi: 10.18653/v1/D19-1016.
[19] S. Li *et al.*, "Contrast and Generation Make BART a Good Dialogue Emotion Recognizer," *AAAI*, vol. 36, no. 10, pp. 11002–11010, Jun. 2022, doi: 10.1609/aaai.v36i10.21348.
[20] L. Zhu *et al.*, "Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, 2021, pp. 1571–1582. doi: 10.18653/v1/2021.acl-long.125.
[21] Y. Li *et al.*, "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset," presented at the Proceedings of the Eighth International Joint Conference on Natural Language Processing, Taipei, Taiwan: Asian Federation of Natural Language Processing, 2017, pp. 986–995.
[22] V. Kalra *et al.*, "Importance of Text Data Preprocessing & Implementation in RapidMiner," presented at the The First International Conference on Information Technology and Knowledge Management, Jan. 2018, pp. 71–75. doi: 10.15439/2017KM46.
[23] P. Ekman *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion.," *Journal of Personality and Social*

Psychology, vol. 53, no. 4, pp. 712–717, 1987, doi: 10.1037/0022-3514.53.4.712.
[24] A. Grattafiori *et al.*, "The Llama 3 Herd of Models," Nov. 23, 2024, *arXiv*: arXiv:2407.21783. doi: 10.48550/arXiv.2407.21783.