ユーザ発話を用いたシステム応答に対する不満検出

齋藤 由佳 [1]　沼屋 征海 [1]　赤間 怜奈 [1,2,3]　鈴木 潤 [1,3,4]
[1] 東北大学　[2] 国立国語研究所　[3] 理化学研究所　[4] 国立情報学研究所 LLMC
is-failab-research@grp.tohoku.ac.jp

概要

本研究では，LLM とのコミュニケーションを通じてユーザが感じる不満が，発話テキストにどの程度表出しているかを明らかにすることを目的とする．ユーザ自身が付与した不満ラベルを正解として二値分類を行い，入力に用いる文脈を変えた複数条件において不満検出を評価する．実験の結果，F1 スコアがラベル比率に基づく無作為抽出を上回り，不満感情が一定程度検出可能な形で表出していることが示唆された．

1　はじめに

大規模言語モデル（LLM）は，ChatGPT[1)] をはじめとしたチャット形式のコミュニケーションを通じて様々な用途に活用されており [1]，モデルが生成する応答を適切に評価することに加え，ユーザの嗜好を考慮した評価の重要性が高まっている．先行研究では，システム応答の内容自体に着目した評価手法が広く用いられている [2, 3, 4]．一方で，高度な文脈理解による一貫した応答生成が可能となったことを背景に [5]，ユーザの意図や目的，嗜好などを考慮し，応答内容だけでなくユーザ発話に着目する試みが活発化している．実際に，ユーザ発話から推定した感情を用いることでシステム応答を評価する手法が提案されている [6]．不満などのユーザの負の感情は，応答が期待に沿っていないことの兆候となり得ると考えられるため，その時点で検知できれば，応答方針の修正による利用体験の向上やユーザに合わせた応答生成に寄与する可能性がある [7]．

ユーザの感情を発話から推定する既存研究では，ユーザ本人ではないアノテータが付与した正解ラベルを用いて性能評価を行うことが多い [8, 9, 10]．しかしながら，第三者がテキスト情報から知覚する感情は，実際のユーザ心理と必ずしも一致しない可能性が考えられる [11]．したがって，まずはユーザの感情がテキスト上にどの程度観測可能な形で表れているかを明らかにする必要がある．そこで本研究では，感情表出を「ユーザの感情が発話テキストを通して顕在化したもの」と定義し，ユーザの不満に関する感情表出が検出できるかを検証する．具体的には，図 1 に示すように，ユーザ発話に対して本人が付与した不満ラベルを正解とした二値分類を行う．実験の結果，一定の割合で不満が正しく検出され，ユーザの抱く不満感情が発話テキストに一部表出している可能性が確認された．一方で，発話テキストのみからは不満感情を捉えにくい事例も見られた．

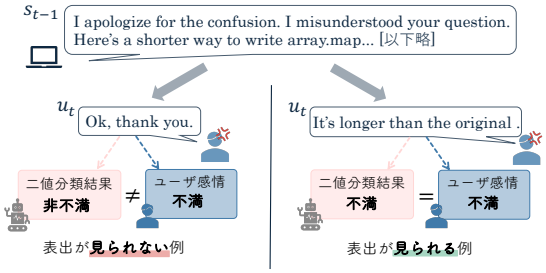

図 1　システム応答に対するユーザの不満感情を，直後のユーザ発話（該当発話に対する反応）から検出できるか？

2　関連研究

システム応答に着目した自動評価　先行研究では，LLM とのコミュニケーションにおいて，システム応答に着目することでモデル性能を自動評価する手法が広く用いられている．例えば，システム応答に対する人手評価データを用いて学習する評価器である ADEM [2] や，LLM を用いた MT-bench [3]，LLM-Eval [4] などの手法が提案されている．また，応答の良さだけでなく，システム応答に対する不満のような負の側面を注釈付けしたデータセットも提案されている [12]．本研究では，応答を評価する手がかりとしてユーザ発話に表れる不満に着目する．

ユーザ反応・フィードバックに基づく評価　ユーザ発話にはシステム応答に対する要求や感情などが含まれており [13]，暗黙的にユーザ本人の評価が内

1）https://openai.com/chatgpt

在する可能性が考えられる．このことから，ユーザ発話から推定した感情を手がかりにシステム応答を間接的に評価する手法も提案されている [6, 13]．本研究では，先行研究からユーザ発話をシステム応答に対する暗黙的評価として捉え，とくに不満感情の表出に着目してその検出可能性を検証する．

感情推定　対話品質の向上を目的として，各発話における感情を推定する研究が進められている．先行研究では，DialogueRNN [14] のようにユーザの心理状態を追跡して，対話の流れから感情を推定する手法のほか，BERT や RoBERTa などのモデルを fine-tuning する手法 [15] や LLM を用いた手法も提案されている [16, 17]．これらの研究は，実際に対話を行ったユーザ本人ではないアノテータがつけた感情を正解とするものが多い [8, 9, 10]．しかしながら，第三者が発話から知覚する感情は，ユーザの感情と乖離し得る可能性が報告されており [11]，感情推定が真にユーザの感情を捉えているとは限らない．そこで本研究では，ユーザ自身が付与した不満ラベルを用いることで，ユーザ発話に不満感情が表出しているかを検証する．さらに心理学の観点からは，対話相手が LLM のような非人間である場合，人間相手よりも社会的遠慮が働きにくく，負の感情が顕在化することが示唆されており [18]，不満の表出を捉えやすい可能性がある．

3　実験

3.1　実験設定

タスク定義　本研究では，感情表出を「ユーザの感情がテキストを通して顕在化したもの」と定義し，ユーザの不満に関する感情表出を検出できるかを検証する．具体的にはシステム応答直後のユーザ発話を対象に，本人の不満ラベルを正解として「不満」「非不満」の二値分類を行った．

データセット　ユーザの発話における不満表出を分析するため，ChatGPT Dissatisfaction Dataset [12] に含まれる英語データを用いた．このデータセットは，ユーザが ChatGPT との対話で不満に感じた応答を選択し，ラベル付けしたものである．使用データは，分析の信頼性を担保するために，不備データの除外等の前処理を行った（付録 A.1）．前処理後のデータは 107 対話，824 発話で構成され，不満発話を約 17.8%の割合で含む不均衡データであった．また，評価に用いるデータ数を担保するため，学習，検証，テストデータを 6:1:3 の割合で分割した．

表 1　入力条件の定義．$[a, b]$ は文字列の連結を表す．

条件	入力
(i)	対象ユーザ発話 u_t
(ii)	(i) + 直前システム応答 $[s_{t-1}, u_t]$
(iii)	(i) + 直前ユーザ発話 $[u_{t-1}, u_t]$

モデル　不満表出の検出には，先行研究で広く用いられている BERT 系モデルおよび GPT 系モデルを採用した．BERT 系モデルは実験に利用可能なデータが小規模であることを考慮し [19, 20, 21]，BERT-small[2], BERT-mini[3] を用いた [22]（付録 B.1）．GPT 系モデルは GPT-5.1 を使用し，zero-shot と few-shot の双方で比較を行った．GPT 系モデルには，ユーザ発話に明示的な感情語が含まれない場合も考慮した上で不満の有無を判定し，併せて判断理由を出力するよう指示した（付録 B.3）．

入力条件　比較分析のため，入力として与える情報が異なる 3 条件（表 1）で検出性能を測定する．t ターン目のユーザ発話を u_t，u_t に対するシステム応答を s_t とし，対話を発話列 $\{(u_t, s_t)\}_{t=1}^{T}$ で表す．条件 (i) は u_t のみを入力とする．条件 (ii) は，ユーザ発話が直前のシステム応答への反応として生じる点を踏まえ，直前のシステム応答 s_{t-1} と u_t を入力とする．条件 (iii) は，ユーザ発話の表現の変化を捉えることを目的とし，直前のユーザ発話 u_{t-1} と u_t を入力とする．BERT 系モデルで学習を行う際，ターンの異なる発話を区別するため，発話間に特殊トークンを挿入したテキストを入力とした（付録 A.2）．

評価方法　本実験では，不満発話の割合が少ない不均衡データを用いる．分類タスクの評価指標として一般的に用いられる正解率は多数派クラスの影響を強く受けるため，本研究では正例に対する適合率，再現率および F1 スコアを主要指標とする．またベースラインとして，データのクラス比率に従って無作為抽出を行う試行を 5,000 回繰り返した平均を算出した．なお，比較するモデル評価の最終スコアは，対象データを異なる無作為分割で 5 回実行した平均を用いた．

3.2　結果

表 2 に，異なる入力条件下での各モデルの検出性能（平均スコア ± 標準偏差）を示す．なお，条件 (i) で性能の高かった BERT-small と GPT-5.1 (few-shot 設

2）https://huggingface.co/prajjwal1/bert-small
3）https://huggingface.co/prajjwal1/bert-mini

表 2　入力条件 (i)–(iii) における不満検出性能（平均スコア ± 標準偏差）．条件 (ii)/(iii) の第 2 行は条件 (i) からの差分．

	条件 (i) 対象ユーザ発話			条件 (ii) (i) + 直前システム応答			条件 (iii) (i) + 直前ユーザ発話		
モデル	適合率	再現率	F1 スコア	適合率	再現率	F1 スコア	適合率	再現率	F1 スコア
BERT-small	23.6 ± 3.0	76.4 ± 13.5	35.7 ± 3.0	27.4 ± 4.5 (+3.8)	57.7 ± 20.9 (−18.7)	35.6 ± 4.5 (−0.1)	29.2 ± 5.5 (+5.6)	48.6 ± 16.1 (−27.8)	35.7 ± 7.1 (+0.0)
BERT-mini	23.5 ± 1.6	68.6 ± 11.4	34.7 ± 1.3	–	–	–	–	–	–
GPT-5.1 (zero-shot)	40.2 ± 5.5	35.9 ± 3.0	37.9 ± 3.8	–	–	–	–	–	–
GPT-5.1 (few-shot)	40.3 ± 4.4	36.4 ± 6.6	38.1 ± 5.4	23.1 ± 1.3 (−17.2)	54.1 ± 4.7 (+17.7)	32.3 ± 2.0 (**−5.8**)	36.2 ± 4.1 (−4.1)	43.2 ± 6.6 (+6.8)	39.3 ± 4.9 (**+1.2**)
無作為抽出	17.9 ± 5.3	17.8 ± 5.8	17.8 ± 5.4	17.9 ± 5.3	17.8 ± 5.8	17.8 ± 5.4	17.9 ± 5.3	17.8 ± 5.8	17.8 ± 5.4

表 3　不満表出に関する定性分析のユーザ発話の例．

分類	例（ユーザ発話）
TP（例 1）	*still combo doesn't clear after clicking no-mole*
FN（例 2）	*something with fire swan phrase*
FP（例 3）	*I don't want to give me a list.* *Do you think I should attend gym 2–3 a week?(*略*)*
FP（例 4）	*print all js*

表 4　条件 (ii) において非不満発話を誤検出している例．

発話種類	例（システム応答 → 対象ユーザ発話）
s_{t-1}	*To implement the search() method on the client side, you'll need to communicate with the server ... (*略*)*
u_t	*displaySearchResults(results)*

定) のみを条件 (ii), (iii) の結果として示す．全ての条件において，F1 スコアは無作為抽出（17.8%）を一貫して上回った一方で，40% 未満に留まった．また，入力条件間で F1 スコアを比較すると，BERT 系モデルではすべての条件で変化が見られなかったが，GPT 系モデルでは，条件 (i) と比べ条件 (ii) でスコアが低下し，条件 (iii) では改善した．さらに，いずれのモデルにおいても，条件 (i) と比較して適合率と再現率の差が縮小する傾向が確認された．

3.3　不満表出に関する定性分析

結果より，ユーザ発話における不満の表出は一定程度検出が可能であった一方で，表出していない，あるいは検出が困難なケースも存在した．表 3 に，全てのモデルで表出を検出できた事例および表出が乏しく検出できなかった事例を示す．

不満を検出できた例（TP）　例 1 の発話ではユーザの不満を検出することができた．この発話例に含まれる *still* や *doesn't* のように，継続的な不具合を示す表現や修正を要求する否定的な表現などは不満の表出として捉えやすいと考えられる．

不満を見逃した例（FN）　例 2 の発話ではユーザが不満を感じていた一方で検出ができなかった．このように指示が曖昧であるなど，情報量が不足している発話では不満の表出が検出されない場合があることが確認された．

非不満を不満と誤検出した例（FP）　例 3，4 の発話ではユーザが不満を感じていなかった一方でモデルが不満であると誤検出した．例 3 のような否定形を含む発話や例 4 のような命令形の依頼ではモデルが否定的な表現であると過剰反応し，ユーザが不満を持っているとして誤検出したと考えられる．

3.4　入力の違いがスコアに与える影響

第 3.2 節に示した GPT 系モデルの結果では，条件 (ii) で F1 スコアが低下し，条件 (iii) で改善がみられた．ここでは，この原因を実例から考察する．

直前のシステム応答を考慮した場合の影響　条件 (ii) では，対象のユーザ発話 u_t に加えて直前のシステム応答 s_{t-1} を入力して不満の検出を行った．ユーザ発話が生じた状況を考慮することで，対象のユーザ発話の理解に役立つことが期待された一方で，適合率および F1 スコアが低下した．ここで，表 4 に条件 (ii) のみが不満であると誤検出した例を示す．例において，ユーザ発話と直前のシステム応答との文脈的一貫性が乏しく，話題を転換したことが不満として解釈され，適合率および F1 スコアの低下につながった可能性がある．よって，LLM との対話では，ユーザが予告せずに別の要求へ移るなど，自然な文脈の遷移による話題転換が生じやすいことを

表 5　条件 (iii) において不満発話を検出できている例.

発話種類	例（直前ユーザ発話 → 対象ユーザ発話）
u_{t-1}	*What are your primary concerns regarding sending remittances for healthcare? This question is required. Be as detailed as you can!*
u_t	*summarise*

表 6　実験に使用した多言語データの統計情報.

言語	対話数	総発話数	不満発話数 (%)
英語	107	824	147 (17.8)
ポーランド語	20	152	20 (13.2)
スペイン語	6	29	6 (20.7)
日本語	13	74	26 (35.1)

考慮する必要性が示された.

直前ユーザ発話を考慮した場合の影響　条件 (iii) では，対象のユーザ発話 u_t に加えて直前のユーザ発話 u_{t-1} を入力することで不満の検出を行った．比較対象として u_{t-1} を用いることでユーザの発話表現の変化を捉えたことが F1 スコアの改善に寄与した可能性が考えられる．ここで，表 5 に条件 (iii) のみが不満を検出できた例を示す．この例では，長い説明文による指示を行っていたユーザ発話 u_{t-1} から短い指示の発話 u_t へと表現の形式が変化しており，この変化が不満表出として現れたことで正しく検出できた可能性がある．したがって，直前のユーザ発話と対象のユーザ発話との比較によって不満が検出しやすくなる可能性が示された.

4　多言語への拡張

ユーザの不満感情の表出は，言語的・文化的要因の影響を受ける可能性が考えられるため，前節で示した結果が言語横断的に成立するかを検証する.

4.1　実験設定

ChatGPT Dissatisfaction Dataset には多言語の対話が含まれるが，本研究では英語に次いで対話数が多いポーランド語およびスペイン語を使用した．さらに，本データセットの収集方法に従い，Yahoo!クラウドソーシングを用いて新たに日本語の対話データを収集し，ユーザによる不満のラベル付けを行った．各言語についても英語と同一の前処理を適用して二値分類を行い，F1 スコアを指標としてユーザ発話への不満感情の表出を検証した．表 6 に各言語の統計情報を示す．拡張した言語は英語に比べてデータ数が少なく，fine-tuning による学習が困難であるため，GPT 系モデル（few-shot 設定）を用いて多言語データにおける評価を行った.

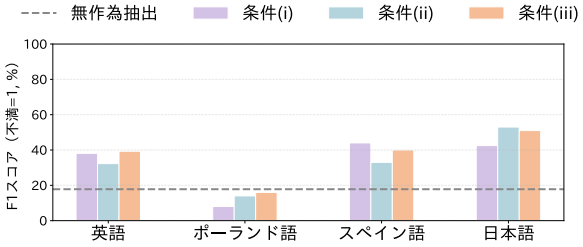

図 2　多言語における条件ごとの不満検出の F1 スコア.

4.2　結果

図 2 に，各言語における条件 (i)–(iii) の F1 スコアを示す．英語と同様に，スペイン語および日本語データにおいて無作為抽出を超える 40% ほどの F1 スコアが確認された．一方で，ポーランド語データのスコアは他の言語と比較して低い値を示し，無作為抽出を下回った.

4.3　分析

実験では，ポーランド語のデータにおいて低い F1 スコアが得られ，ユーザの不満を検出することが困難であった．この要因として，ポーランド語では不満感情を他言語と異なる形で表出した，または明示的な手がかりが少なかった可能性が考えられる．さらに，ポーランド語以外の言語データにおけるスコアは無作為抽出を一貫して超える傾向が見られたが，今回は言語によってデータ数が異なるため，更なる検討が必要とされる.

5　おわりに

本研究では，LLM とのコミュニケーションで生じるユーザの不満がテキストにどの程度表出しているかを，ユーザが付与した不満ラベルを正解とした二値分類から検証した．結果として，多くの条件で無作為抽出を上回り，システム応答に対する不満感情をユーザ発話から一定程度検出できることが示された．したがって，ユーザ発話における不満表出を捉えることで，応答評価や応答の修正に一部活用できる可能性が示唆される．一方で，不満を誤検出する事例も確認された．本設定のようなシステムとの対話では，話題転換が生じやすいことなどの特徴が見られ，人同士の対話とは異なる傾向を考慮することが今後の課題である.

謝辞

本研究を進めるにあたってご指導ならびにご助言を賜りました Tohoku NLP グループの皆様に感謝申し上げます．また，データ収集にご協力いただいた皆様にも感謝申し上げます．なお，本研究の一部は，JSPS 科研費 JP25K21263, JST BOOST JPMJBY24A1, JST ムーンショット型研究開発事業 JPMJMS2011-35 (fundamental research) および，文部科学省の補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の助成を受けたものです．

参考文献

[1] Aaron Chatterji, Tom Cunningham, David Deming, Zoë Hitzig, Christopher Ong, Carl Shan, and Kevin Wadman. How people use chatgpt. Technical report, OpenAI, 2025.

[2] Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. Towards an automatic Turing test: Learning to evaluate dialogue responses. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, 2017.

[3] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, 2023.

[4] Yen-Ting Lin and Yun-Nung Chen. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In **Proceedings of the 5th Workshop on NLP for Conversational AI**, 2023.

[5] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, 2020.

[6] Sarik Ghazarian, Behnam Hedayatnia, Alexandros Papangelis, Yang Liu, and Dilek Hakkani-Tur. What is wrong with you?: Leveraging user sentiment for automatic dialog evaluation. In **Findings of the Association for Computational Linguistics**, 2022.

[7] Ying-Chun Lin, Jennifer Neville, Jack Stokes, Longqi Yang, Tara Safavi, Mengting Wan, Scott Counts, Siddharth Suri, Reid Andersen, Xiaofeng Xu, Deepak Gupta, Sujay Kumar Jauhar, Xia Song, Georg Buscher, Saurabh Tiwary, Brent Hecht, and Jaime Teevan. Interpretable user satisfaction estimation for conversational systems with large language models. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics**, 2024.

[8] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. **Language Resources and Evaluation**, Vol. 42, No. 4, pp. 335–359, 2008.

[9] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, 2019.

[10] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In **Proceedings of the Eighth International Joint Conference on Natural Language Processing**, 2017.

[11] Kazunori Komatani, Ryu Takeda, and Shogo Okada. Analyzing differences in subjective annotations by participants and third-party annotators in multimodal dialogue corpus. In **Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, 2023.

[12] Yoonsu Kim, Jueon Lee, Seoyoung Kim, Jaehyuk Park, and Juho Kim. Understanding users' dissatisfaction with chatgpt responses: Types, resolving tactics, and the effect of knowledge level. In **Proceedings of the 29th International Conference on Intelligent User Interfaces**, 2024.

[13] Yuhan Liu, Michael JQ Zhang, and Eunsol Choi. User feedback in human-LLM dialogues: A lens to understand users but noisy as a learning signal. In **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, 2025.

[14] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. In **Proceedings of the AAAI Conference on Artificial Intelligence**, 2019.

[15] Xiangyu Qin, Zhiyu Wu, Tingting Zhang, Yanran Li, Jian Luan, Bin Wang, Li Wang, and Jinshi Cui. BERT-ERC: Fine-tuning bert is enough for emotion recognition in conversation. In **Proceedings of the AAAI Conference on Artificial Intelligence**, 2023.

[16] Shutong Feng, Guangzhi Sun, Nurul Lubis, Wen Wu, Chao Zhang, and Milica Gasic. Affect recognition in conversations using large language models. In **Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, 2024.

[17] Mireia Hernandez Caralt, Ivan Sekulic, Filip Carevic, Nghia Khau, Diana Nicoleta Popa, Bruna Guedes, Victor Guimaraes, Zeyu Yang, Andre Manso, Meghana Reddy, Paolo Rosso, and Roland Mathis. "stupid robot, I want to speak to a human!" user frustration detection in task-oriented dialog systems. In **Proceedings of the 31st International Conference on Computational Linguistics: Industry Track**, 2025.

[18] Annabell Ho, Jeff Hancock, and Adam S. Miner. Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. **Journal of Communication**, Vol. 68, No. 4, pp. 712–733, 2018.

[19] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models, 2019. https://arxiv.org/abs/1908.08962.

[20] Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets llm finetuning: The effect of data, model and finetuning method. In **The Twelfth International Conference on Learning Representations**, 2024.

[21] Daniel Grie
sshaber, Johannes Maucher, and Ngoc Thang Vu. Fine-tuning BERT for low-resource natural language understanding via active learning. In **Proceedings of the 28th International Conference on Computational Linguistics**, 2020.

[22] Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. Generalization in nli: Ways (not) to go beyond simple heuristics, 2021. https://arxiv.org/abs/2110.01518.

[23] Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**, 2020.

[24] Shun Katada, Kiyoaki Shirai, and Shogo Okada. Incorporation of contextual information into bert for dialog act classification in japanese. In **2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing**, 2021.

[25] Bingkun Chen, Shaobing Dai, Shenghua Zheng, Lei Liao, and Yang Li. Dsbert:unsupervised dialogue structure learning with bert, 2021. https://arxiv.org/abs/2111.04933.

A　データ詳細

A.1　前処理

本研究で使用するデータに対し，分析の信頼性を担保するために以下の手順で前処理を行った：

1. **不備データの除外．**発話全体がマスキングされた対話や，データセット中の対話ログに欠損があるデータを除外した．
2. **タスク非適合データの除外．**「システム応答に対する不満」を対象とするため，不満ラベルが付与されたシステム応答の直後にユーザ発話が存在せず，対話が終了しているケースを除外した．この対話は英語には 157 対話中 50 対話，ポーランド語は 27 対話中 7 対話，スペイン語は 11 対話中 5 対話存在した．
3. **文脈情報の確保．**入力条件によっては対象のユーザ発話以前の発話を必要とするため，参照すべき文脈が存在しない 1 ターン目の発話を除外した．

A.2　使用データの実例

BERT 系モデルに使用するデータは，入力テキストを最大長 `max_length` でトークナイズした．また，末尾にある対象のユーザ発話の情報を保持するため，切り詰め方向を左（`truncation_side=left`）に設定した．なお，本実験では異なる 3 つの入力条件で二値分類を行った．条件 (ii) では話者境界を明確にするために直前システム応答の前に `[ASSISTANT]`, 対象ユーザ発話の前に `[USER]` を，条件 (iii) では発話間関係を特殊トークンで捉えるために直前ユーザ発話の前に `[PREV]`, 対象ユーザ発話の前に `[CUR]` を用いた [23, 24, 25]．実例を表 7 に示す．

表 7　入力条件 (i)–(iii) の実例．

条件 (i)	It's longer than the original.
条件 (ii)	`[ASSISTANT]` I apologize for the confusion. . . .
	`[USER]` It's longer than the original.
条件 (iii)	`[PREV]` That is even longer in terms of characters.
	`[CUR]` It's longer than the original.

B　モデル詳細

B.1　使用モデルの選択

スケーリング則において，性能はモデルサイズとデータサイズの双方に依存し，データ制約下は大きなモデルが必ずしも最適とはならないと報告されている [20]．また，コンパクトな BERT 系モデルも事前学習を施した上で fine-tuning すると，サイズの大きなモデルと比べて遜色ない性能を達成し得ることが示されている [19]．さらに，データ数が 1000 例未満では fine-tuning が難しい状況が想定される [21]．本研究では本文中の表 6 に示す通りデータ数が小規模であるため，パラメータ数が異なる複数のモデルにおいて，今回のデータを用いて学習が適切に行われているかを調査した．ここでは，比較的大規模な RoBERTa-base と，より軽量な BERT-small で，検証データにおける予測確率の振る舞いを図 3 および図 4 に示す．実際に，図 3 に示すように，RoBERTa-base で二値分類を行った場合，検証データでの予測確率はほぼ全てのデータで 0 付近に集中しており，学習ができていないことが確認できた．一方で，図 4 に示す通り，BERT-small での結果は予測確率が分散している様子が確認できる．したがって本研究では，データ数を考慮して，BERT-small および BERT-mini を用いて不満表出検出を行った．

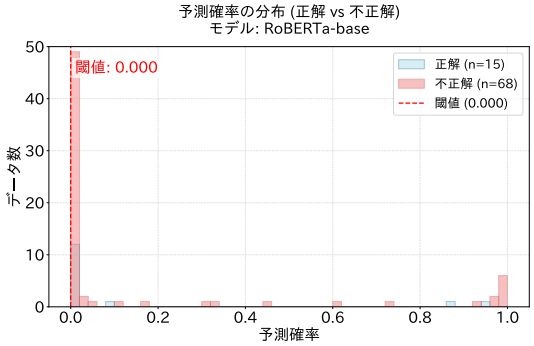

図 3　RoBERTa-base で二値分類した際の予測確率の分布．

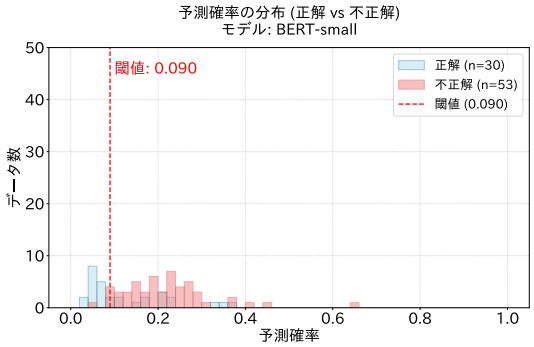

図 4　BERT-small で二値分類した際の予測確率の分布．

B.2　BERT 系モデルにおける閾値選択

学習後に不均衡データへの対策として，検証データに対して各例の正例確率を算出し，閾値 $t \in [0, 1]$ を走査することで F1 スコアが最大となる閾値 t を選択した．

B.3　GPT 系モデルの使用プロンプト

図 5 に，本実験で使用したプロンプトのうち，zero-shot の条件 1 を示す．条件の拡張は入力として与える説明のみを変更し，few-shot への拡張は与える入力例を追加した．プロンプトは，不満が言葉に明示的に表れていないと判断されても，ユーザが不満を感じていると推測できる場合には 1 を選択するように示している．

```
[System]
You are an assistant that infers
whether a user is feeling dissatisfied from dialogue logs.
Given a single user utterance, determine
whether the user is continuing the conversation while feeling dissatisfied.
Even if the utterance does not contain explicit keywords
indicating dissatisfaction, label it as dissatisfied
if the user might be feeling dissatisfied at that moment.
Output must be in JSON: set "label" to 1 for dissatisfaction and
0 for no dissatisfaction, and write a brief "reasoning" in Japanese.

[User]
Classify the following utterance.
---
{text}
---
Return only JSON.
```

図 5　GPT-5.1（zero-shot）で用いたプロンプト（条件 (i)：対象ユーザ発話のみ）．