Painsight: An Extendable Opinion Mining Framework for Detecting Pain Points Based on Online Customer Reviews

Yukyung Lee[1], Jaehee Kim[1], Doyoon Kim[1], Yookyung Kho[1], Younsun Kim[2], Pilsung Kang[1]

[1]Korea University, Seoul, Republic of Korea
[2]LG Electronics Inc., Seoul, Republic of Korea

[1]{yukyung_lee,jaehee_kim, doyooni303, yookyung_kho, pilsung_kang}@korea.ac.kr
[2]younsunny.kim@gmail.com

Abstract

As the e-commerce market continues to expand and online transactions proliferate, customer reviews have emerged as a critical element in shaping the purchasing decisions of prospective buyers. Previous studies have endeavored to identify key aspects of customer reviews through the development of sentiment analysis models and topic models. However, extracting specific dissatisfaction factors remains a challenging task. In this study, we delineate the pain point detection problem and propose Painsight, an unsupervised framework for automatically extracting distinct dissatisfaction factors from customer reviews without relying on ground truth labels. Painsight employs pre-trained language models to construct sentiment analysis and topic models, leveraging attribution scores derived from model gradients to extract dissatisfaction factors. Upon application of the proposed methodology to customer review data spanning five product categories, we successfully identified and categorized dissatisfaction factors within each group, as well as isolated factors for each type. Notably, Painsight outperformed benchmark methods, achieving substantial performance enhancements and exceptional results in human evaluations.

1 Introduction

The thriving e-commerce market has rendered online customer reviews an indispensable factor in influencing the purchasing decisions of potential consumers (Zhu and Zhang, 2010; Kwahk and Kim, 2017; Dellarocas et al., 2007). These reviews offer invaluable insights for businesses, empowering them to refine their products and services (Alibasic and Popovic, 2021). The analysis of copious customer reviews to comprehend customer needs and pain points is of paramount importance for augmenting service quality and heightening customer satisfaction (Plotkina and Munzel, 2016; Eslami and Ghasemaghaei, 2018; Wu and Chang, 2020; Berger et al., 2020). Pain points pertain to specific difficulties or problems encountered by customers while utilizing a product or service (Lee, 2014). These encompass emotional challenges arising from psychological demands and the incongruities between customers' actual perceptions and their expectations of products and services. Identifying and addressing pain points serves as a crucial initial step in enhancing the quality of products and services (East et al., 2008; Ho-Dac et al., 2013; Geetha et al., 2017).

The increasing importance of customer reviews has spurred a wealth of research into comprehending their effects through the application of natural language processing methodologies, encompassing opinion mining, sentiment analysis, topic modeling, and keyword extraction (Eslami and Ghasemaghaei, 2018; Heng et al., 2018; De Geyndt et al., 2022). Nonetheless, the identification of specific attributes, such as pain points in customer review data, necessitates annotation tailored to each product and service. Owing to this constraint, prior studies have predominantly concentrated on devising methods for extracting keywords, a notion more expansive than pain points. Wang et al. (2018) extracted attributes from product specifications, while Klein et al. (2022) employed a BERT-based BIO tagging (Ramshaw and Marcus, 1995) model to extract aspect and opinion terms. More recently, researchers have shifted their focus toward aspect-based sentiment analysis (Bu et al., 2021), striving to analyze customers' opinions at the granular aspect level, as opposed to mere classification into being positive or negative (Zhang et al., 2022). Moreover, Wu and Chang (2020) utilized Latent Dirichlet Allocation (LDA) to differentiate review types and explore topic-related negative emotions.

Existing research has mainly focused on extracting keywords and aspect terms from reviews or simply analyzing review types; nonetheless, such methodologies exhibit limitations. Firstly, pain

points are not only more specific but also contingent upon the product or service in question, in contrast to keywords or aspect terms. This is because customers may articulate disparate pain points utilizing identical keywords. Consequently, a comprehensive definition of pain points must be established prior to their detection (Forman et al., 2008; De Bonte and Fletcher, 2014; Wang et al., 2016). Secondly, even when equipped with a definition for pain points, the unique nature of pain points across products and services necessitates the ongoing annotation of new products, an endeavor that is both labor-intensive and costly (Saura et al., 2021). To overcome these limitations, an automated framework for the definition, extraction, and analysis of pain points from customer reviews is requisite, which is applicable to any product or service. In this study, we propose Painsight, an extendable opinion-mining framework for pain point detection, composed of a series of modules. Painsight implements a pipeline that conducts sentiment analysis and topic modeling using pre-trained language models, subsequently extracting pain points based on gradient-based attribution. When applied to customer reviews encompassing five categories of home appliances, Painsight effectively classified pain points emerging in diverse product groups. The extracted pain points exhibited substantial enhancements in performance, both quantitatively and qualitatively, in comparison to the results procured by the baseline model. The main contributions of this study can be summarized as follows:

- We propose Painsight, an automated and scalable opinion-mining framework explicitly tailored for pain point analysis.
- Painsight encompasses a comprehensive pipeline that executes sentiment analysis, topic modeling, and task-specific gradient-based attribution, drawing on a pre-trained language model.
- Painsight demonstrates both quantitatively and qualitatively exceptional performance in the accurate identification of pain points concerning sentiment and topic across different product groups.

2 Problem Statement

We aim to detect pain points in user-generated content (UGC), such as customer reviews, to identify customer discomforts and challenges (Cheng et al.,

2021). However, the absence of standardized criteria for determining what constitutes pain points, contingent on the target product or service, poses a challenge because pain points exhibit variation in scale (Humphreys and Wang, 2018). As a result, the precise definition of pain points pertinent to the target product or service is crucial.

2.1 Definition of Pain Point

Pain points arise from emotions customers experience while utilizing products and services, typically characterized by keywords reflecting negative opinions. However, not all negative keywords constitute pain points; rather, they signify complaints addressable through functional or procedural enhancements (Homburg and Fürst, 2007; Rawson et al., 2013). For instance, while "I tried vacuum A, and it's not good" conveys dissatisfaction without actionable insight, "I used vacuum B, and the battery drains too fast" distinctly identifies "battery" as an area for potential improvement. The scope of pain points can be determined based on their prevalence among users, with a focus on addressing common concerns to develop product and service improvement strategies, while simultaneously uncovering infrequent, personalized demands for diversification. In this study, we concentrate on detecting pain points affecting the majority of users, with the prospect of future methodology advancements catering to minority-specific pain points.

2.2 Task Formulation

We first formulate pain point detection, a novel task proposed in this study. Given n customer review texts $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$, the model aims to identify k pain points $p_1, \ldots, p_k$ within a specific product group review dataset $\mathcal{X}$ ($\mathbf{x} \in \mathcal{X}$). The set of pain points in $\mathcal{X}$ is denoted as $\mathcal{P}(\mathcal{X})$, with each pain point $p_k \in \mathcal{P}(\mathcal{X})$ comprising contiguous tokens $x_i, \ldots, x_{i+t}$ that form part of review text $\mathbf{x}$. In real-world settings, ground truth pain points for each product group are not predefined. Consequently, extracting suitable pain point candidates from review data and selecting product-specific pain points is essential. However, executing pain point extraction in a fully unsupervised manner presents significant challenges. Therefore, we assume the existence of relevant tasks with strong dependencies on pain point detection, with supervision y provided for learning these tasks on the dataset. These relevant tasks serve as weak supervision, approximating pain point detection by mod-
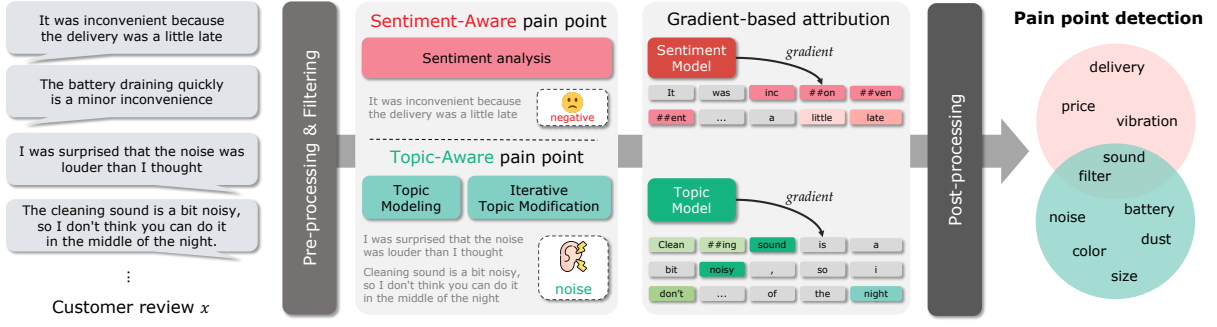
Figure 1: Architecture of Painsight. Painsight receives customer reviews as input and constructs a sentiment-aware and topic-aware pain point model. The trained models are analyzed by a gradient-based attribution method to calculate the importance of each word in the input sentence, where darker tokens indicate a greater contribution to the prediction. The final pain points are detected through a series of post-processing steps based on high attribution scores.

eling contexts of product-specific review and incorporating valuable knowledge. Given that each review is grounded in customer perception, words with substantial influence on relevant task predictions can be understood as pain points of actual customers.

In this study, we utilized sentiment analysis and iterative topic modification (ITM) based on topic modeling as relevant tasks for pain point detection. Moreover, we aimed to extract keyword-based pain points using token attribution scores derived from these tasks. In both tasks, token attribution scores are computed through the following procedure. Given review $\mathbf{x}$ and its corresponding embedding sequence $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{|\mathbf{x}|})$, we define the relevant task classifier f^{task}, which takes the embedding sequence as input. The input gradient of each token can be used to evaluate the influence of the token for the target task, represented as a normalized gradient attribution vector $\mathbf{a} = (a_1, a_2, \ldots, a_{|\mathbf{x}|})$. As described by Wang et al. (2020), the attribution at position i can be expressed as:

$$a_i = \frac{|\nabla_{\mathbf{x}_i} \mathcal{L} \cdot \mathbf{x}_i|}{\sum_j |\nabla_{\mathbf{x}_j} \mathcal{L} \cdot \mathbf{x}_j|}, \qquad (1)$$

where $\mathcal{L}$ denotes the loss generated by classifier f^{task}, and a_i is calculated through the dot product between the gradient of $\mathcal{L}$ and the embedding $\mathbf{e}_i$. Gradient-based attribution $\mathbf{a}$ (Sundararajan et al., 2017; Ross et al., 2017) represents each token's influence on the final prediction and can thus approximate word importance (Feng et al., 2018).

3 Painsight

The primary objective of the proposed framework is to model the entire process of automatically detecting pain points in real-world scenarios. Specifically, this study aims to address the following two practical research questions:

- Q1: How do customers perceive products in general?
- Q2: What types of discomfort do customers experience?

Considering the significance of both perspectives from prior works, a framework capable of generating accurate and diverse output, covering a wide range of distinct pain points, is necessary. The architecture of Painsight, depicted in Figure 1, features a parallel structure to incorporate these two research questions.

3.1 Pre-processing & Filtering

We constructed a dataset by collecting various customer reviews from the web. As real-world customer reviews contain diverse noise, such as grammatical errors and outliers (i.e., irrelevant reviews), we applied three pre-processing steps for data refinement: i) spacing correction, ii) keyword dictionary construction, and iii) data filtering. Review data often consists of colloquial styles and may contain grammatical and spacing errors. We first utilized a pre-trained spacing model[1] to correct spacing errors and employed regular expressions to fix common typos. Furthermore, as mentioned in Section 2.1, filtering out sentences without pain

---

[1] We employed a character-level convolution neural network-based spacing model.
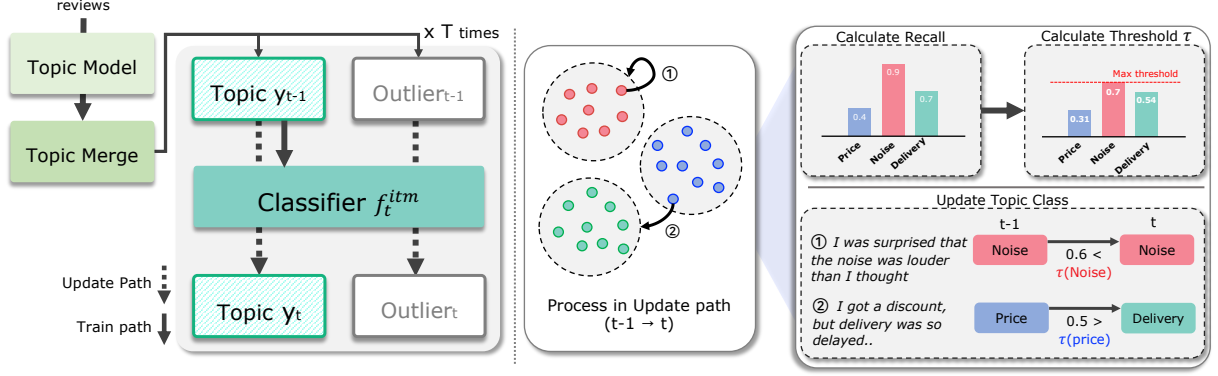
217

Figure 2: Pipeline of topic-aware pain point. Review data undergoes topic modeling and merging to obtain initial topics. Then, a classifier f_t^{itm} is trained for T iterations. At each time step, recall is calculated based on predicted probabilities. Topic classes are updated when the confidence of class at time $t-1$ exceeds the threshold at time t.

points is essential. To achieve this, we constructed a stopwords list and a keyword dictionary to filter out sentences expressing sentiments irrelevant to pain point extraction. We considered words frequently appearing in negative reviews and not included in the stopwords list as important keywords. Additionally, we filtered out duplicate reviews or those with fewer than ten tokens.

3.2 Sentiment-Aware Pain Point

The Sentiment-Aware pain point approach focuses on negative emotions to address the question of ***"How do customers perceive products in general?"***. To achieve this goal, we conducted sentiment analysis on customer reviews using a classifier denoted as f^{sent}. The sentiment label y is defined as $y \in \{positive, negative\}$, and $\hat{y}$ represents the predicted class, expressed as $\hat{y} = \arg\max_{\hat{y}} f^{sent}(\hat{y} \mid \mathbf{x})$. Reviews predicted as 'negative' by the trained f^{sent} were considered likely to contain pain points.

3.3 Topic-Aware Pain Point

The Topic-Aware pain point approach focuses on identifying specific types of discomfort that customers experience, addressing the question ***"What types of pain points do customers encounter?"***. We began by analyzing negative reviews to identify various types of complaints. To detect pain points by topic, we further refined sentences in outlier topics to properly segment them. The Topic-Aware pain point approach consists of three stages: i) topic modeling, ii) topic merging, and iii) ITM, as illustrated in Figure 2.

Topic modeling The construction of Topic-Aware pain points necessitates establishing ground truth for each product group, which includes the number of topics and review-topic matching information. However, such labels are typically absent in customer reviews available on the web. Even when predefined pain points exist, detecting new pain points from real-time collected review data remains challenging. To address this issue, we first employed topic modeling to determine the initial topic class. Specifically, we utilized BERTopic (Grooten-dorst, 2022) to identify significant clustered topics. However, we observed that over 50% of the data were classified as outliers, defined in this study as reviews devoid of any distinct topic, with more than 100 topics extracted. Misclassified topic modeling results can impede interpretability and provide users with incorrect pain points. To tackle this problem, we merged topics to relabel misclassified ones and performed ITM to accurately distinguish outlier reviews with low relevance to pain points.

Topic merging In the above paragraph, we addressed the issue of reviews with similar topics being assigned to different clusters in the BERTopic output. To resolve this challenge, we propose a procedure for selecting representative words for each topic and determining which topics should be merged. Initially, we extracted nouns with the highest c-TF-IDF scores for each topic and designated them as representative words. Words with scores at least s were considered keywords for the product group.[2] Subsequently, we searched for topics containing these keywords and merged them into a single topic. Following the topic merging process, we observed that semantically similar topics were combined, resulting in roughly 20 to 30 merged topics.

---

[2] We set the hyperparameter s to 0.1 based on experimental results.

We also adjusted minor topic classes, which consisted of 5 to 10 reviews. However, the described topic merging approach primarily focuses on lexical matching assigned to the topics. Therefore, incorporating additional improvements that consider semantic aspects is crucial for refining the results.

Iterative topic modification Despite the improvements in topic modeling results through merging, over 50% of reviews still remain outliers. To address this problem, it is essential to assign initial topics that may be misclassified and to distinguish between reviews containing pain points and outlier reviews in the data. Consequently, we propose an ITM algorithm to enhance topic modeling results by updating the predicted topic with a confidence threshold above a certain level if it differs from the existing topic class during the training process. Our ITM algorithm is inspired by curriculum pseudo labeling (CPL) (Zhang et al., 2021), a method designed for semi-supervised learning that flexibly adjusts the threshold based on the learning difficulty for each class. Similar to CPL, the ITM algorithm adjusts the number of updated data points by varying the threshold according to the classification difficulty for each topic during the training process. As a result, this approach enables the modification of misclassified topic classes and induces additional topic merging.

Let the classifier for ITM be denoted as f_t^{itm}, where t represents the current time-step of the classifier, and $f_t^{itm}(y_t \mid \mathbf{x})$ represents the prediction probability of y_t. y_t is the topic class at time t for review $\mathbf{x}$ and belongs to the topic set $\mathcal{C} = topic_1, topic_2, \ldots, topic_M$. We used the merged topic output assigned to each review as the initial topic y_0. τ denotes a pre-defined threshold[3], and the threshold, $\mathcal{T}(y_t)$, for label modification based on f_t^{itm} can be defined as:

$$\mathcal{T}(y_t) = \mathcal{R}(y_t) \cdot \tau. \qquad (2)$$

$\mathcal{T}(y_t)$ is a flexible threshold for topic class yt at time-step t, and $\mathcal{R}(y_t)$ is a relative recall value representing topic class difficulty. In CPL, accuracy is used as a measure of difficulty; however, accuracy can lead to biased estimates due to class imbalance in our topic modeling. As an alternative, we utilize recall, a sensitivity measure, to define the difficulty of the topic class. High recall indicates an easy class with high $\mathcal{T}(y)$, while low recall implies a

difficult class with low $\mathcal{T}(y)$. Class difficulty is defined as:

$$\mathcal{R}(y_t) = \frac{\text{recall}(y_t)}{\max_{y_t \in \mathcal{C}}(\text{recall}(y_t))}. \qquad (3)$$

We train f_t^{itm} to maximize log-likelihood based on the topic at $t-1$ and calculate the difficulty of each class $\mathcal{T}(y_t)$ at every time-step[4]. If the predicted probability $f_t^{itm}(y_t \mid \mathbf{x})$ is greater than $\mathcal{T}(y_t)$, we modify the topic to y_t. y_t is defined as:

$$y_t = \begin{cases} y_t & \text{if } f_t^{itm}(y_t \mid \mathbf{x}) > \mathcal{T}(y_t), \\ y_{t-1}, & \text{otherwise.} \end{cases} \qquad (4)$$

Upon initializing the training with initial topics, we terminated the process when topics no longer merged or reached a satisfactory state[5]. We then considered the final prediction of ITM as the topic for each review.

3.4 Gradient-based Attribution

In this study, we aim to extract word importance related to pain points from two interdependent tasks: sentiment analysis and topic modeling. We employed gradient-based attribution, using token attribution scores for each task's prediction. Normalized attribution vectors a_i^{sent} and a_i^{itm} for individual tokens are derived from trained classifiers f^{sent} and f^{itm}, as shown in Eq. (1):

$$a_i^{sent} = \frac{|\nabla_{\mathbf{x}_i} \mathcal{L}_{sent} \cdot \mathbf{x}_i|}{\sum_j |\nabla_{\mathbf{x}_j} \mathcal{L}_{sent} \cdot \mathbf{x}_j|}, \qquad (5)$$

$$a_i^{itm} = \frac{|\nabla_{\mathbf{x}_i} \mathcal{L}_{topic} \cdot \mathbf{x}_i|}{\sum_j |\nabla_{\mathbf{x}_j} \mathcal{L}_{topic} \cdot \mathbf{x}_j|}. \qquad (6)$$

We computed a_i^{sent} for 'negative' reviews and a_i^{itm} for reviews in $M-1$ topics, excluding outlier topics. We then selected the top-g a_i^{sent} and a_i^{itm} with the highest attribution vectors and extracted words with index i. These words form the pain point candidate set $\mathcal{C}(\mathcal{X})$. The pain point set $\mathcal{P}(\mathcal{X})$ is a subset of $\mathcal{C}(\mathcal{X})$. We employed Captum (Kokhlikyan et al., 2020)'s integrated gradient method for axiomatic attribution calculation.

3.5 Post-Processing

The derived $\mathcal{C}(\mathcal{X})$, when used as the final result for pain points, has certain limitations. Since a_i^{sent} and a_i^{itm} are token attribution scores contributing to

---

[3]τ is a hyperparameter set experimentally within the search space [0.4, 0.5, 0.6, 0.7].

[4]In this study, we set each epoch as a time-step.

[5]Meaning two out of three evaluation metrics no longer show improvement.

(a) Captum (top 3 attribution score)
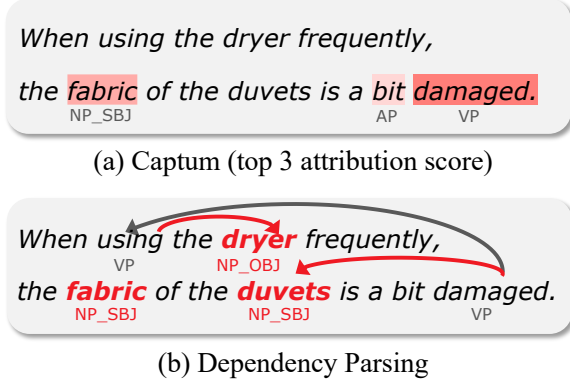

(b) Dependency Parsing

Figure 3: Example of post-process. (a) represents the top three words with high attribution scores, where darker colors indicate a greater contribution to the prediction. (b) shows the result of dependency parsing, based on the part-of-speech information obtained in (a), to extract pain point candidates. The words highlighted in red belong to $\mathcal{C}(\mathcal{X})$.

each class prediction, they tend to focus on words describing "states," such as verbs and adjectives like "not good" and "bad." To include both the subject and object related to the predicate's action, additional post-processing is necessary. We addressed this issue by refining the results using a dependency parsing model (Heo et al., 2021), a method that captures relationships between words in a sentence.

First, we sorted the top g tokens with the highest attribution scores from f_{task} and extracted the corresponding words w_i. Then, we analyzed word relationships within sentences using dependency parsing. If w_i belongs to a noun phrase (NP), we define it as $\mathcal{C}(\mathcal{X})'$; if it belongs to a verb phrase (VP), we add the related NP to $\mathcal{C}(\mathcal{X})'$. If the related word is a VP, we search for an NP to supplement $\mathcal{C}(\mathcal{X})'$. We calculate word frequencies in the stopword-filtered $\mathcal{C}(\mathcal{X})'$ and define the top N words as the pain points $\mathcal{P}(\mathcal{X})$ for each product group[6]. Examples of post-processing can be found in Figure 3.

4 Experiment setup

4.1 Dataset

We utilized Korean customer reviews on five home appliance categories gathered from various web sources between January 2020 and November 2021. The target categories included dryers, stylers, washing machines, vacuum cleaners, and robotic vacu-

---
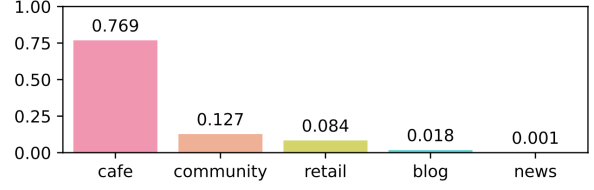
Figure 4: The overall data distribution by site type.

	Dryer	Styler	WM	VC	RV
Positive	48,249	26,453	111,603	159,222	36,091
Negative	16,608	9,034	23,510	49,919	9,566
Total	64,857	35,487	135,113	209,141	45,657

Table 1: Data statistics for product groups.

ums, obtained from 17 sites spanning five source types: retail, blog, cafe, community, and news. The dataset, originally collected by a home appliance company in Korea, had sentiment labels assigned based on the company's internal sentiment analysis logic within their voice of customer (VOC) analysis system. Data distribution and statistics for each product group are illustrated in Figure 4 and Table 1. Detailed distribution of product categories can be found in Appendix A.

4.2 Baselines

4.2.1 Sentiment Analysis

TextCNN (Kim, 2014) utilizes 1D convolutions to capture variable-length local features and n-grams. **Hierarchical Attention Network (HAN)** (Yang et al., 2016) consists of a single-layer Gated Recurrent Unit (GRU) that utilizes attention mechanisms at the word, sentence, and document levels to extract important information effectively. **BERT** (Devlin et al., 2019) is a bidirectional pre-trained language model [7]. **BERT + AVocaDo** (Hong et al., 2021) is an extension of BERT with AVocaDo technique. It is designed to handle product-specific review data by performing vocab expansion on a domain-specific corpus. The expansion process is based on tokenizer training and calculation of the sub-word segmentation ratio for the domain corpus, with contrastive learning between the original tokenizer and the expanded one during fine-tuning. **BERT + Vocab Expansion** manually adds frequently occurring vocabulary from the review to the BERT model. High-frequency words were iden-

---

[6]We set g to 3, the number of $\mathcal{P}(\mathcal{X})_{sent}$ to 30, and the number of $\mathcal{P}(\mathcal{X})_{topic}$ to 10 for each topic.

[7]The 'klue/bert-base' version was employed.

Model	Dryer		Styler		WM		VC		RV		Avg.	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
TextCNN	94.35	92.76	90.20	84.32	96.75	94.31	96.47	95.15	95.28	93.06	94.61	91.92
HAN	94.52	92.77	92.39	89.86	96.59	94.13	96.14	94.74	95.54	93.42	95.04	92.98
BERT	98.40	97.93	**97.57**	**96.80**	98.80	97.92	**98.36**	**97.70**	98.03	97.02	98.23	97.47
BERT+AVocaDo	97.52	96.79	96.23	95.01	98.55	97.50	98.13	97.46	98.39	97.58	97.76	96.87
BERT+Vocab Expansion	**98.44**	**97.96**	97.23	96.35	**98.81**	**97.94**	98.32	**97.70**	**98.52**	97.63	**98.26**	**97.52**

Table 2: Results on sentiment analysis in accuracy and f1 score. we utilize the following abbreviations: "WM" for Washing Machine, "VC" for Vacuum Cleaner, and "RV" for Robotic Vacuum. The average performance for the five datasets is presented in the rightmost column, with the highest performance indicated in bold.

tified using a count-based approach, and those not present in BERT vocabulary were added.

4.2.2 Topic Modeling

LDA (Blei et al., 2003) is a method inferring topics from word occurrence patterns in a corpus. While LDA provides multiple topics for each review, for comparison with Painsight, we considered the highest probability topic as the representative.
BERTopic is a BERT-based topic model [8] to generate document embeddings, clustering and creates representations using the c-TF-IDF procedure.

4.2.3 Pain Point Detection

spaCy (Honnibal and Montani, 2017) is baseline using part-of-speech tagging to extract nouns within sentences. The same post-processing method as Painsight was applied.

5 Experiment Results

5.1 Sentiment Analysis

Table 2 presents the sentiment analysis performance across the five product groups. Given that review data is domain-specific, the language model can benefit from a vocabulary expansion process (Hong et al., 2021). We aimed to enhance performance by applying AVocaDo and Vocab Expansion to the original BERT model. The experimental results showed that the BERT + Vocab Expansion model achieved superior performance and was thus selected as the final model for calculating attribution scores.

5.2 Topic Modeling

Table 3 presents the topic modeling performance for the five product groups. We evaluated the performance of LDA, BERTopic, and ITM using

Dataset	Model	NPMI ↑	Outlier (%)
Dryer	LDA	-0.0327	-
	BERTopic	**0.0558**	59.39%
	ITM (Ours)	0.0208	**30.10%**
Styler	LDA	-0.0470	-
	BERTopic	0.0379	55.81%
	ITM (Ours)	**0.0737**	**39.39%**
WM	LDA	-0.0055	-
	BERTopic	0.0406	50.53%
	ITM (Ours)	**0.0487**	**29.92%**
VC	LDA	0.0244	-
	BERTopic	0.0512	71.92%
	ITM (Ours)	**0.0628**	**26.67%**
RV	LDA	-0.0232	-
	BERTopic	**0.0488**	56.39%
	ITM (Ours)	0.0450	**37.81%**

Table 3: Results on topic modeling in NPMI, The rightmost column shows the proportion of outlier topics for each method, except for LDA, which does not extract outlier topics separately. The highest performance and low outlier ratio are marked in bold.

coherence-based clustering metrics such as NPMI. LDA generates the desired number of topics without considering outliers, resulting in significantly lower performance compared to BERTopic and ITM. Conversely, BERTopic designates, on average, 57% of the reviews as outliers and assigns topics only to the remaining data. This results in BERTopic being evaluated with a smaller number of reviews compared to other methods, which could create a more favorable environment for BERTopic's performance measurements. In contrast, our ITM algorithm effectively assigns reviews initially deemed as outliers to suitable topics, achieving comparable or superior performance to BERTopic in most cases. These results demonstrate that ITM successfully reduced the average proportion of outliers to 32%, even under relatively

---

[8]In this study, we used the "sentence-transformers/xlm-r-100langs-bert-base-nli-stsb-mean-tokens" model.

challenging experimental conditions.

5.3 Pain Point Detection

As discussed in Section 2.2, real-world customer reviews lack gold labels for pain points. Thus, to validate the effectiveness of Painsight in detecting sentiment-aware and topic-aware pain points, we conducted human evaluations. We designed experiments for three tasks, each addressing the following research questions:

- Task 1: Can the sentiment-aware pain point module extract appropriate pain point candidates from each sentence?
- Task 2: Can the sentiment-aware pain point module identify suitable pain points for each product category?
- Task 3: Can the topic-aware pain point module detect appropriate pain points for specific topics within each product category?

We conducted human evaluations for the three tasks across five product groups, with eight unique human judges participating in each experiment. In Task 1, we randomly selected 100 sentences per product category and used spaCy and Painsight to identify pain point candidates in each sentence. Judges assessed the appropriateness of the extracted candidates on a scale from 1 to 5, where 1 signifies 'all words are extracted incorrectly' and 5 indicates 'all words are extracted correctly'. In Task 2, we provided 100 randomly sampled example sentences for each product category and examined the adequacy of the final pain points identified for each category. Judges evaluated each pain point word, assigning a score of 0 for unsuitable and 1 for suitable. For Task 3, we randomly selected 20 example sentences per product category according to topics and assessed the appropriateness of the detected pain points for the corresponding topics. Judges assigned a score of 0 for unsuitable and 1 for suitable pain point words.

Table 4 presents the human evaluation results across the three tasks. First, in Task 1 (Table 4 - (a)), which concentrates on extracting pain point candidates from each review sentence, Painsight's approach — extracting words contributing to 'negative' predictions based on high attribution scores — outperforms spaCy's noun extraction, with an average improvement of 0.65 points. In Task 2 (Table 4 - (b)), the performance of the final sentiment-aware pain points is evaluated. These pain points

4 - (a)	Task 1 (1~5 scale)					
	Dryer	**Styler**	**WM**	**VC**	**RV**	**Avg.**
spaCy	3.02	2.93	2.67	2.99	3.17	2.95
Painsight	**3.64**	**3.57**	**3.41**	**3.58**	**3.81**	**3.60**

4 - (b)	Task 2 (0 or 1)					
	Dryer	**Styler**	**WM**	**VC**	**RV**	**Avg.**
spaCy	0.53	0.61	0.45	0.54	0.60	0.55
Painsight	**0.72**	**0.77**	**0.67**	**0.69**	**0.75**	**0.72**

4 - (c)	Task 3 (0 or 1)					
	Dryer	**Styler**	**WM**	**VC**	**RV**	**Avg.**
LDA	0.13	0.16	0.18	0.14	0.10	0.12
BERTopic	0.29	0.31	0.40	0.33	0.32	0.33
Painsight	**0.54**	**0.51**	**0.49**	**0.39**	**0.51**	**0.47**

Table 4: Results of human evaluation for Tasks 1, 2, and 3. Task 1 was evaluated on a scale of 1 to 5, while Tasks 2 and 3 were assessed with scores of 0 or 1. The average performance across the five datasets is displayed in the rightmost column, with the highest performance indicated in bold. Paired t-tests were conducted comparing baselines with Painsight, and all experiments exhibited significant differences with p-values < 0.001.

are obtained by sorting the pain point candidates from Task 1 by frequency across all product categories. Assessing the output for each word reveals an average improvement of 0.17 points across the five product categories. Lastly, Task 3 (Table 4 - (c)) examines the final results of the topic-aware pain points by extracting pain points for each topic within the product categories. This assessment considers the relevance of the final pain points for each topic, and Painsight records the highest performance across the five product categories. Our method exhibits an average performance improvement of 0.35 points over LDA and 0.14 points over BERTopic. The results in Table 4 demonstrate the effectiveness of ITM, which could not be solely assessed using the NPMI metrics in Table 3. By employing Tasks 1, 2, and 3, the Painsight pipeline, which extracts pain points based on relevant tasks, also records higher performance compared to the baseline in human evaluation results. This validates the appropriateness of Painsight as an automatic framework for pain point detection.

6 Conclusion

In this study, we propose Painsight, a novel framework for automatically extracting and evaluating pain points from customer reviews. We address the under-explored problem of pain point detection and

present a practical pipeline for real-world scenarios. By employing sentiment analysis and topic modeling, we identify sentiment-aware and topic-aware pain points that reflect customer perceptions and various types of discomfort. The final output is obtained by extracting the most important words or features from the data using a gradient-based attribution score. This score enables us to determine which words or features are most critical in influencing the model's decision-making process and utilize this information in the post-process to recognize more accurate and meaningful pain points. Experimental results demonstrate that Painsight outperforms existing models on five product group reviews, with human evaluation results indicating a high level of agreement compared to the baseline. Future work could involve incorporating diverse customer feedback and constructing a high-quality benchmark dataset to further validate and enhance the proposed approach.

Limitations

In prior research (Salminen et al., 2022), several challenges have been identified in this field, such as noisy or low-quality data, semantic ambiguity, absence of standards, social desirability bias, and the requirement for human intervention. Our study aimed to tackle the challenge of detecting pain points and devised various strategies for managing noisy real-world reviews. Nonetheless, to fully unlock the potential of the Painsight, additional research is necessary to explore the wide range of emotional polarities beyond the generic 'negative' sentiment. Furthermore, customer reviews often show mixed sentiments, which calls for addressing semantic ambiguity. Lastly, the performance of Painsight assessment was constrained to five product categories, highlighting the need for a comprehensive, high-quality benchmark encompassing diverse domains and performance evaluations across distinct categories.

Ethics Statement

Throughout our human evaluation, we collected demographic details such as name, age, gender, and highest education level, after securing participants' consent and assuring them that their information would be exclusively utilized for research purposes. The results from the human evaluation were anonymized to protect participant confidentiality. The authors meticulously examined all customer reviews employed in the assessment, verifying the absence of any offensive or biased material. Participants took part in the evaluation for an estimated 40 minutes. They were compensated with a 5,000 KRW (equivalent to 3.7 USD) gift card, which was marginally above the Korean minimum wage during that period.

Acknowledgements

References

Armin Alibasic and Tomo Popovic. 2021. Applying natural language processing to analyze customer satisfaction. In *2021 25th International Conference on Information Technology (IT)*, pages 1–4. IEEE.

Jonah Berger, Ashlee Humphreys, Stephan Ludwig, Wendy W Moe, Oded Netzer, and David A Schweidel. 2020. Uniting the tribes: Using text for marketing insight. *Journal of marketing*, 84(1):1–25.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Jiahao Bu, Lei Ren, Shuang Zheng, Yang Yang, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. ASAP: A Chinese review dataset towards aspect category sentiment analysis and rating prediction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2069–2079, Online. Association for Computational Linguistics.

Li-Chen Cheng, Kuanchin Chen, Ming-Chu Lee, and Kua-Mai Li. 2021. User-defined swot analysis– a change mining perspective on user-generated content. *Information Processing & Management*, 58(5):102613.

Austina De Bonte and Drew Fletcher. 2014. *Scenario-Focused Engineering: A toolbox for innovation and customer-centricity*. Microsoft Press.

Ellen De Geyndt, Orphee De Clercq, Cynthia Van Hee, Els Lefever, Pranaydeep Singh, Olivier Parent, and Veronique Hoste. 2022. SentEMO: A multilingual adaptive platform for aspect-based sentiment and emotion analysis. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 51–61, Dublin, Ireland. Association for Computational Linguistics.

Chrysanthos Dellarocas, Xiaoquan Zhang, and Neveen F Awad. 2007. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive marketing*, 21(4):23–45.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Robert East, Kathy Hammond, and Wendy Lomax. 2008. Measuring the impact of positive and negative word of mouth on brand purchase probability. *International journal of research in marketing*, 25(3):215–224.

Seyed Pouyan Eslami and Maryam Ghasemaghaei. 2018. Effects of online review positiveness and review score inconsistency on sales: A comparison by product involvement. *Journal of Retailing and Consumer Services*, 45:74–80.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

Chris Forman, Anindya Ghose, and Batia Wiesenfeld. 2008. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information systems research*, 19(3):291–313.

M Geetha, Pratap Singha, and Sumedha Sinha. 2017. Relationship between customer sentiment and online customer ratings for hotels-an empirical analysis. *Tourism Management*, 61:43–54.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Yan Heng, Zhifeng Gao, Yuan Jiang, and Xuqi Chen. 2018. Exploring hidden factors behind online food shopping from amazon reviews: A topic mining approach. *Journal of Retailing and Consumer Services*, 42:161–168.

Hoon Heo, Hyunwoong Ko, Soohwan Kim, Gunsoo Han, Jiwoo Park, and Kyubyong Park. 2021. Pororo: Platform of neural models for natural language processing. https://github.com/kakaobrain/pororo.

Nga N Ho-Dac, Stephen J Carson, and William L Moore. 2013. The effects of positive and negative online customer reviews: do brand strength and category maturity matter? *Journal of marketing*, 77(6):37–53.

Christian Homburg and Andreas Fürst. 2007. See no evil, hear no evil, speak no evil: a study of defensive organizational behavior towards customer complaints. *Journal of the Academy of Marketing Science*, 35:523–536.

Jimin Hong, TaeHee Kim, Hyesu Lim, and Jaegul Choo. 2021. AVocaDo: Strategy for adapting vocabulary to downstream domain. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4692–4700, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Ashlee Humphreys and Rebecca Jen-Hui Wang. 2018. Automated text analysis for consumer research. *Journal of Consumer Research*, 44(6):1274–1306.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Ayal Klein, Oren Pereg, Daniel Korat, Vasudev Lal, Moshe Wasserblat, and Ido Dagan. 2022. Opinion-based relational pivoting for cross-domain aspect term extraction. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 104–112, Dublin, Ireland. Association for Computational Linguistics.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch.

Kee-Young Kwahk and Byoungsoo Kim. 2017. Effects of social media on consumers' purchase decisions: evidence from taobao. *Service Business*, 11:803–829.

Sampson Lee. 2014. Pig (pain is good) strategy: Make customer centricity obsolete and start a resource revolution. imatchpoint limited. *Hong Kong*.

Daria Plotkina and Andreas Munzel. 2016. Delight the experts, but never dissatisfy your customers! a multi-category study on the effects of online review source on intention to buy a new product. *Journal of Retailing and Consumer Services*, 29:1–11.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Alex Rawson, Ewan Duncan, and Conor Jones. 2013. The truth about customer experience. *Harvard business review*, 91(9):90–98.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2662–2670.

Joni Salminen, Mekhail Mustak, Juan Corporan, Soon-gyo Jung, and Bernard J Jansen. 2022. Detecting pain points from user-generated social media posts using machine learning. *Journal of Interactive Marketing*, 57(3):517–539.

Jose Ramon Saura, Domingo Ribeiro-Soriano, and Daniel Palacios-Marqués. 2021. From user-generated data to data-driven innovation: A research agenda to understand user privacy in digital markets. *International Journal of Information Management*, 60:102331.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961, Jeju Island, Korea. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Binda Wang, Yunwen Miao, Hongya Zhao, Jian Jin, and Yizeng Chen. 2016. A biclustering-based method for market segmentation using customer pain points. *Engineering Applications of Artificial Intelligence*, 47:101–109.

Junlin Wang, Jens Tuyls, Eric Wallace, and Sameer Singh. 2020. Gradient-based analysis of NLP models is manipulable. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 247–258, Online. Association for Computational Linguistics.

Yuren Wang, Xin Lu, and Yuejin Tan. 2018. Impact of product attributes on customer satisfaction: An analysis of online reviews for washing machines. *Electronic Commerce Research and Applications*, 29:1–11.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jia-Jhou Wu and Sue-Ting Chang. 2020. Exploring customer sentiment regarding online retail services: a topic-based approach. *Journal of Retailing and Consumer Services*, 55:102145.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. 2021. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*.

Feng Zhu and Xiaoquan Zhang. 2010. Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of marketing*, 74(2):133–148.

A Data Distribution

The distribution of site types used for collecting review data for five product categories (Dryer, Styler, Washing Machine, Vacuum Cleaner, and Robotic Vacuum) is presented in Figure 5. The data collection involved a diverse range of site types across product categories, with the highest number of reviews collected from cafe and community sites. The sentiment-aware pain point detection experiments utilized train, valid, and test datasets in an 8:1:1 ratio. However, the topic modeling in topic-aware pain point detection did not involve splitting the dataset.

B Training Details

The training process of Painsight was conducted using NVIDIA RTX 2080 Ti (for sentiment analysis and topic modeling) and RTX A6000 (for gradient-based attribution score and ITM). The PyTorch [9] library was performed for model training, and post-processing was carried out using Pororo [10] and Captum [11]. The hyperparameters used in Painsight are described in detail below:

---

[9] https://pytorch.org/
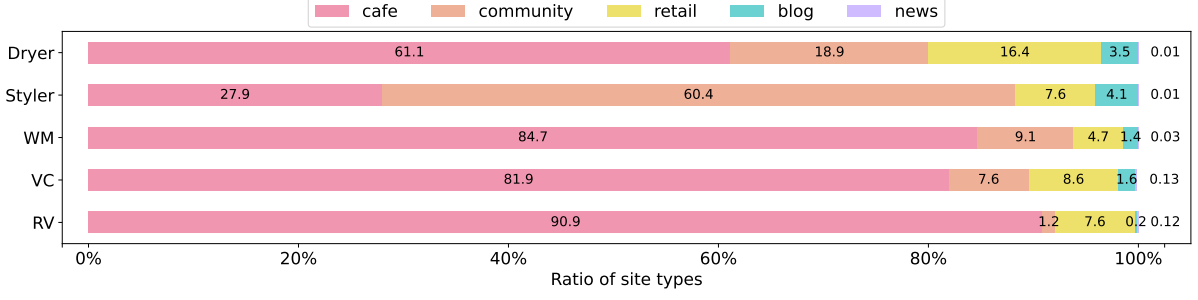[10] https://github.com/kakaobrain/pororo
[11] https://captum.ai/

Figure 5: Site type distribution of review data by product category.

B.1 BERT + Vocab Expansion

The model was trained using the transformers library (Wolf et al., 2019) with a BERT classifier. The batch size was set to 32, and the optimizer used was AdamW with a learning rate of 2e-05. The model was trained for 10 epochs, with 62 new vocabularies added. The maximum length of input sequences was set to 128.

B.2 ITM

The initial topic labels were generated using BERTopic, and the transformers library with a BERT classifier was used for classification. The optimizer used was Adam, and the model was trained for 1 epoch with a validation iteration of 10 and a maximum iteration of 100. Early stopping was used with patience 2. The maximum length of input sequences was set to 128, and the batch size was 64 with a learning rate of 3e-5.

C Evaluation Metric

We employed Normalized Pointwise Mutual Information (NPMI) (Stevens et al., 2012) as an evaluation metric to measure the performance of ITM. NPMI is a widely used measure of the correlation between two words, which is computed by normalizing Pointwise Mutual Information (PMI). PMI measures the probability of two words occurring together, taking into account the frequency of their individual occurrences. However, PMI tends to overestimate the importance of infrequent words. To address this issue, NPMI normalizes PMI by considering the probability of the respective words. Through this normalization process, NPMI can more accurately measure the correlation between two words and ranges between -1 and 1. NPMI is often used in topic modeling and is computed using the following formula:

$$\text{PMI}(w_i, w_j) = \frac{P(w_i, w_j)}{P(w_i)P(w_j)}, \quad (7)$$

$$\text{NPMI}(w_i, w_j) = \frac{\text{PMI}(w_i, w_j)}{-\log(P(w_i, w_j))}. \quad (8)$$

Here, $P(w_i, w_j)$ denotes the probability of words w_i and w_j co-occurring, while $P(w_i)$ and $P(w_j)$ represent their individual probabilities. The numerator normalizes the probability of the two words occurring together by dividing it by the product of their individual occurrence probabilities. The denominator uses the log value of their co-occurrence probability to obtain PMI. Using NPMI, we can extract sets of related words in topic classes classified through ITM and evaluate if each topic has coherence. Therefore, we evaluated the consistency of each topic in LDA, BERTopic, and ITM with a set of related reviews using NPMI.

D Human Evaluation

The evaluation instructions provided to annotators for each task in human evaluation are as follows, and the example is shown for the dryer.

D.1 Task 1: Sentiment-aware pain point evaluation (Pain point candidates)

Each sheet contains three items for each product group:

- Consumer reviews of appliances for each product group collected online
- Keywords extracted for each review (2-4 per review)

Please rate how well the keywords were extracted from each customer review:

- 1: All keywords were extracted incorrectly.
- 2: Keywords were generally not extracted.

- "I felt bad because the delivery driver was rude."
  (delivery, driver - pain point)
- "It's inconvenient to store the dryer
  because there's nowhere to put it." (storage - pain point)
- "It takes too long to dry my hair, much longer than
  a Product D." (time, long - pain point)
- "The dryer is heavy and the sound is loud, causing pain in
  my wrist and ears." (heavy, wrist, sound - pain point)

(a) Hair Dryer

- "It's too big and doesn't even fit on the balcony."
  (big, balcony - pain point)
- "The smell doesn't come out easily." (smell - pain point)

(b) Refrigerator

Figure 6: Examples of Pain Points: We explained the concept of pain points using examples of home appliances that were not used in the evaluation, specifically hair dryers and refrigerators.

- 3: Keywords were extracted at an average level (50% of all keywords).
- 4: Keywords were generally well extracted.
- 5: All keywords were extracted well.

D.2 Task 2: Sentiment-aware pain point evaluation (Final pain points)

Evaluate the suitability of pain points identified as frequently appearing in negative reviews of the product category after reviewing 100 example sentences. To clarify the concept of "pain point," we provide example sentences (Figure 6):

Scoring criteria:

- 0: This pain point cannot be considered a pain point for the product category.
- 1: This pain point can be considered a pain point for the product category.

Notes:

- Pain points may be composed of morphemes or spaced units rather than conventional word structures, selected based on their meaning.
- There may be cases where a pain point cannot be extracted according to the logic. In such cases, you can give a score of 1.

D.3 Task 3: Topic-aware pain point evaluation

The following example contains three elements:

- Consumer reviews on dryers collected from online community
- Review examples for each topic class
- Key pain points for each topic class

20 reviews are provided for each topic class. Please assess the reviews in detail and evaluate whether the pain point can be used as a representative pain point for the topic class, using the following scores:

- 0: The pain point cannot represent the topic class.
- 1: The pain point can represent the topic class.

Notes:

- Pain points can be constructed at the morpheme level, rather than a common word structure, depending on their meaning.
- Do not consider overlaps in meaning or form between pain points. Please only judge the representativeness of each pain point for the reviews on the topic.
- If the topic modeling performance is low, the topics of each review may not match. If the meaning of the topic cannot be identified through the sampled sentences for each topic class, you can assign 0 points to all the pain points.
- Each topic often includes more than 500-1,000 sentences. Some pain points may be not contained in the sampled review. If similar words to that pain points, however, they could be suitable pain points for its cluster. ([ex] "I tried it myself and my wrist hurts", pain point ('arms': considered correct))