"Stupid robot, I want to speak to a human!" User Frustration Detection in Task-Oriented Dialog Systems

Mireia Hernandez Caralt[*], **Ivan Sekulić**[*], **Filip Carević**[*], **Nghia Khau,**
Diana Nicoleta Popa, Bruna Guedes, Victor Guimarães, Zeyu Yang,
Andre Manso, Meghana Reddy, Paolo Rosso, Roland Mathis
Telepathy Labs GmbH, Zürich, Switzerland
`{firstname}.{lastname}@telepathy.ai`

Abstract

Detecting user frustration in modern-day task-oriented dialog (TOD) systems is imperative for maintaining overall user satisfaction, engagement, and retention. However, most recent research is focused on sentiment and emotion detection in academic settings, thus failing to fully encapsulate implications of real-world user data. To mitigate this gap, in this work, we focus on user frustration in a deployed TOD system, assessing the feasibility of out-of-the-box solutions for user frustration detection. Specifically, we compare the performance of our deployed keyword-based approach, open-source approaches to sentiment analysis, dialog breakdown detection methods, and emerging in-context learning LLM-based detection. Our analysis highlights the limitations of open-source methods for real-world frustration detection, while demonstrating the superior performance of the LLM-based approach, achieving a 16% relative improvement in F1 score on an internal benchmark. Finally, we analyze advantages and limitations of our methods and provide an insight into user frustration detection task for industry practitioners.

1 Introduction

Berkowitz (1989) defines frustration as an emotional state that is a result of the occurrence of an obstacle that prevents the satisfaction of a need. As such, in the context of task-oriented dialog (TOD) systems, detection of user's frustration is an essential component in ensuring the fulfillment of the user's goal (Hinrichs and Le, 2018). The importance stems from the fact that frustrated users often abruptly terminate their conversation with a TOD system, leading to a low likelihood of their return. Thus, timely detection of user frustration has many benefits, as the system can employ dialog flow repair techniques or transfer the user to a
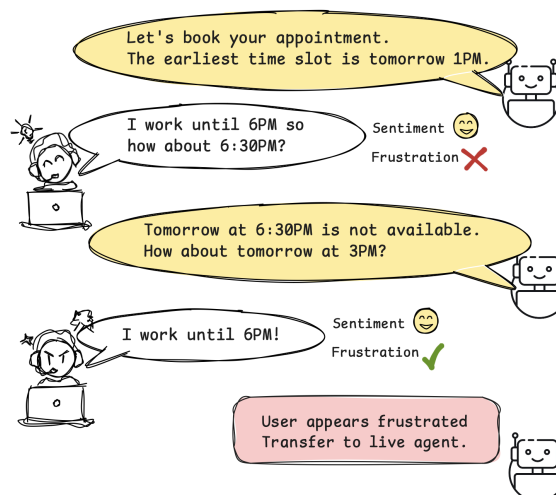

Figure 1: Example of user frustration in a deployed TOD system. The user can only come after 6PM due to work, but the system misses this and suggests the next available slot. Traditional sentiment models often fail to detect such nuances, as there is no explicit mention of negative sentiment.

human agent, in order to improve the user experience (Zhang et al., 2023a).

While a large body of work on the topic of emotion detection in dialog (Pereira et al., 2022; Zhang et al., 2023b; Wang et al., 2024) and dialog breakdown detection (Li et al., 2020; Terragni et al., 2022) exists, research on user frustration detection in real-world TOD systems is scarce. Therefore, in this paper, we present a unique perspective from the industry, analyzing user frustration in conversations from a deployed TOD system.

Specifically, we showcase the gap between research-oriented approaches and real-world applications. To this end, we compare emotion detection datasets constructed for academic research, namely EmoWoZ (Feng et al., 2022), to our internal data gathered from real users conversing with a deployed TOD system, finding several differences, discussed in Sect. 5.1. We hypothesize that the differences arise mainly from the fact that there is

---

[*]These authors contributed equally to this work.

276

a real sense of urgency and importance in completing the real-world tasks, whereas users in academic benchmarks are more cooperative and may even tolerate mistakes from the TOD system.

Additionally, we assess established methods for sentiment analysis (Hartmann et al., 2023), emotion detection (Huang, 2024), and dialog breakdown detection (Bodigutla et al., 2020) on our internal dataset, concluding they are mostly insufficient for successfully detecting user frustration. In an attempt to mitigate this gap, we propose two approaches, stemming from two different angles: i) currently deployed keyword-based user frustration detection method, grounded in sentiment analysis; ii) novel, emerging in-context learning LLM-based method.

We conclude that our rule-based approach, although precise, fails to detect user frustration in a large number of cases. Furthermore, LLM-based approach outperforms all of the aforementioned approaches on our internal frustration detection benchmark. Finally, we outline promising directions for future work through the industry-relevant perspective.

2 Related Work

Early detection of user frustration is essential for improving the quality of TOD systems. Causes of user frustration include poor performance, poor utility and poor usability of the systems they interact with (Hertzum and Hornbæk, 2023), such as the inability of the system to correctly understand user requests, a mismatch between user expectations and obtained results or an overall dissatisfaction with the provided results.

Much of the existing work on user frustration explores the problem from a non-technical view (Goetsu and Sakai, 2020; Brendel et al., 2020; Hertzum and Hornbæk, 2023), focuses on the broader scope of human-machine interaction (Weidemann and Rußwinkel, 2021) or explores mitigating breakdowns in such interactions (Li et al., 2020; Terragni et al., 2022), yet without explicitly targeting user frustration detection. Additionally, some studies have focused on user satisfaction estimation and overall dialog quality assessment (Rach et al., 2017; Bodigutla et al., 2019a, 2020; Sun et al., 2021) Since determining user frustration is a more targeted goal, it can be seen as a subset of such studies, making the nuances of mapping user satisfaction levels to frustration both challenging and error-prone. Similarly, emotion detection (Pereira et al., 2022; Zhang et al., 2023b; Wang et al., 2024) could be seen as an important prerequisite for user frustration assessment. The main idea behind such studies is that frustration is present if emotions such as dissatisfaction or anger can be detected in the given textual content.

As detecting user frustration is both subtle and complex, hand-crafted feature engineering may also not be enough (Hinrichs and Le, 2018; Ang et al., 2002). Meanwhile, LLMs have recently yielded impressive performance on a variety of tasks, encoding much general world knowledge (Touvron et al., 2023; OpenAI, 2023). We thus investigate to what extent such an approach could be suitable for detecting user frustration. To the best of our knowledge, this is the first work targeting unsupervised user frustration detection using LLMs within deployed TOD systems.

3 User Frustration Detection

In this section, we first formalize the task of user frustration detection. Next, we describe our deployed rule-based approach and propose a novel method based in in-context learning with LLMs. Finally, we describe competitive baselines used as comparison to our approaches and details regarding our internal benchmark data, sourced from a deployed TOD system.

We define $\mathcal{U}$ as the domain of all textual utterances. Then, given an ordered list of tuples (i.e., a dialog history) $H = [(s_i, u_i) \mid i \in \{1, \ldots, t\}]$, where $s_i \in \mathcal{U}$ and $u_i \in \mathcal{U}$ denote system and user utterance at dialog turn i, respectively, the goal is to find such function $f : \mathcal{U} \times \mathcal{U} \to \{0, 1\}$ that for presence of frustration in the dialog outputs positive label, and negative otherwise.

3.1 Rule-Based Approach

Our deployed user frustration detection system relies on keyword match in user utterances. Specifically, we have curated a set of keywords $K = \{kw_1, \ldots, kw_N\}$, mainly composed of profanity, words explicitly indicating negative emotion, and insults, indicating potential frustration. In practice, for each keyword $kw_i \in K$, we check if it appears in the current user utterance $kw_i \in u_t$; if a match is found, the conversation is marked as frustrated.

3.2 LLM-Based Approach

In-context learning (ICL) paradigm with LLMs has demonstrated strong performance across a

wide range of tasks, without the need for time- and compute-expensive fine-tuning (Brown et al., 2020). We hypothesize that LLMs are well-suited to identifying the nuanced indicators of user frustration when given an appropriate description and context. Therefore, we propose a novel approach for frustration detection in TOD systems, based on ICL with LLMs.

Specifically, we design an ICL prompt $\mathcal{P}(T, D, H)$ which includes: **i) task description (T)** of user frustration in the context of TOD systems and common cues for its identification, **ii) domain (D)** of the conversation (e.g., booking appointments), which helps the language model understand the expected interaction patterns and **iii) conversation history (H)** formatted as a string with *"USER:"* and *"SYSTEM:"* prefixes to distinguish between roles. The LLM processes this context and generates the corresponding binary frustration label $UF = f_{LLM}(\mathcal{P}(T, D, H))$.

The prompt strings used are detailed in App. A. In the experiments shown in Sect. 4 we also provide the results obtained when further augmenting the prompt with few-shot examples, and compare them with this zero-shot approach.

3.3 Baselines

We compare our aforementioned in-house methods for user frustration detection with baselines from three different streams of approaches: i) sentiment analysis; ii) emotion detection; iii) dialog breakdown detection (DBD).

Sentiment Analysis and Emotion Detection. Sentiment analysis baselines aim to classify the sentiment of a text as either *positive* or *negative*. All samples having *negative* sentiment are considered as entailing user frustration. As the representative of this group, we leverage *RoBERTa-large* model, fine-tuned on a large variety of sentiment classification datasets (Hartmann et al., 2023). We dub this method *RoBERTa-Sent*. On the other hand, emotion detection involves identifying a specific emotion from a predefined set in a given text. We employ two models: a distilled version of *BERT* model trained on the conversational EmoWOZ (Feng et al., 2022) dataset, and and a distilled version of RoBERTA model trained on various emotion classification datasets (Hartmann, 2022). The first model predicts the emotions of *satisfaction*, *dissatisfaction*, *abuse*, *apology*, *excitement*, *fear* and *neutrality*, while the latter one

classifies the states of *anger*, *fear*, *disgust*, *joy*, *neutrality*, *sadness* and *surprise*. We dub these models *DistilBERT-EmoWoZ* and *DistilRoBERTa–Emo*, respectively. In this work, *dissatisfaction* and *abuse*, as well as *anger* and *disgust* are considered as indicators of users' frustration.

As the above-described methods might be designed for either single-sentence or full-conversation input formats, following Feng et al. (2022), we evaluate them with two different input types: i) only the last user utterance u_N as input (*-LU*); ii) full conversation H as input (*-FC*). The input format is indicated by appending the abbreviation to the method dub.

Feature-Based Dialog Breakdown Detection. This method leverages hand-crafted features in order to estimate the amount of user satisfaction on different conversation levels (Bodigutla et al., 2019a,b). We adapt this baseline to our use-case by applying a simple classifier on top of the subset of features presented in (Bodigutla et al., 2019a). The list of used features is portrayed in App. B.

3.4 Data

Our data consists of real user conversations with our currently deployed TOD system. The conversations generated by this system are often lengthy and span multiple dialog phases. In this work, we focus on two specific dialog phases that are particularly prone to user frustration: i) booking negotiations, where the system attempts to schedule a suitable time slot for a user seeking an appointment; and ii) receptionist, where the system attempts to route the user to the appropriate department or agent based on their needs.

We collect conversations with more than one turn from a week of production data, resulting in a dataset of 270 booking negotiations and 285 receptionist transfers. Further details and comparison to EmoWoz (Feng et al., 2022) is shown in Table 3.

Although previous work has explored automated signals to detect user frustration, such as hang-ups or requests for a live operator (Terragni et al., 2022), these methods are susceptible to noise, since a frustrated user may choose to continue the conversation, or a user might hang up for reasons unrelated to frustration. Therefore, we conduct manual annotation of the collected data: we employ three in-house experts to annotate each sample with a binary label indicating whether the user is frustrated or not. The annotators were provided with

	UF = 0 (not frustrated)			UF = 1 (frustrated)			
	P	R	F_1	P	R	F_1	Macro-F_1
Sentiment Analysis (Hartmann et al., 2023)							
RoBERTa-Sent-FC	0.73	0.11	0.18	0.34	0.92	0.50	0.34
RoBERTa-Sent-LU	0.77	0.72	0.75	0.51	0.58	0.54	0.65
Emotion Detection							
DistilBERT-EmoWoZ-FC (Huang, 2024)	0.70	0.77	0.73	0.42	0.34	0.38	0.56
DistilBERT-EmoWoZ-LU	0.74	0.68	0.71	0.45	0.52	0.48	0.60
DistilRoBERTa-Emo-FC (Hartmann, 2022)	0.67	1.00	0.80	0.00	0.00	0.00	0.40
DistilRoBERTa-Emo-LU	0.68	1.00	0.81	0.88	0.04	0.07	0.44
Dialog Breakdown Detection							
DBD+LogReg (Bodigutla et al., 2020)	0.78	0.93	0.85	0.78	0.46	0.58	0.71
Rule-Based Approach							
Keyword Matching	0.67	**1.00**	0.80	**1.00**	0.01	0.01	0.41
LLM-based ICL Approach							
GPT-4o-zero-shot	0.98	0.83	0.90	0.74	0.96	0.83	0.86
GPT-4o-two-shot	0.85	0.97	**0.91**	0.92	0.66	0.77	0.84
Llama-3.1-405B-zero-shot	**0.99**	0.74	0.85	0.67	**0.99**	0.79	0.83
Llama-3.1-405B-two-shot	0.97	0.84	0.90	0.75	0.96	**0.84**	**0.87**

Table 1: Results of various approaches for user frustration detection on our deployed TOD system benchmark.

guidelines, reaching an inter-rater agreement, as measured by Fleiss' κ, of 0.48. This agreement indicates moderate reliability (Landis, 1977), while also suggesting a degree of subjectivity in the annotation task. Disagreements were manually resolved in a post-processing phase.

4 Results

Table 2 presents the results of our experiments. While we primarily focus on macro-F1 for performance comparison, we additionally focus on recall for UF=1 (frustrated), since it highlights the proportion of frustrated conversations correctly identified, which has the greatest impact on user satisfaction. We make several observations from the results, with a follow-up discussion presented in Sect. 6.

Sentiment analysis models and emotion detection models fine-tuned for the TOD domain perform poorly compared to dialog breakdown or LLM-based ICL approaches. Notably, these baselines perform the best when only the final user utterance is considered, with the performance of the best model dropping from 66% to 34% when analyzing the full conversation history. This suggests that these models capture only the emotion in isolated utterances, but fail to detect frustration cues embedded in the broader conversational context.

Features derived from the dialog breakdown detection domain are effective for detecting user frustration, outperforming sentiment and emotion detection baselines. However, although DBD outperforms them in terms of Macro-F1, it does seem inclined towards the UF=0 class, as suggested by a relatively low recall in the UF=1 class.

Our currently deployed keyword-based approach achieves 100% precision, but suffers from an extremely low recall of only 1% for frustrated conversations, resulting in a very low Macro-F1 score of 41%. This indicates that poor conversation handling does not always manifest as overtly negative language. While keyword-based methods may be inexpensive, they are inadequate for capturing the full range of frustrated scenarios.

ICL with LLMs outperforms all other approaches, both in zero- and two-shot settings. Specifically, in terms of Macro-F1, we observe more that $+33\%$ relative improvement over sentiment and emotion detection methods and $+22\%$ relative improvement over the DBD method. We further note the comparable performance of both LLaMA-3.1-405B and GPT-4o in both zero- and few-shot settings, suggesting that our ICL prompt generalizes well across different LLMs and number of shots. However, adding few-shot examples to the prompt does not yield substantial performance improvement, and even slightly degrades performance for GPT-4o.

5 Qualitative Analysis

This section compares academic benchmarks with real-world data and qualitatively analyzes open-source and in-house methods.

5.1 Academic Data vs. Real-World Data

We compare our data to an academic benchmark for emotion detection in TOD systems, EmoWoZ (Feng et al., 2022), built in a controlled lab environment through Wizard-of-Oz and crowd-

	Macro-F_1
Sentiment Analysis (Hartmann et al., 2023)	
RoBERTa-Sent-FC	0.34
RoBERTa-Sent-LU	0.65
Emotion Detection	
DistilBERT-EmoWoZ-FC (Huang, 2024)	0.56
DistilBERT-EmoWoZ-LU	0.60
DistilRoBERTa-Emo-FC (Hartmann, 2022)	0.40
DistilRoBERTa-Emo-LU	0.44
Dialog Breakdown Detection	
DBD+LogReg (Bodigutla et al., 2020)	0.71
Rule-Based Approach	
Keyword Matching	0.41
LLM-based ICL Approach	
GPT-4o-zero-shot-LU	0.67
GPT-4o-two-shot-LU	0.75
Llama-3.1-405B-zero-shot-LU	0.63
Llama-3.1-405B-two-shot-LU	0.75
GPT-4o-zero-shot-FC	0.86
GPT-4o-two-shot-FC	0.84
Llama-3.1-405B-zero-shot-FC	0.83
Llama-3.1-405B-two-shot-FC	**0.87**

Table 2: Comparison of UF detections performed on last user utterance (*LU*) and full conversations (*FC*).

	EmoWoZ	Internal data
# Dialogues	11,438	555
# Unique tokens	28,417	995
Avg. tokens / user turn	10.6	3.1
Avg. user tokens/dialogue	55.6	7.8
% Repeated Utt. (fuzzy)	2.1%	8%
% Repeated Utt. (cosine)	4%	9.6%

Table 3: Internal dataset vs. EmoWOZ benchmark.

sourcing techniques. In contrast, our internal dataset derives from real user interactions with a deployed TOD system, resulting in differences in emotional complexity, dialogue flow, frustration triggers, and user behavior. A key distinction is that real-world data includes dialogues where unfulfilled tasks can have tangible negative effects, creating a more urgent and authentic environment than academic benchmarks.

Through our analysis, which included two experts analyzing hundreds of conversations of both datasets, we observe several patterns indicating differences between them: i) different frustration triggers; ii) real-world urgency vs lenient lab environment; iii) number of user requests for human assistance; iv) user familiarity with the system.

Frustration in our data stems from task-specific issues, like unavailable appointment times and the system's persistence with non-preferred options, while in EmoWoZ, it arises from a broader range of factors like misunderstandings or delays. Real users aim to complete tasks in real-time, so system failures lead to immediate dissatisfaction. In con-

trast, users in a lab environment face less pressure to complete tasks and can often move on to another goal within the same dialogue, giving the system a chance to recover or simply accepting the mistake.

Moreover, requests for human assistance are frequent and explicit in real-world, while almost non-existent in the EmoWoZ data. While this is understandable in academic benchmarks, where dialogues are acquired through WoZ techniques, such requests are nonetheless an important part of deployed TOD systems and should be looked at through the lens of potential frustration. Finally, we observe cases where user familiarity with the capabilities of the system impacts their ability to fulfill their tasks without being frustrated. Such scenarios are not possible in academic benchmarks, as users in the lab are given specific instructions on how to interact with the system (Kim et al., 2024).

As shown in Table 3, frustration in real-world data is frequently conveyed through context and repetition rather than explicit wording. This aligns with the findings of qualitative analysis presented above. User responses are typically concise and to the point due to the nature of human-system interactions, with an average token count per user utterance reflecting this directness. Also the choice of words tends to be simpler and with less variability as indicated by a lower number of unique tokens present in the dialogues.

5.2 Open-Source vs. In-House Methods

We observe a limitation of sentiment and emotion detection methods in presence of system's apologetic behavior (e.g., responses containing phrases like *"I'm sorry"* or *"I apologize"*), as shown on Fig. 2a. Such system utterances tend to dominate the overall perception of sentiment/emotion in the conversation, thereby diminishing the impact of expressed frustration in one or more user utterances. Both the LLM- and DBD-based method are more robust to this phenomena.

Sentiment analysis and emotion detection approaches fail to recognize the repetition of users' requests, in cases where these requests were not initially met by the system, as an indication of frustration (Fig. 2b). On the other hand, LLM-and DBD-based methods more successfully capture such repetition pattern.

However, all of the approaches encounter difficulties in long negotiation scenarios (Fig 2c), with a significant number of false positives arising from misinterpreting the negative sentiment and repe-
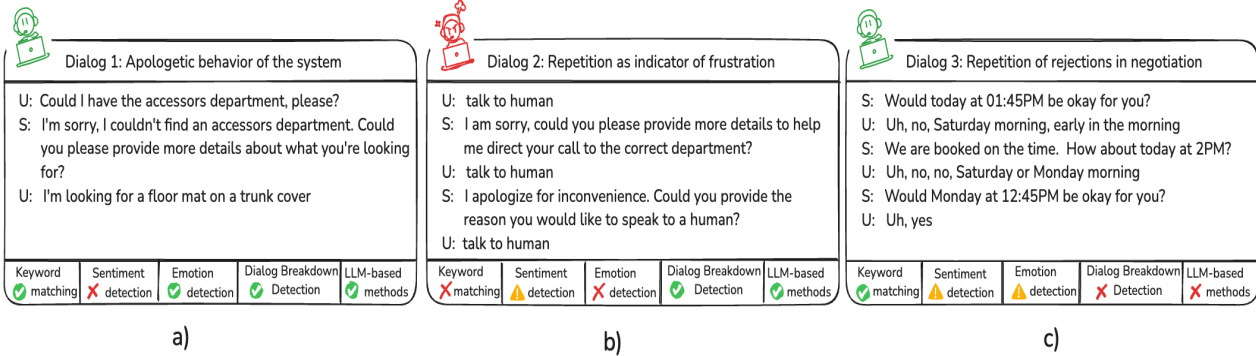
Figure 2: a) Negative sentiment prediction influenced by the apologetic behavior of the system b) Frustration caused by the system's failure to transfer the user to live agent. Yellow exclamation sign indicates that the example has been correctly classified by the FC-based sentiment approach only. c) Non-frustrated repetition of rejections in the process of time slot negotiations. Yellow exclamation sign indicates that the example has been correctly classified by LU-based and incorrectly by FC-based sentiment/emotion detection methods.

titions of users' rejections (e.g., responding "no" to system's question about user's availability at a certain time slot) as frustration. Moreover, emotion-based methods exhibit high sensitivity to interjections in the text, such as *"uh"* or *"ah"*.

Finally, we observe poor performance of LLM-based methods in short conversations, which often lack significant contextual information. These conversations typically consist of up to three turns.

6 Discussion

User frustration detection is an important component in real-world TOD systems. We argue that, especially with the rise of popularity of conversational interfaces (McTear, 2017), the task tackled in this study is essential for maintaining user satisfaction and engagement. However, as pointed out in Sect. 5.1, current academic benchmarks for similar tasks are too sterile, as the dialogues were created in a controlled lab setting. The differences between the real-world and academic benchmarks stem mainly from the real sense of urgency of fulfilling the task in the real world, while the simulated lab environment lacks the unpredictability and pressure of real-world scenarios. Therefore, we call for additional attention to user frustration detection both from academia and industry practitioners.

Frustration manifests in many ways, not limited to negative language. As detailed in Sect. 4, our currently deployed keyword-based approach, that relies on identifying profane and negative language in user utterances, suffers from extremely low recall. This, together with qualitative analysis presented in Sect. 5.2, indicates that poor conversation handling does not always manifest as overtly negative language. Thus, while keyword-based

methods may be inexpensive, they are inadequate for capturing the full range of frustrated scenarios. On the other hand, our method fitted on the dialog breakdown detection features performs fairly well. Although this approach is not directly comparable to the out-of-the-box or zero-shot methods, it shows how dialog breakdown relates to our user frustration task; features such as repetition and negation serve as strong indicators of frustration.

General-domain emotion and sentiment models are insufficient for real-world TOD systems. The poor performance might stem from the facts that the manifestation of frustration varies from person to person (Bandura, 1973) and includes a wide variety of emotions, e.g., depression (Berkowitz, 1989), thus going beyond the fixed set of emotions typically covered by pre-trained methods. We hypothesize that their performance would increase if fine-tuned on domain data of a TOD system. However, such approach introduces maintenance overhead for systems operating across diverse and evolving domains. Further, a drop in performance when full conversations were used, indicates that such methods are over-sensitive to non-emotion-related text, as the overall emotion expressed by the user gets diluted by system's utterances.

ICL with LLMs is an emerging method for frustration detection. LLM-based methods outperform all baselines, suggesting that they capture both semantic- and dialog structure-related signals. Moreover, similar performance of GPT-4o and Llama-3.1 demonstrates that our ICL prompt generalizes well across different LLMs. Another advantage of this approach is that it can be adapted to any domain as long as the domain is adequately described in the ICL prompt (Feng et al., 2024).

7 Conclusions

In this study, we investigated the feasibility of user frustration detection with out-of-the-box methods, including open-source sentiment and emotion detection, as well as deployed rule- and LLM-based methods. We conclude that open-source methods are not fit for production TOD systems, likely due to the nature of the data, which is vastly different from real-world data, they were trained on. Moreover, we find an LLM-based approach promising, as it tends to capture both emotion and potential dialog breakdowns, thus significantly outperforming other methods. Future work encapsulates a promising direction of multi-modal (speech + text) methods for user frustration detection (Ang et al., 2002). Finally, we aim to expand our detection across multiple user calls, therefore creating a user profile that can help with frustration detection.

References

Jeremy Ang, Rajdip Dhillon, Ashley Krupski, Elizabeth Shriberg, and Andreas Stolcke. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. *ICSLP 2002*.

Albert Bandura. 1973. *Aggression: A social learning analysis*. Prentice-Hall.

Leonard Berkowitz. 1989. Frustration-aggression hypothesis: examination and reformulation. *Psychological bulletin*, 106(1):59.

Praveen Kumar Bodigutla, Lazaros Polymenakos, and Spyros Matsoukas. 2019a. Multi-domain conversation quality evaluation via user satisfaction estimation. *The 3rd Conversational AI workshop: Today's practice and tomorrow's potential at NeurIPS 2019*, abs/1911.08567.

Praveen Kumar Bodigutla, Aditya Tiwari, Spyros Matsoukas, Josep Valls-Vargas, and Lazaros Polymenakos. 2020. Joint turn and dialogue level user satisfaction estimation on multi-domain conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3897–3909, Online. Association for Computational Linguistics.

Praveen Kumar Bodigutla, Longshaokan Wang, Kate Ridgeway, Joshua Levy, Swanand Joshi, Alborz Geramifard, and Spyros Matsoukas. 2019b. Domain-independent turn-level dialogue quality evaluation via user satisfaction estimation. *CoRR*, abs/1908.07064.

Alfred Benedikt Brendel, Maike Greve, Stephan Diederich, Johannes Bührke, and Lutz M Kolbe. 2020. You are an idiot!-how conversational agent communication patterns influence frustration and harassment. In *AMCIS*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gasic. 2022. EmoWOZ: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4096–4113, Marseille, France. European Language Resources Association.

Shutong Feng, Guangzhi Sun, Nurul Lubis, Wen Wu, Chao Zhang, and Milica Gasic. 2024. Affect recognition in conversations using large language models. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 259–273, Kyoto, Japan. Association for Computational Linguistics.

Shiyoh Goetsu and Tetsuya Sakai. 2020. Different types of voice user interface failures may cause different degrees of frustration. *CoRR*, abs/2002.03582.

Jochen Hartmann. 2022. Emotion english distilroberta-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

Morten Hertzum and Kasper Hornbæk. 2023. Frustration: Still a common user experience. *ACM Transactions of Computer-Human Interaction*, 30(3).

Hauke Hinrichs and Nguyen-Thinh Le. 2018. Which text-mining technique would detect most accurate user frustration in chats with conversational agents? In *Proceedings of the 32nd International BCS Human Computer Interaction Conference*, HCI '18, Swindon, GBR. BCS Learning & Development Ltd.

Mary Huang. 2024. Distilbert fine-tuned emowoz. https://huggingface.co/mh0122/distilbert-finetuned-emowoz.

Takyoung Kim, Jamin Shin, Young-Ho Kim, Sanghwan Bae, and Sungdong Kim. 2024. Revealing user familiarity bias in task-oriented dialogue via interactive evaluation. *Preprint*, arXiv:2305.13857.

JR Landis. 1977. The measurement of observer agreement for categorical data. *Biometrics*.

Chi-Hsun Li, Su-Fang Yeh, Tang-Jie Chang, Meng-Hsuan Tsai, Ken Chen, and Yung-Ju Chang. 2020. A conversation analysis of non-progress and coping strategies with a banking task-oriented chatbot. CHI '20, page 1–12, New York, NY, USA. Association for Computing Machinery.

Michael F McTear. 2017. The rise of the conversational interface: A new kid on the block? In *Future and Emerging Trends in Language Technology. Machine Learning and Big Data: Second International Workshop, FETLT 2016, Seville, Spain, November 30–December 2, 2016, Revised Selected Papers 2*, pages 38–49. Springer.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Patrícia Pereira, Helena Moniz, and Joao Carvalho. 2022. Deep emotion recognition in textual conversations: A survey.

Niklas Rach, Wolfgang Minker, and Stefan Ultes. 2017. Interaction quality estimation using long short-term memories. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 164–169, Saarbrücken, Germany. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. SIGIR '21, page 2499–2506, New York, NY, USA. Association for Computing Machinery.

Silvia Terragni, Bruna Guedes, Andre Manso, Modestas Filipavicius, Nghia Khau, and Roland Mathis. 2022. BETOLD: A task-oriented dialog dataset for breakdown detection. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 23–34, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Yan Wang, Bo Wang, Yachao Zhao, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. Emotion recognition in conversation via dynamic personality. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5711–5722, Torino, Italia. ELRA and ICCL.

Alexandra Weidemann and Nele Rußwinkel. 2021. The role of frustration in human–robot interaction – what is needed for a successful collaboration? *Frontiers in Psychology*, 12.

Xuanming Zhang, Rahul Divekar, Rutuja Ubale, and Zhou Yu. 2023a. GrounDialog: A dataset for repair and grounding in task-oriented spoken dialogues for language learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 300–314, Toronto, Canada. Association for Computational Linguistics.

Yazhou Zhang, Mengyao Wang, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2023b. Dialoguellm: Context and emotion knowledge-tuned llama models for emotion recognition in conversations. *ArXiv*, abs/2310.11374.

A In-Context Learning Prompt

Fig. 3 shows the different components of our in-context learning prompt for user frustration detection with Large Language Models (LLMs). The task definition and domain description are human-generated, and the conversation history is a variable which is sample-dependent. We also attach output instructions which inform the LLM to respond in the desired format.

B Dialog Breakdown Detection Features

Table 4 lists the set of hand-crafted features utilized in the baseline from the *Dialog Breakdown Detection* domain. The original set is given in Bodigutla et al. (2020). In our experiments, embeddings leveraged in the calculation of cosine similarity are created by the MPNet-like model (*all-mpnet-base-v2*) trained using the process described in Reimers and Gurevych (2019).

C Ethical Considerations

As our study relies on data gathered from real users, we take several steps towards ensuring users' rights, privacy, and fair use of their data, in accordance with the US law. First, prior to their conversation with our TOD system, we obtain an informed consent on the recording of the conversation and using the recording for any types of advancements of our system. Second, we ensure privacy by performing anonymization of any potentially identifying user information. Finally, we do not report any real user

---

Task Description (T)

In this task, you are given a conversation between a user and a task-oriented dialog system. Your goal is to determine if the user is frustrated during the conversation.
User frustration is often expressed through negative emotions, such as anger, irritation, or dissatisfaction. Some cues indicating frustration include:

- Profanity or abusive language directed at the system.

- Hostility or irritation toward the assistant.

- Direct expressions of frustration, such as complaints about the system's performance or the conversation itself.

Frustration can also be more subtle and does not always involve negative language. A user may become frustrated when the system is unable to handle the conversation effectively or help the user accomplish their task. This may lead to breakdowns in the dialogue, where the user either disengages or expresses a desire to stop the interaction.
To classify frustration, consider the following signs:

- Repetition of requests or questions due to the system's failure to resolve the user's issue.

- Use of negation, where the user rejects the system's suggestions or responses.

- Long, unresolved conversations where the user's task remains incomplete.

- The user's general dissatisfaction with the system's responses, even without overt hostility.

Domain (D)

The conversation you are analyzing occurs in one of these two domains:

- Receptionist system responsible for transferring calls to the appropriate department or agent. The system is expected to ask clarifying questions to help find the correct target.

- Booking agent that is negotiating the time slot for an appointment. We expect some back and forth between the user and the system to find a slot that works well for the user.

Conversation History (H)

CONVERSATION: {$chat\_history$}

Output Instructions

Return a single number:

- **0** if the user is **not frustrated**

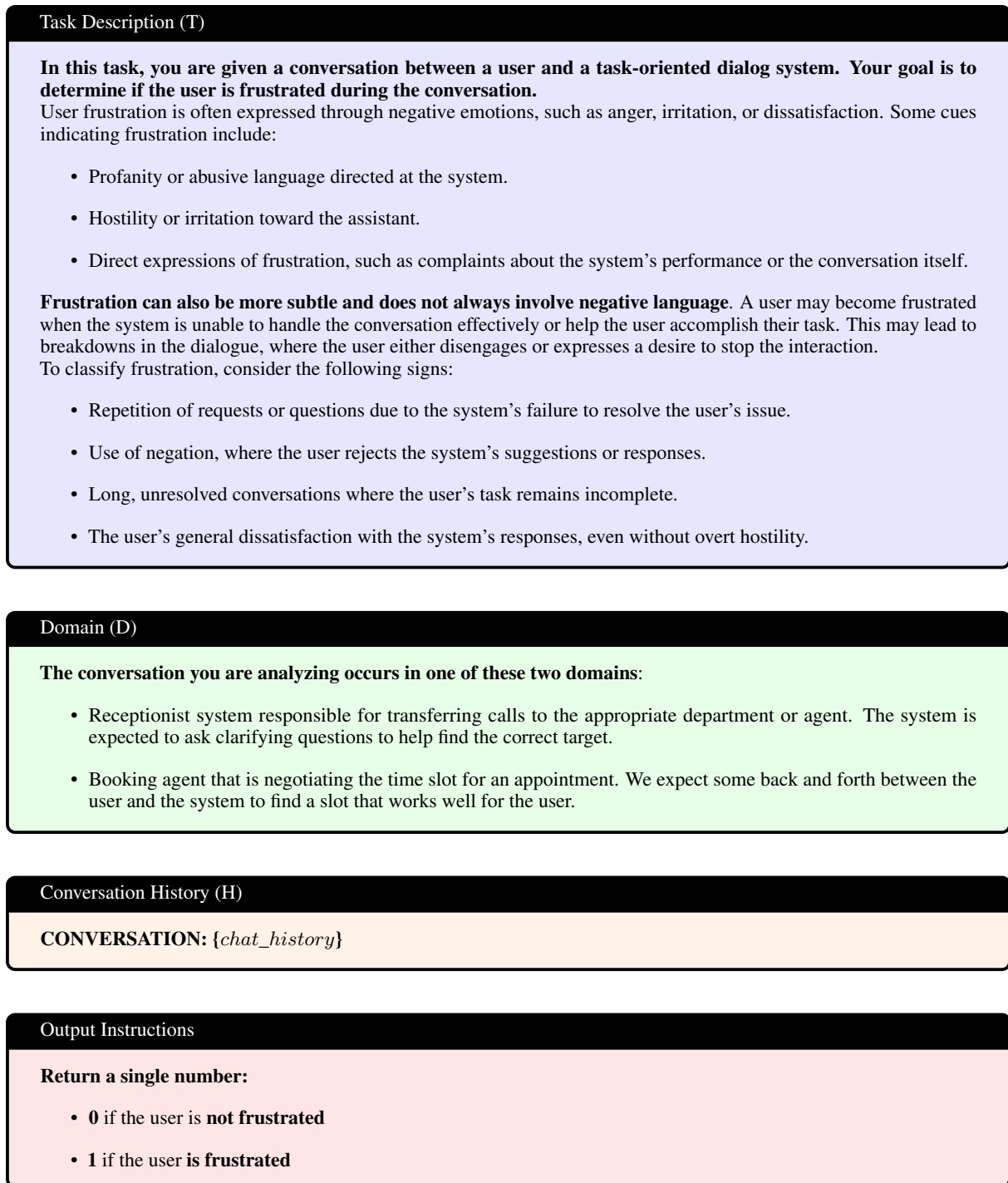- **1** if the user **is frustrated**

Figure 3: In-Context Learning Prompt for User Frustration Detection in Task-Oriented Dialog Systems. The context is comprised of the description of the task (T), the domain of the conversation (D) and the conversation history (H). Our prompt also includes output instructions to generate binary user frustration labels.

data in this paper. Reported examples are manually augmented or rephrased in a way that preserves the right context, while ensuring no user utterance exactly matches the original user utterance.

Feature name	Computation methodology
Semantic paraphrase of user's req.	MA of the cosine similarity between pairs of utterances u_{t-1} and u_t
Semantic repetition of system's resp.	MA of the cosine similarity between pairs of utterances s_{t-1} and s_t
Semantic coherence of user's req. and system's resp.	MA of the cosine similarity between utterances s_{t-1} and u_t
Syntactic paraphrase of user's req.	MA of the Jaccard index between sets of words of utterances u_{t-1} and u_t
Syntactic repetition of system's resp.	MA of the Jaccard index between sets of words of utterances s_{t-1} and s_t
Syntactic coherence of user's req. and system's resp.	MA of the Jaccard index between sets of words of utterances s_{t-1} and u_t
Length of user's requests	MA of the utterance length u_t
Length of system's response	MA of the utterance length s_t
Length of full conversation	Number of characters in observed dialog H
Number of turns in conversation	Number of pairs (u_t, s_t)

Table 4: The set of features used in *Dialog Breakdown Detection* approach. *MA* represents the moving average across the consecutive pairs in the observed dialog H.