A Japanese Dataset for Aspect-based Sentiment Polarity Classification and Emotion Intensity Estimation

Kentaro Hanafusa[†] **Kota Manabe**[†] **Yuki Maeda**[†] **Daisuke Maekawa**[†]
Tomoyuki Kajiwara[†] **Hideaki Hayashi**[‡] **Yuta Nakashima**[‡] **Hajime Nagahara**[‡]
[†]Ehime University [‡]The University of Osaka
{hanafusa@ai., manabe@ai., maeda@ai., maekawa@ai., kajiwara@}cs.ehime-u.ac.jp
{hayashi, n-yuta, nagahara}@ids.osaka-u.ac.jp

Abstract

We manually construct and publicly release a Japanese dataset for Aspect-based Sentiment Analysis (ABSA), annotated with both sentiment polarity and the emotional intensities for Plutchik's eight emotions. Existing datasets for Japanese ABSA only handle sentiment polarity classification. Therefore, we manually annotated Plutchik's eight emotions with a four-point scale and sentiment polarity with a five-point scale to words in the Japanese sentiment analysis corpus WRIME. Analysis of this corpus revealed that word-level emotions more strongly reflect the reader's objective impression than the writer's subjective perspective. Furthermore, the results of evaluation experiments on word-level emotion estimation quantitatively demonstrated that while Large Language Models achieve high performance, they struggle with the estimation of the "trust" emotion. Additionally, we demonstrated that multi-task learning, utilizing both word and sentence levels, can improve performance on difficult-to-estimate subjective emotions.

Keywords: Aspect-based Sentiment Analysis, Dataset Construction, Japanese

1. Introduction

Aspect-Based Sentiment Analysis (ABSA) (Zhang et al., 2023) is being studied to estimate fine-grained sentiment/emotion for each term that cannot be captured by sentence-level sentiment analysis. When both positive and negative aspects are mixed within a single sentence, as shown in Figure 1 and Table 1, ABSA is crucial.

Sentence-level sentiment analysis datasets, such as SemEval-2007 Task-14 (Strapparava and Mihalcea, 2007) in English and WRIME (Kajiwara et al., 2021; Suzuki et al., 2022) in Japanese, are annotated with labels for both sentiment polarity classification and emotional intensity estimation. However, existing ABSA datasets (Pontiki et al., 2014; Dong et al., 2014; Saeidi et al., 2016; Jiang et al., 2019; Nakayama et al., 2022) focus solely on sentiment polarity classification, such as positive and negative, and do not cover diverse emotions, such as joy and sadness.

In this study, we release the first ABSA dataset[1] covering both sub-tasks of sentiment polarity classification and emotional intensity estimation. We provide additional annotation of aspect-based sentiment/emotion labels for WRIME (Kajiwara et al., 2021; Suzuki et al., 2022), a dataset for sentiment/emotion analysis on social media posts in Japanese. Our annotation scheme follows WRIME, employing a five-point sentiment polarity label ranging from -2 to $+2$ and a four-point emotional intensity label ranging from 0 to 3 for each of Plutchik's eight basic emotions (Plutchik, 1980). This enables our dataset to have consistent sentiment/emotion annotations between the word and sentence levels, thereby providing additional benefits such as improving the performance of sentence-level sentiment/emotion analysis by considering word-level information.

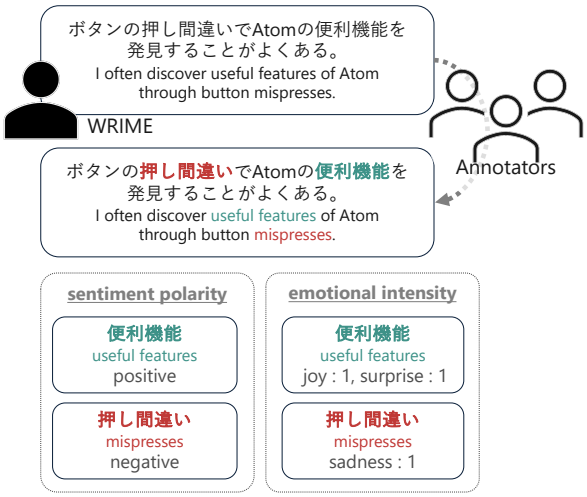

Figure 1: Overview of our annotation. We extracted aspect terms from posts in the WRIME dataset and annotated each aspect term with both sentiment polarity and emotional intensity labels. Our sentiment polarity labels range from -2 to $+2$ on a five-point scale, while emotional intensity labels employ a four-point scale from 0 to 3 for each of Plutchik's eight basic emotions.

[1] https://github.com/ids-cv/wrime

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8076–8084
11–16 May 2026.

仕事早く終わったけど**ひとりぼっち**！ (Got off **work** early, but I'm all **alone**!)									
	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Sentiment
仕事 (work)	1	0	0	1	0	0	1	0	1
ひとりぼっち (alone)	0	1	0	0	0	0	0	0	-1
Entire sentence	0	2	0	0	0	0	1	0	-1

Table 1: Example annotations. In this example, two aspect terms highlighted in bold have been annotated with contrasting labels. The labels on the entire sentence provided from the original WRIME dataset are negative, and cannot capture fine-grained sentiment/emotion without aspect-based annotation.

Experimental results for aspect term extraction and aspect-based sentiment/emotion analysis revealed that the encoder model performed better for the former task, while the decoder model performed better for the latter task.

2. Related Work

2.1. Subtasks of ABSA

The ABSA task consists of multiple subtasks: (1) aspect term extraction, (2) aspect category detection, (3) opinion term extraction, and (4) sentiment analysis. (1) Aspect term extraction is the task of identifying the targets of opinions that appear in the given text. (2) Aspect category detection is the task of classifying each aspect term into predefined categories. (3) Opinion term extraction is the task of extracting expressions used to describe the sentiment toward each aspect term. (4) Finally, sentiment polarity is estimated for each aspect term or category. Note that aspect categories are defined for each target domain. In open domain settings (Dong et al., 2014), defining aspect categories is challenging and therefore may not be considered.

2.2. ABSA Dataset

Table 2 lists the existing ABSA datasets. The datasets provided in the series of SemEval shared tasks (Pontiki et al., 2014, 2015, 2016) are widely used as standard benchmarks for the ABSA task. These datasets consist of review texts from two domains: laptops and restaurants. Subsequently, ABSA datasets for other domains such as Twitter (Dong et al., 2014) and financial reports[2] have also been proposed.

These existing ABSA datasets focus solely on sentiment polarity classification, such as positive and negative, and do not cover emotional intensity estimation, such as the basic emotions defined by Ekman (Ekman, 1992) and Pultchik (Plutchik, 1980), which are studied in sentence-level sentiment/emotion analysis. While dimABSA (Xu et al., 2024) evaluates aspects beyond sentiment polarity based on the valence-arousal space (Russell, 1980), it does not cover diverse types of emotions.

[2] https://github.com/chakki-works/chABSA-dataset/

3. Dataset Construction

In this study, four Japanese native-speaking university students annotated each text in the WRIME corpus, which consists of SNS posts.

3.1. Aspect Term Identification

In this study, we first identified the aspect terms within the sentences that are associated with emotions. The specific extraction guidelines are as follows. In addition to the guidelines above, expressions such as maxims or common proverbs that do not refer to a specific target were excluded from extraction.

1. **Extract nouns and demonstrative pronouns associated with emotions**
   - Sentence: 新しいPC買ったんだけど、これマジでサクサクで最高！ (I bought a new PC, and this is seriously fast and amazing!)
   - Identified aspect terms: PC, これ (this)
2. **Extract the formal noun (fact/act of):** This targets cases where the formal noun follows the attributive form of a verb or adjective, thereby nominalizing the action or state to serve as the target of the emotion.
   - Sentence: 昨日ライブに行けたことがもう最高. (The fact that I could go to the concert yesterday is just the best.)
   - Identified aspect term: こと (fact)
3. **Extract nominalized adjectives:**
   - Sentence: この映画の悲しさは異常. (The sadness of this movie is abnormal.)
   - Identified aspect term: 悲しさ (sadness)

		Annotations					Domains					
	Lang	Term	Category	Opinion	Sentiment	Emotion	R	L	H	E	F	S
SemEval-2014 (Pontiki et al., 2014)	En	✓	✓	-	✓	-	✓	✓	-	-	-	-
Twitter (Dong et al., 2014)	En	✓	-	-	✓	-	-	-	-	-	-	✓
SemEval-2015 (Pontiki et al., 2015)	En	✓	✓	-	✓	-	✓	✓	-	-	-	-
SemEval-2016 (Pontiki et al., 2016)	Multi	✓	✓	-	✓	-	✓	-	✓	✓	-	-
SentiHood (Saeidi et al., 2016)	En	✓	-	-	✓	-	-	-	-	-	-	-
ASC-QA (Wang et al., 2019)	Zh	✓	✓	-	✓	-	-	-	-	✓	-	-
MAMS (Jiang et al., 2019)	En	✓	✓	-	✓	-	✓	-	-	-	-	-
ARTS (Xing et al., 2020)	En	✓	-	-	✓	-	✓	✓	-	-	-	-
ASTE-Data-V2 (Xu et al., 2020)	En	✓	-	✓	✓	-	✓	✓	-	-	-	-
ASAP (Bu et al., 2021)	Zh	-	✓	-	✓	-	✓	-	-	-	-	-
ACOS (Cai et al., 2021)	En	✓	✓	✓	✓	-	✓	✓	-	-	-	-
ABSA-QUAD (Zhang et al., 2021)	En	✓	✓	✓	✓	-	✓	-	-	-	-	-
dimABSA (Xu et al., 2024)	Zh	✓	✓	✓	✓	V/A	✓	-	-	-	-	-
chABSA-dataset[3]	Ja	✓	-	-	✓	-	-	-	-	-	✓	-
Rakuten Travel (Nakayama et al., 2022)	Ja	-	✓	-	✓	-	-	-	✓	-	-	-
Ours	**Ja**	✓	-	-	✓	P8	-	-	-	-	-	✓

Table 2: Dataset comparison. In the Emotion column, V/A represents the valence-arousal space (Russell, 1980), and P8 represents Plutchik's basic eight emotions (Plutchik, 1980). Domain abbreviations: R (Restaurant), L (Laptop), H (Hotel), E (Electronics), F (Financial), S (Social Networking Service).

Pair	Jaccard	QWK	
	Words	Emotion	Sentiment
A-B	0.650	0.634	0.789
A-C	0.655	0.699	0.862
A-D	0.651	0.615	0.764
B-C	0.659	0.611	0.835
B-D	0.605	0.601	0.777
C-D	0.663	0.682	0.779
Avg.	0.647	0.640	0.801

Table 3: Inter-annotator agreement rates. QWK denotes Quadric Weighted Kappa.

	-2	-1	0	+1	+2
Size	464	5,910	4,277	5,096	329

Table 4: Distribution of sentiment polarity labels

3.2. Emotion Label Annotation

For the aspect terms extracted in Section 3, four annotators assigned (strong negative, negative, neutral, positive, and strong positive) and four-point scale intensity labels (none, weak, medium, and strong) for each of Plutchik's eight emotions.

To maintain annotation consistency, we established the following guidelines. For example, if a sentence frequently used interjections or contained exclamation marks (e.g., "!"), annotators were instructed to judge that the emotion was strongly expressed and assign an intensity of "Medium (2)" or higher for the corresponding emotion label. The specific guidelines are as follows.

1. **Assign emotions based on context:** The emotion associated with an aspect term is determined by its surrounding context.

   - Sentence: 新しいPC、サクサクで良い感じ. (My new PC is fast and feels good.)
   - Guideline: For the aspect term "PC," assign labels such as "Polarity: +1, Joy: 1, Trust: 1" based on the positive evaluation "feels good".

2. **Increase intensity based on emphasizing adverbs:** If adverbs that intensify emotion (e.g., "very", "extremely", "seriously/really" are used, add +1 to the base emotion intensity.

   - Sentence: このケーキ, マジで美味しい！ (This cake is seriously delicious!)
   - Guideline: If the base intensity for "delicious" is "Joy: 1," add +1 due to the emphasizing adverb "seriously", resulting in a higher intensity assignment like "Polarity: +2, Joy: 2."

3. **Assign higher intensity for interjections or exclamation marks:** If interjections or multiple exclamation marks ("!") are used in the sentence, judge that the emotion is strongly expressed and assign an intensity of "Medium (2)" or higher.

   - Sentence: え、ヤバい！ この景色、最高すぎる...！！ (Eh, whoa! This view is too amazing...!!)
   - Guideline: Due to the use of interjections ("Eh, whoa!") and multiple exclamation marks ("!"), assign high values to the aspect term "view", such as "Polarity: +2, Joy: 3, Surprise: 2."

	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust
0	11,707	11,762	10,652	13,637	15,549	14,152	13,704	13,985
1	1,954	2,494	2,748	1,333	205	1,184	1,480	1,104
2	2,122	1,702	2,590	1,042	246	697	758	891
3	293	118	86	64	76	43	134	96

Table 5: Distribution of emotion intensity labels

4. **Assign higher intensity based on the severity of real-world events:** For events that generally evoke strong negative emotions, such as natural disasters or accidents, assign an intensity of "Medium (2)" or higher.

   - Sentence: 地震こわい．震度 5 はさすがに焦る．(The earthquake is scary. A seismic intensity of 5 is really alarming.)
   - Guideline: Since the aspect term "earthquake" evokes strong negative emotions for most people, assign high values such as "Polarity: -2, Fear: 2."

3.3. Annotation Tool

We employed a custom web-based annotation tool (Figure 2). The target text is displayed at the top of the screen as character-level buttons, allowing annotators to select an aspect span by clicking its start and end positions. This button-based selection prevents input errors and improves efficiency.

After selecting a span and its polarity, pressing the "add" button records the entry in the "Annotation result" section. If either the span or polarity is missing, the system displays a warning instead of adding the entry. To reduce cognitive load, nouns are automatically highlighted in green using the morphological analyzer MeCab (Kudo et al., 2004) with the MeCab-ipadic-NEologd[3] dictionary. These interface designs help ensure consistent and high-quality annotations.

3.4. Evaluation of Annotations

To evaluate the reliability of the annotations, we calculated the inter-annotator agreement on 50 data samples. For the task of identifying words, we used the Jaccard index, shown in Equation (1), for all annotator pairs ($p_1, p_2 \in P$). For the emotion label assignment task, we employed the Quadratic Weighted Kappa (QWK) (Cohen, 1968), which is suitable for ordinal scales. Table 3 shows the agreement scores for each task per annotator pair.

The agreement for the word identification task achieved high concordance, with $J > 0.6$ for all pairs. Furthermore, for the QWK on emotion label

[3] https://github.com/neologd/mecab-ipadic-neologd

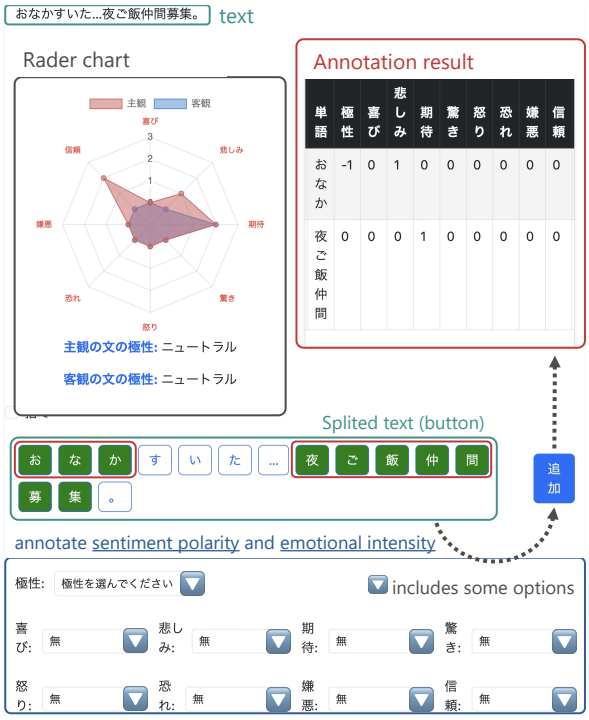

Figure 2: Screen of the developed annotation tool.

assignment, substantial agreement ($\kappa > 0.6$) was confirmed for both the eight emotions (Emotion) and the sentiment polarity (Sentiment). In particular, a very high agreement of $\kappa > 0.76$ was obtained for polarity, demonstrating that the annotations in this dataset possess high reliability.

$$\frac{1}{|P|} \sum_{p_1, p_2 \in P} \frac{|p_1 \cap p_2|}{|p_1 \cup p_2|} \quad (1)$$

4. Analysis

4.1. Distribution of Emotion Labels

Table 4 presents the distribution of the five-point sentiment polarity labels, while Table 5 presents the distribution of the intensity labels for Plutchik's eight emotions. Regarding sentiment polarity, "-1 (Negative)" and "+1 (Positive)" are the most frequent, followed by "0 (Neutral)." Conversely, the extreme labels "-2 (Strong negative)" and "+2 (Strong positive)" are relatively scarce, indicating a distribution

	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Polarity
Writer	0.671	0.590	0.476	0.502	0.511	0.410	0.532	**0.285**	0.627
Reader	**0.736**	**0.642**	**0.632**	**0.646**	**0.672**	**0.570**	**0.622**	0.253	**0.765**

Table 6: Correlation between word sentiment labels and sentence-level sentiment labels

	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Polarity
BERT	0.663	0.515	0.539	0.226	0.036	0.346	0.197	0.058	0.683
ModernBERT	0.704	0.594	0.604	0.416	0.240	**0.376**	0.276	0.269	0.730
Swallow	**0.741**	**0.678**	**0.652**	**0.511**	**0.483**	0.352	**0.536**	**0.437**	**0.775**

Table 7: Results for sentiment intensity estimation and sentiment polarity classification (QWK)

trend similar to that of previous research (Suzuki et al., 2022)

Regarding emotion intensity, ”0 (None)” is the most frequent category for all emotions, and the counts decrease as the intensity level increases. Analyzed by emotion, the labels ”Anticipation”, ”Joy”, and ”Sadness” are frequently assigned, whereas ”Anger” tends to be assigned infrequently. This also indicates a distribution trend similar to that of prior work (Kajiwara et al., 2021).

4.2. Relationship Between Word and Sentence Emotion Labels

To investigate the relationship between the word-level emotion labels of this dataset and the sentence-level emotion labels of the original WRIME corpus, we calculated Pearson's correlation coefficient. Specifically, we computed the average sentiment polarity and the average intensity of each emotion for the words contained within each sentence. We then calculated the correlation of these averages with the subjective and objective labels assigned to the entire sentence. The results are shown in Table 6.

Analysis revealed an overall positive correlation between word-level sentiment averages and sentence-level sentiment for many emotions. This indicates that sentiment labels assigned to individual words align with the overall sentiment of the sentence. Furthermore, the correlation with objective labels exceeded that with subjective labels for nearly all items. This suggests that emotions expressed by individual words may reflect the impressions of third-party readers more strongly than the writer' s internal emotions, consistent with prior studies showing that estimating a writer' s emotions is more difficult than estimating a reader' s emotions (Kajiwara et al., 2021; Suzuki et al., 2022). By emotion, “trust” showed particularly low correlations in both subjective and objective assessments, likely because it is formed indirectly from the overall sentence context and multiple factual relationships.

4.3. Analysis of Emotion Label Overlap

Focusing on the degree of mixed sentiment polarity among annotated words within sentences, the dataset was divided into the following two groups:

- **Sentences with single sentiment polarity**: Sentences where words within the sentence exhibit only positive or negative sentiment.
- **Sentences with mixed sentiment polarity**: Sentences where words within the sentence exhibit both positive and negative sentiment.

For each group, we calculated the agreement between subjective and objective sentence-level polarity labels using QWK. Sentences with single sentiment showed a high agreement of 0.668, whereas sentences with mixed sentiment showed a significantly lower agreement of 0.511. This result indicates that the coexistence of positive and negative emotions in a sentence is a key factor causing a mismatch between the writer and the reader.

5. Experiments

5.1. Experimental Settings

Dataset We conducted evaluation experiments using the dataset described in Section 3. The dataset was split into training, validation, and evaluation sets at an 8:1:1 ratio, corresponding to 5,600, 700, and 700 instances, respectively.

Tasks We conducted evaluations on the following two tasks:

1. **Aspect Term Sentiment Analysis (ATSA)**: A task of estimating five-point sentiment polarity and four-point emotion intensity for each of eight emotions for a given aspect term.
2. **Sentiment Analysis of Sentences Using Word Information (SA)**: A task to verify the effectiveness of the word-level emotion information, which was annotated in this study, for sentence-level sentiment classification.

Subjective labels	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Polarity
BERT	0.453	0.265	0.156	**0.284**	**0.202**	**0.200**	0.148	0.309	0.387
BERT w/ word	**0.532**	**0.321**	**0.321**	0.212	0.149	0.127	**0.183**	**0.322**	**0.445**
ModernBERT	0.651	0.411	0.384	0.344	**0.297**	0.244	0.300	0.257	**0.586**
ModernBERT w/ word	**0.688**	**0.427**	**0.408**	**0.388**	0.295	**0.246**	**0.347**	**0.284**	0.532
Swallow	**0.648**	0.374	**0.469**	0.401	0.230	0.096	**0.447**	**0.223**	0.543
Swallow w/ word	0.632	**0.502**	0.461	**0.411**	**0.259**	**0.277**	0.415	0.202	**0.608**
Objective labels	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Polarity
BERT	0.554	**0.341**	**0.570**	0.356	0.113	**0.435**	**0.297**	**0.133**	0.620
BERT w/ word	**0.604**	0.326	0.428	**0.372**	**0.120**	0.263	0.177	0.067	**0.645**
ModernBERT	0.708	0.323	**0.755**	**0.636**	**0.475**	0.491	0.469	**0.248**	0.764
ModernBERT w/ word	**0.769**	**0.551**	0.699	0.503	0.474	**0.615**	**0.541**	0.239	**0.784**
Swallow	0.676	0.564	**0.701**	**0.621**	0.400	**0.615**	**0.615**	0.066	0.758
Swallow w/ word	**0.737**	**0.582**	0.666	0.514	**0.502**	0.468	0.598	**0.232**	**0.777**

Table 8: Results for sentence-level sentiment classification tasks (QWK). "w/ word" indicates the multi-task learning model that additionally utilizes word-level sentiment information.

Models We used BERT (Devlin et al., 2019) and ModernBERT (Warner et al., 2025) as Encoder models, and Swallow (Fujii et al., 2024; Okazaki et al., 2024) as a Large Language Model (LLM). For the Encoder models, we utilized the pre-trained BERT[4] and ModernBERT[5]. For the LLM, we employed Swallow (Fujii et al., 2024)[6], a Japanese LLM that was continually pre-trained on Japanese data (Okazaki et al., 2024) from an English LLM[7] (Llama Team, 2024).

Implementation of ATSA For the Encoder models, we implemented the task as a classification task. The emotion label is estimated as $y = \text{softmax}(hW)$, where h is the feature vector obtained from the [CLS] token for BERT, or the '<s>' token for ModernBERT. Following (Song et al., 2019), the input sequence for the Encoder models was formatted as: [CLS] + sentence + [SEP] + aspect term + [SEP]. For the LLM, which lacks special tokens such as [SEP], we formulated the task as text generation in response to a prompt, based on Instruction-tuning (Wei et al., 2022).

Implementation of SA For the sentence-level emotion classification task, we compared two models to validate the effectiveness of word-level emotion information.

- **Baseline Model**: This model takes only the sentence as input and is trained solely on the sentence-level emotion label (either subjective or objective). For the Encoder models (BERT, ModernBERT), we obtain the feature vector h from the embedding of the [CLS] or <s> token, respectively. For the LLM, h is derived using average pooling of the hidden state vectors from the final layer. Classification is then performed as $y = \text{softmax}(hW)$.
- **Multi-task Learning Model (w/ word)**: This model learns both sentence-level and word-level emotions simultaneously. In addition to the sentence-level classification layer (identical to the baseline), this model includes an aspect term (word) level classification layer (identical to the ATSA task). During training, we compute the sentence-level classification loss $L_{sentence}$ and the word-level classification loss L_{word} for each aspect term contained in the sentence. The entire model is optimized by minimizing the total loss: $L_{total} = L_{sentence} + \sum L_{word}$.

Hyperparameters For fine-tuning all models, we set the learning rate to $2e-5$, used Adam (Kingma and Ba, 2015) as the optimizer, and applied early-stopping with a patience of 3 epochs. The batch size was set to 16 for the Encoder models and 8 for the LLM. For fine-tuning the LLM, we employed LoRA (Low-Rank Adaptation) (Hu et al., 2022) with a rank (r) of 16, a scaling factor (α) of 16, and a dropout rate of 0.05.

Metrics We adopted QWK, a metric commonly employed for this type of task, because the task

[4] https://huggingface.co/tohoku-nlp/bert-large-japanese-v2

[5] https://huggingface.co/sbintuitions/modernbert-ja-310m

[6] https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5

[7] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

Text	喉が痛い, , どうしたら口開けて寝なくなるの？？ My throat hurts, , How can I stop sleeping with my mouth open??								
Annotation	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Polarity
喉 (throat)	0	0	0	0	0	0	1	0	-1
口 (mouth)	0	0	1	0	0	1	0	0	-1
Subjective labels	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Polarity
Writer	0	1	0	0	0	**1**	1	0	-1
Swallow	0	1	0	0	0	**0**	1	0	-1
Swallow w/ word	0	1	0	0	0	**1**	1	0	-1

Table 9: Case study demonstrating improved subjective emotion estimation through multi-task learning. “w/ word” indicates the multi-task learning model that utilizes word-level emotional information.

addressed in this study is an ordinal classification problem.

5.2. Results

5.2.1. Aspect Term Sentiment Analysis

Table 7 presents the results for the sentiment polarity classification and the eight-emotion intensity estimation tasks. As an overall trend, Swallow outperformed ModernBERT, which in turn outperformed the baseline BERT. This suggests that model scale and the diversity of pre-training data are crucial for capturing subtle emotions. In particular, Swallow, which was continually pre-trained on a Japanese corpus, achieved the best performance for most emotions.

Analyzing by emotion, relatively high QWK scores were obtained for ”Joy,” ”Sadness,” and the polarity classification. These emotions are presumably conveyed using explicit language in the text, making them easier for the models to learn. On the other hand, emotions such as ”Anger,” ”Fear,” and ”Trust” yielded low scores, particularly for the BERT model. This suggests that these emotions are more context-dependent and are often implied through indirect expressions, demanding advanced contextual comprehension abilities. Swallow demonstrated improved performance even for these complex emotions, which confirms the efficacy of LLM.

5.2.2. Sentiment Analysis of Sentences Using Word Information

We investigated whether word-level emotion labels improve sentence-level emotion classification. We compared a baseline model using only sentences with a model (w/ word) that also uses annotated word-level emotion labels. Experiments on subjective and objective sentence emotions (Table 8) show that word-level information affects these two types of emotions differently.

For objective emotion estimation (bottom half of Table 8), adding word-level information consistently improved polarity classification performance across all models. This indicates that the objective impression perceived by a reader is often directly reflected in the emotions of the words within the sentence; thus, the word-level labels functioned as effective auxiliary information.

Conversely, the estimation of subjective emotion (top half of Table 8) is an inherently difficult task, as also noted by (Suzuki et al., 2022), because it often involves the writer’s own irony or complex emotions. In fact, we observed cases where adding word-level information led to performance degradation. However, we also confirmed instances where this information contributed to performance improvement, such as the polarity classification for the Swallow model (0.543 $\rightarrow$ 0.608).

Table 9 presents a case study where multi-task learning improved subjective emotion estimation. In this example, the baseline Swallow model correctly estimated ”Disgust” and ”Sadness” but failed to detect ”Fear”. In contrast, the multi-task Swallow model leveraged the ”Fear” label annotated to the word ”mouth” (口) as auxiliary information and correctly estimated ”Fear” at the sentence level. This result demonstrates that detailed word-level emotion information helps estimate the writer’s latent emotions, which are difficult to capture from the sentence alone.

6. Conclusion

We released[1] a Japanese ABSA dataset extending WRIME with word-level sentiment and Plutchik emotion labels. Word emotions correlate more with objective reader impressions than subjective writer emotions, and using them as auxiliary input in multi-task learning may improve subjective emotion estimation and reduce the subjective-objective gap.

Acknowledgments

This work was supported by JST BOOST Program Japan Grant Number JPMJBY24036821.

Bibliographical References

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.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-Category-Opinion-Sentiment Quadruple Extraction with Implicit Aspects and Opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350.

Jacob Cohen. 1968. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin*, 70(4):213–220.

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*, pages 4171–4186.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54.

Paul Ekman. 1992. An Argument for Basic Emotions. *Cognition and Emotion*, 6(3–4):169–200.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In *Proceedings of the First Conference on Language Modeling*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the Tenth International Conference on Learning Representations*.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A Challenge Dataset and Effective Models for Aspect-Based Sentiment Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6280–6285.

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2021. WRIME: A New Dataset for Emotional Intensity Estimation with Subjective and Objective Annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Llama Team. 2024. The Llama 3 Herd of Models. *arXiv:2407.21783*.

Yuki Nakayama, Koji Murakami, Gautam Kumar, Sudha Bhingardive, and Ikuko Hardaway. 2022. A Large-Scale Japanese Dataset for Aspect-based Sentiment Analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7014–7021.

Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. 2024. Building a Large Japanese Web Corpus for Large Language Models. In *Proceedings of the First Conference on Language Modeling*.

Robert Plutchik. 1980. A General Psychoevolutionary Theory of Emotion. *Theories of Emotion*, pages 3–33.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier

Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35.

James A. Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. 2016. SentiHood: Targeted Aspect Based Sentiment Analysis Dataset for Urban Neighbourhoods. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1546–1556.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional Encoder Network for Targeted Sentiment Classification. *CoRR*, abs/1902.09314.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective Text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, pages 70–74.

Haruya Suzuki, Yuto Miyauchi, Kazuki Akiyama, Tomoyuki Kajiwara, Takashi Ninomiya, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2022. A Japanese Dataset for Subjective and Objective Sentiment Polarity Classification in Micro Blog Domain. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7022–7028.

Jingjing Wang, Changlong Sun, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2019. Aspect Sentiment Classification Towards Question-Answering with Reinforced Bidirectional Attention Network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3548–3557.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned Language Models Are Zero-Shot Learners. In *Proceedings of the Tenth International Conference on Learning Representations*.

Xiaoyu Xing, Zhijing Jin, Di Jin, Bingning Wang, Qi Zhang, and Xuanjing Huang. 2020. Tasty Burgers, Soggy Fries: Probing Aspect Robustness in Aspect-Based Sentiment Analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3594–3605.

Hongling Xu, Delong Zhang, Yice Zhang, and Ruifeng Xu. 2024. HITSZ-HLT at SIGHAN-2024 dimABSA Task: Integrating BERT and LLM for Chinese Dimensional Aspect-Based Sentiment Analysis. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 175–185.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-Aware Tagging for Aspect Sentiment Triplet Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. Aspect Sentiment Quad Prediction as Paraphrase Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023. A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35:11019–11038.