Advances in Technology Innovation, vol. 10, no. 4, 2025, pp. 383-394

Analysis of Characteristics of Complaints on Parenting Q&A Sites Using pLSA and Data Augmentation

Tomoki Yoshimi[1,*], Takashi Namatame[2]

[1]Graduate School of Science and Engineering, Chuo University, Tokyo, Japan

[2]Faculty of Science and Engineering, Chuo University, Tokyo, Japan

Received 30 December 2024; received in revised form 08 August 2025; accepted 11 August 2025

DOI: https://doi.org/10.46604/aiti.2025.14700

Abstract

This study investigates the classification and clustering of complaints on a Japanese parenting Q&A site, aiming to identify meaningful patterns from limited labeled data. To address data scarcity, generative AI was utilized for data augmentation through prompts that reflected authentic parenting frustrations, with synthetic data validated by comparing classification performance under varying proportions of generated content. Complaint texts were vectorized using Bag-of-Words, Doc2Vec, and Sparse Composite Document Vectors, providing multiple levels of semantic representation. LightGBM was used as the classifier, and F1 scores measured performance. Clustering of predicted complaints employed probabilistic Latent Semantic Analysis, with topic numbers selected via Bayesian Information Criterion. Six distinct themes emerged, including childcare stress and family conflict. Incorporating generated data improved the F1 score from 0.824 to 0.865. The findings highlight the potential of generative AI to augment low-resource datasets and demonstrate the effectiveness of context-aware embeddings and probabilistic clustering in structuring real-world text data.

Keywords: Q&A site, complaint, NLP, data augmentation

1. Introduction

In recent years, social networking services (SNS) and Q&A sites have become platforms where users frequently post their daily frustrations. On SNS, venting complaints is often seen as a stress-relieving activity, providing users with an emotional outlet [1]. On Q&A sites, however, the nature of complaints differs due to their primary function as platforms for solving problems and sharing practical knowledge. Despite this, complaints are often posted on Q&A sites, which diverges from their intended purpose. This phenomenon can sometimes reduce the effectiveness of these platforms as reliable sources of information. In Japan, many complaints on Q&A sites revolve around family and child-related issues, such as child-rearing and housework. These topics represent common sources of frustration, particularly for mothers. Previous studies have also noted that family-related issues are a recurring theme in complaints shared on online platforms [2].

Research on complaints has been limited, particularly in the context of the Japanese language. Regarding Japanese complaints, few studies have utilized data-driven approaches. For instance, aside from Ito et al. [3], collected complaints from the former version of Twitter and classified them into 11 categories based on their characteristics. Moreover, Ito et al. [4] developed models that incorporate the subject of the complaint to improve classification accuracy. Most studies have focused on linguistic or psychological perspectives rather than leveraging data for analysis.

In contrast, substantial research has been conducted on complaints in other languages. For example, Preotiuc-Pietro et al. [5] developed a model to automatically classify the existence and types of complaints. Jin and Aletras [6] created a model that

* Corresponding author. E-mail address: a19.pb66@g.chuo-u.ac.jp

categorized complaints into four types: “no explicit reproach,” “disapproval,” “accusation,” and “blame.” Fang et al. [7] proposed using Best-Worst Scaling to represent the intensity of complaints, pioneering quantitative evaluation in this area. Tian et al. [8] developed an automated system for processing customer complaints in the water utility sector by leveraging word vector-based similarity, sentiment analysis, and intent recognition using NLP tools such as SpaCy and Rasa.

More recently, Alarifi et al. [9] employed machine learning approaches, including logistic regression and support vector machines, to effectively predict and analyze customer complaint data, demonstrating their utility in understanding consumer grievances. Additionally, Jin et al. [10] introduced a method for enhancing the representation of complaints through word embeddings and utilized a deep learning approach to improve classification accuracy. Complaints have also been utilized in product development to identify customer needs and improve product offerings. In this context, recent work by Wang et al. [11] explored how customer complaints can be used as a valuable source of feedback for product innovation and development. However, these studies primarily focus on the severity or subject of complaints, with less attention given to complaints specifically related to child-rearing.

Based on these situations, this study aims to classify complaints related to child-rearing using data derived from a Japanese Q&A site. To address the challenges associated with limited labeled training data, this study explores the use of generative AI, specifically ChatGPT by OpenAI, as a tool for augmenting datasets. By generating additional data aligned with specific criteria, the study evaluates the impact of AI-generated data on classification performance and compares it to traditional data augmentation methods. Through this approach, the study seeks to expand the available dataset, improve classification accuracy, and provide new insights into the potential of generative AI in natural language processing tasks.

2. Data Set

This section describes the datasets used for analyzing complaints on parenting Q&A sites, including their sources, preprocessing steps, and key characteristics. The datasets were selected to provide text data specifically related to parenting, containing a substantial number of posts expressing dissatisfaction or complaints, making them suitable for the objectives of this study.

(1) Data source: This study analyzes data provided by a Japanese Q&A website focused on parenting and household matters. The data spans two years, covering 2022 and 2023. All users have been anonymized. This website is primarily used by mothers and serves as both a Q&A platform and a resource offering various columns and articles related to parenting.

(2) Dataset Description: The dataset consists of a total of 5,451,588 entries and is composed of question data and corresponding answer data. In this study, only question data was used for analysis. The dataset includes the following columns:

- Question ID: Each post is uniquely identified and linked to its corresponding answer.
- User ID: A unique value is assigned to each user.
- Category ID: Each post can be assigned to a category selected by the poster from among 15 available categories.
- Content: The actual content of the post written by the user is included.
- Created date: The date and time when the post was submitted are recorded.

(3) Preprocessing: The raw text data underwent several preprocessing steps in preparation for analysis:

- Segmentation using morphological analysis: Since Japanese does not separate words with spaces, it is necessary to divide the text into individual words through morphological analysis.
- Removal of unnecessary numbers and symbols: Numbers and symbols do not carry meaningful information on their own and need to be removed for analysis.

(4) Summary of categories: Table 1 summarizes the 15 content categories available on the parenting Q&A site, along with their respective proportions. The largest category, "Parenting & Goods" (34.3%), includes questions and complaints related to child-rearing techniques, recommended items, and everyday parenting experiences. "Pregnancy & Childbirth" (14.1%) is another major category, reflecting frequent discussions about physical and emotional challenges during pregnancy, preparation for childbirth, and postpartum recovery. The "Chat & Tweets" category (9.7%) includes casual conversations and emotionally charged messages that do not necessarily fit into practical question-and-answer formats, and this category often contains venting-type complaints.

Table 1 Category of contents

Category	(%)	Category	(%)
Pregnancy & Childbirth	14.1	Work	2.5
Parenting & Goods	34.3	Fashion & Cosmetics	2.4
Supplements & Health	1.9	Family & Husband	5.7
Mental Health & Worries	6.0	Outings	2.6
Trying to Conceive	3.4	Gynecology & Pediatrics	2.0
Household Chores & Cooking	3.8	Housing	1.4
Money & Insurance	3.4	Others	6.7
Chat & Tweets	9.7		

3. Analysis method

The methodology used in this study is introduced below. The purpose of this section is to outline how the dataset was prepared, how additional data were generated, and how the resulting text was analyzed to identify complaint patterns. By presenting each step in detail, the process highlights both the methodological rigor and the rationale for combining manual labeling, generative augmentation, classification, and clustering in a single workflow. The steps of the study are as follows:

(1) From the posted comments, 2,000 entries were randomly sampled and manually labeled to determine whether they were complaints or not.
(2) The posted comments were used to train ChatGPT to generate similar complaints. However, rather than directly training on the data provided, the subjects were given constructed prompts about the topic of dissatisfaction. The prompts used for this process will be described later.
(3) The content of the posts was vectorized using various methods. In this study, three methods were employed: Bag-of-Words (BoW), Doc2Vec, and Sparse Composite Document Vectors (SCDV).
(4) Classification was performed using Light Gradient Boosting Machine (LightGBM) while varying the proportion of generated complaints.
(5) The classifier with the best performance was used to extract a large number of complaint data from the original dataset. Clustering was then performed to analyze the types of complaints being posted. The clustering method used was probabilistic Latent Semantic Analysis (pLSA).

Fig. 1 illustrates the overall analysis workflow adopted in this study, which consists of five major steps. First, 2,000 posts were randomly sampled from the parenting Q&A dataset and manually labeled to identify whether they were complaints or not. Second, generative AI (ChatGPT) was used to augment the dataset by generating synthetic complaints based on carefully designed prompts. Third, the text data—both real and synthetic—were vectorized using three different methods: BoW, Doc2Vec, and SCDV, each capturing different levels of semantic and contextual information. Fourth, LightGBM was applied to classify the posts, and the performance was evaluated using the F1 score to determine the optimal proportion of generated data. Finally, complaint posts identified by the best-performing model were clustered using pLSA, which uncovered latent complaint themes and provided insights into the underlying structure of the data.

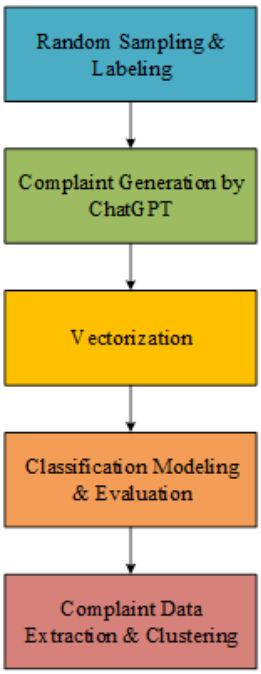

Fig. 1 Analysis procedure

3.1. Data labeling

The original dataset did not contain labels indicating whether each post was a complaint. Therefore, the first step was to manually label the data based on whether the content constituted a complaint. For this purpose, 2,000 posts were randomly sampled from the full dataset. Each post was carefully read and evaluated by two researchers using predefined criteria: a post was labeled as a complaint if it expressed dissatisfaction, frustration, or a change request, often accompanied by negative emotions or references to problems encountered. In cases where the judgment was ambiguous, the two researchers discussed the content to reach an agreement. This labeled subset was later used as the ground truth for training and evaluating the complaint classification models in the machine learning experiments described in subsequent sections.

3.2. Data augmentation

Due to the limited amount of complaint data, data augmentation was performed using ChatGPT. In this study, the following prompt was used for the analysis:

I am a researcher on the complaint. I am currently researching what mothers think about complaining. So, please write 10 complaints about child-rearing and housework from a mother's point of view, about 40 words each. It is OK if the contents are the same. “My husband has not done any housework at all since he left. I’m tired too, but... can’t you just give me a break?” , “I am constantly frustrated with raising my children. I’m always crying... even though I want to cry too...”, “I am frustrated with my mom's friends at daycare. I always give them gifts and such, but I never receive anything in return. Do they have no common sense?”

As a result of data augmentation using the above prompt, a variety of synthetic complaints were generated. For example, one post described the difficulty in satisfying every family member's food preferences, stating that preparing meals was exhausting. Another synthetic entry mentioned a child’s reluctance to begin homework, causing frequent frustration for the parent. A third example highlighted the burden placed on mothers due to the unequal division of household chores, despite both partners being employed. These generated complaints resemble real-world parenting concerns and were used to augment the training dataset.

3.3. Vectorizing method

To process the textual data for classification, the posts were converted into numerical representations using the following vectorization methods:

(1) Bag-of-Words

The BoW method is one of the most useful techniques for representing text data numerically. This approach treats a piece of text as an unordered collection of words, disregarding grammar, syntax, and word order while retaining the frequency of each word's occurrence within the text. Each unique word in the dataset forms a feature in the representation, resulting in a high-dimensional, sparse vector for each document. Despite its simplicity, BoW is effective in capturing the presence or absence of keywords and is widely used in basic text classification and retrieval tasks [12-13]. However, it cannot capture semantic relationships between words or contextual meaning, which can limit its effectiveness in complex natural language processing tasks.

(2) Doc2Vec

Doc2Vec [14] is an advanced text embedding technique that builds on the principles of Word2Vec, extending the algorithm to represent entire documents rather than individual words. By capturing the context in which words appear and their relative positions within the text, Doc2Vec creates dense, low-dimensional vector representations that encode semantic information. This method is particularly well-suited for tasks where understanding the overall meaning or sentiment of a document is important, such as classification, clustering, or recommendation systems. Unlike BoW, Doc2Vec considers terms' sequence and co-occurrence patterns of words, allowing it to preserve the structure and meaning of the text more effectively. Its ability to generate embeddings that reflect both word-level and document-level semantics makes it a powerful tool for text representation in various machine-learning applications [15]. Fig.2 visually represents Doc2Vec.

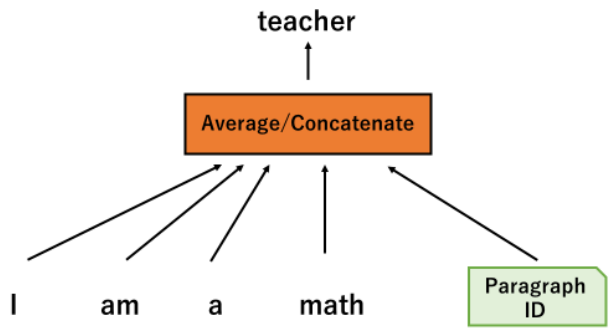

Fig. 2 Image of Doc2vec

(3) Sparse Composite Document Vectors

SCDV [16] combines the strengths of Word2Vec [17] embeddings and soft clustering techniques to create a hybrid approach for text representation. SCDV begins by generating word vectors using Word2Vec and then applies a soft clustering algorithm, such as Gaussian Mixture Models, to group these vectors into clusters. Each word is assigned a probability of belonging to multiple clusters, capturing its polysemous (multi-sense) nature. These clustered word vectors are then aggregated to form sparse and dense document embeddings, allowing for a more nuanced and context-aware representation of text. SCDV effectively balances between capturing the global structure of a document and the localized meaning of individual words, making it a robust choice for tasks that require both semantic depth and interpretability. Its flexibility and ability to handle multi-sense words provide significant advantages over traditional vectorization methods like BoW or even Doc2Vec in certain contexts [18].

3.4. Classification method

In this study, LightGBM [19] was employed as the classification algorithm to differentiate between complaints and non-complaints in the parenting Q&A data. LightGBM is a state-of-the-art supervised learning technique based on gradient boosting, a method that builds an ensemble of decision trees sequentially to minimize prediction errors. LightGBM is known for its high computational efficiency and scalability, capable of handling large datasets and high-dimensional features with ease. Its design optimizes both speed and memory usage, making it suitable for handling sparse data, such as that generated by BoW or SCDV. By leveraging techniques such as histogram-based learning and leaf-wise tree growth, LightGBM achieves superior performance in terms of classification accuracy and model interpretability while maintaining fast training and prediction times. This makes it a highly effective tool for real-world machine-learning tasks where both precision and computational efficiency are critical.

In this study, under-sampling was applied to address the class imbalance in the dataset. Since complaints accounted for only about 9% of the total data, directly training a classifier could lead to bias toward the majority class (non-complaints). Under-sampling reduces the number of non-complaint samples to create a more balanced dataset, ensuring that the classifier pays adequate attention to the minority class. By doing so, the training process became more focused on identifying complaint-related patterns, ultimately improving the model's ability to distinguish between the two classes effectively.

3.5. pLSA

pLSA [20] is a statistical method for clustering and topic modeling, designed to uncover latent structures in co-occurrence data, such as documents and their associated terms. The method assumes the existence of a shared latent variable z, which represents a probabilistic class that links two observed dimensions: x (e.g., documents) and y (e.g., terms or vector features). In pLSA, the joint probability P(x,y)P(x, y)P(x,y) is modeled as a mixture over the latent classes, where each observation is generated by first sampling a latent topic and then drawing a term conditioned on that topic. This approach enables the simultaneous clustering of both documents and terms, capturing the underlying semantic structure of the corpus. The probabilistic formulation of the model is given as follows:

$$P(x, y) = \sum_z P(x \mid z)P(y \mid z)P(z) \tag{1}$$

pLSA is particularly well-suited for sparse matrices, making it an excellent choice for feature analysis on BoW data. By modeling the co-occurrence relationships between rows and columns, pLSA effectively identifies latent topics or patterns within the data. This capability makes it a powerful tool for understanding the underlying structure of complaint-related content in the dataset [21].

4. Result

This section describes the results of the data augmentation and the analysis of the complaints extracted based on the data augmentation. Specifically, it presents the differences in complaint classification outcomes resulting from data augmentation, as well as the selection of the number of classes using pLSA and the types of complaint classes that emerged from this analysis.

4.1. Data Augmentation

In this study, the effect of data augmentation using generative AI was examined by incrementally adding 10% of generated data to the original dataset and performing classification to determine whether a post was a complaint. The performance of the classification was evaluated using the F1 score, which considers the balance between Precision and Recall.

Table 2 shows the results of F1 scores, which are the harmonic mean of precision and recall, for each vectorization method when varying the amount of generated complaint data. These results provide insights into the impact of different vectorization

approaches on the classification performance under augmented data conditions. From this table, it appears that accuracy improves as more generated data is added. Among the vectorization methods, Doc2Vec achieved the highest F1 score, indicating its superior performance in this context.

Table 2 F1 scores using all data

	BoW	Doc2Vec	SCDV
0%	0.824	0.824	0.775
10%	0.711	0.836	0.801
20%	0.797	0.865	0.822
30%	0.813	0.865	0.783
40%	0.857	0.867	0.789
50%	0.859	0.869	0.830

On the other hand, Table 3 presents the F1 scores calculated using only the original dataset, without any generated data. These results suggest that while adding generated data initially improves the performance metrics, the scores start to decline beyond a certain point. It is observed that BoW has the lowest performance metrics compared to Doc2Vec and SCDV, indicating that it is more heavily influenced by the generated complaints. The best performance was observed when 20% of generated complaints were added using Doc2Vec. Based on this model, pLSA will be applied to analyze the latent classes within the complaints.

Table 3 F1 scores without generated data

	BoW	Doc2Vec	SCDV
0%	0.824	0.824	0.775
10%	0.673	0.828	0.780
20%	0.757	0.840	0.779
30%	0.75	0.833	0.733
40%	0.777	0.809	0.717
50%	0.787	0.824	0.791

4.2. pLSA

Using the previously mentioned model, 10,000 randomly selected unlabeled data entries were classified. As a result, 2,642 entries were identified as complaints. To analyze this data, the number of latent classes was determined by running pLSA with class numbers ranging from 2 to 30 while calculating the Bayesian Information Criterion (BIC) values [22]. The results are shown in Fig. 3. While the BIC values decreased continuously as the number of classes increased, the class count was set to 6 based on interpretability and the rate of decline in BIC values. BIC is given by the following equation:

$$BIC = -2\ln L + k\ln N \tag{2}$$

$$\ln L = \sum_i \sum_j N(i, j) \ln p(x_i, y_j) = \sum_i \sum_j \left\{ N(i, j) \ln \left\{ \sum_k p(z_k)\, p(y_j \mid z_k)\, p(x_i \mid z_k) \right\} \right\} \tag{3}$$

Where L is the likelihood of the model, k is the number of free parameters, and n is the number of observations. The likelihood L measures how well the model explains the observed data, while k represents the degree of model flexibility, and n indicates the data size used for estimation. The BIC introduces a penalty term for model complexity, which increases with the number of parameters, to prevent overfitting. Lower BIC values, therefore, indicate a better balance between model fit and parsimony, meaning that the model both explains the data well and avoids unnecessary complexity.

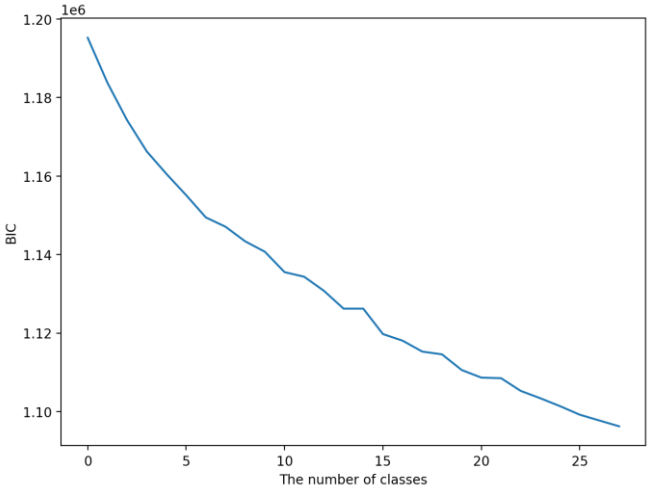

Fig. 3 BIC curve

As a result of performing pLSA with 6 classes, the sizes of the latent classes are summarized in Table 4, and the representative words for each class are shown in Table 5. From Tables 4 and 5, it is evident that a significant portion of the complaints pertain to issues involving husbands and family members. However, it is also noteworthy that terms unrelated to family, such as "hospital" and "New Year," appear as representative words in the clustering analysis. These terms highlight additional areas of concern, such as difficulties in securing medical appointments or the stress associated with traditional holiday gatherings. A more detailed discussion of these findings will be provided later.

Table 4 Size of the latent classes

Class	Size
Class1	0.1688
Class2	0.1740
Class3	0.1908
Class4	0.1308
Class5	0.1218
Class6	0.2136

Table 5 Representative words

Class1	Class2	Class3	Class4	Class5	Class6
Time	Parents' home	Children	Schedule	Today	Husband
Work	Aunt	Myself	Childbirth	Tomorrow	Son
Pregnancies	Family	Together	Friends	Yesterday	Unreasonable
Everyday	New year	Divorce	Contact	The New Year's holiday	Usual
Lives	Parents	Quarrel	Unstable	Physical condition	Bath
Raising children	This year	Household	Normal	Hospital	Room
Nursery school	Marriage	Opponent	Worry	Physical condition	Prepare

5. Discussion

The results of data augmentation and pLSA are discussed, respectively. The discussion first examines the effects of data augmentation using generative AI, highlighting both its benefits in improving classification performance and the challenges it presents, along with an evaluation of which vectorization methods are most suitable based on their representational characteristics. It then considers the complaint classes identified through pLSA, describing how complaints on parenting Q&A sites were segmented into distinct thematic categories.

5.1. Data augmentation using Generative AI

From Tables 2 and 3, it was observed that increasing the proportion of generated complaints improved the performance metrics up to a certain point. However, beyond that point, further increasing the amount of generated data led to a decline in the metrics.

The decline in performance when adding too much generated data can be attributed to subtle differences in the characteristics of the generated complaints compared to the original data. Although specific examples were provided during the generation process, some of the generated complaints diverged from the themes present in the original dataset. For instance, while the original data did not contain complaints like "my child refuses to do homework," such examples were generated during data augmentation.

One notable observation is that the BoW method was most affected by the excess generated data. Since BoW relies on a direct, sparse representation of word frequencies without considering the context or semantics of the words, it is highly sensitive to variations in the dataset. As a result, BoW was more prone to overfitting and being influenced by the unique patterns in the generated complaints, leading to a significant decrease in performance metrics.

All three methods—BoW, Doc2Vec, and SCDV—aim to convert text into numerical vectors so that machine learning models can process them. However, the way they create these vectors differs significantly. BoW represents a document by counting the frequency of each word in the text, ignoring grammar, word order, and context. As a result, two sentences with the same words in different orders—such as “The child is crying” and “Crying is the child”—would be represented identically in BoW. In contrast, Doc2Vec captures the order and surrounding context of words, producing dense vectors that reflect the semantic meaning of the entire document. SCDV further builds on this by combining word embeddings with soft clustering to account for words with multiple meanings. Although all these methods perform vectorization, context-aware methods like Doc2Vec and SCDV are better suited for handling subtle differences in sentence structure and meaning. This capability made them less susceptible to performance degradation when synthetic data was introduced, as they could focus on the overall context rather than isolated word occurrences. To address these discrepancies, the following improvements could be considered:

(1) Providing More Specific Examples:

Increasing the number and diversity of examples from the original data during the prompt creation process can help align the generated data more closely with the original dataset's characteristics.

(2) Summarizing Frequent Themes:

Including an overview of the most common types of complaints in the dataset as part of the prompt can guide the generative model to produce data that better reflects the distribution of original themes.

5.2. Complaints in parenting Q&A sites

As a result of the analysis, clustering based on representative words, as shown in Table 5, was achieved for the parenting Q&A site data. From the representative words in each class, the following characteristics and themes can be inferred:

Class 1

Complaints in Class 1 focus on the daily stress and struggles of parents with small children, managing household chores, and childcare. These posts reflect the exhaustion and challenges of maintaining a work-life balance while handling everyday responsibilities.

Class2

Class 2 represents complaints related to family gatherings, particularly during the New Year holidays when people visit their parents' homes. The word “Aunt” highlights a common stress point in Japanese culture, where tension between mothers and their mothers-in-law often arises. This suggests that meeting extended family members during such events causes emotional strain.

Class3

Class 3 pertains to family and marital issues. The inclusion of words like “husband” and “divorce” indicates that the complaints revolve around conflicts within the household, some of which escalate to discussions about separation or divorce. Posts in this class likely detail arguments and the emotional toll of strained family relationships, including the impact on children.

Class4

Class 4 captures anxieties related to the arrival of a new child. The presence of terms like “schedule” and “birth” suggests concerns about planning and preparation for childbirth. As the term “maternity blues” implies, childbirth is a significant life event that often brings emotional challenges for mothers, highlighting their vulnerability during this period.

Class5

Complaints in Class 5 focus on health-related anxieties. Words referencing dates and “hospital” suggest frustrations about the availability of medical care on desired days. These posts reflect concerns about managing appointments, dealing with illness, and accessing timely healthcare, which are common stressors for parents, especially when caring for children.

Class6

Similar to Class 3, Class 6 revolves around domestic issues, specifically those involving family dynamics. The mention of “husband” and “son” indicates challenges in managing relationships and roles within the household. The overlap with Class 3 suggests that family-related complaints are a dominant theme, possibly reflecting ongoing or recurring issues within domestic life.

Overall, the clustering results reveal that complaints on parenting Q&A sites are not limited to child-rearing itself but are closely tied to a wide range of social and emotional stressors. The dominant presence of themes related to family relationships, such as marital conflict, in-law tensions, and domestic responsibilities, suggests that many users turn to the platform to express frustrations rooted in interpersonal dynamics. Furthermore, the appearance of health-related and seasonal concerns, such as hospital access and holiday stress, underscores the influence of temporal and situational factors on parental stress. These findings highlight the complex and multifaceted nature of parenting-related complaints and demonstrate the effectiveness of topic-based clustering in uncovering latent patterns in user-generated content.

6. Conclusion

This study addressed the classification problem of determining whether a post is a complaint, using data from a parenting Q&A site. Due to the insufficiency of complaint data, data augmentation was performed using ChatGPT. Additionally, feature analysis was conducted by applying pLSA to the complaints identified by the classification model.

(1) Using generative AI to augment data improved classification performance up to a certain point, with the optimal results achieved when 20% of the dataset consisted of generated complaints. This highlights the potential of generative AI in addressing data scarcity, though care must be taken to ensure the generated data closely aligns with the original dataset to avoid performance degradation.

(2) The feature analysis using pLSA successfully achieved distinctive clustering related to parenting, shedding light on the causes of dissatisfaction among mothers. The findings revealed that the concerns extended beyond childcare itself, encompassing issues such as difficulties with hospital appointments and challenges in relationships with extended family, particularly with parents or in-laws.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number 24H00370.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] T. Kozue and A. Munakata, "Monologues on Twitter: Does the Relieved Stress Exceed the Social Networking Fatigue," Proceedings of the Annual Meeting of the Japanese Psychological Association, no. 83, article no. 71, 2019.

[2] T. Shimada and A. Sakurai, "Recognition of Questions Seeking Sympathy in Community QA Sites," Journal of Japanese Society for Fuzzy Theory and Intelligent Informatics, vol. 29, no. 4, pp. 611-618, 2017.

[3] I. Ito, H. Muranoi, and N. Shibata, "Mega Data Analysis of Grumbles on Social Networking Service," Bulletin of the Faculty of Education, Ibaraki University (Educational Science), Special Issue, pp. 389-406, 2014.

[4] K. Ito, T. Murayama, S. Yada, S. Wakamiya, and E. Aramaki, "Construction of a Japanese 'Guchi' Dataset Considering Targets," Proceedings of the 28th Annual Meeting of the Association for Natural Language Processing, Paper No. F8-4, 2022.

[5] D. Preoţiuc-Pietro, M. Gaman, and N. Aletras, "Automatically Identifying Complaints in Social Media," Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5008-5019, 2019.

[6] M. Jin and N. Aletras, "Modeling the Severity of Complaints in Social Media," Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2264-2274, 2021.

[7] M. Fang, S. Zong, J. Li, X. Dai, S. Huang, and J. Chen, "Analyzing the Intensity of Complaints on Social Media," Findings of the Association for Computational Linguistics: NAACL 2022, pp. 1742-1754, 2022.

[8] X. Tian, I. Vertommen, L. Tsiami, P. van Thienen, and S. Paraskevopoulos, "Automated Customer Complaint Processing for Water Utilities Based on Natural Language Processing—Case Study of a Dutch Water Utility," Water, vol. 14, no. 4, article no. 674, 2022.

[9] G. Alarifi, M. F. Rahman, and M. S. Hossain, "Prediction and Analysis of Customer Complaints Using Machine Learning Techniques," International Journal of E-Business Research, vol. 19, no. 1, pp. 1-25, 2023.

[10] M. Jin and N. Aletras, "Complaint Identification in Social Media with Transformer Networks," Proceedings of the 28th International Conference on Computational Linguistics, pp. 1765-1771, 2020.

[11] J. Wang, J. Lai, and Y. Lin, "Social media analytics for mining customer complaints to explore product opportunities," Computers & Industrial Engineering, vol.178, article no. 109104, 2023.

[12] Y. Zhang, R. Jin, and Z. H. Zhou, "Understanding Bag-of-Words Model: A Statistical Framework," International Journal of Machine Learning and Cybernetics, vol. 1, pp. 43-52, 2010.

[13] S. Martinčić-Ipšić, T. Miličić, and L. Todorovski, "The Influence of Feature Representation of Text on the Performance of Document Classification," Applied Sciences, vol. 9, no. 4, article no. 743, 2019.

[14] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," Proceedings of the 31st International Conference on Machine Learning (ICML'24), vol. 32, no. 2, pp. 1188-1196, 2014.

[15] Q. Chen and M. Sokolova, "Specialists, Scientists, and Sentiments: Word2Vec and Doc2Vec in Analysis of Scientific and Medical Texts," SN Computer Science, vol. 2, article no. 414, 2021.

[16] D. Mekala, V. Gupta, B. Paranjape, and H. Karnick, "SCDV: Sparse Composite Document Vectors Using Soft Clustering over Distributional Representations," Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 659–669, 2017.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Proceedings of the International Conference on Learning Representations (ICLR) Workshop, 2013

[18] V. Gupta, A. Saw, P. Nokhiz, H. Gupta, and P. Talukdar, "Improving Document Classification with Multi-Sense Embeddings," Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020), pp. 324-331, 2020.

[19] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 3146-3154, 2017.

[20] T. Hofmann, "Probabilistic Latent Semantic Analysis," Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI1999), pp. 289-296, 1999.

[21] T. Yang, G. Kumoi, H. Yamashita, and M. Goto, "Transfer Learning Based on Probabilistic Latent Semantic Analysis for Analyzing Purchase Behavior Considering Customers' Membership Stages," Journal of Japan Industrial Management Association, vol. 73, no. 2E, pp. 160-175, 2022.

[22] G. Schwarz, "Estimating the Dimension of a Model," The Annals of Statistics, vol. 6, no. 2, pp. 461-464, 1978.