第7回とめ研究所若手研究者懸賞論文

# アスペクトベースの感情極性分類と感情強度推定のための日本語感情分析データセット

† 愛媛大学 大学院理工学研究科
‡ 大阪大学 D3センター

花房 健太郎†，眞鍋 光汰†，前田 雄生†，前川 大輔†，
梶原 智之†‡，早志 英朗‡，中島 悠太‡，長原 一‡

応募日：2026年9月4日

# 概要

アスペクトベース感情分析 (Aspect-Based Sentiment Analysis: ABSA) は、文や文書を単位とする感情分析では捉えられない語句ごとの詳細な感情を推定するタスクである。通常の感情分析は、所与のテキスト全体に対して感情ラベルを推定するが、それではテキスト中の異なる評価対象に対する感情を区別できないため、1 つのテキストの中に肯定的な側面と否定的な側面の両方が混在する場合に ABSA の技術が有用である。

文単位の感情分析データセットには、感情極性分類および感情強度推定のための 2 種類の感情ラベルが付与されている。しかし、既存の ABSA データセットは、ポジティブまたはネガティブの感情極性分類のみを対象としており、喜びや悲しみといった多様な感情は扱われていない。さらに、アスペクト単位の感情とテキスト全体の感情との関係や、主観的な感情と客観的な感情との関係を分析することも難しく、アスペクトベース感情分析はまだ十分に研究されているとは言えない状況である。

本研究では、感情極性分類および感情強度推定の両サブタスクを対象とするアスペクトベース感情分析データセットを初めて公開する。具体的には、日本語 SNS 投稿を対象とする感情分析データセットである WRIME に対して、単語単位の感情極性・感情強度ラベルを新たに付与した。

アスペクト抽出およびアスペクトベース感情極性分類・感情強度推定に関する実験の結果、前者のタスクではエンコーダモデルが高い性能を示し、後者のタスクではデコーダモデルが高い性能を示すことが明らかとなった。また、大規模言語モデルは高い性能を示す一方で、「信頼」の感情の推定には苦戦することが定量的に示された。加えて、単語レベルと文レベルの両方を活用するマルチタスク学習によって、推定が難しい主観的な感情推定に対する性能を向上させられることを実証した。

# 第 1 章　はじめに

アスペクトベース感情分析 (Aspect-Based Sentiment Analysis: ABSA) (Zhang et al. 2023) は，文や文書を単位とする感情分析では捉えられない語句ごとの詳細な感情を推定するタスクである．通常の感情分析は，所与のテキスト全体に対して感情ラベルを推定するが，それではテキスト中の異なる評価対象に対する感情を区別できないため，図 1 および表 1 に示すような 1 つのテキストの中に肯定的な側面と否定的な側面の両方が混在する場合に ABSA の技術が有用である．

英語の SemEval-2007 Task 14 (Strapparava and Mihalcea 2007) や日本語の WRIME (Kajiwara et al. 2021; Suzuki et al. 2022) などの文単位の感情分析データセットには，感情極性分類および感情強度推定のための 2 種類の感情ラベルが付与されている．しかし，既存の ABSA データセット (Pontiki et al. 2014; Dong et al. 2014; Saeidi et al. 2016; Jiang et al. 2019; Nakayama et al. 2022) は，ポジティブまたはネガティブの感情極性分類のみを対象としており，喜びや悲しみといった多様な感情は扱われていない．さらに，アスペクト単位の感情とテキスト全体の感情との関係や，主観的な感情と客観的な感情との関係を分析することも難しく，アスペクトベース感情分析はまだ十分に研究されているとは言えない状況である．

本研究では，感情極性分類および感情強度推定の両サブタスクを対象とするアスペクトベース感情分析データセット[1]を初めて公開する．具体的には，日本語 SNS 投稿を対象とする感情分析データセットである WRIME (Kajiwara et al. 2021; Suzuki et al. 2022) に対して，単語単位の感情極性・感情強度ラベルを新たに付与した．アノテーション方針は WRIME に準拠しており，感情極性については-2 から+2 の 5 段階ラベルを，感情強度については Plutchik の基本 8 感情 (Plutchik 1980) のそれぞれに対して 0 から 3 の 4 段階ラベルを付与している．これにより，本データセットでは単語レベルと文レベルのアノテーションの整合性が保たれており，単語レベルの情報を活用することで文レベ

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

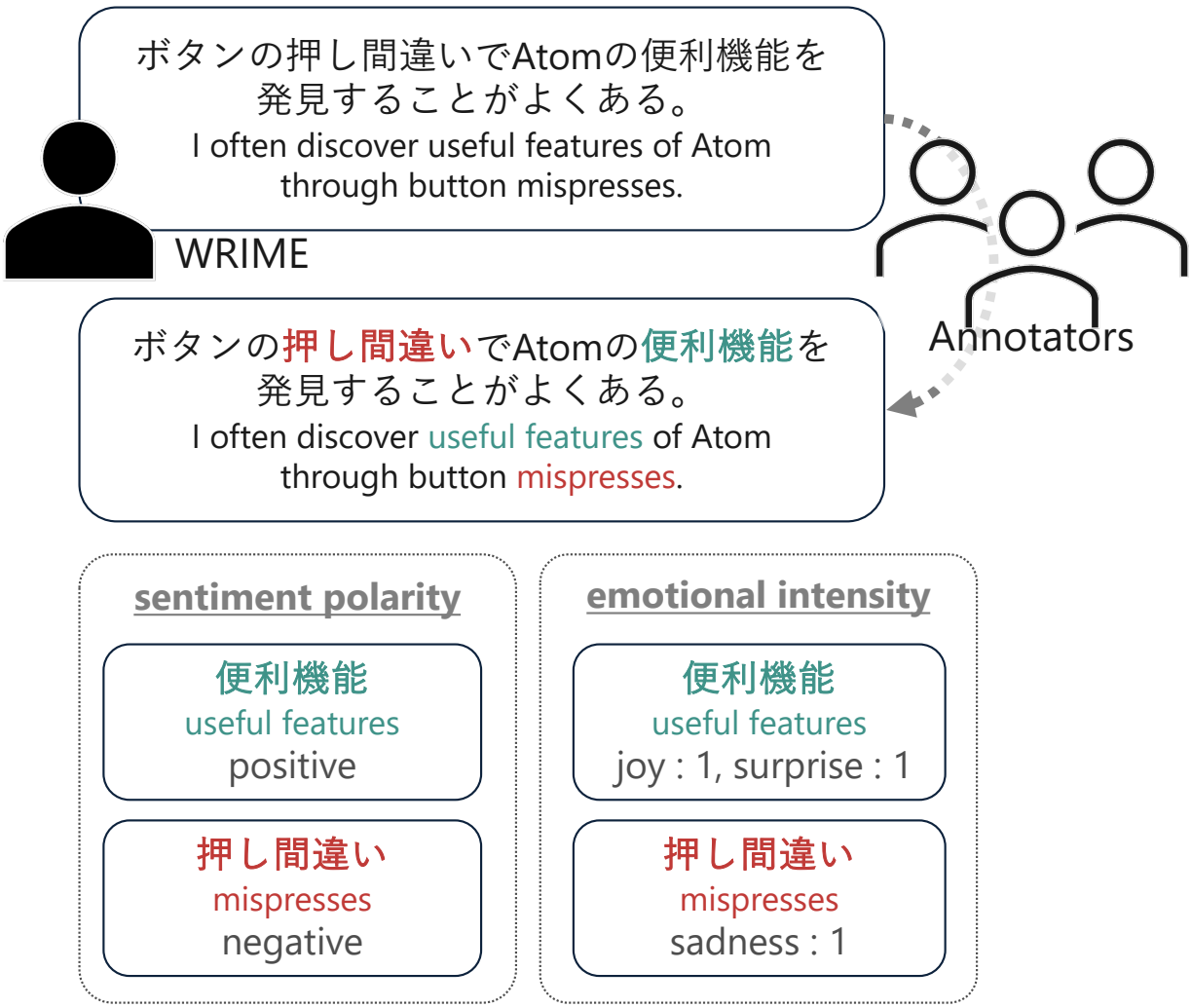


**図 1**　本例では，太字で示した 2 つのアスペクト項に対して，対照的なラベルが付与されている. 一方，元の WRIME データセットにおいて文全体に付与されたラベルはネガティブであり，アスペクト単位のアノテーションなしでは，このような細かい感情を捉えることはできない.

ルの感情分析性能の向上といった利点も期待できる.

本コーパスの分析結果から，アスペクト単位の感情は書き手の主観感情よりも読み手の客観感情との相関が高く，単語レベルの感情表現が客観的な印象を強く反映していることが明らかとなった. また，複数のアスペクト表現を含む文を分析した結果，アスペクト表現の位置によって，文全体の感情解釈に大きな影響を与えることが確認され，

**表 1**　アノテーション例．本例では，太字で示した 2 つのアスペクト表現に対して，対照的なラベルが付与されている．一方，元の WRIME データセットにおいて文全体に付与されたラベルはネガティブであり，アスペクト単位のアノテーションなしでは，このような細粒度な感情・情緒を捉えることはできない．

| | **仕事**早く終わったけど**ひとりぼっち**！ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 喜び | 悲しみ | 期待 | 驚き | 怒り | 恐れ | 嫌悪 | 信頼 | 極性 |
| **仕事** | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| **ひとりぼっち** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | -1 |
| 文ラベル | 0 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | -1 |

アスペクト表現の配置や極性が主観感情と客観感情の乖離に関係している可能性が示唆された．

アスペクト抽出およびアスペクトベース感情極性分類・感情強度推定に関する実験の結果，前者のタスクではエンコーダモデルが高い性能を示し，後者のタスクではデコーダモデルが高い性能を示すことが明らかとなった．また，大規模言語モデル (Large Language Model: LLM) は高い性能を示す一方で，「信頼」の感情の推定には苦戦することが定量的に示された．加えて，単語レベルと文レベルの両方を活用するマルチタスク学習によって，推定が難しい主観的な感情に対する性能を向上させられることを実証した．

# 第 2 章　関連研究

## 2.1　ABSA のサブタスク

ABSA タスクは複数のサブタスクから構成される:(1) アスペクト表現抽出，(2) アスペクトカテゴリ検出，(3) オピニオン表現抽出，(4) 感情極性分類である.(1) アスペクト表現抽出は，入力テキスト中に出現する意見の対象となる表現を抽出するタスクである.(2) アスペクトカテゴリ検出は，各アスペクト表現をあらかじめ定義されたカテゴリに分類するタスクである.(3) オピニオン表現抽出は，各アスペクト表現に対する感情を表現する語句を抽出するタスクである.(4) 最後に，感情極性分類はアスペクト表現またはカテゴリごとに感情極性を推定するタスクである. なお，アスペクトカテゴリは対象ドメインごとに定義される. そのため，オープンドメイン設定 (Dong et al. 2014) においてはカテゴリの定義自体が困難であり，これを扱わない場合もある.

## 2.2　ABSA のデータセット

表 2 に既存の ABSA データセットを示す.SemEval 共有タスクシリーズ (Pontiki et al. 2014, 2015, 2016) で提供されたデータセットは，ABSA タスクの標準的なベンチマークとして広く利用されている. これらのデータセットは，laptop および restaurant の 2 つのドメインにおけるレビュー文から構成されている. その後，Twitter (Dong et al. 2014) や財務報告書[2]など，他ドメインを対象とした ABSA データセットも提案されている. また，日本語を対象とした ABSA データセットとしては，有価証券報告書を対象とした chABSA-dataset[2]や，宿泊施設のレビューを対象とした Rakuten Travel (Nakayama et al. 2022) などが存在する.

しかし，これらの対象ドメインは金融やホテルなどに限定されており，日本語において多様なドメインを扱うデータセットが不足しているという課題がある. さらに，これ

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

**表 2**　データセット比較．強度の列において，V/A は感情価–覚醒度空間 (Valence-Arousal Space) を，P8 は Plutchik の基本 8 感情を表す．また，ドメインの略称は，R：Restaurant（レストラン），L：Laptop（ノート PC），H：Hotel（ホテル），E：Electronics（電子機器），F：Financial（金融），S：Social Networking Service（SNS）を表す．

| | | アノテーション | | | | | ドメイン | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 言語 | アスペクト表現 | アスペクトカテゴリ | オピニオン表現 | 極性 | 強度 | 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[2] | Ja | ✓ | - | - | ✓ | - | - | - | - | - | ✓ | - |
| Rakuten Travel (Nakayama et al. 2022) | Ja | - | ✓ | - | ✓ | - | - | - | ✓ | - | - | - |
| **Ours** | **Ja** | ✓ | - | - | ✓ | P8 | - | - | - | - | - | ✓ |

らの既存 ABSA データセットは，ポジティブ・ネガティブといった感情極性分類のみを対象としており，文単位の感情分析で扱われる Ekman (Ekman 1992) や Plutchik の基本 8 感情 (Plutchik 1980) に基づく感情強度推定には対応していない. また，dimABSA (Xu et al. 2024) は感情極性にとどまらず，感情価 (Valence) および覚醒度 (Arousal) からなる感情空間 (Russell 1980) に基づいてアスペクトを評価するが，多様な感情の種類までは扱っていない.

本研究では，既存データセットの課題を踏まえ，新たに日本語の SNS ドメインを対象として，従来のポジティブ・ネガティブといった単純な感情極性分類にとどまらず，「喜び」や「悲しみ」などの詳細な感情を単語単位で扱えるデータセットを構築する.

# 第 3 章　データセットの構築

本研究では, 日本語母語話者である 4 名の大学生が, SNS 投稿から構成される WRIME コーパス (Kajiwara et al. 2021; Suzuki et al. 2022) の各テキストに対してアノテーションを実施した.

## 3.1　アスペクト表現を特定

本研究ではまず，文中に含まれる感情と関連するアスペクト表現を特定した. 具体的な抽出ガイドラインを以下に示す. また，上記のガイドラインに加え，格言やことわざなど，特定の対象を指さない表現については抽出対象から除外した.

(1) **感情の対象となる名詞および指示代名詞を抽出する**

文：新しい PC 買ったんだけど、これマジでサクサクで最高！

抽出したアスペクト項：PC、これ

(2) **形式名詞「こと」を抽出する**

動詞または形容詞の連体形に後続する形式名詞が、その動作や状態を名詞化し、感情の対象となっている場合を対象とする。

文：昨日ライブに行けたことがもう最高。

抽出したアスペクト項：こと

(3) **名詞化された形容詞を抽出する**

文：この映画の悲しさは異常。

抽出したアスペクト項：悲しさ

## 3.2　感情ラベルのアノテーション

(1) **文脈に基づいて感情を付与する**

アスペクト項に対する感情は，その周辺文脈に基づいて判断する.

文：新しい PC、サクサクで良い感じ。

感情ラベル：「良い感じ」という肯定的な評価に基づき，アスペクト項「PC」に対して「極性：+1，喜び：1，信頼：1」のようにラベルを付与する．

(2) **強調表現に応じて感情強度を高くする**

「とても」「非常に」「マジで」など，感情を強調する副詞が用いられている場合は，基本となる感情強度に対して 1 段階高い値を付与する．

文：このケーキ，マジで美味しい！

感情ラベル：「美味しい」に対する基本的な感情強度が「喜び：1」である場合，「マジで」という強調表現を考慮して 1 段階引き上げ，「極性：+2，喜び：2」などのラベルを付与する．

(3) **感動詞や感嘆符が用いられている場合は高い感情強度を付与する**

文中に感動詞や複数の感嘆符 (!) が用いられている場合は，感情が強く表現されていると判断し，対応する感情ラベルに対して「中 (2)」以上の強度を付与する．

文：え、やばい！この景色、最高すぎる...!!

感情ラベル：感動詞（え、やばい）および複数の感嘆符（!）が用いられるため，アスペクト項「景色」に対して「極性：+2，喜び：3，驚き：2」などの高い値を付与する．

(4) **現実世界における事象の重大性を考慮して感情強度を付与する**

自然災害や事故など，一般的に強い負の感情を喚起する事象については，「中 (2)」以上の感情強度を付与する．

文：地震こわい、震度 5 はさすがに焦る。

感情ラベル：アスペクト項「地震」は多くの人に強い負の感情を喚起するため，「極性：-2，恐れ：2」などの高い値を付与する．

## 3.3 アノテーションの品質評価

アノテーションの信頼性を評価するため，50 件のデータを対象としてアノテータ間一致率を算出した．アスペクト表現の推定タスクについては，全てのアノテータ組 $(p_1, p_2 \in P)$ に対して，式 1 に示す Jaccard 係数を用いた．また，感情ラベル付与タスク

**表 3**　アノテータ間一致率．QWK は Quadratic Weighted Kappa を表す．

| | Jaccard | QWK | |
|---|---|---|---|
| Pair | 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 |

については，順序尺度の評価に適した Quadratic Weighted Kappa (QWK) (Cohen 1968) を用いた．表 3 に，各アノテータ組におけるタスクごとの一致率を示す．

アスペクト表現の抽出タスクでは，全ての組み合わせにおいて $J > 0.6$ を達成しており，高い一致が確認された．さらに，感情ラベル付与タスクにおける QWK についても，8 感情および感情極性のいずれにおいても $\kappa > 0.6$ を達成しており，十分な一致が確認された．特に感情極性については，全ての組み合わせで $\kappa > 0.76$ という非常に高い一致率が得られた．これらの結果から，本データセットのアノテーションは高い信頼性を有していることが示された．

$$\frac{1}{|P|} \sum_{p_1, p_2 \in P} \frac{p_1 \cap p_2}{p_1 \cup p_2} \tag{1}$$

# 第 4 章　本データセットの分析

## 4.1　感情ラベルの分布

表 4 に 5 段階の感情極性ラベルの分布を，表 5 に Plutchik の基本 8 感情の強度ラベルの分布を示す．感情極性では，「-1（ネガティブ）」と「+1（ポジティブ）」が多く，「0（ニュートラル）」がそれに続く．一方で，「-2（強いネガティブ）」と「+2（強いポジティブ）」の極端なラベルは比較的少なく，先行研究 (Kajiwara et al. 2021; Suzuki et al. 2022) と同様の分布傾向を示している．

感情強度については，すべての感情で強度「0（無）」が最も多く，高い強度ほど件数が少ない．感情別に見ると，「期待」「喜び」「悲しみ」のラベルが多く付与されている一方，「怒り」の付与数は少ない傾向にある．これも (Kajiwara et al. 2021; Suzuki et al. 2022) と同様の分布傾向を示している．

## 4.2　単語と文の感情ラベルの関係

本データセットの単語単位の感情ラベルと，元コーパスである WRIME の文単位の感情ラベルとの関連性を調査するため，ピアソンの相関係数を算出した．具体的には，各文に含まれる単語の感情極性および各感情強度と，文全体に付与された主観および客観のラベルとの相関を計算した．結果を表 6 に示す．

分析の結果，まず全体的な傾向として，多くの感情において単語単位の感情と文全体の感情との間に正の相関が確認された．これは，個々の単語に付与した感情ラベルが文全体の感情と一致していることを示す．また，主観と客観のラベル間では，ほぼすべての項目で，客観ラベルとの相関係数が主観ラベルとの相関係数を上回った．この結果は，個々の単語が持つ感情の集合は，書き手本人の内面的な感情よりも，第三者である読み手が受け取る客観的な印象をより強く反映している可能性を示唆している．これは，書き手の感情推定が読み手より困難であるという先行研究 (Kajiwara et al. 2021; Suzuki et al. 2022) の指摘とも一致する．感情別に見ると，「信頼」は主観・客観の両方

**表 4**　感情ラベルの分布

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

**表 5**　感情強度ラベルの分布

| | 喜び | 悲しみ | 期待 | 驚き | 怒り | 恐れ | 嫌悪 | 信頼 |
|---|---|---|---|---|---|---|---|---|
| 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 |

**表 6**　単語レベルとの感情ラベルと文レベル感情ラベルとの相関

| | 喜び | 悲しみ | 期待 | 驚き | 怒り | 恐れ | 嫌悪 | 信頼 | 極性 |
|---|---|---|---|---|---|---|---|---|---|
| 書き手 | 0.671 | 0.590 | 0.476 | 0.502 | 0.511 | 0.410 | 0.532 | **0.265** | 0.627 |
| 読み手 | **0.736** | **0.642** | **0.632** | **0.646** | **0.672** | **0.570** | **0.622** | 0.253 | **0.765** |

で特に低い相関を示した．これは，「信頼」という感情が，文全体の文脈や複数の事実関係から間接的に形成される感情であることが要因と考えられる．

また，単語単位で特定の感情が出現した場合に，文全体の主観および客観の感情強度がどのように分布するかを調査した．この結果を図 2 に示す．両ヒートマップにおいて，対角成分が全体的に高い値を示していることが確認できる．これは，本データセットにおいて，特定の感情を表す単語の存在が，文全体の同じ感情の強度と強く関連していることを示している．

主観と客観を比較すると，客観の方が全体的に対角成分の値が高い傾向が見られる．特に，「喜び」や「期待」といったポジティブな感情においてその傾向が顕著である．これは，先行する極性の分析と同様に，単語単位の感情表現が，書き手の内面よりも読み手の印象により直接的に結びついている可能性を示唆している．また，両ヒートマッ

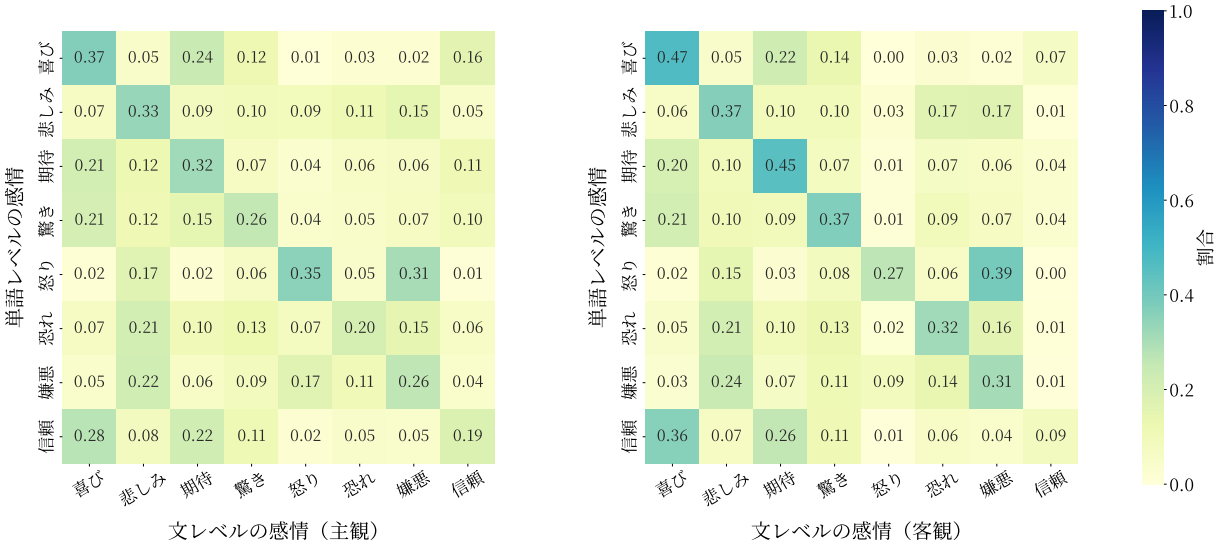


**図 2**　単語単位の感情と文単位の感情の関係.

プで単語単位の「怒り」と文単位の「嫌悪」の間に比較的高い対応関係が見られる．これは，これらの感情が単語単位で共起しやすいだけでなく，文全体の感情としても関連して認識されやすいことを示している．

## 4.3　単語の感情ラベルが主観と客観に与える影響

文中にアノテーションされた単語の感情極性の混在度に着目し，データセットを以下の 2 グループに分割した．

- **感情が単一の文**：文中の単語がポジティブまたはネガティブのいずれか一方のみを持つ文
- **感情が混在した文**：文中の単語がポジティブとネガティブの両方の感情を含む文

それぞれのグループで，文全体の主観極性ラベルと客観極性ラベルの一致率を QWK で算出した結果，感情が単一の文では 0.668 と高い一致率を示したのに対し，感情が混在した文では 0.511 へと大幅に低下した．

この結果は，文中にポジティブおよびネガティブの感情が混在することが，書き手と読み手の間に乖離を生じさせる主要な要因の一つであることを定量的に示している．

# 第 5 章　評価実験

## 5.1　実験設定

**データセット**　本研究では，3 節で構築したデータセットを用いて評価実験を行った．データセットは，訓練データ，検証データ，評価データの比率が 8:1:1 となるように分割し，それぞれ 5,600 件，700 件，700 件を割り当てた．

**タスク**　以下の 3 つのタスクについて評価をした．

(1) **アスペクト表現抽出**文中からアスペクト表現を抽出するタスクである．

(2) **アスペクトベース感情分析**与えられたアスペクト項に対して，5 段階の感情極性および 8 感情それぞれの 4 段階感情強度を推定するタスクである．

(3) **単語情報を用いた文の感情分析**本研究で付与した単語レベルの感情情報が，文レベルの感情分類において有効であるかを検証するタスクである．

**モデル**　エンコーダモデルとして事前学習済みの BERT[3] (Devlin et al. 2019) および ModernBERT[4] (Warner et al. 2025) を用いた．また，LLM として，英語 LLM[5] (Llama Team 2024) を基に日本語データ (Okazaki et al. 2024) による継続事前学習が施された日本語 LLM である Swallow[6] (Okazaki et al. 2024; Fujii et al. 2024) を使用した．

**アスペクト表現抽出の実装方法**　エンコーダモデルの場合，BIO 方式による系列ラベリングとして実装した．LLM の場合，Instruction-tuning (Wei et al. 2022) に基づき，プロンプトに対するテキスト生成としてタスクを定式化した．

**アスペクトベース感情分析の実装方法**　エンコーダモデルに対しては，本タスクを分類タスクとして実装した．感情ラベルは，$h$ を BERT では [CLS] トークン，ModernBERT では¡s¿トークンから得られる特徴ベクトルとして，

$$y = \mathrm{softmax}(hW)$$

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

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

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

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

により推定する．また，先行研究 (Nakayama et al. 2022) に従い，エンコーダモデルへの入力列は [CLS] + 文 + [SEP] + アスペクト表現 + [SEP] のような形式とした．LLM の場合は，Instruction-tuning (Wei et al. 2022) に基づき，プロンプトに対するテキスト生成としてタスクを定式化した．

**文の感情分析の実装**　文レベルの感情分類タスクにおいては，単語レベルの感情情報の有効性を検証するため，2 種類のモデルを比較した．

- **ベースライン：**本モデルは文レベルの感情ラベル（主観または客観）のみを用いて学習する．エンコーダモデル（BERT, ModernBERT）では, [CLS] または<s>トークンの埋め込みから特徴ベクトル $h$ を取得する．LLM に対しては，最終層の隠れ状態ベクトルを平均プーリングすることで $h$ を算出する．その後, $y = \mathrm{softmax}(hW)$ により分類する．
- **マルチタスク学習モデル：**文レベルと単語レベルの感情を同時に学習する．文レベルの分類層はベースラインモデルと同一であり，これに加えて，単語に対する分類層を備える．学習時には，文レベルの分類損失 $L_{\mathrm{sentence}}$ と，単語レベル分類損失 $L_{\mathrm{word}}$ を計算する．モデル全体は式 1 に示す総損失を最小化することで最適化される．

$$L_{\mathrm{total}} = L_{\mathrm{sentence}} + \sum L_{\mathrm{word}} \tag{1}$$

**ハイパーパラメータ**　全モデルのファインチューニングでは，学習率を $2e-5$，最適化手法に Adam (Kingma and Ba 2015) を使用し，3 エポックの Early Stopping を適用した．バッチサイズは 16 に設定した．LLM のファインチューニングには LoRA (Hu et al. 2022) を使用し，ランクを $r = 16$，スケーリング係数を $\alpha = 16$，dropout 率を 0.05 に設定した．

**評価指標**　アスペクト表現抽出タスクは系列ラベリングによる固有表現抽出問題として扱うため，評価指標には適合率（Precision），再現率（Recall），および F1 スコアを用いた．一方，アスペクトベース感情分析タスクおよび文の感情分析タスクは順序分類問題であるため，この種のタスクで一般的に用いられる評価指標である QWK（Quadratic

**表 7**　アスペクト表現抽出の実験結果

| | Precision | Recall | F1 |
|---|---|---|---|
| BERT | 0.646 | 0.701 | 0.672 |
| ModernBERT | 0.677 | **0.762** | **0.717** |
| Swallow 8B | **0.723** | 0.681 | 0.702 |

Weighted Kappa (Cohen 1968) を採用した.

## 5.2　実験結果

### 5.2.1　アスペクト表現抽出

アスペクト表現抽出タスクの実験結果を表 7 に示す．F1 スコアでは, ModernBERT が最も高く 0.717 を記録し，次いで Swallow 8B が 0.702，BERT が 0.672 となった．Swallow 8B は，Precision において 0.723 と最も高いスコアを示したが，Recall が最も低く，F1 スコアでは ModernBERT に及ばなかった．一方で，ModernBERT は適合率と再現率のバランスが最も良く，結果として最高の F1 スコアを達成した．これらの結果は，LLM が高い適合率で正確なアスペクト表現を抽出する能力を有する一方で，本タスクにおいては，最適化されたエンコーダモデルが適合率と再現率とのバランスにおいて優位であることを示唆している.

### 5.2.2　アスペクトベース感情分析

感情極性分類および 8 感情の強度推定タスクの結果を表 8 に示す．全体的な傾向として，Swallow が ModernBERT を上回り，ModernBERT がベースラインの BERT を上回る結果となった．これは，モデルの規模や事前学習データの多様性が，単語単位の微細な感情を捉える上で重要であることを示している.

特に，日本語コーパスで継続事前学習を行った Swallow は，ほとんどの感情で最高性能を達成した．感情別に見ると，「喜び」や「悲しみ」，そして「極性」では比較的高い

**表 8**　感情強度推定および感情極性分類の実験結果（QWK）

| | 喜び | 悲しみ | 期待 | 驚き | 怒り | 恐れ | 嫌悪 | 信頼 | 8 感情 | 極性 |
|---|---|---|---|---|---|---|---|---|---|---|
| BERT | 0.663 | 0.515 | 0.539 | 0.226 | 0.036 | 0.346 | 0.197 | 0.058 | 0.478 | 0.683 |
| ModernBERT | 0.704 | 0.594 | 0.604 | 0.416 | 0.240 | **0.376** | 0.276 | 0.269 | 0.555 | 0.730 |
| Swallow | **0.741** | **0.678** | **0.652** | **0.511** | **0.483** | 0.352 | **0.536** | **0.437** | **0.625** | **0.775** |

スコアが得られた．これらの感情は，テキスト中で明確な表現を用いて語られることが多く，モデルが学習しやすかったと考えられる．一方で，「怒り」「恐れ」「信頼」といった感情は，特に BERT モデルにおいて低いスコアとなった．これらの感情は，より文脈依存性が高く，間接的な表現で示唆されることが多いため，高度な文脈理解能力が求められることを示している．また，「怒り」や「恐れ」のような複雑な感情では，BERT から ModernBERT への性能向上が確認されたことから，エンコーダモデルにおいても事前学習の改善が感情理解能力の向上に有効であることが示された．さらに，Swallow はこれらの複雑な感情においても性能が向上しており，大規模言語モデルの有効性が確認できた．一方で，「信頼」は Swallow においても他の感情と比較して低い性能に留まっており，本データセットにおける最も難しい感情の一つであることが示唆された．

### 5.2.3　単語情報を用いた文の感情極性分類

本データセットに付与された単語単位の感情ラベルが，文全体の感情分類タスクの性能向上に寄与するかを検証した．具体的には，文のみを入力とするベースラインモデルと，文に加えて本研究で付与した単語の感情ラベルを入力とするモデルの性能を比較した．この実験は，文の主観的感情と客観的感情それぞれについて実施し，結果を表 9 と表 10 に示す．単語情報の追加は，主観と客観の感情推定に対して異なる影響を与えることが明らかになった．客観感情の推定（表 9）では，すべてのモデルで単語情報を追加することで極性分類の性能が一貫して向上した．これは，読み手が受け取る客観的な印象は文中の単語の感情に素直に反映されやすいため，単語単位の感情ラベルが有効な補助情報として機能したことを示している．

**表 9**　客観感情で評価した結果（QWK）．「w/ word」は，単語レベルの感情情報を併用したマルチタスク学習モデルを示す．

| 客観 | 喜び | 悲しみ | 期待 | 驚き | 怒り | 恐れ | 嫌悪 | 信頼 | 8 感情 | 極性 |
|---|---|---|---|---|---|---|---|---|---|---|
| BERT | 0.554 | 0.341 | 0.570 | 0.356 | 0.113 | 0.435 | 0.297 | 0.133 | 0.449 | 0.620 |
| BERT w/ word | **0.604** | 0.326 | 0.428 | **0.372** | **0.120** | 0.263 | 0.177 | 0.067 | 0.411 | **0.645** |
| ModernBERT | 0.708 | 0.323 | 0.755 | 0.636 | 0.475 | 0.491 | 0.469 | 0.248 | 0.565 | 0.764 |
| ModernBERT w/ word | **0.769** | **0.551** | 0.699 | 0.503 | 0.474 | **0.615** | **0.541** | 0.239 | **0.631** | **0.784** |
| Swallow | 0.676 | 0.564 | 0.701 | 0.621 | 0.400 | 0.615 | 0.615 | 0.066 | 0.649 | 0.758 |
| Swallow w/ word | **0.737** | **0.582** | 0.666 | 0.514 | **0.502** | 0.468 | 0.598 | **0.232** | 0.629 | **0.777** |

**表 10**　主観感情で評価した結果（QWK）．「w/ word」は，単語レベルの感情情報を併用したマルチタスク学習モデルを示す．

| 主観 | 喜び | 悲しみ | 期待 | 驚き | 怒り | 恐れ | 嫌悪 | 信頼 | 8 感情 | 極性 |
|---|---|---|---|---|---|---|---|---|---|---|
| BERT | 0.453 | 0.265 | 0.156 | 0.284 | 0.202 | 0.200 | 0.148 | 0.309 | 0.280 | 0.387 |
| BERT w/ word | **0.532** | **0.321** | **0.321** | 0.212 | 0.149 | 0.127 | **0.183** | **0.322** | **0.312** | **0.445** |
| ModernBERT | 0.651 | 0.411 | 0.384 | 0.344 | 0.297 | 0.244 | 0.300 | 0.257 | 0.372 | 0.586 |
| ModernBERT w/ word | **0.688** | **0.427** | **0.408** | **0.388** | **0.295** | **0.246** | **0.347** | **0.284** | **0.407** | 0.532 |
| Swallow | 0.648 | 0.374 | 0.469 | 0.401 | 0.230 | 0.096 | 0.447 | 0.223 | 0.476 | 0.543 |
| Swallow w/ word | 0.632 | **0.502** | 0.461 | **0.411** | **0.259** | **0.277** | 0.415 | 0.202 | **0.478** | **0.608** |

一方，主観感情の推定（表 10）は，既存研究 (Kajiwara et al. 2021; Suzuki et al. 2022) も指摘するように，書き手本人の皮肉や複雑な感情を含むため本質的に困難なタスクである．実際，通常の学習では客観感情（表 9）のスコアに比べて主観感情のスコアは低い水準に留まっている．（表 10）の結果を見ると，極性分類において単語情報を追加したマルチタスク学習（w/ word）は，Swallow（0.543→0.608）や BERT（0.387→0.445）で性能を向上させた一方，ModernBERT（0.586→0.532）のように性能が低下するケースも見られた．

しかし，8 感情全体を評価した QWK に着目すると，極性分類で低下が見られた ModernBERT（0.372 → 0.407）も含め，すべてのモデルにおいて単語情報の追加による性

**表 11**　マルチタスク学習による主観感情推定の改善事例.「w/ word」は，単語レベルの感情情報を併用したマルチタスク学習モデルを示す.

| テキスト | 喉が痛い,, どうしたら口開けて寝なくなるの？？ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| アノテーション | 喜び | 悲しみ | 期待 | 驚き | 怒り | 恐れ | 嫌悪 | 信頼 | 極性 |
| 喉 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | -1 |
| 口 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | -1 |
| ラベル | 喜び | 悲しみ | 期待 | 驚き | 怒り | 恐れ | 嫌悪 | 信頼 | 極性 |
| 書き手 | 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 |

能向上が確認された．この結果は，個々の単語が持つ細粒度な感情情報が，モデルが文全体の感情（喜びや悲しみなど）を捉える上で有効な補助情報として機能していることを示唆している．特に，マルチタスク学習を適用した Swallow (Swallow w/ word) は，8 感情および極性分類の双方において最高性能を達成した．これにより，難度の高い主観感情推定タスクにおいて，LLM に対するマルチタスク学習の適用が極めて有力なアプローチであることが確認された.

表 11 にマルチタスク学習によって主観感情の推定結果が改善した具体的な事例を示す．この例では，ベースラインの Swallow モデルは，「嫌悪」と「悲しみ」は推定できたものの，「恐れ」の感情を見逃している．一方で，マルチタスク学習をした Swallow は，「口」という単語に付与された「恐れ」のラベルを補助情報として活用できたため，文全体の感情として「恐れ」を正しく推定できた．この事例は，単語単位の詳細な感情情報が，文全体からは捉えにくい書き手の隠れた感情を推定する上で，有用であることを示している.

# 第 6 章　おわりに

本研究では，日本語 ABSA における単語に対して 5 段階の感情極性ラベルと Plutchik の基本 8 感情に基づく 4 段階の感情強度ラベルを付与したデータセットを構築し，公開した．本データセットは，既存の文単位感情分析データセット WRIME を拡張したものであり，SNS というオープンドメインなテキストを対象としている．単語単位の感情ラベルと文単位の感情ラベルの相関を調査した結果，単語の感情の集合は書き手の主観的な感情よりも，読み手が受け取る客観的な印象をより強く反映していることが明らかになった．これは，既存研究が文単位で指摘してきた主観と客観のギャップが，より詳細な単語単位の分析においても存在することを示している．

さらに，本データセットを用いて，アスペクト表現抽出，アスペクトベース感情分析，単語情報を用いた文の感情分析の 3 つの評価実験を行った．アスペクト表現抽出では ModernBERT が最も高い F1 スコアを示し，系列ラベリングに基づくエンコーダモデルの有効性が確認された．一方，アスペクトベース感情分析では Swallow が多くの感情および感情極性において最も高い性能を示し，単語単位の細粒度な感情を捉える上で LLM が有効であることが示された．その一方で，「信頼」や「怒り」などの文脈依存性の高い感情では，他の感情と比較して推定性能が低く，依然として推定が困難であることも確認された．また，本研究で付与した単語単位の感情情報を補助的に用いたマルチタスク学習では，8 感情全体の推定性能がすべてのモデルで向上した．特に Swallow では，単語情報を用いることで主観感情の 8 感情および感情極性の双方において高い性能を示し，単語単位の細粒度な感情情報が，文全体からは捉えにくい書き手の感情を推定するための有効な補助情報となることが示された．

今後は，「信頼」や「怒り」など推定が困難な感情について誤り分析を行い，文脈依存性やデータ分布などの観点から性能低下の要因を明らかにする必要がある．また，レビューや対話など異なるドメインへの適用や，単語単位と文単位の感情情報をより効果的に統合する学習手法について検討する．

# 参考文献


[1] 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**, pp. 11019–11038.

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

[3] 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*, pp. 2095–2104.

[4] 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*, pp. 7022–7028.

[5] 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)*, pp. 27–35.

[6] 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)*, pp. 49–54.

[7] 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*, pp. 1546–1556.

[8] 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)*, pp. 6280–6285.

[9] 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*, pp. 7014–7021.

[10] Robert Plutchik. 1980. "A General Psychoevolutionary Theory of Emotion." *Theories of Emotion*, pp. 3–33.

[11] 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)*, pp. 486–495.

[12] 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)*, pp. 19–30.

[13] 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*, pp. 3548–3557.

[14] 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)*, pp. 3594–3605.

[15] 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)*, pp. 2339–2349.

[16] 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*, pp. 2069–2079.

[17] 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)*, pp. 340–350.

[18] 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*, pp. 9209–9219.

[19] 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)*, pp. 175–185.

[20] Paul Ekman. 1992. "An Argument for Basic Emotions." *Cognition and Emotion*, **6**

(3–4), pp. 169–200.

[21] James A. Russell. 1980. "A Circumplex Model of Affect." *Journal of Personality and Social Psychology*, **39** (6), pp. 1161–1178.

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

[23] 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*, pp. 4171–4186.

[24] 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)*, pp. 2526–2547.

[25] Llama Team. 2024. "The Llama 3 Herd of Models." *arXiv:2407.21783*.

[26] 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*.

[27] 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*.

[28] 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*.

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

[30] 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*.