JAIST Repository
https://dspace.jaist.ac.jp/

Title	感情表出の顕現性と書き手特性を考慮した書き手の感情強度推定
Author(s)	岡留, 司真
Citation	
Issue Date	2026-03
Type	Thesis or Dissertation
Text version	author
URL	https://hdl.handle.net/10119/20551
Rights	
Description	Supervisor:白井　清昭, 先端科学技術研究科, 修士(情報科学)

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

修士論文

感情表出の顕現性と書き手特性を考慮した書き手の感情強度推定

岡留 司真

主指導教員 白井 清昭

北陸先端科学技術大学院大学
先端科学技術研究科
(情報科学)

令和 8 年 3 月

Abstract

In recent years, emotion recognition in text has become an active area of research in the field of natural language processing. Among various tasks, emotion intensity estimation has attracted particular attention as an important and challenging problem. Emotion intensity estimation aims not only to classify the type of emotion expressed in a text, but also to estimate how strongly the writer feels that emotion. This task plays a crucial role in practical applications such as dialogue systems, sentiment analysis, and opinion mining, where understanding the degree of emotional involvement is essential. However, many existing studies assume that emotions are explicitly expressed in text. As a result, they often overlook cases in which a writer's emotions are not directly reflected in linguistic expressions, as well as discrepancies between the emotions actually felt by the writer and those interpreted by readers. Consequently, most conventional approaches rely heavily on surface-level cues, such as emotion words or intensifiers, to estimate emotion intensity. These methods tend to struggle with texts that contain implicit emotional expressions, making it difficult to accurately estimate the true emotion intensity of the writer.

This study aims to estimate the writer's emotion intensity more accurately by considering both the salience of emotion and the writer's characteristics. Salience of emotion refers to the degree to which a writer's emotion is explicitly expressed in the text. In this study, it is measured by the difference between the emotion intensity annotated by a writer and that annotated by a reader. When the discrepancy between the writer's and reader's emotion intensity scores is large, the emotion is considered to be weakly expressed, indicating low salience of emotion. Conversely, a small difference suggests that the emotion is clearly expressed in the text. In addition to salience of emotion, this study also focuses on the writer ' s characteristics. Some writers tend to express their emotions directly and explicitly, while others prefer indirect or implicit expressions. Such individual differences in emotional expression are important factors in estimating emotion intensity. By modeling these tendencies as writer's characteristics, we can capture writer-specific patterns of how emotions are expressed in text. This study proposes an approach that incorporates

both salience of emotion and writer's characteristics into an emotion intensity estimation model.

To construct the emotion intensity estimation model, we use WRIME, a Japanese emotion dataset that includes annotations from both writers and multiple readers. A distinctive feature of WRIME is that it includes emotion intensity scores assigned by the writers themselves as well as by readers for the same text. This allows for a detailed analysis of discrepancies in emotion recognition between writers and readers. Furthermore, WRIME provides a writer ID for each text, enabling the modeling of writer-specific characteristics. This study adopts two experimental settings when splitting WRIME into training and test data: "Open Task" and "Closed Task." In the Open Task setting, texts written by the authors in the test data do not appear in the training data, whereas in the Closed Task setting, they do. By comparing these two settings, we examine how much writer's characteristics and salience of emotion contribute to improving emotion intensity estimation performance when past texts written by the target writer are available or unavailable during training.

This study proposes two models for emotion intensity estimation: a late fusion model and a multi-task learning model. The late fusion model estimates emotion intensity by integrating multiple types of information derived from text. Specifically, it extracts three kinds of feature vectors representing the semantic meaning of the text, the salience of emotion, and the writer's characteristics. These feature vectors are concatenated and fed into a Fully Connected Layer (FCL) to predict the emotion intensity. The semantic feature vector is obtained from a fine-tuned pre-trained language model, BERT, which captures contextual and semantic information from the input text. The feature vector representing salience of emotion is obtained from a model referred to as SE-BERT. In this study, the "salience of emotion estimation task" is formulated as a binary classification problem that determines whether there is a difference between the emotion intensity annotated by the writer and that annotated by the reader. A BERT-based model is fine-tuned to solve this task, and the resulting model is defined as SE-BERT. The feature vector representing the writer's characteristics is obtained from another BERT-based model, referred to as ID-BERT. In this case, the "writer estimation task" is de-

fined as the task of predicting the writer ID from a given text. A BERT model is fine-tuned to solve this task, and the resulting model is used as ID-BERT. During the training of the late fusion model, the parameters of SE-BERT and ID-BERT are fixed and not updated. Only the parameters of the BERT model used to obtain the semantic feature vector and the parameters of the FCL for emotion intensity estimation are updated.

In the multi-task learning model, emotion intensity estimation is treated as the main task, while salience of emotion estimation task is treated as an auxiliary task. A pre-trained BERT model is used as a common base model for the two tasks, and FCL for obtaining the outputs of the main task and the auxiliary task are respectively placed on top of it. The parameters of the entire model are updated using a loss function that combines the losses of the two tasks.

Several experiments were conducted to evaluate the effectiveness of the proposed methods. Using the WRIME dataset, emotion intensity for each of the eight emotion classes are estimated. First, we discuss the results of the late fusion model. In the Open Task setting, incorporating the salience of emotion into the model led to improved performance in emotion intensity estimation. In contrast, the writer's characteristics did not show effectiveness under this setting. Furthermore, combining salience of emotion and writer's characteristics did not result in additional performance gains. This result indicates that, in situations where texts written by the same writer cannot be used during training, it is difficult to sufficiently capture writer-specific tendencies, and that information about the extent to which emotions are explicitly expressed in the text plays a major role in estimating the writer's emotion intensity.

On the other hand, in the Closed Task setting, it was found that writer's characteristics work effectively for emotion intensity estimation. Furthermore, combining salience of emotion with writer's characteristics resulted in higher performance than using either feature alone. This result demonstrates that, under conditions where the writer is known, the model is able to learn writer-specific emotional expression tendencies to some extent, and that writer's characteristics plays a complementary role with the salience of emotion in emotion intensity estimation.

For the multi-task learning model, no clear performance improvement

was observed by incorporating salience of emotion in either the Open Task or Closed Task settings. This result suggests that although emotion intensity estimation and salience of emotion estimation are related tasks, the features required for each task may not be fully aligned. In other words, simultaneous optimization of these two tasks may interfere with the learning of task-specific representations, potentially hindering performance improvements for emotion intensity estimation.

概要

近年，自然言語処理分野において，テキストから感情を推定する感情認識の研究が盛んに行われている．中でも感情強度推定は重要な課題として注目されている．感情強度推定とは，テキストに表される感情の種類を分類するだけでなく，テキストの書き手がその感情を「どの程度強く感じているか」を推定するタスクであり，対話システムやオピニオンマイニングなど，実応用においても重要な役割を果たす．しかし，多くの既存研究では，感情表現がテキスト中に明示的に表出していることを前提としており，書き手の感情が必ずしも言語表現に直接反映されない場合や，書き手の実際の感情と読み手が解釈する感情の差異については十分に考慮されていなかった．その結果，感情語やそれを強調する語など表層的な手がかりに基づく手法が主流であり，暗黙的な感情表現を含むテキストに対しては感情強度を適切に推定できないという課題がある．

本研究は，感情顕現性および書き手特性を考慮することで書き手の感情強度をより適切に推定することを目的とする．感情顕現性とは，書き手の感情がテキストにどの程度明示的に表出されているかを表す指標であり，書き手と読み手の感情強度の差として定量化する．すなわち，書き手と読み手の感情強度の差が大きいときは，書き手の感情が明示的に示されておらず，感情顕現性が低いとみなす．一方，本研究では書き手の特性も考慮する．書き手には感情を直接的に表す人もいれば，間接的または婉曲的に感情を表す人もいる．このような書き手の特性は感情強度を推定する上での重要な特徴量になる．本研究では，感情顕現性ならびに書き手特性を感情強度推定モデルに組み込むアプローチを提案する．

感情強度推定モデルを構築するにあたり，書き手と複数の読み手によるアノテーションを含む日本語感情データセット WRIME を用いる．WRIME は，同一テキストに対して書き手自身の感情強度と読み手による感情強度が付与されている点に特徴があり，書き手・読み手間の感情認識の差を分析できる貴重なデータセットである．また，WRIME にはテキストに対して書き手の ID も付与されている．本研究では，WRIME を訓練データとテストデータに分割するにあたり，Open Task と Closed Task の 2 通りの設定を用いる．Open Task はテストデータの書き手が書いたテキストが訓練データに存在しない設定とし，Closed Task では存在する設定とする．これにより，評価対象の書き手が書いた過去のテキストをモデルの訓練時に利用できるときとできないときとで，書き手特性や感情顕現性が感情強度の推定性能の向上にどの程度貢献するかを検証する．

本研究では Late Fusion モデルとマルチタスク学習モデルの 2 つの感情強度推定モデルを提案する．Late Fusion モデルでは，テキストの意味を表す特徴ベクトル，感情顕現性を表す特徴ベクトル，書き手特性を表す特徴ベクトルを抽出し，これらを連結したベクトルを全結合層に渡して感情強度を推定する．意味の特徴ベクトルはファインチューニングされた事前学習済み言語モデル BERT から得る．感情顕現性を表す特徴ベクトルは SE-BERT から得る．「感情顕現度推定タスク」を書き手と読み手の感情強度に差があるかないかの二値分類問題とし，これを解くモデルを BERT のファインチューニングによって構築する．このモデルを SE-BERT とする．書き手特性を表す特徴ベクトルは ID-BERT から得る．「書き手推定タスク」をテキストの書き手 ID を予測するタスクとし，これを解くモデルを BERT のファインチューニングによって構築する．このモデルを ID-BERT とする．Late Fusion モデルの学習は，SE-BERT と ID-BERT のパラメタは事前学習時のものに固定し，意味の特徴ベクトルを得る BERT とクラス分類のための結合層のみパラメタを更新する．

マルチタスク学習モデルでは，感情強度推定タスクを主タスク，感情顕現性推定タスクを補助タスクとするマルチタスク学習を行う．2 つのタスクの共通の基盤モデルとして事前学習済みの BERT を使用し，その上に主タスクと補助タスクの出力を得るための全結合層をそれぞれ置く．2 つのタスクの損失を合わせた損失関数によってモデル全体のパラメタを更新する．

提案手法の評価実験を行った．WRIME を使用し，8 つの感情クラスのそれぞれについて感情強度を推定した．まず，Late Fusion モデルの結果を考察する．Open Task においては，感情顕現性をモデルに組み込むことで感情強度推定の性能が向上することが確認された．一方で，書き手の特徴は，書き手情報が未知である条件下では有効性を示さず，感情顕現性と書き手特徴を統合した場合においても性能の向上は確認されなかった．この結果は，同じ書き手が書いたテキストを学習時に利用できない状況では，書き手固有の傾向を十分に捉えることが困難であること，感情がどの程度テキスト上に顕在化しているかという情報が書き手の感情強度を推定する上で大きな役割を果たすことを示している．

一方，Closed Task においては，書き手特性が感情強度推定に対して有効に機能することが確認された．さらに，感情顕現性と書き手特徴を組み合わせることで，これらを単独で用いる場合よりも性能が向上することも確認された．この結果は，書き手が既知である条件下では，書き手

固有の感情表現傾向をある程度学習することができていること，書き手特性は感情顕現性と相補的に働くことを示している．

一方で，マルチタスク学習モデルにおいては，Open Task，Closed Task のいずれにおいても，感情顕現度をモデルに組み込むことによる明確な性能向上は確認されなかった．この結果は，感情強度推定と感情顕現性推定が互いに関連するタスクである一方で，モデルが学習すべき特徴表現が必ずしも一致していない可能性を示唆している．すなわち，複数のタスクを同時に最適化することで，それぞれのタスクに特化した表現学習が阻害されている可能性がある．

目 次

第 1 章　はじめに **1**
1.1　背景 1
1.2　目的 3
1.3　本論文の構成 4

第 2 章　関連研究 **6**
2.1　感情強度推定 6
2.2　書き手の感情推定 7
2.3　暗黙的な感情推定 8
2.4　書き手と読み手の感情の差に着目した研究 10
2.5　本研究の特色 10

第 3 章　提案手法 **13**
3.1　タスク設定 13
3.2　感情顕現性と書き手を考慮した Late Fusion モデル 14
3.2.1　概要 14
3.2.2　感情顕現性の特徴抽出 16
3.2.3　書き手の特徴抽出 18
3.2.4　実装の詳細 19
3.3　感情顕現性を考慮したマルチタスク学習モデル 21

第 4 章　評価 **24**
4.1　実験設定 24
4.1.1　データセット 24
4.1.2　Open Task/Closed Task 設定 25
4.1.3　評価指標 26
4.2　感情顕現性判定モデルの評価 26
4.2.1　Late Fusion モデルにおける感情顕現性判定モデルの評価 26

4.2.2 マルチタスク学習モデルにおける感情顕現性判定モデルの評価 28
4.3 Late Fusion モデルの評価 29
4.3.1 Open Task における Late Fusion モデルの実験結果 30
4.3.2 Closed Task における Late Fusion モデルの実験結果 31
4.3.3 Open Task と Closed Task の結果の比較 32
4.4 マルチタスク学習モデルの評価 33
4.4.1 Open Task におけるマルチタスク学習モデルの実験結果 33
4.4.2 Closed Task におけるマルチタスク学習モデルの実験結果 33
4.4.3 考察 34
4.5 Late Fusion モデルとマルチタスク学習モデルの比較 . . . 35

第 5 章 おわりに **37**
5.1 本論文のまとめ 37
5.2 今後の課題 38

図 目 次

3.1 Late Fusion モデル 15
3.2 マルチタスク学習モデル 22

表 目 次

4.1 Open Task における書き手感情ラベル分布 25
4.2 Closed Task における書き手感情ラベル分布 25
4.3 Open Task と Closed Task における書き手の数 26
4.4 感情顕現性ラベルの分布 (二値分類，Open Task) 27
4.5 感情顕現性ラベルの分布 (二値分類，Closed Task) 27
4.6 感情顕現性判定モデルの評価 (二値分類) 28
4.7 感情顕現性ラベルの分布 (マルチクラス分類，Open Task) 29
4.8 感情顕現性ラベルの分布 (マルチクラス分類，Closed Task) 30
4.9 感情顕現性判定モデルの評価 (マルチクラス分類) 30
4.10 Open Task における Late Fusion モデルの評価 31
4.11 Closed Task における Late Fusion モデルの評価 31
4.12 Open Task におけるマルチタスク学習モデルの評価 33
4.13 Closed Task におけるマルチタスク学習モデルの評価 . . . 34
4.14 Open Task における Late Fusion モデルとマルチタスク学習モデルの比較 36
4.15 Closed Task における Late Fusion モデルとマルチタスク学習モデルの比較 36

第1章　はじめに

1.1　背景

近年，ソーシャルメディアやオンラインコミュニケーションの急速な普及により，テキストデータが日常的かつ大量に生成・蓄積されている．多くのテキストには人間の感情が含まれる．例えば，ブログ記事，SNS 投稿，レビュー文，オンライン対話ログなどには，利用者の意見や態度だけでなく，喜び，不安，怒り，悲しみといった多様な感情が自然言語として表現されている．このようなテキストから感情的情報を自動的に分析する感情分析 (Emotion Analysis) は，自然言語処理分野における重要な研究課題の一つとして位置づけられており，マーケティング分析や世論分析に加え，メンタルヘルス支援，対話システム，ユーザ理解など，人間中心の応用分野において高い有用性が期待されている．

感情分析研究の初期段階では，テキストが肯定的か否定的かを判定する極性分析が主に扱われてきた [1]．これは感情を二値的な極性として捉えるものであり，計算機による扱いやすさの観点から広く研究されてきた．その後，単純な極性判断では人間の感情の多様性を十分に表現できないという認識が広まり，喜び・怒り・悲しみなどの感情カテゴリ分類へと研究対象が拡張された [2]．これらの研究は，感情の「種類」という側面を捉える上で大きな進展をもたらした一方で，感情の「強さ」や「程度」といった側面については十分に扱われていなかった．

しかし，実際の人間の感情は二値的あるいは離散的なカテゴリに単純に還元できるものではなく，連続的・段階的な性質を持つ．例えば,「少し悲しい」状態と「非常に悲しい」状態とでは，心理的影響や行動への影響が大きく異なる．このような感情の強さを捉えるため，近年では感情強度推定 (Emotion Intensity Estimation) に注目が集まっている．Mohammad らによる WASSA-2017 Emotion Intensity Shared Task は，感情強度推定を明確なタスクとして定義した最初期の取り組みであり，信頼性の高いアノテーションデータに基づく評価基盤を確立した点で重要である [3]．

感情強度推定を議論する上では，感情をどのような枠組みで定義するかという問題も考慮する必要がある．心理学および感情科学の分野では，Ekman による基本感情モデル [4]，Plutchik の感情の輪モデル [5]，Valence Arousal Dominance(VAD) モデル [6] など，さまざまな感情モデルが提案されてきた．Ekman モデルは少数の基本感情に着目する簡潔な枠組みである一方，Plutchik モデルは感情の強度や感情間の関係性を重視しており，感情強度推定との親和性が高い．また，VAD モデルは感情を連続的な心理次元として捉える点に特徴があり，感情の量的評価を理論的に支える枠組みとして広く用いられている．

一方で，感情分析，特に感情強度推定を困難にしている要因として，感情の主観性が挙げられる．心理学的研究においても，感情は個人の経験，価値観，文脈に強く依存する主観的な現象であることが指摘されている [7]．自然言語処理分野においても，感情ラベルはテキストに内在する客観的属性ではなく，アノテータの解釈に基づく評価であることが早くから認識されている [8, 9, 10]．

従来の多くの感情分析研究では，書き手とは異なる読み手 (アノテータ) による感情ラベルを正解として用いることが一般的であった．本研究では，このようなラベルを読み手ラベルと呼ぶ．一方で，テキストを書いた本人が感じていた感情は，必ずしもテキスト表層に明示的に表現されるとは限らず，読み手による解釈と乖離する可能性がある．この書き手自身の暗黙的な感情に基づくラベルは，主観感情ラベルとして読み手ラベルとは区別して扱う必要がある．

この書き手と読み手の感情認識のズレに着目した研究として，WRIME データセットが提案されている [11]．WRIME は，日本語で書かれた SNS 投稿に対して，書き手自身による感情強度ラベルと，複数の読み手による感情強度ラベルを同時に付与した大規模データセットであり，感情の主観性や書き手と読み手の視点の差を定量的に分析可能な点に大きな特徴がある．WRIME に基づく分析では，書き手感情の推定は読み手感情の推定よりも難易度が高く，両者の間に体系的なズレが存在することが示されている．

さらに，Suzuki らは，書き手の性格特性を考慮することで，書き手の主観的感情強度推定を改善できる可能性を示している [12]．このことは感情強度がテキスト表層情報のみによって決定されるものではなく，書き手固有の傾向や属性に強く依存することを示唆している．

以上のように，感情強度推定においては，感情の強さの認識に加えて，

書き手と読み手の感情認識のズレをどのように扱うかが重要な課題である．しかし，書き手と読み手の両方が揃ったデータセットの少なさから，このズレは明示的にモデル化されてこなかった．書き手の暗黙的な感情がどの程度テキストを通じて他者に伝達されているかという観点からこのズレを捉え，感情強度推定に統合する試みは，依然として十分に検討されていない．

1.2 目的

本研究では，書き手の感情がテキストに明示的に表現されている度合を感情顕現性と定義する．感情顕現性が高い場合には，書き手自身がの感情強度と，読み手によって知覚される感情強度との間の差は小さくなる．一方，感情顕現性が低い場合には，両者の間に大きなズレが生じると考えられる．

このズレは，書き手が感情を直接的な感情語として表現していない場合や，文脈依存的・暗黙的な表現を用いている場合に顕著となる．したがって，感情顕現性は書き手と読み手の感情のズレを表す指標となる．ただし，感情顕現性は感情の強さや種類とは独立した概念であり，書き手の感情がどの程度言語表現として顕現されているかを表す指標と位置づけられる．

従来の感情分析研究では，このような書き手と読み手の感情認識のズレは主にアノテーションにおけるノイズとして扱われてきたが，本研究では，このズレ自体を感情表現の特性として捉え，感情顕現性を明示的にモデル化する．

しかし，書き手の感情強度推定にはいくつかの課題が存在する．第一に，感情強度は段階的であり，かつ主観性の高い概念であるため，テキスト表層に明示的に現れない場合が多い．特に，感情を直接的な感情語として表現しない暗黙的な感情表現においては，単純な単語ベースや表層的特徴に基づく手法では正確な推定が困難である．

第二に，同じ感情強度であっても，書き手ごとに表現の仕方には大きな個人差が存在する．過去の発話を通じて書き手の表現傾向を学習できる場合と，未知の書き手に対して推定を行う場合とでは，問題の性質が大きく異なる．この点は，実運用を想定した際に特に重要な課題である．

これらの課題に対処するため，本研究では，感情強度推定を単独のタスクとして扱うのではなく，感情がテキスト中に明示的に表現されてい

るか否かを判定する感情顕現性判定タスクを補助的に導入する．感情顕現性を明示的にモデル化することで，感情が暗黙的に表現されている場合であっても，書き手の感情状態をより適切に捉えることを目指す．

さらに，本研究では，WRIME データセットに付与されている書き手 ID および書き手の感情強度ラベルを活用し，書き手ごとの感情表現傾向を考慮した学習を行う．

以上を踏まえ，本研究の具体的な目的は以下の三点にまとめられる．

- 与えられたテキストに対し，(読み手ではなく) 書き手の感情強度を推定する．WRIME データセットの仕様にしたがい，感情強度は 0,1,2,3 の離散値として表す．
- 感情顕現性および書き手情報を考慮した感情強度推定モデルを提案し，暗黙的感情表現を含むテキストに対する感情強度推定の性能を向上させる．
- WRIME データセットを用いて提案手法の有効性を実験的に検証する．

本研究により，感情強度推定における主観性や個人差という課題に対する一つの解決策を提示するとともに，書き手の暗黙的な感情状態をより精緻に推定するための基盤を提供することを目指す．

1.3 本論文の構成

本論文の構成は以下のとおりである．

2 章では，感情強度推定に関連する既存研究について概観する．感情分析および感情強度推定に関する先行研究を整理するとともに，書き手と読み手の感情ラベルを扱った研究や，主観性・個人差を考慮した手法について論じ，本研究の位置づけを明確にする．

3 章では，本研究で提案する感情強度推定手法について詳述する．まず，タスク設定および使用するデータセットについて説明する．次に，感情強度推定の全体的な枠組みを示し，感情顕現性および書き手ラベルを考慮したモデル設計について説明する．

4 章では，提案手法の有効性を検証するための評価実験について述べる．実験設定および評価指標を説明した後，感情顕現性判定モデルおよ

び感情強度推定モデルの評価結果を示す．さらに，得られた結果について考察を行い，提案手法の効果と限界について議論する．

5 章では，本論文のまとめを行い，本研究で得られた知見を整理する．また，今後の課題および展望について述べる．

第2章　関連研究

2.1　感情強度推定

感情強度推定 (Emotion Intensity Estimation) とは，テキストに含まれる感情の種類を判別するだけでなく，その感情がどの程度の強さで表現されているかを推定するタスクである．従来の感情分類が「怒り」「喜び」といった離散的なカテゴリの予測を目的としていたのに対し，感情強度推定では感情の連続性や段階性を捉える点に特徴がある．このため，より人間の主観的判断に近い感情理解を目指す課題として注目されてきた．

感情強度推定研究の体系的な出発点として広く認識されているのが，Mohammad らの WASSA-2017 と呼ばれる評価型ワークショップ (Shared Task) である [3]．この研究では，Twitter 上の短文を対象として,「怒り」「恐れ」「喜び」「悲しみ」の 4 感情について，0 から 1 の連続値で感情強度が付与されたデータセットが構築された．評価指標としては Pearson 相関係数が用いられ，予測値と人手アノテーションとの一致度が測定された．この取り組みにより，感情強度推定が感情分類とは異なる独立した研究課題として明確に位置づけられた点は重要である．

この Shared Task に参加した代表的な研究として，Lakomkin らの研究がある [13]．この研究では，文字レベルおよび単語レベルの双方向 RNN を組み合わせたモデルが提案され，感情辞書や評価語といった語彙的特徴を積極的に活用した点が特徴である．特に,「very」「extremely」などの強調表現や感情語の出現が感情強度と強く相関することを前提とした設計となっており，感情が明示的に表現されている文に対して高い性能を示した．一方で，感情語を含まない暗黙的表現への対応には限界があることも示唆されている．

その後，深層学習モデルの発展に伴い，Transformer[14] を基盤とした手法が感情強度推定にも導入されるようになった．Ghosh らは，感情強度推定を主タスクとし，VAD の分類を補助タスクとするマルチタスク学習を提案した [15]．この研究は，感情を単一の尺度で捉えるのではなく，

複数の感情次元を併用することで，モデルがより豊かな感情表現を認識できることを示した．

Ghosh らは，27 種類の感情カテゴリをテキストに付与した GoEmotions と呼ばれる大規模データセットを構築した [16]．GoEmotions では一つのテキストに複数の感情ラベルが付与されている．マルチラベル感情分類という設定は，感情が単一ではなく複数同時に表出する可能性が考慮されている．GoEmotions は感情強度推定を直接の目的とはしていないものの，後続の感情強度推定研究においても，感情表現の複雑性を考慮する上で重要な基盤となっている．

以上の研究に共通する特徴として，唯一の感情強度が存在すると仮定し，その推定を試みている点が挙げられる．しかし実際には，感情が明示的に表現されている場合と，文脈的に推測される場合とでは，モデルが利用可能な手がかりの性質が大きく異なる．また，感情の解釈は書き手と読み手で必ずしも一致するとは限らない．このような問題点は従来の感情強度推定研究では十分に考慮されていなかった．

2.2 書き手の感情推定

感情強度推定に関する多くの研究では，テキストに付与された感情ラベルを単一の正解として扱い，その感情が誰の視点に基づくものかを明示的に区別しない場合が多い．しかし実際には，感情はテキストそのものに内在する客観的な属性ではなく，書き手がどのような感情状態でそのテキストを生成したか，あるいは読み手がどのように解釈したかという 2 つの観点から見れば異なるものであり，本質的に主観的な概念である．この観点から，書き手の感情に着目した研究が徐々に進められてきた．

書き手の感情を扱う代表的なデータセットの一つが EmoBank である [17]．EmoBank は，VAD という連続的な感情次元に基づいてアノテーションが行われているが，特徴的なのは，同一テキストに対して書き手 (author) と読み手 (reader) の両方の感情評価が付与されている点である．このデータセットにより,「書き手が意図した感情」と「読み手が受け取った感情」が必ずしも一致しないことが示された．しかし，EmoBank における感情ラベルは書き手のラベルも読み手のラベルも同じアノテータによって付けられており，すなわち読み手が書き手の感情を推測して書き手の感情を付与しているため，真の書き手感情ラベルが付与されていないという問題点がある．

書き手と読み手の感情強度を明示的に区別したデータセットとしてWRIME[11]が挙げられる．WRIMEでは，ソーシャルメディアに投稿された日本語の短文に対し，書き手自身による感情強度ラベルと，複数の読み手による感情強度ラベルが付与されている．さらに，Plutchikの感情モデルに基づく8感情それぞれについて，0,1,2,3の離散値で感情強度が与えられている．このような設計により，書き手と読み手の両方の感情強度ラベルが付与された数少ないデータセットである．

書き手の属性を明示的にモデルに組み込む試みとしては，Suzukiらの研究がある[12]． この研究では，書き手の性格特性(Big Fiveなど)を補助情報として利用し，感情強度推定の精度向上を目指している．結果として，性格特性を考慮することで，一部の感情カテゴリにおいて推定性能が向上することが示されており，書き手固有の情報をモデル化することの有効性が示されている．

BostanとKlingerは，異なる感情アノテーション方式(離散ラベル，連続値，複数感情など)を用いたデータセット間で感情分析の性能を比較している[18]．感情データの設計やラベル形式がモデル性能に大きな影響を与えることを示しており，特に書き手自身が付与したラベルと第三者が付与したラベルで予測モデルの性能に差異が存在する点を指摘している．

これらの研究から，書き手の感情強度を推定する際には，単にテキスト表層の情報だけでなく，誰が書いたか，どのような感情傾向を持つかといった要因が重要であることが明らかになっている．しかしながら，多くの研究では，書き手情報は補助的に扱われるか，あるいは明示的にモデル化されていない．書き手の感情強度推定に関する研究は一定の進展を見せているものの，書き手と読み手の感情を同時に考慮する枠組みや，書き手の感情表現の顕現性を考慮する方法については，十分に検討されているとは言い難い．

2.3 暗黙的な感情推定

感情分析の多くの研究は，テキスト中に明示的に現れる感情語や評価表現を手がかりとして感情を推定することを前提としてきた．しかし実際の自然言語使用においては，感情が必ずしも直接的な単語として表出されるとは限らず，文脈や出来事の記述を通じて暗黙的に示される場合も多い．このような暗黙的感情表現を扱う研究は，明示的感情認識とは異なる難しさを伴う．

この問題を体系的に扱った最初期の取り組みの一つが，Klinger らによって提案された Implicit Emotions Shared Task(IEST) である [19]． IEST では，ツイート中の感情語をマスクした状態で，背後にある感情カテゴリを推定するタスクが設定された．これにより，モデルは感情語そのものではなく，前後の文脈や出来事の意味理解に基づいて感情を推測する必要がある．評価結果からは，暗黙的感情認識が明示的感情分類に比べて著しく難易度が高いことが示され，感情推定における文脈理解の重要性が明確に示された．

Boutouta らはアラビア語テキストを対象とした暗黙的感情認識に取り組んでいる [20]．この研究では，Transformer と RNN を組み合わせたハイブリッドモデルを提案し，単語系列の長期依存関係と局所的な文脈情報の双方を捉えることを目指している．アラビア語のように形態的に複雑な言語においても，暗黙的感情の予測が可能であることを示した．

CEFER は，暗黙的感情と明示的感情を統一的に扱うことを目的とした多層フレームワークである [21]．文脈レベルの埋め込み表現と，語彙レベルの感情特徴を統合し，さらに明示的な感情語の有無を考慮した特徴融合を行っている．これにより，暗黙的な手がかりと明示的な表現を同時に利用するモデルが構築され，BERT やその派生モデルをベースとした分類器を上回る性能が報告されている．

近年では，対話文脈における暗黙的感情に注目した研究も進められている．Koga らは，対話履歴から聞き手に喚起される暗黙的感情を直接予測するタスクを提案している [22]．従来の対話感情研究では，次発話に現れる感情表現をラベルとして用いることが多かったが，本研究では感情が明示的に表出されない場合にも着目している．実験結果から，暗黙的感情の予測は明示的感情の予測よりも難易度が高く，文脈理解能力が強く要求されることが示されている．

これらの研究から，暗黙的感情認識は，感情語に依存しない高度な意味理解を必要とするタスクであることが明らかになっている．一方で，多くの研究では感情カテゴリの推定を目的としており，感情の強度や，書き手・読み手といった視点の違いが暗黙的感情の認識に与える影響については，十分に検討されていない．

また，暗黙的感情表現と感情強度との関係は，感情がどの程度言語化されているかという「感情の顕現性」と密接に関係している．

2.4　書き手と読み手の感情の差に着目した研究

Nakagawa らは，書き手と読み手で感情認識が異なる文を検出する BERT ベースの分類器を提案した [23]．このモデルは標準的な BERT の CLS 表現を入力に用いる二値分類器で、書き手と読み手の感情が異なる文を検出し，さらにそのような文に共通する暗黙的表現パターンを統計的に抽出することで，書き手と読み手の主観ズレの原因となる表現を明らかにしている．

Buechel と Hahn らは，感情ラベルは誰の視点でアノテーションするのかによって変わる可能性があるとして，アノテータに書き手 (Writer)，読み手 (Reader)，テキストそのもの (text) の感情を区別して感情ラベルをアノテートさせた [24]．そして，異なる視点が感情アノテーションの結果に与える影響を検討し，実際に視点が異なれば感情ラベルのアノテーション結果も異なることを示した．この研究は感情分析が主観性を扱うタスクであることを表す例である．

Wegge らは，テキスト中に複数存在しうる主体ごとに感情カテゴリを推定する枠組みを提案した [25]．単一のテキストに対しても異なる主体が異なる感情を経験する可能性があるという点を明示し，従来の感情推定の限界を示している．

Wan らは，主観的なラベルが付与されるタスクにおいて，複数のアノテーター間で生じる意見の不一致を単なるノイズではなくアノテーターの背景情報に基づく差異として捉える枠組みを提案した [26]．これにより，複数のアノテーター間のラベル不一致そのものを予測可能な対象として扱うことで，主観的感情評価の多様性を扱った．

2.5　本研究の特色

本研究は，感情強度推定という課題に対して，感情表現の顕現性および書き手という主体の主観性に着目し，これらを明示的にモデル化する点に特色を有する．

従来の感情強度推定研究では，主にテキスト表層に現れる感情語や評価表現を手がかりとして，感情の強さを推定することが中心的な課題とされてきた [3, 13, 15, 16]．これらの研究は，感情強度推定を独立したタスクとして確立し，高性能なモデルの構築に大きく貢献している．一方で，多くの場合，感情がテキスト中に十分に表現されていることを暗黙

の前提としており，感情が明示的に現れない文や，表現が控えめな書き手に対する頑健性については十分に検討されていない．

また，EmoBank，WRIME といったデータセットの登場により，感情分析研究はより多様な感情カテゴリや視点を扱えるようになった．しかし，これらのデータを用いた多くの研究では，書き手と読み手の視点差を利用した明示的なモデル化はされておらず，書き手の暗黙的な感情強度推定そのものを主目的として，視点差を補助的に利用する枠組みは限定的である．

暗黙的感情認識に関する研究では，感情語に依存しない文脈理解の重要性が示されているが，これらは主に感情カテゴリの推定を対象としており，感情の強度や主観的評価の難しさについては十分に扱われていない [19, 20, 21, 22]．特に，感情が暗黙的に表現されている文において，書き手自身の感情強度を推定することは，読み手による解釈とのズレが生じやすく，従来手法では困難な課題である．

さらに，視点による感情のズレを扱った研究は，感情分析における主観性や多様性を明らかにしているが，多くの場合，ズレの検出や分析に焦点が当てられており，ズレの存在を前提とした上で，書き手感情強度推定を改善するモデル設計には踏み込んでいない [23, 24, 25, 26]．

これらの先行研究を踏まえ，本研究の特色は以下のようにまとめられる．

- 第一に，感情がテキスト中に明示的に表現されているか否かを表す感情顕現性という概念を導入し，これを補助タスクとして学習する点に特徴がある．感情顕現性を明示的に扱うことで，モデルが「感情がどの程度言語化されている文なのか」を考慮しながら感情強度を推定できる．

- 第二に，書き手 ID を用いて書き手固有の感情表現傾向を特徴量として獲得し，感情強度推定モデルに統合する点である．これにより，表現が控えめな書き手や，感情を誇張しやすい書き手といった個人差を感情強度推定モデルが考慮できる．

- 第三に，書き手感情強度推定タスクと感情顕現性予測タスクのマルチタスク学習や，各補助タスクを個別に学習し，得られた特徴表現を感情強度推定モデルに統合するモデルを構築する．視点差 (書き手と読み手の差) や表現特性を補助信号として活用する点に新規性がある．これは，書き手と読み手の感情強度の差そのものを目的変数とするのではなく，書き手感情強度推定を主タスクとして改善す

るための手段として感情強度のズレや書き手の感情の顕現性を利用するアプローチである．

以上のように，本研究は，感情強度推定における「主観性」「暗黙性」「視点差」というこれまで個別に扱われてきた要因を統合的に捉え，書き手の暗黙的な感情強度をより安定的に推定する枠組みを提案する．

第3章　提案手法

3.1　タスク設定

本研究では，テキストに含まれる感情表現を対象として，書き手が暗黙的に感じている感情の強度を推定するタスクを扱う．ここでの感情とは，Plutchik の感情モデル [5] に基づく 8 つの感情クラスとする．具体的には，Joy，Sadness，Anticipation，Surprise，Anger，Fear，Disgust，Trust，の 8 つである．上記の感情クラスのそれぞれについて，感情強度を独立に推定する．感情強度は，WRIME データセットの仕様に従い，0,1,2,3 の 4 段階の離散値として定義する．これらの値は，数値が大きいほど該当する感情がより強く表出していることを意味する．強度ラベル 0 は書き手が当該感情をほとんど感じていない，あるいは中立的な状態であることを示し，強度ラベル 3 は書き手が強い感情を抱いている状態を表す．

本タスクにおける入力は，ソーシャルメディアに投稿された単一のテキストであり，トークン列

$$x = w_1, w_2, \ldots, w_n$$

として表現される．一方，出力は感情強度ラベル

$$y \in \{0, 1, 2, 3\}$$

である．すなわち，本タスクは 8 感情それぞれに対して独立に感情強度を推定する多クラス分類問題として定式化される．

書き手に関するタスク設定　本研究では，分類対象となるテキストの書き手が訓練データに含まれているか否かによって，タスクの性質および難易度が大きく異なる点に着目する．実際の応用場面においては，特定の書き手に関する十分な情報が得られる場合と，初めて観測される書き手のテキストを扱う場合の両方が想定される．そのため，本研究では以下の 2 種類のタスク設定を導入する．

「Closed Task」では，テストデータに含まれるテキストの書き手が，訓練データにも出現していると仮定する．この設定では，同一書き手による過去の投稿およびそれに付随する感情強度ラベルを学習に利用できるため，モデルは書き手固有の感情表現の傾向や強度の付け方を学習することが可能である．

「Open Task」では，テストデータの書き手は未知であり，同一書き手によるテキストは訓練データに含まれないと仮定する．この設定では，モデルは書き手に関する事前情報を利用できず，テキスト表層および文脈情報のみに基づいて感情強度を推定する必要がある．そのため，Open Task は，書き手の情報を得られない場合の設定であり，より困難な問題設定であるといえる．

本研究では，これら 2 種類のタスク設定の下でモデルを構築および評価することにより，書き手情報が利用可能な場合と利用できない場合の双方において，提案手法がどの程度有効に機能するかを検証する．

3.2 感情顕現性と書き手を考慮したLate Fusionモデル

3.2.1 概要

入力テキストに対し，テキストの意味情報，感情顕現性，書き手の情報を独立に抽象表現に変換し，それを結合したものを特徴量として書き手の感情強度を推定するモデルを提案する．このモデルを「Late Fusion モデル」と呼ぶ．その概要を図 3.1 に示す．

本モデルは事前学習済み言語モデル BERT[27] をベースとする．入力テキストは BERT を用いてベクトル表現へと変換される．このベクトルはテキストの意味的特徴を表すものとみなせる．

次に，BERT による文の意味的特徴に加え，感情顕現性特徴と書き手特徴の二つの意味的特徴を用いる．感情顕現性特徴とは，書き手の感情がどれだけテキストに明示的に表現されているかを表すものとする．感情顕現性特徴により，感情語や感情表現が読み手にとってわかりにくい場合と，明示的に表現されている場合とをモデル内部で区別できるようになる．図 3.1 における SE-BERT は後述する感情顕現性予測タスクで事前学習された BERT モデルであり，これを用いてテキストから感情顕現特徴を抽出する．

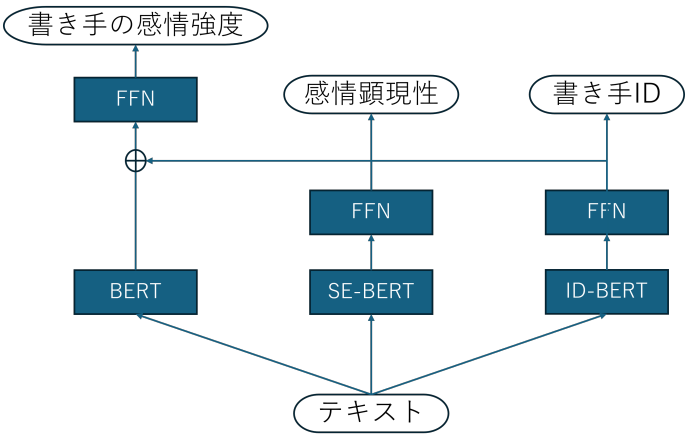

図 3.1: Late Fusion モデル

一方，書き手特徴は，どういった書き手かを表す書き手の特徴である．書き手特徴は，書き手固有の感情傾向であり，感情顕現性と組み合わせることで，書き手が暗黙的な感情表現を好む傾向か明示的な感情表現を好む傾向かを区別し，感情強度推定を補完する役割を果たす．図 3.1 における ID-BERT は後述する書き手 ID 予測タスクで事前学習された BERT モデルであり，これを用いてテキストから書き手特徴を抽出する．

Late Fusion モデルでは，文の意味的特徴，感情顕現特徴，書き手特徴の 3 つのベクトルを統合し，フィードフォワードネットワーク (Feed Forwrd Network;FFN) を通じて書き手の感情強度を推定する．モデルを学習する際には，SE-BERT と ID-BERT は事前学習し，BERT モデルと書き手の感情強度クラスを出力する FFN のパラメタのみ更新する．

Late Fusion モデルは，感情が明示的に表現される場合と暗黙的に表現される場合の双方に対応し，また書き手の特性を考慮した感情強度推定を実現するために設計されている．

なお，このモデルは書き手が既知である Closed Task と未知である Open Task の双方に適用可能であり，書き手情報が利用可能な場合と利用できない場合における感情強度推定を同一の枠組みで扱うことができる．この点は，実運用環境を想定した感情分析において重要な特徴である．

3.2.2　感情顕現性の特徴抽出

既に述べたように，感情顕現性とは，テキスト中において書き手の感情が読み手に伝わりやすいかどうか，つまり感情が明示的に表現されているか否かを表す概念とする．具体的には，感情語や評価語，感情を直接示す形容詞や副詞などが明示的に用いられている場合，感情は顕在化していることが多い．例えば,「本当に腹が立つ」「とても嬉しい」といった表現は，怒りや喜びといった感情が言語表現として直接的に表出されており，読み手にとっても感情の存在や強度を比較的容易に把握できる文である．一方で,「もう二度と会わない」「今日は特別な日だった」のような文では，感情を示す語が明示的には含まれていないものの，文脈や状況に基づいて感情が推測される．このような文では，感情は暗黙的に表現されていると考えられる．

このように，自然言語における感情表現は，明示的な場合と暗黙的な場合が混在しており，両者は質的に異なる情報を含んでいる．明示的感情表現では，モデルは感情語や評価語といった表層的な手がかりを直接利用できるのに対し，暗黙的感情表現では，出来事の描写，語用論的含意，常識的知識などを踏まえた推論が必要となる．この違いは感情強度推定の難易度に大きく影響する．

従来の感情強度推定研究では，感情がどの程度顕在化しているかという観点は十分に考慮されてこなかった．しかし，感情が明示的に表現されている文と，暗黙的に示唆されている文とでは，モデルが利用可能な情報量や情報の性質が大きく異なる．特に暗黙的な感情表現では，感情の存在自体が曖昧であるため，感情強度の推定がより難しい．したがって，感情顕現性を考慮して感情強度を推定することによって，推定精度やモデルの頑健性の改善につながる可能性がある．

感情顕現性のモデル化

本研究では，感情顕現性は書き手と読み手の感情解釈のズレによって定量化できると考える．書き手は，自身の暗黙的感情を前提としてテキストを生成するため，必ずしも感情を明示的に言語化するとは限らない．特に，感情を抑制する傾向のある書き手や，文脈共有を前提としたコミュニケーションでは，感情が暗黙的に表現されることが多い．一方で，読み手は書き手の暗黙的状態を直接知ることができず，テキスト表面の情報や自身の解釈に基づいて感情を推定する．そのため，感情が暗黙的に

表現されている文ほど，書き手と読み手の間で感情の有無や強度に関する解釈のズレが生じやすい．

WRIME データセットにおいても，書き手ラベルと読み手ラベルの間には大きな乖離が存在することが報告されており，特に感情が明示的に表現されていない文では，読み手が感情を過大評価または過小評価する傾向が見られる．このことは，感情顕現性が書き手–読み手の主観的ズレを生み出す重要な要因であることを示唆している．以上から，本研究では，書き手と読み手の感情強度の差が感情顕現性を表すものとし，これにしたがって感情顕現性を表す特徴ベクトルを獲得する．

感情顕現性予測モデルの学習

テキストから感情顕現性を表す抽象表現を取得するモデルを学習する．このモデルを学習するために，入力をテキストとし，そのテキストに対する書き手と読み手の感情強度の違いを予測するタスクを定義する．以下，これを「感情顕現性予測タスク」を呼ぶ．

同タスクにおける出力は，書き手の感情強度と読み手の感情強度とに差がある (ラベル 1) かない (ラベル 0) かの二値のクラスとする．ここで，感情顕現性のクラス 0 は感情強度ラベル 0(中立) とは異なる．すなわち，感情強度が 0 であっても感情が明示的に言及されている場合があり得る一方で，感情強度が正の値を持つ場合でも感情が暗黙的に表現されている可能性がある．

感情顕現性予測モデルは事前学習済み BERT をファインチューニングすることで得る．WRIME データセットを利用し，書き手と読み手の感情強度ラベルを前述の感情顕現性のラベルに変換し，これを訓練データとして BERT をファインチューニングする．以下，このように学習された感情顕現性予測モデルを「SE-BERT」と呼ぶ．SE は Salience of Emotion の略である．BERT モデルでは入力テキストの先頭に CLS という特殊トークンを追加するが，感情顕現性予測モデルから得られる CLS トークンに対応する埋め込み表現は，感情の表出様式に関する高次の特徴を内包しているとみなす．BERT モデルでは CLS トークンの埋め込みとして 768 次元のベクトルが得られるが，本研究ではこれを FFN によって 256 次元に圧縮した上で，感情強度推定モデルの特徴量と結合する．

感情顕現性の導入により，書き手の感情が明示的に表現されない場合においても，感情強度をより正確に推定できることが期待できる．

3.2.3　書き手の特徴抽出

感情表現は，本質的に主観的な現象であり，同一の出来事や状況に対しても，書き手によって感情の表出の仕方や強度は大きく異なる．人は自身の性格，価値観，経験，社会的背景などに基づいて感情を解釈し，それを言語として表現するため，感情表現には必然的に個人差が生じる．例えば，ある書き手は暗黙的な感情を積極的に言語化し,「非常に腹が立った」「本当に嬉しかった」といった強い表現を頻繁に用いる一方で，別の書き手は感情を直接的に表現することを避け,「少し困った」「まあ良かった」といった控えめな表現にとどめる傾向がある．このような表現スタイルの違いは，感情の有無や種類だけでなく，感情の強度がどの程度言語化されるかにも大きな影響を与える．

さらに，感情強度は連続的な尺度であるため，離散的な感情カテゴリ分類と比べて，書き手ごとの主観的基準の違いがより顕著に現れる．すなわち，同じ「怒り」や「喜び」という感情カテゴリであっても，どの程度の強さをもってその感情を表現するかは書き手によって大きく異なる．このことは，感情強度推定が単にテキスト中の感情語を検出する問題ではなく,「その書き手にとって，この表現がどの程度の感情を意味するのか」を推定する問題であることを示している．

しかしながら，従来の感情強度推定研究の多くは，入力テキストのみを手がかりとして感情推定を行うことが多く，書き手固有の感情傾向や表現スタイルは明示的に考慮されていなかった．すなわち,「同一の表現は常に同一の感情強度を持つ」という暗黙の前提に基づいており，書き手間の個人差はモデル化されていない．その結果，同一文であっても，異なる書き手が書いた場合に生じる感情強度の違いを十分に捉えられないという課題が存在する．

特にSNS投稿のような短文テキストでは，文脈情報が限定的であるため，書き手の表現傾向が感情強度解釈に与える影響は相対的に大きくなる．にもかかわらず，書き手情報を考慮しないモデルでは，これらの個人差がすべてノイズとして扱われてしまい，感情強度推定の性能低下を招く可能性がある．この点は，感情強度推定における性能向上のみならず，モデルの解釈性や一貫性の観点からも重要な問題である．

以上の背景から，本研究では感情表現に内在する書き手ごとの主観的差異を明示的に捉える必要があると考え，書き手情報をモデルに組み込むことを試みる．

書き手予測モデルの学習

WRIME データセットでは，それぞれのテキストに対し，書き手と読み手の感情強度ラベルの他に，書き手の ID が付与されている．ここでは，入力をテキストとし，書き手の ID を予測するタスクを「書き手予測タスク」と定義する．また，テキストから書き手 ID を予測するモデルを「書き手予測モデル」と呼ぶ．書き手予測モデルは BERT のファインチューニングによって学習する．これ以降，BERT による書き手予測モデルを「ID-BERT」と記す．CSL トークンに対応する ID-BERT の埋め込みを書き手の特徴を表す抽象表現とみなす．また，BERT によって得られる 768 次元のベクトルを FFN を用いて 256 次元のベクトルに縮退することで，書き手の情報を表す特徴ベクトルを得る．Late Fusion モデルでは，この特徴ベクトルを通常の BERT と SE-BERT から得られる特徴ベクトルと連結し，最終的な特徴ベクトルを得る．

Closed Task と Open Task における書き手特徴の違い

書き手特徴の役割はタスク設定によって異なる．Closed Task においては，テストデータの書き手が訓練データにも出現するため，同一書き手による過去の投稿とそれに対応する書き手 ID をモデルの学習に利用できる．この場合，モデルは書き手固有の感情表現の傾向や感情強度の付与の仕方を学習することが可能であり，感情強度推定の精度向上が期待される．

一方，Open Task においては，テストデータの書き手は未知であり，同一書き手によるデータは訓練データに存在しない．この設定では，訓練データに含まれる多数の書き手のラベルを通じて学習された一般的な書き手の傾向が，未知の書き手に対する推定に間接的に使われると考えられる．すなわち，書き手特徴は Closed Task では個別書き手の情報として，Open Task では集団的な傾向として機能する．

3.2.4 実装の詳細

本項では Late Fusion モデルの実装の詳細について述べる．まず，入力として与えられるのは，SNS 投稿や短文テキストなどの自然言語文である．これらの入力文は，BERT のトークナイザによってサブワード単位に分割され，CLS トークンを先頭に付与した系列としてモデルに入力さ

れる．BERT のエンコーダ層を通過することで，各トークンに対応するベクトル表現が獲得される．また，CLS トークンに対応するベクトルが文全体の意味を表すベクトル表現として得られる．

3.2.1 項で述べたように，Late Fusion モデルでは 3 種類の BERT モデルを使用する．ただし，後述するマルチタスク学習モデルと条件を同じにするため，SE-BERT と ID-BERT については，BERT の出力次元 (768 次元) を一度 256 次元の意味特徴空間に射影し，その後分類 (感情強度の推定，感情顕現性の予測，書き手 ID の予測) を行う．この次元圧縮層により，感情推定に有用な情報を集約した表現を得ることを意図している．

本研究では Late Fusion モデルとして以下の 4 つのモデルを実装した．

- Base モデル

  BERT のみを用いて感情強度を推定するモデルである．BERT によって得られた CLS 表現を用いて，書き手自身の感情強度ラベル (0,1,2,3 の離散値) を直接予測する．このモデルは，テキストの表層的・文脈的情報のみに基づいて感情強度推定を行うベースラインと位置づけられる．

- SE only モデル

  感情顕現性予測モデル (SE-BERT) を別途事前学習し，その出力を感情強度推定に利用する．SE-BERT から得られた CLS ベクトルを 256 次元に圧縮する．この圧縮された感情の顕現性表現を，Base モデルで用いられる CLS 表現 (768 次元) と結合することで，合計 1024 次元の特徴ベクトルを構成する．これにより，文が感情をどの程度明示的に表現しているかという情報を感情強度推定に明示的に利用することが可能となる．

- ID only モデル

  書き手 ID 判定モデル (ID-BERT) を別途事前学習し，その出力を感情強度推定に利用する．ID-BERT から得られた CLS ベクトルを 256 次元に圧縮する．この圧縮された書き手固有の表現を，Base モデルで用いられる CLS 表現 (768 次元) と結合することで，合計 1024 次元の特徴ベクトルを構成する．これにより，文がどの程度書き手固有の表現を反映しているかという情報を感情強度推定に明示的に反映させることが利用できる．

- SE+ID モデル

  SE-BERT と ID-BERT の両方を補助モデルとして用いる．これら補助情報タスクのモデルから得られた CLS ベクトルは 256 次元に圧縮され，これを Base モデルの 768 次元特徴ベクトルと結合することで，最終的に 1280 次元のベクトル表現が得られる．このベクトル表現には，文内容，感情顕現性，および書き手の個人差という三つの異なる観点の情報が含まれている．

上記 4 つのいずれのモデルにおいても，最終的に，統合された特徴ベクトルは，全結合層を通じて感情強度ラベル (0,1,2,3) へと写像される．本研究では感情強度推定を離散値分類問題として定式化しているため，出力層にはソフトマックス関数を用い，クロスエントロピー損失によって学習を行う．

3.3 感情顕現性を考慮したマルチタスク学習モデル

本節では，Late Fusion モデルとは異なるアプローチとして，マルチタスク学習に基づく感情強度推定モデルを提案する．Late Fusion モデルでは，感情顕現性や書き手情報をエンコードするモデルを個別に学習し，それらのモデルのエンコード結果を BERT の特徴量と統合することで書き手感情強度推定の性能向上を図った．一方，マルチタスク学習モデルでは，感情強度推定を主タスク，それと密接に関連する感情顕現性予測タスクを補助タスクとしたマルチタスク学習を実施することで，感情強度予測モデルの性能を高めることを目的とする．

感情強度推定は本質的に主観的なタスクであり，書き手と読み手の解釈の違いや，感情表現の顕現性といった要因が推定結果に大きく影響する．マルチタスク学習は，これらの関連要因を補助タスクとして明示的に与えることで，単一タスク学習では捉えにくい構造的情報をモデル内部の表現に反映できる点に利点がある．言い換えれば，感情顕現性の情報をマルチタスク学習の枠組で感情強度推定モデルに反映させることを試みる．

マルチタスク学習モデルの概要を図 3.2 に示す．モデルの基盤には事前学習済み BERT を用い，入力テキストを符号化する．BERT の最終層から得られる CLS トークンに対応する隠れ状態を文全体の意味表現として

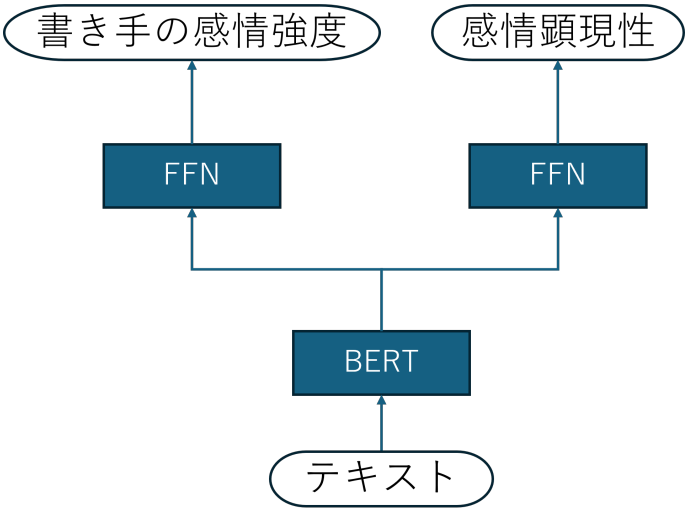

図 3.2: マルチタスク学習モデル

用いる．この意味表現は，書き手感情タスクおよび感情顕現性タスクの両方に入力され，各タスクごとに独立した出力ヘッドを通じて感情強度ならびに感情顕現性の分類を行う．

書き手の感情と感情顕現性は，同一のテキストに対する異なる視点からの感情評価であり，必ずしも一致しない．そのため，本モデルでは，両タスクで中間層を分離することで，それぞれのタスクの特徴表現を学習できるよう設計している．学習時には，書き手の感情強度に対する損失を主タスクとし，感情顕現性に対する損失を補助タスクとして加えた重み付き損失関数を最小化する．テスト時には，感情強度推定用の BERT モデルのみを用いて入力テキストの感情強度を推定する．

Late Fusion モデルとの違い　Late Fusion モデルとマルチタス学習モデルの最大の違いは，補助情報の利用方法にある．Late Fusion モデルでは，感情顕現性や書き手情報を事前に学習した特徴として抽出し，感情強度推定モデルに統合する．一方，マルチタス学習モデルでは，これらの情報を予測対象そのものとして同時に学習し，共有表現を通じて暗黙的に活用する．ただし，マルチタスク学習モデルでは書き手の特性は考慮されていない．

Late Fusion モデルは補助情報を明示的に制御できる一方で，マルチタス学習モデルはタスク間の相互作用を柔軟に学習できるという特性を持つ．本研究では，これら 2 つのアプローチを比較することで，感情強度

推定における補助情報の効果的な利用方法について検討する.

第4章　評価

本章では，本研究で提案する感情顕現性ならびに書き手の情報を考慮した感情強度推定モデルの有効性を定量的に評価するために行った実験について述べる．本研究の目的は，書き手の感情が必ずしもテキスト中に明示的に表現されないという前提のもと，書き手の意図した感情強度をより正確に推定することである．そのため，本章では，従来の単純な感情強度推定手法と比較しつつ，感情顕現性予測および書き手情報の導入が性能に与える影響を多角的に検証する．

4.1　実験設定

本節では評価実験の設定について述べる．具体的には，使用したデータセット，タスク設定，評価指標について順に説明する．

4.1.1　データセット

本研究では，書き手の感情強度がアノテーションされた既存の感情分析データセット WRIME[11] を用いて実験を行った．既に述べたように，本データセットは，Joy，Sadness，Anticipation，Surprise，Anger，Fear，Disgust，Trust の 8 種類の基本感情を対象としており，各テキストに対して感情ごとの強度情報が付与されている．

感情強度のラベルは，書き手によって付与されており，テキスト表層に現れる感情語の有無に依存せず，書き手が意図した感情の強さがアノテーションされている．そのため，感情語が明示的に含まれるテキストだけでなく，皮肉表現や文脈依存的な表現を含むテキストに対しても，書き手の真の感情がラベル付けされている．また，WRIME では書き手とは異なる第 3 者 (読み手) によっても感情強度ラベルが付与されている．書き手と読み手の感情強度ラベルの差は感情顕現性予測モデルの学習に用いる．

表 4.1: Open Task における書き手感情ラベル分布

Writer Emotion	Train					Test				
	0	1	2	3	Total	0	1	2	3	Total
Joy	16783	4205	3704	3308	28000	2272	255	347	626	3500
Sadness	17967	4670	2977	2386	28000	2500	395	346	259	3500
Anticipation	17638	4740	3349	2273	28000	2351	331	496	322	3500
Surprise	20312	3704	2291	1693	28000	2576	309	392	223	3500
Anger	24505	1692	987	816	28000	3109	99	90	202	3500
Fear	22965	2742	1379	914	28000	2966	245	141	148	3500
Disgust	21888	2940	1811	1361	28000	3100	115	90	195	3500
Trust	21942	3090	1878	1090	28000	3253	62	94	91	3500

表 4.2: Closed Task における書き手感情ラベル分布

Writer Emotion	Train					Test				
	0	1	2	3	Total	0	1	2	3	Total
Joy	16934	3988	3575	3503	28000	2113	486	478	423	3500
Sadness	18266	4453	2950	2331	28000	2208	593	397	302	3500
Anticipation	17759	4546	3363	2332	28000	2241	565	433	261	3500
Surprise	20330	3549	2367	1754	28000	2545	450	310	195	3500
Anger	24531	1616	970	883	28000	3079	185	121	115	3500
Fear	23044	2657	1372	927	28000	2889	336	167	108	3500
Disgust	22202	2723	1708	1367	28000	2766	372	190	172	3500
Trust	22286	2858	1785	1071	28000	2795	361	219	125	3500

4.1.2 Open Task/Closed Task 設定

本実験では，特に書き手の情報の有効性を調べるために Open Task と Closed Task の 2 種類の設定を用いた．それぞれのタスクの定義を再度説明する．

Closed Task では，学習データおよび評価データにおいて同一の書き手が含まれる．すなわち，モデルは学習時に観測した書き手の情報を推論時にも利用できる設定である．この設定は，特定のユーザを対象とした感情分析や，パーソナライズされた感情推定を想定している．

一方，Open Task では，評価時に現れる書き手は学習時には観測されていない未知の書き手である．この設定では，書き手固有の情報に依存しすぎたモデルは性能が低下する可能性がある．そのため，Open Task における結果はモデルの汎化性能を評価する重要な指標となる．

本研究では，提案手法が未知書き手に対しても書き手情報や感情顕現性が有効に機能するかを検証するため，両設定での比較を行う．データ

表 4.3: Open Task と Closed Task における書き手の数

Task	Train	Test
Open	54	6
Closed	60	60

セットの統計を表 4.1,4.2,4.3 に示す．本実験では WRIME データセットを 8:1:1 に分割し，訓練，開発，テストデータとした．これらの表における Train と Test は訓練データとテストデータの統計を示す．開発データの量及びラベルの分布はテストデータと大きな差はないため，省略する．

4.1.3 評価指標

評価指標としては Macro F1 を用いた．分類問題では Accuracy も評価指標としてよく用いられる．Accuracy は全体の正解率を示す指標であり，モデルの総合的な性能を把握するのに有用である．一方で，感情クラス間のデータ数に偏りがある場合，高頻度の感情クラスを常にもしくは頻繁に予測するモデルの正解率が不当に高く評価されるという欠点がある．そのため，本研究では，各感情クラスに対する分類性能を均等に評価する Macro F1 を評価指標とする．感情強度推定においては，特定の感情強度のみを高精度に推定するのではなく，全ての感情強度に対して安定した推定性能を発揮することが重要なためである．

4.2 感情顕現性判定モデルの評価

本節では感情強度推定モデルの補助モデルである感情顕現性判定モデルの性能を評価する．

4.2.1 Late Fusion モデルにおける感情顕現性判定モデルの評価

本項では Late Fusion モデルにおける感情顕現性判定モデルの性能評価を行う．感情顕現性判定モデルは，テキスト中において特定の感情が明示的に表現されているか否かを判定することを目的としており，後段の感情強度推定モデルにおいて感情表現の解釈を補助する役割を担う．

表 4.4: 感情顕現性ラベルの分布 (二値分類，Open Task)

Emotion	Train			Test		
	0	1	Total	0	1	Total
Joy	20588	7412	28000	2782	718	3500
Sadness	22331	5669	28000	2978	522	3500
Anticipation	21734	6266	28000	2827	673	3500
Surprise	23778	4222	28000	3063	437	3500
Anger	27264	736	28000	3441	59	3500
Fear	25388	2612	28000	3200	300	3500
Disgust	24990	3010	28000	3336	164	3500
Trust	27253	747	28000	3481	19	3500

表 4.5: 感情顕現性ラベルの分布 (二値分類，Closed Task)

Emotion	Train			Test		
	0	1	Total	0	1	Total
Joy	20755	7245	28000	2600	900	3500
Sadness	22526	5474	28000	2754	746	3500
Anticipation	21817	6183	28000	2763	737	3500
Surprise	23860	4140	28000	2973	527	3500
Anger	27295	705	28000	3413	87	3500
Fear	25381	2619	28000	3180	320	3500
Disgust	25191	2809	28000	3137	363	3500
Trust	27276	724	28000	3410	90	3500

表 4.4，4.5 に Open Task，Closed Task における感情顕現性ラベルの分布を示す．表 4.6 に 8 つの感情クラスのそれぞれにおける感情顕現性判定モデルの Macro F1 を示す．

全体として Joy，Sadness，Anticipation といった感情では比較的高い性能が得られていることが分かる．特に Joy においては，Open Task で 0.810，Closed Task で 0.817 と安定した高い性能を示している．一方で，Anger，Fear，Trust といった感情では性能が相対的に低く，感情顕現性の判定が難しいことが確認された．

Open Task と Closed Task を比較すると，多くの感情において Closed Task の方が性能が向上している．特に，Fear では Open Task の 0.549 に対し Closed Task では 0.656 と大きな向上が見られる．これは，書き手固

表 4.6: 感情顕現性判定モデルの評価 (二値分類)

Task	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust
Open	0.810	0.714	0.758	0.700	0.572	0.549	0.626	0.499
Closed	0.817	0.730	0.781	0.664	0.548	0.656	0.585	0.575

有の感情表現傾向がモデルに学習され，感情が明示的に表出しているかどうかの判定がしやすくなったと考えられる．一方で，Anger，urprise，Disgust では Closed Task で性能が低下しており，必ずしも書き手固有の感情傾向が一貫して有効とは限らないことが示された．

また，Trust においては Open Task において性能が低く，感情顕現性判定が特に困難であることが分かる．Trust は，他の感情と比較して明確な感情語が存在しにくく，また，書き手特徴を必要とされることが多い感情であるためだと考えられる．

以上の結果から，感情顕現性判定モデルは，感情の種類によって性能に大きな差が生じることが確認された．特に Anger，Fear，Disgust，Trust については書き手と読み手で感情のズレが発生しやすく感情強度推定も難しいことが予想される．

4.2.2 マルチタスク学習モデルにおける感情顕現性判定モデルの評価

本項ではマルチタスク学習モデルにおける感情顕現性判定モデルの性能評価を行う．マルチタスク学習モデルにおいて，当初は感情顕現性を Late Fusion モデルと同じように二値分類として扱うことを検討した．しかし，この設定では学習が安定せず，損失が十分に低下しないことが確認された．

そこでマルチタスク学習モデルでは，書き手の感情と読み手の感情の差の絶対値 (0,1,2,3) を感情顕現度のラベルとして学習を行った．これは書き手と読み手の感情が一致しているほど 0 に近くなり，乖離しているほど大きくなる．

ここではマルチタスク学習終了後，2 つのタスクで共有する BERT モデルと感情顕現度クラスを出力する FFN を用いて感情顕現度を推定したときの評価結果を報告する．表 4.7，4.8 に感情顕現性ラベル (書き手と読み手の感情強度の差の絶対値) の分布を示す．表 4.9 に 8 つの感情クラスのそれぞれにおける感情顕現度推定モデルの Macro F1 を示す．

表 4.7: 感情顕現性ラベルの分布 (マルチクラス分類，Open Task)

Emotion	Train					Test				
	0	1	2	3	Total	0	1	2	3	Total
Joy	18254	6944	2243	559	28000	2290	530	492	188	3500
Sadness	18594	6319	2325	762	28000	2498	615	301	86	3500
Anticipation	17340	7498	2547	615	28000	2246	698	466	90	3500
Surprise	19037	6373	2053	537	28000	2383	702	350	65	3500
Anger	24577	1999	944	480	28000	3096	133	119	152	3500
Fear	20919	5467	1279	335	28000	2687	592	165	56	3500
Disgust	21178	4684	1645	493	28000	2899	362	139	100	3500
Trust	21579	3711	1891	819	28000	3207	111	101	81	3500

Open Task においては，Joy(0.317)，Anticipation(0.315)，Sadness(0.299) といった感情で比較的高い性能が得られている．一方で，Anger(0.235) は最も低い値を示しており，書き手・読み手間の感情の認識の差を捉えることが特に難しいことが分かる．

Closed Task では，全体的に Open Task と同程度，あるいは一部の感情で性能向上が見られる．特に Trust では 0.342 と最も高い F1 を示しており，書き手情報が既知である条件において，感情顕現度の推定が比較的安定して行われていることが分かる．Late Fusion モデルにおいても Trust は Closed Task と Open Task の乖離が大きかったことから，この感情の認識には書き手の特徴が重要であると考えられる．

一方で，Trust を除く多くの感情においては，Open Task と Closed Task の性能差は比較的小さく，マルチタスク学習により書き手ラベルと感情顕現性を同時に学習することが，書き手情報の有無による影響を一定程度緩和していると考えられる．

4.3 Late Fusion モデルの評価

本節では，Late Fusion モデルについて，Open Task および Closed Task のそれぞれにおける実験結果を示す．本実験では，3.2.4 項で述べた Base，SE only，ID only，SE+ID の 4 つのモデルを比較した．

表 4.8: 感情顕現性ラベルの分布 (マルチクラス分類，Closed Task)

Emotion	Train					Test				
	0	1	2	3	Total	0	1	2	3	Total
Joy	18254	6944	2243	559	28000	2290	530	492	188	3500
Sadness	18594	6319	2325	762	28000	2498	615	301	86	3500
Anticipation	17340	7498	2547	615	28000	2246	698	466	90	3500
Surprise	19037	6373	2053	537	28000	2383	702	350	65	3500
Anger	24577	1999	944	480	28000	3096	133	119	152	3500
Fear	20919	5467	1279	335	28000	2687	592	165	56	3500
Disgust	21178	4684	1645	493	28000	2899	362	139	100	3500
Trust	21579	3711	1891	819	28000	3207	111	101	81	3500

表 4.9: 感情顕現性判定モデルの評価 (マルチクラス分類)

Task	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust
Open	0.317	0.299	0.315	0.291	0.235	0.290	0.302	0.297
Closed	0.313	0.303	0.301	0.283	0.269	0.289	0.307	0.342

4.3.1 Open Task における Late Fusion モデルの実験結果

表 4.10 に Open Task における各モデルの感情クラス別 Macro F1 を示す.「Average」は 8 つの感情クラスの平均である.

まず，書き手感情ラベルのみで学習されたベースラインモデル (Base) では，Joy において F1 が 0.367 と比較的高い値を示した一方で，Fear においては 0.252，Trust においては 0.241 など一部の感情クラスでは低い値にとどまった．全体として，感情クラス間で性能のばらつきが大きく，テキストのみを用いた推定には限界があることが確認された.

次に，書き手 ID 特徴のみを統合したモデル (ID only) では，Anticipation においては 0.333，Trust においては 0.264 といったように，一部の感情クラスでベースラインを上回る性能が得られた．一方で，Anger においては 0.235，Disgust においては 0.259 など，いくつかの感情クラスでは性能低下が見られ，書き手特徴がすべての感情に対して一様に有効であるわけではないことが示された.

感情顕現性 (SE) のみを統合したモデル (SE only) では，多くの感情クラスで安定した性能向上が確認された．特に，Joy においては 0.380，Surprise においては 0.300，Fear においては 0.268，Disgust においては 0.280 といったように，ベースラインを上回る F1 を達成している．これらの感情クラスは，必ずしも明示的な感情語を伴わない表現が多い点で共通しており，

表 4.10: Open Task における Late Fusion モデルの評価

Model	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Average
Base	0.367	0.332	0.305	0.286	0.270	0.252	0.276	0.241	0.291
SE only	0.380	0.319	0.326	0.300	0.260	0.268	0.280	0.248	0.298
ID only	0.368	0.321	0.333	0.285	0.235	0.262	0.259	0.264	0.291
ID+SE	0.382	0.320	0.325	0.297	0.266	0.253	0.274	0.250	0.296

表 4.11: Closed Task における Late Fusion モデルの評価

Model	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Average
Base	0.382	0.361	0.334	0.311	0.292	0.301	0.270	0.319	0.321
SE only	0.416	0.369	0.364	0.325	0.262	0.292	0.316	0.347	0.336
ID only	0.431	0.405	0.329	0.311	0.247	0.296	0.292	0.335	0.331
ID+SE	0.433	0.393	0.391	0.347	0.280	0.336	0.326	0.350	0.357

感情顕現性が有効に働いたと推測できる．

最後に，書き手ID特徴と感情顕現性特徴の双方を統合したモデル (SE+ID) では，Joy においては 0.382，Surprise においては 0.297 と，最も高い F1 を示した．一方で，Sadness においては 0.320，Anger においては 0.266 といったように，SE 単独モデルを下回る結果も得られており，全感情クラスにおいて一貫した改善は確認されなかった．

さらに，8 つの感情クラスの F1 を平均した結果を見ると，ベースラインモデルは 0.291，ID only モデルは 0.291 とほぼ同等であったのに対し，SE only モデルは 0.298，SE+ID モデルは 0.296 となり，ベースラインモデルを上回った．SE only モデルが全体平均として最も高い性能を示した．

この結果から，Open Task においては，書き手ラベルに基づく特徴よりも，書き手に依存しない感情顕現性情報の方が，全体として安定した性能向上に寄与することが示されたといえる．一方で，SE と ID を同時に用いた場合でも，SE only モデルを大きく上回る性能は得られておらず，両特徴の単純な結合による相乗効果は限定的であることが確認された．

4.3.2 Closed Task における Late Fusion モデルの実験結果

表 4.11 に，Closed Task における各モデルの感情クラス別 Macro F1 を示す．まず，書き手感情ラベルのみで学習されたベースラインモデル (Base) では，Joy においては 0.382，Sadness においては 0.361，Anticipation においては 0.334 といったように，比較的出現頻度が高く，表現が明示的に

なりやすい感情クラスで比較的高い F1 スコアが得られた．一方で，Anger においては 0.292，Disgust においては 0.270 といったように，一部の感情クラスについては性能が低く，書き手情報のみでは感情の判定が難しいことが示唆される．

次に，SE のみを追加したモデル (SE only) では，Joy においては 0.416，Sadness においては 0.369，Trust においては 0.347 といったように，一部の感情クラスでベースラインを上回る結果が得られた．一方，Anger においては 0.262，Fear においては 0.292 といったように，性能の低下が見られた感情クラスもあり，SE の導入が必ずしも全感情に対して一貫した改善をもたらすわけではないことが確認された．

ID のみを追加したモデル (ID only) では，Joy においては 0.431，Sadness においては 0.405 といったように，主要感情で大きな性能向上が見られた．一方，Anticipation においては 0.329，Surprise においては 0.311 といったように，これらの感情カテゴリではベースラインと同程度，あるいはそれ以下の性能に留まっている．

最後に，SE と ID の両方を結合したモデル (SE+ID) では，Joy においては 0.433，Sadness においては 0.393，Anticipation においては 0.391，Surprise においては 0.347 といったように，複数の感情カテゴリにおいて最も高い性能を示した．ただし，Anger(0.280) ではベースラインよりも F1 が低下し，全感情カテゴリでの一様な改善には至っていない．

さらに 8 つの感情クラスの F1 の平均を見ると，ベースラインモデルは 0.321，SE only モデルは 0.336，ID only モデルは 0.331 であったのに対し，ID+SE モデルは 0.357 と最も高い値を示した．この結果から，Closed Task においては，書き手ラベルと感情顕現性の双方を統合することで，感情強度推定の性能を全体として大きく向上させられることが確認された．

4.3.3 Open Task と Closed Task の結果の比較

本項では Open Task と Closed Task の実験結果を比較する．表 4.10 と表 4.11 より，Closed Task は Open Task と比較して，全体的に高い F1 を示す傾向が確認された．この傾向は，全感情クラスの平均値においても確認できる．推定対象のテキストの書き手が書いた他のテキストが訓練データに存在することで，書き手固有の情報が適切にモデルに反映され，より正確に感情強度が推定できたと考えられる．また，Open Task では感情クラス毎の Macro F1 のばらつきが大きいが，Closed Task では相対

表 4.12: Open Task におけるマルチタスク学習モデルの評価

Model	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Average
Base	0.367	0.332	0.305	0.286	0.270	0.252	0.276	0.241	0.291
Multi Task	0.356	0.321	0.310	0.271	0.275	0.258	0.271	0.257	0.290

的にばらつきは小さく，感情強度推定モデルはどの感情クラスに対しても安定した性能を発揮した.

4.4 マルチタスク学習モデルの評価

本節ではマルチタスク学習モデルの実験結果を報告し，考察を行う.

4.4.1 Open Task におけるマルチタスク学習モデルの実験結果

表 4.12 に Open Task におけるマルチタスク学習モデルの各感情クラスの Macro F1 を示す.

Joy では 0.356，Sadness では 0.321，Anticipation では 0.310 といったように，これらの感情クラスについては比較的高い F1 が得られた．一方で，Surprise では 0.271，Fear では 0.258，Anger では 0.275，Trust においては 0.257 といったように，F1 が低い感情クラスもあった.

ベースラインモデルと比較すると，8 つの感情クラスのうち 4 つについては，マルチタスク学習モデルはベースラインを上回るものの，8 つの感情クラスの平均ではベースラインにやや劣る結果が得られた

全体として，マルチタスク学習モデルは，感情強度推定において一定の性能を示したものの，感情クラスごとに安定して高性能を達成するには至らず，感情顕現性を考慮しないモデルからの顕著な改善も見られなかった.

4.4.2 Closed Task におけるマルチタスク学習モデルの実験結果

表 4.13 に Closed Task におけるマルチタスク学習モデルの各感情クラスにおける Macro F1 を示す.

表 4.13: Closed Task におけるマルチタスク学習モデルの評価

Model	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Average
Base	0.382	0.361	0.334	0.311	0.292	0.301	0.270	0.319	0.321
Multi Task	0.382	0.364	0.366	0.299	0.284	0.295	0.286	0.349	0.328

マルチタスク学習モデルは，Joy において 0.382，Sadness において 0.364，Anticipation において 0.366，Trust では 0.349 といったように，多くの感情クラスで比較的高い性能を示した．一方で，Surprise では 0.299，Anger では 0.285，Disgust では 0.286，Fear では 0.295 といったように，一部の感情クラスでは性能が伸びなかった．

ベースラインモデルと比較すると，8 つの感情クラスのうち 4 つのクラス (Sadness，Anticipation，Disgust，Trust) で改善が見られ，8 つの感情クラスの平均でも F1 が 0.007 ポイント向上した．したがって，マルチタスク学習によって感情顕現性を考慮することの有効性が認められた．とはいえ，ベースラインモデルとマルチタスク学習モデルの差が小さく，感情顕現性予測タスクを補助タスクとするマルチタスク学習の効果は限定的であった．

4.4.3 考察

マルチタスク学習モデルがベースラインモデルと比べて感情強度推定の性能が大きく改善しなかった要因として，書き手の感情強度推定タスクと感情顕現性推定タスクの間で，有効な特徴が異なり，その結果として勾配競合が発生した可能性が考えられる．

書き手の感情強度推定は，文全体に内在する主観的感情の強さを捉えることを目的とするのに対し，感情顕現性推定は，感情がどの程度明示的に表現されているかという表層的・表現的側面に注目するタスクである．このように，両タスクは関連性を持つ一方で，モデルが捉えるべき特徴が必ずしも一致していない可能性がある．本研究のマルチタスク学習の設定では，CLS トークンに対するベクトル表現を共有表現として用い，その後にタスクごとのヘッドによって分類クラスを決定している．この構成では，両タスクの損失が同時に逆伝播されるため，一方のタスクにとって有益な更新が，もう一方のタスクにとっては性能低下を引き起こす方向に働く可能性がある．特に，感情顕現性タスクは，感情語の有無や明示的な表現パターンに強く依存する傾向があり，文脈全体から感

情の強度を推定しようとする書き手の感情強度推定タスクとは異なる特徴に着目している可能性がある．

以上をまとめると，感情強度と感情顕現性という関連はあるが異質なタスクを単純に同時学習させるだけでは，必ずしも性能向上につながらないことがわかった．

4.5 Late Fusionモデルとマルチタスク学習モデルの比較

本節では，Late Fusion モデルとマルチタスク学習モデルの違いについて論じる．

Open Task におけるベースライン，Late Fusion モデル，マルチタスク学習モデルの感情クラス別 Macro F1 を表 4.14 に再掲する．Late Fusion モデルとしては Macro F1 が最も高かった SE only モデルの結果を載せている．

マルチタスク学習モデルは多くの感情クラスにおいて Late Fusion モデルよりも F1 スコアが低かった．特に，Joy や Sadness においては，マルチタスク学習モデルが一定の F1 を示しているものの，SE を特徴として統合した Late Fusion モデルが達成した性能と同等，あるいはそれを下回る結果となった．

一方で Anger や Trust のような，感情顕現性判定のデータセットにおいて「書き手と読み手と感情強度に差がある」ラベルが付与されたサンプルが極端に少ない感情クラスにおいては，SE only モデルよりもマルチタスク学習モデルの方が高い性能を示した．この結果は，明示的な特徴統合が常に有利であるとは限らず，感情顕現性判定のデータセットにおけるラベルの偏りが大きい感情クラスにおいてはマルチタスク学習モデルが有効に機能しうることを示唆している．

次に，Closed Task における各手法の感情クラス別 Macro F1 を表 4.15 に再掲する．Late Fusion モデルとしては Macro F1 が最も高かった SE+ID モデルの結果を載せている．

SE+ID モデルは，Joy で 0.433，Sadness で 0.393，Anticipation で 0.391，Surprise で 0.347，Fear で 0.268，Disgust で 0.280 と，いずれもマルチタスク学習モデルを大きく上回る F1 を示した．特に，Joy および Surprise では 0.05 ポイント近い差が確認され，書き手情報と感情顕現性情報を明示的に統合する効果が顕著に現れている．

表 4.14: Open Task における Late Fusion モデルとマルチタスク学習モデルの比較

Model	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Average
Base	0.367	0.332	0.305	0.286	0.270	0.252	0.276	0.241	0.291
SE only	0.380	0.319	0.326	0.300	0.260	0.268	0.280	0.248	0.298
Multi Task	0.356	0.321	0.310	0.271	0.275	0.258	0.271	0.257	0.290

表 4.15: Closed Task における Late Fusion モデルとマルチタスク学習モデルの比較

Model	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Average
Base	0.382	0.361	0.334	0.311	0.292	0.301	0.270	0.319	0.321
SE+ID	0.433	0.393	0.391	0.347	0.280	0.336	0.326	0.350	0.357
Multi Task	0.382	0.364	0.366	0.299	0.284	0.295	0.286	0.349	0.328

一方で，Open Task と同じように，Anger や Trust といった感情顕現性判定のデータセットにおけるラベルの偏りが大きい感情クラスにおいては，Late Fusion モデルが同等か下回る結果となっている．Trust については書き手特徴の導入により一定の改善が見られたものの，Anger においては書き手特徴を加えても大きな改善は得られなかった．これは，Anger においては感情顕現性，書き手情報のいずれも感情強度推定に対して有効な手がかりとなりにくい可能性を示唆している．

以上の結果から，Closed Task においても，書き手感情強度推定と感情顕現性判定を同時に最適化するマルチタスク学習より，補助タスクで獲得した特徴を主タスクに統合する Late Fusion モデルの方が，多くの感情クラスにおいて安定した性能向上をもたらすことが確認された．ただし，データセットにおいて感情顕現性ラベルの分布に偏りがある感情クラスでは，マルチタスク学習が相対的に有利となる場合があることも示され，感情クラスの特性に応じた学習戦略の選択が重要であるといえる．

第5章　おわりに

5.1　本論文のまとめ

本研究では，書き手と読み手の主観性の違いが感情理解に影響を与えるという問題意識のもと，感情強度推定タスクにおいて，書き手依存情報と書き手非依存情報をどのように活用すべきかを検討した．特に，感情がテキスト中にどの程度明示的に表現されているかを表す「感情顕現性」に着目し，これを補助特徴として導入することで，感情強度推定の性能向上を図った．

本研究では，書き手感情ラベルを正解とする感情強度推定タスクを対象とし，書き手特徴および感情顕現性特徴 (SE) を用いた複数のモデルを構築した．さらに，推論時のテキストの書き手のサンプルが訓練データに含まれる Closed Task と，未知の書き手が書いたテキストを対象とする Open Task の二つの設定を導入し，特徴の有効性がタスク条件によってどのように変化するかを分析した．

実験の結果，Open Task においては，感情顕現性特徴が最も安定して有効であることが確認された．未知書き手を対象とする状況では，書き手特徴は十分に汎化できず，感情クラスによっては性能低下を引き起こすことが示された．これに対し，感情顕現性特徴は書き手に依存しない表現レベルの情報を捉えるため，一貫した性能向上をもたらした．

一方，Closed Task では，書き手特徴が感情強度推定において重要な役割を果たすことが明らかとなった．学習時に観測された書き手の感情表現傾向を推論時にも利用できるため，書き手特徴を用いたモデルは，多くの感情クラスにおいて高い性能を示した．さらに，書き手特徴と感情顕現性特徴を組み合わせたモデルでは，書き手固有の感情表現傾向と，感情の明示性という異なる側面を同時に考慮することの有効性が確認された．

これらの結果から，感情強度推定においては，書き手情報が利用可能な環境と利用不可能な環境とで，有効な特徴が本質的に異なることが示された．また，本研究は書き手に依存しない汎用的な補助特徴として感

情顕現性を導入し，タスク条件に応じた特徴設計の重要性を実証的に示したすなわち，未知の書き手を対象とする Open Task においても頑健に機能する特徴を発見したことは，実運用を想定した感情理解モデルの設計に対して有用な知見を提供するものであった．

5.2 今後の課題

本研究では，感情顕現性特徴および書き手特徴を用いることで，感情強度推定における性能向上の可能性を示したが，いくつかの課題も残されている．本節では，本研究の限界を踏まえ，今後の発展に向けた課題について述べる．

第一に，Open Task における書き手特徴の活用方法の改善が挙げられる．本研究の結果から，未知の書き手を対象とする Open Task では，書き手特徴が十分に汎化せず，感情クラスによっては性能低下を引き起こすことが明らかとなった．これは，書き手を離散的なラベルとして扱う現在の表現方法が，未知の書き手の感情表現傾向を適切に捉えられていないことに起因すると考えられる．

第二に，感情顕現性判定モデル自体の精度向上と，感情強度推定への影響分析が課題として挙げられる．本研究では，感情顕現性を二値分類問題として定式化し，得られた中間特徴を感情強度推定モデルに統合したが，顕現性判定の誤りが下流タスクにどの程度影響を及ぼしているかについては十分に分析できていない．今後は，感情顕現性の推定精度と感情強度推定性能の関係を定量的に分析し，感情顕現性推定の信頼度を考慮した統合手法を検討することが望まれる．

第三に，読み手の主観性を考慮した拡張が挙げられる．本研究では，書き手が意図した感情強度の推定に焦点を当てたが，実際のコミュニケーションにおいては，読み手がどのように感情を受け取るかも重要な要素である．今後は，書き手感情と読み手感情の相互作用を明示的にモデル化し，感情同士の相乗効果から双方の視点へと拡張することで，より実用的かつ理論的に意義のある感情理解モデルの構築を検討するべきである．

第四に，特徴統合方法の高度化も今後の課題である．本研究では，補助タスクから得られた特徴を単純に連結する手法を採用したが，特徴間の重要度や相互関係を動的に調整する仕組みは導入していない．Attention 機構や重み付け学習を用いることで，感情クラスや入力文脈に応じて，書き手情報と感情顕現性情報の寄与度を制御する手法を検討する余地がある．

最後に，使用するデータセットの拡大が挙げられる．本研究では，特定のデータセットを用いて評価を行ったが，異なるデータセットを用いて書き手特徴モデルを構築することで，Open Task においても頑健に働く汎用的な書き手特徴の獲得が期待される．

参考文献

[1] Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. *Emotion Classification Using Massive Examples Extracted from the Web.* Coling 2008 Organizing Committee, Manchester, UK, August 2008.

[2] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[3] Saif Mohammad and Felipe Bravo-Marquez. *WASSA-2017 Shared Task on Emotion Intensity.* Association for Computational Linguistics, Copenhagen, Denmark, September 2017.

[4] Paul Ekman. *An argument for basic emotions*, Vol. 6. Routledge, 1992.

[5] ROBERT PLUTCHIK. Chapter 1 - a general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pp. 3–33. Academic Press, 1980.

[6] James A. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, Vol. 110, No. 1, pp. 145–172, 2003.

[7] Klaus R. Scherer and Harald G. Wallbott. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, Vol. 66, No. 2, pp. 310–328, 1994.

[8] Saima Aman and Stan Szpakowicz. Identifying expressions of emotion in text. pp. 196–205, 09 2007.

[9] Carlo Strapparava and Rada Mihalcea. SemEval-2007 task 14: Affective text. In Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors, *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 70–74, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[10] Saif Mohammad and Felipe Bravo-Marquez. Emotion intensities in tweets. In Nancy Ide, Aurélie Herbelot, and Lluís Màrquez, editors, *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pp. 65–77, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[11] Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2095–2104, Online, June 2021. Association for Computational Linguistics.

[12] Haruya Suzuki, Sora Tarumoto, Tomoyuki Kajiwara, Takashi Ninomiya, Yuta Nakashima, and Hajime Nagahara. Emotional intensity estimation based on writer's personality. In Yan Hanqi, Yang Zonghan, Sebastian Ruder, and Wan Xiaojun, editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pp. 1–7, Online, November 2022. Association for Computational Linguistics.

[13] Egor Lakomkin, Chandrakant Bothe, and Stefan Wermter. Gradascent at emoint-2017: Character and word level recurrent neural net-

work models for tweet emotion intensity detection. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, 2017.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[15] Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. Vad-assisted multitask transformer framework for emotion recognition and intensity prediction on suicide notes. *Information Processing & Management*, Vol. 60, No. 2, p. 103234, 2023.

[16] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054, Online, July 2020. Association for Computational Linguistics.

[17] Sven Buechel and Udo Hahn. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 578–585, Valencia, Spain, April 2017. Association for Computational Linguistics.

[18] Laura-Ana-Maria Bostan and Roman Klinger. An analysis of annotated corpora for emotion classification in text. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2104–2119, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[19] Roman Klinger, Orphée De Clercq, Saif M. Mohammad, and Alexandra Balahur. Iest: Wassa-2018 implicit emotions shared task, 2018.

[20] Hanane Boutouta, Abdelaziz Lakhfif, Ferial Senator, and Chahrazed Mediani. A transformer-based hybrid model for implicit emotion recognition in arabic text. *Engineering, Technology & Applied Science Research*, Vol. 15, No. 3, pp. 23834–23839, Jun 2025.

[21] Fereshteh Khoshnam, Ahmad Baraani-Dastjerdi, and M. J. Liaghatdar. Cefer: A four facets framework based on context and emotion embedded features for implicit and explicit emotion recognition, 2022.

[22] Yurie Koga, Shunsuke Kando, and Yusuke Miyao. Forecasting implicit emotions elicited in conversations. In Saad Mahamood, Nguyen Le Minh, and Daphne Ippolito, editors, *Proceedings of the 17th International Natural Language Generation Conference*, pp. 145–152, Tokyo, Japan, September 2024. Association for Computational Linguistics.

[23] Tsubasa Nakagawa, Shunsuke Kitada, and Hitoshi Iyatomi. Expressions causing differences in emotion recognition in social networking service documents. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM ' 22, p. 4349–4353. ACM, October 2022.

[24] Sven Buechel and Udo Hahn. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In Nathan Schneider and Nianwen Xue, editors, *Proceedings of the 11th Linguistic Annotation Workshop*, pp. 1–12, Valencia, Spain, April 2017. Association for Computational Linguistics.

[25] Maximilian Wegge, Enrica Troiano, Laura Oberländer, and Roman Klinger. Experiencer-specific emotion and appraisal prediction, 2023.

[26] Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. Everyone's voice matters: Quantifying annotation disagreement using demographic information, 2023.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and

Thamar Solorio, editors, *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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.