E-034

日本語 SNS のためのテキスト正規化および感情分析のデータセット
Japanese Dataset for Text Normalization and Sentiment Analysis in Social Media

近藤 里咲* Risa Kondo　寺面 杏優* Ayu Teramen　堀口 航輝* Koki Horiguchi　梶川 怜恩* Reon Kajikawa　鈴木 陽也* Haruya Suzuki
宮内 裕人* Yuto Miyauchi　山内 洋輝* Hiroki Yamauchi　秋山 和輝* Kazuki Akiyama　梶原 智之*† Tomoyuki Kajiwara　二宮 崇* Takashi Ninomiya
Chenhui Chu‡ Chenhui Chu　武村 紀子§ Noriko Takemura　早志 英朗† Hideaki Hayashi　中島 悠太† Yuta Nakashima　長原 一† Hajime Nagahara

1. はじめに

テキストから書き手の感情を推定する感情分析は，対話システム [1] やソーシャルメディアマイニング [2] など多くの自然言語処理タスクに応用できる．感情分析は，テキストからポジティブまたはネガティブな感情を判定する感情極性の推定と，喜びや悲しみなどのより詳細な感情を判定する基本感情の推定に大別される．さらに，これらの感情は，書き手自身による「主観的な感情」と，書き手の感情を読み手が推測する「客観的な感情」に分けられる．

先行研究では，基本感情として Ekman の 6 感情（喜び・悲しみ・驚き・怒り・恐れ・嫌悪）[3] や 6 感情に期待と信頼を加えた Plutchik の 8 感情 [4] が主に用いられてきた [5-8]．感情分析の既存研究はほとんどが基本感情または感情極性，主観感情または客観感情の一方のみを扱った研究である．基本感情と感情極性が両方付与されたデータセット [7,8] も存在しているものの，主観的な感情は考慮されていない．基本感情と感情極性，主観感情と客観感情のすべてを同時に扱った研究は存在していないため，基本感情と感情極性，主観感情と客観感情間の関係は明らかにされていない．

これらのデータセットは SNS などのユーザーが作成したテキストから収集されている．このようなテキストは誤字脱字に加えて造語やネットスラングなどのノイズ表現を多く含んでいる．日本語でもこれに対処するためにテキスト正規化の試みがある[9-12]．しかし，これまでに公開されている日本語テキスト正規化データセットは 1,000 文対ほどの小規模なデータ [9,10,13,14] や非公開データ [15] に基づくため，使用できるデータセットに限りがある．

本研究では，日本語における基本感情と感情極性の関係や，書き手の主観的な感情と読み手の客観的な感情の差異を明らかにするために，主観と客観の各観点において基本感情と感情極性を包括的に扱う感情分析データセットを構築する（表 1）．まず，クラウドソーシングを用いて 60 人のアノテータを雇用し，アノテータ自身の過去の SNS 投稿テキストに対する基本感情と感情極性の主観的な感情強度ラベルを収集する．その後，新たに雇用した 3 人のアノテータが，それぞれ全ての投稿に対して書き手の感情を推測することで，基本感情と感情極性の客観的な感情強度ラベルを収集する．最終的に計 35,000 件の SNS 投稿テキストに対して，Plutchik の 8 感情に関する 4 段階の強度と，5 段階の感情極性を主観と客観の各観点から付与した日本語感情分析データセット [1] を構築する．さらに，SNS 特有のノイズ表現が感情分析モデルの性能に及ぼす影響を調査するために，本データセットの一部を分析し，6 つの分類を設けて人手でテキスト正規化を実施する．

表 1　感情ラベルと正規化の例

元の文	雪降ってるー。寒いーー家帰りたいーーー。								
正規化後	雪が降っています。寒いです。家に帰りたいです。								
	喜び	悲しみ	期待	驚き	怒り	恐れ	嫌悪	信頼	極性
主観	0	1	0	1	0	0	0	0	-2
客観 1	0	0	1	1	0	1	0	0	-1
客観 2	0	1	0	0	0	0	1	0	-1
客観 3	0	2	0	0	0	0	0	0	-2

構築したデータセットを分析した結果，Plutchik の 8 感情のうち，喜び・期待・信頼はポジティブな感情であり，悲しみ・怒り・恐れ・嫌悪はネガティブな感情，驚きはニュートラルな感情であった．また，読み手は書き手の感情を過小評価する傾向にあり，特に怒りの感情を上手く推定できていなかった．正規化と感情極性の分析から，ノイズ表現の出現率は感情の強さと相関があることがわかった．

感情分析モデルを用いた評価実験の結果，書き手による主観ラベルの推定は読み手による客観ラベルの推定よりも難しいことが明らかとなった．読み手とみなせるモデルにとって，主観ラベルの推定は難しいことが示唆される．さらに，構築したデータセットのノイズ表現を人手で正規化することで，ほとんどの感情において性能を改善できた．

2. 関連研究

2.1 感情分析

2.1.1 基本感情が付与されたデータセット

基本感情を扱ったデータセットには，Ekman の 6 感情 [3] を対象としたもの [5-7] と，Plutchik の 8 感情 [4] を対象

＊愛媛大学　Ehime University
† 大阪大学　The University of Osaka
‡ 京都大学　Kyoto University
§ 九州工業大学　Kyushu Institute of Technology

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

表 2　アノテータ間の一致率（書き手を W，読み手 1～3 を R1～3 と表記）

	喜び	悲しみ	期待	驚き	怒り	恐れ	嫌悪	信頼	8 感情	極性
R1 vs. R2	0.580	0.474	0.528	0.468	0.610	0.430	0.399	0.208	0.508	0.608
R1 vs. R3	0.664	0.507	0.588	0.417	0.625	0.412	0.490	0.209	0.540	0.688
R2 vs. R3	0.657	0.589	0.596	0.471	0.638	0.468	0.378	0.234	0.562	0.530
W vs. R1	0.481	0.309	0.378	0.340	0.274	0.343	0.320	0.137	0.367	0.564
W vs. R2	0.587	0.434	0.441	0.402	0.297	0.357	0.452	0.114	0.450	0.493
W vs. R3	0.544	0.483	0.429	0.352	0.313	0.357	0.286	0.133	0.427	0.605
W vs. Avg. R	0.579	0.453	0.463	0.417	0.303	0.415	0.425	0.121	0.458	0.621

としたもの [8] とその他の感情を対象としたもの [16,17] があり，ほとんどは客観的な感情を対象としている．ISEAR [5] は，各感情に関連する過去の経験を報告するようにアノテータに求め，主観的な感情ラベルを収集したコーパスである．基本感情を対象とした研究の中で書き手の主観的な感情が付与された唯一のデータセットであるが，客観的な感情は考慮されていない．

2.1.2 感情極性が付与されたデータセット

感情極性を扱ったデータセット [7, 8, 18-20] は，ほとんどがレビュー文に対して感情極性が付与されたデータセットである．SemEval-2007 [7]，StudEmo [8]，SST [19]，Tsukuba sentiment-tagged corpus [2]，suzuki らのデータセット [20] は，レビュー文に対して第三者であるアノテータが推測した客観的な感情極性をラベル付けしている．一方で，IMDB [18] は書き手がレビューに付与したスコアを感情極性にマッピングすることで主観的な感情極性を収集しているが，客観的な感情は考慮されていない．

基本感情と感情極性の両方が付与されたデータセットとして，英語を対象とした SemEval-2007 と StudEmo がある．これらのデータセットは客観的な感情のみを対象としており，主観的な感情を考慮していない．さらに，日本語の感情分析では基本感情と感情極性の両方が付与されたデータセットは構築されておらず，主観的な感情を考慮した研究も存在しない．

2.2 テキスト正規化

SNS 投稿テキストに含まれるノイズ表現は，単語分割をはじめとする自然言語処理の性能を悪化させる．この課題に対処するために，日本語では系列ラベリング [9, 10] や系列変換 [11, 12] などのテキスト正規化アプローチが試みられてきた．しかし，これらを含む先行研究は，1,000 文対ほどの小規模なパラレルコーパス [9, 10, 13, 14] や自動生成されたデータ [11]，非公開データ [15] に基づいており，日本語のテキスト正規化のための高品質かつ大規模なパラレルコーパスの公開が望まれている． 本研究では，構築したデータセットのうちの 6,000 投稿（約 11,000 文）を人手でテキスト正規化し，テキスト正規化が感情分析の性能に及ぼす影響を調査する．

3. データセットの構築

本研究では，基本感情と感情極性，主観感情と客観感情の関係を明らかにするために，これらを包括するデータセットを構築する．さらに，テキストに含まれるノイズ表現が感情分析モデルの性能に与える影響を調査するために，構築したデータセットの一部を人手で正規化する．

2 http://www.nlp.mibel.cs.tsukuba.ac.jp/~inui/SA/corpus/

3.1 主観感情のアノテーション

まず，主観的な感情ラベルを収集するために，クラウドソーシング「ランサーズ [3]」を通じて 60 人のアノテータを雇用した．アノテータの性別の内訳は男性が 21 人，女性が 39 人であり，年齢の内訳は 20 代が 28 人，30 代が 22 人，40 歳以上が 10 人である．アノテータは 100 件から 1,000 件までの自身の SNS 投稿テキストを 100 件単位で提供した．このとき，感情分析の対象としてテキストのみを考慮するために，画像や URL が添付された投稿は対象外とした．投稿時期については特に制限を設けなかった．その結果，2010 年 8 月から 2020 年 11 月までの約 10 年分の投稿が集まった．このように収集した過去の投稿に対し，アノテータ自身が Plutchik の 8 感情 [4] の感情強度を 4 段階（0:無，1:弱，2:中，3:強），感情極性を 5 段階（-2:強いネガティブ，-1:ネガティブ，0:ニュートラル，+1:ポジティブ，+2:強いポジティブ）で付与した．最終的に，計 35,000 件の主観感情ラベルが付与された SNS 投稿テキストを収集した．

アノテーションの品質を評価するために，各アノテータから 30 件の投稿を無作為抽出し，基本感情の感情強度と感情極性のアノテーション品質を評価した．品質評価はそれぞれ 1 人の評価者が，以下の 4 段階で実施した．

- 3: 付与されたラベルに完全に同意できる
- 2: 付与されたラベルに概ね同意できる
- 1: 付与されたラベルに同意しにくい
- 0: 付与されたラベルに全く同意できない

基本感情の感情強度の評価結果について，アノテータごとに平均すると，最も評価が低いアノテータは 1.8 点，最も評価が高いアノテータは 2.5 点であり，アノテータ全体の平均は 2.1 点であった．アノテータのうち 5 人は 2 点を下回っていたものの，著しく低品質なアノテータはいなかった．感情極性の評価結果をアノテータごとに平均すると最低 1.9 点，最高 2.8 点であり，平均は 2.4 点であった．2 点を下回るアノテータは 4 人いたものの，著しく低品質なアノテータはいなかった．なお，基本感情においても感情極性においても 0 点と評価された投稿は存在しなかった．

3.2 客観感情のアノテーション

主観感情のアノテーションと同様に，ランサーズを用いて 3 人のアノテータを雇用した．アノテータの内訳は，30 代の女性が 2 人と 40 代の女性が 1 人である．3.1 節で収集した 35,000 件の SNS 投稿テキストに対して，各アノテータが Plutchik の 8 感情の感情強度と感情極性を付与した．本研究で扱う読み手の感情は，テキストを受け取った読み

3 https://www.lancers.jp/

表 3　基本感情と感情極性のピアソン相関

	喜び	悲しみ	期待	驚き	怒り	恐れ	嫌悪	信頼
Writer	0.585	-0.526	0.381	0.052	-0.353	-0.298	-0.467	0.296
Avg. Readers	0.665	-0.539	0.400	0.037	-0.229	-0.410	-0.470	0.252

手が抱く感情ではなく，読み手が推測する書き手の感情であることに注意されたい．

アノテーションの品質を評価するために，読み手間の一致率を Quadratic Weighted Kappa (QWK) [21] で評価した．読み手間の一致率を表 2 に示す．基本感情においては，喜び・期待・怒りが 0.5 を超える高い一致率を示した．8 感情の中では信頼が 0.2 程度の低い一致率を示したものの，基本感情全体においては 0.5 ほどの高い一致が見られた．感情極性においても 0.5〜0.6 ほどの高い一致率を示した．

3.3 テキスト正規化のアノテーション

テキスト正規化による感情分析の性能変化を観察するために，構築したデータセットに含まれるノイズ表現を人手で正規化する．先行研究で用いられた日本語テキスト正規化 [10, 14, 15, 22] と，本データセット分析して得た分類から 6 種類の正規化分類を定義する．

以降では，分類ごとの定義や具体例について説明する．なお，1 文中に複数箇所が同時に正規化される場合や，1 箇所の表現に対して複数の正規化が適用される場合があることに注意されたい．

3.3.1 誤字脱字

先行研究でも考慮されている誤字脱字は明らかにノイズであるため，本研究でも正規化する．ら抜き言葉などの活用形の誤りや脱字，先行研究 [10, 14, 15, 22] でも正規化の対象となった漢字の誤用やタイプミスも正しい表記に修正する．また，句読点の有無のような些細な差も感情分析モデルの性能に影響を与えるため，不足する句読点は補完し，書籍や映画等の作品名は『』で囲うように統一する．

3.3.2 方言

SNS 投稿テキストには，ネットスラングや伏字などの SNS 特有の表現に加え，地域方言や個性的な語尾などの書き手の個性を反映した表現が多数含まれる [10, 15]．本研究でもこれらをアノテータが検出できる範囲で，Web 検索などを用いつつ正規化する．また，構築したデータセット内には常体と敬体が混同していたため，これらを敬体に統一して多様性を吸収する．

3.3.3 異表記

先行研究 [10, 14, 15, 22] でも使用されている表記揺れは，本データセットにも頻出する．本研究では，「こりゃ（これは）」のような発音の崩れや「行けそーな（行けそうな）」のような同音異表記，「ウィルス」のような大文字／小文字は，正規の表現に修正する．また，句読点の連続は 3 点リーダに置換し，文末の読点は句点に置換する．鍵括弧の用法は，発話に対しては「」，作品名に対しては『』で統一する．

英字やカタカナ表記の外来語は，翻訳または翻字によって同等以上の流暢性が得られる場合には，それらの日本語表記に置換する．また，略語や同義語および平仮名・片仮名・漢字の表記揺れについては，高頻度語に置換する．単語頻度は，大規模 Web コーパスである CC-100 [4] [23] の日本語データを単語分割 [5] して数える．そのため，高頻度な略語は置換しないことに注意されたい．

3.3.4 強調表現

先行研究 [10, 14, 15, 22] でも使用されている促音や長音をはじめとした音の挿入や記号の挿入および文字や記号の繰り返しは，強調を目的として SNS 投稿テキストでもよく用いられる．冗長性の除去およびコーパス全体の表現を統一するために，本研究でも冗長な音や記号，繰り返しを削除することで正規化する．記号や文字や語句の繰り返しは，先行研究 [10] に従って 2 回までに削減する．

また，一部の投稿には箇条書きによる並列要素の列挙や，倒置表現などの一般的ではない語順や読みにくい語順への並べ替えが見られた．本研究では，箇条書きは文に展開し，流暢性が改善される場合は語句の順序を並べ替える．

3.3.5 平易化

新規の正規化分類として，新語や造語などの SNS 投稿テキストに特有の表現や，冗長な表現および情報の欠落，難解な表現などを可読性の高い表現に編集する．本研究では，新語や造語などを一般的な表現に平易化し，冗長な表現の削除や不足する情報の補完を実施する．補完はアノテータが推測できる範囲で実施する．さらに，複数文で表現すべき長い複文は分割し，本来 1 文で表現すべき短すぎる文は前後の文と結合することで，可読性の向上を図る．

3.3.6 感情表現

SNS 投稿テキストでは，書き手の感情を表現するために顔文字や絵文字が頻繁に使用される．これらは感情分析のための強力な手がかりとなり得るが，同じ喜びの感情を表現するにしても，「(笑)」や「w」など多様な表現がある．本研究では，新規の正規化分類として，絵文字や顔文字，その他の感情記号を Plutchik の 8 感情に集約し，表現する感情ごとに共通の特殊トークンに置換 [6] する．基本感情の信頼を 1，恐れを 2，驚きを 3，悲しみを 4，嫌悪を 5，怒りを 6，期待を 7，喜びを 8 とし，<1> 〜 <8> の特殊トークンを用いる．

また，数値表現も同様に，特殊トークンに集約する．数値表現は大小を区別せず，<num> の特殊トークンに置換する．ただし，「五十歩百歩」などの熟語やことわざの一部である数値表現は，意味を損なうため編集しない．

4. アノテーションに対する分析

4.1 基本 8 感情と感情極性

感情強度と感情極性のピアソン相関を表 3 に示す．Plutchik の 8 感情 [4] のうち，悲しみ・怒り・恐怖・嫌悪はネガティブな感情であり，喜び・期待・信頼はポジティブな感情であった．驚きは感情極性との相関が 0 に近く，ニュートラルな感情であることがわかる．特に喜びはポジテ

4 https://data.statmt.org/cc-100/
5 https://github.com/neologd/mecab-ipadic-neologd
6 [14] も顔文字を考慮するが，正規化はしていない．

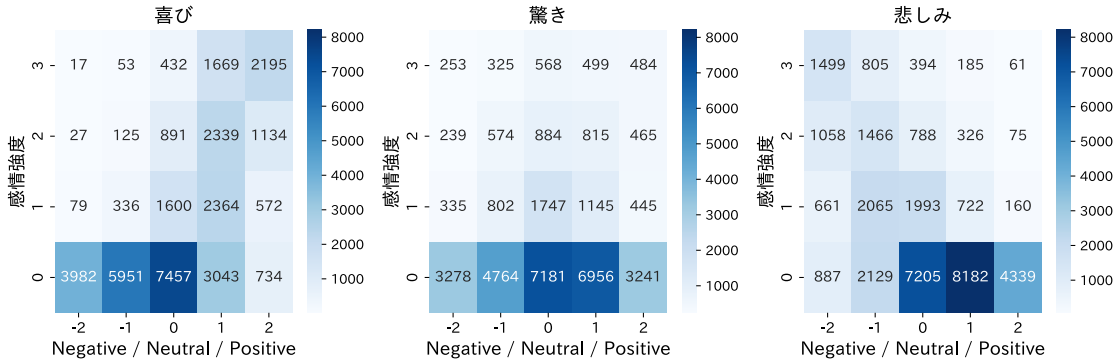

図 1　喜び・驚き・悲しみにおける感情強度と感情極性の分布

表 4　各基本感情のラベル数

	喜び				悲しみ				期待			
強度	W	R1	R2	R3	W	R1	R2	R3	W	R1	R2	R3
0	21,167	26,578	23,728	26,987	22,742	29,788	26,945	25,896	22,204	25,026	22,435	26,957
1	4,951	4,064	3,210	3,088	5,601	2,578	3,411	4,180	5,671	6,346	6,502	3,609
2	2,798	6,593	2,927	3,088	3,713	1,878	4,023	3,336	4,253	2,167	3,734	2,656
3	4,366	1,560	1,469	1,998	2,944	756	621	1,588	2,872	1,461	2,329	1,778
	驚き				怒り				恐れ			
強度	W	R1	R2	R3	W	R1	R2	R3	W	R1	R2	R3
0	25,420	26,790	25,693	30,033	30,664	34,040	34,096	33,977	28,824	27,618	27,648	30,744
1	4,474	4,248	5,336	2,954	2,006	475	317	389	3,321	4,584	3,975	2,265
2	2,977	2,560	2,492	1,321	1,211	288	348	378	1,688	1,875	2,415	1,510
3	2,129	1,402	1,479	692	1,119	197	239	256	1,167	923	962	481
	嫌悪				信頼				全体			
強度	W	R1	R2	R3	W	R1	R2	R3	W	R1	R2	R3
0	27,733	30,535	27,457	32,322	27,892	31,978	33,825	33,892	206,646	232,353	221,827	240,808
1	3,428	2,759	3,223	1,388	3,549	2,560	684	535	33,001	27,614	26,658	18,408
2	2,119	1,106	1,950	837	2,228	347	364	391	22,705	13,019	21,919	13,356
3	1,720	600	2,370	453	1,331	115	127	182	17,648	7,014	9,596	7,428

表 5　感情極性のラベル数

極性	W	Avg. R	R1	R2	R3
-2	4,105	1,687	2,254	1,056	9,581
-1	6,465	10,468	10,316	4,029	4,256
0	10,380	11,462	8,741	20,147	10,687
+1	9,415	9,138	11,216	8,510	2,841
+2	4,635	2,245	2,473	1,258	7,635

ィブな感情極性と相関が強く，悲しみはネガティブな感情極性との相関が強かった．喜びと驚きと悲しみにおける感情強度と感情極性の感情ラベルの分布を図 1 に示す．喜びがネガティブな感情極性を持つときに出現することは少なく，ポジティブな感情極性を持つときにはその程度に応じて感情強度が高くなる傾向にある．驚きは，感情極性がニュートラルである場合に多く出現し，ネガティブとポジティブのどちらの感情においても大きな偏りなく出現する傾向にある．悲しみはポジティブな感情極性を持つときに出現することは少なく，ネガティブな感情極性を持つときにはその程度に応じて感情強度が高くなる傾向にある．

4.2 主観感情と客観感情

書き手によるアノテーションと，読み手によるアノテーションの一致率を表 2 の下段に示す．ここで，Avg.R は客観的アノテータ 3 人が付与したラベルの平均である．基本感情においては，喜びが 0.5 ほどの最も高い一致率を示し，3.2 節の客観的アノテータ間の一致率と同様に信頼が 0.1 ほどの特に低い一致率を示した．さらに，感情極性では，0.5〜0.6 の高い一致率を示した．読み手同士の一致率と書き手-読み手間の一致率を比較すると，全体的に書き手-読み手間の一致率の方が低い傾向にあり，客観アノテータの平均を取ることで書き手-読み手間の一致率に改善が見られた．特に怒りでは読み手同士と書き手-読み手間の一致率に大きな差があり，読み手は書き手の怒りの感情を汲み取ることが相対的にできていない．

アノテータごとの感情強度ラベルの分布を表 4 に示す．どのアノテータもすべての感情において 0（無）をラベル付けした数が最も多かった．なかでも，読み手は書き手の怒りと信頼の感情を過小評価する傾向にある．

また，アノテータごとの感情極性ラベルの分布を表 5 に示す．書き手と読み手 2，読み手 3 はニュートラルを最も

表 6　基本感情の書き手と読み手間の相関 (%)

	喜び				悲しみ				期待			
Writer＼Readers	0	1	2	3	0	1	2	3	0	1	2	3
0	91.4	7.5	1.1	0.1	90.7	7.9	1.3	0.1	83.4	13.4	2.9	0.3
1	52.7	33.0	13.3	1.0	52.5	34.0	12.2	1.3	48.6	33.9	15.0	2.5
2	30.4	37.2	27.1	5.3	40.0	37.5	19.7	2.8	37.9	35.9	21.7	4.4
3	18.7	31.3	35.5	14.5	31.8	35.8	26.8	5.6	25.0	32.4	29.1	13.5
	驚き				怒り				恐れ			
Writer＼Readers	0	1	2	3	0	1	2	3	0	1	2	3
0	85.5	12.3	2.0	0.2	99.2	0.7	0.1	0.0	85.8	12.2	1.8	0.2
1	55.5	32.8	10.6	1.1	89.6	7.6	2.3	0.5	56.2	34.1	8.8	0.9
2	43.8	39.2	13.8	3.2	79.3	13.0	6.0	1.7	39.8	38.6	17.7	4.0
3	28.8	38.0	24.8	8.4	62.1	17.2	11.3	9.4	32.2	31.5	26.1	10.1
	嫌悪				信頼				全体			
Writer＼Readers	0	1	2	3	0	1	2	3	0	1	2	3
0	91.0	7.6	1.2	0.2	97.5	2.3	0.2	0.0	90.9	7.7	1.3	0.1
1	61.4	30.3	7.4	0.9	92.9	6.6	0.4	0.1	60.2	28.7	9.9	1.2
2	49.1	35.4	12.1	3.4	86.0	12.9	1.1	0.0	45.7	33.5	17.4	3.5
3	34.3	35.3	19.8	10.6	75.0	20.6	4.1	0.3	32.5	31.7	25.7	10.1

表 7　感情極性の書き手と読み手間の相関 (%)

Writer＼Readers	-2	-1	0	+1	+2
-2	20.3	59.9	15.7	3.8	0.2
-1	9.3	63.7	23.2	3.7	0.2
0	1.8	27.8	49.3	19.4	1.6
+1	0.6	8.2	32.7	48.1	10.5
+2	0.2	5.0	24.3	47.5	23.0

多くラベル付けした．読み手 1 はネガティブとポジティブを多くラベル付けしており，読み手 3 はニュートラルの次に極端な感情極性を多くラベル付けした．書き手と読み手 2 はニュートラルの次にネガティブとポジティブを多くラベル付けしており，次いで強いネガティブと強いポジティブをラベル付けしている．読み手の平均を取ることで書き手と同様の傾向を示すものの，読み手は書き手の強い感情を過小評価する傾向にあり，書き手と比べてネガティブを多くラベル付けしている．

書き手の感情強度ラベルと読み手の平均の感情強度ラベルの混同行列を表 6 に示す．書き手が 0 をラベル付けしたものに対して読み手も 0 をラベル付けする割合は非常に高く，すべての感情において 83%～99%である．書き手が 1～3 とラベル付けしたものに対して，読み手はその感情を過小評価する傾向にある．特に，怒りや信頼においてその傾向が顕著であり，書き手が 3 とラベル付けした際に読み手が 0 とラベル付けする確率がそれぞれ 62.1%，75.0%と最も高かった．読み手は書き手の怒りや信頼の感情をほとんど検出できていないことがわかる．

書き手の感情極性ラベルと読み手の平均の感情極性ラベルの混同行列を表 7 に示す．感情極性でも感情強度と同様に，読み手は書き手が強いネガティブとラベル付けしたものをネガティブ，強いポジティブとラベル付けしたものをポジティブと極性の強さを過小評価する傾向にある．しかし，書き手が強いネガティブとラベル付けしたものに対して，読み手が強いネガティブまたはネガティブとラベル付けした割合は 80.2%，書き手が強いポジティブとラベル付けしたものに対して読み手が強いポジティブまたはポジティブとラベル付けした割合は 70.5%であり，読み手は書き手のおおまかな極性を捉えられることがわかる．したがって，読み手は書き手の大まかな感情極性を推定できるものの，その強度を過小評価する傾向にあることが示唆される．

4.3 ノイズ表現と感情極性

テキスト正規化の結果，方言が最も多くの投稿で出現しており，その数は 6,000 投稿中 5,447 件（全体の 92.8%）であった．次いで誤字脱字（90.8%），異表記（72.5%），平易化（45.1%），強調表現（43.5%），感情表現（42.3%）と続いた．6,000 投稿のうち 99.5%が正規化の対象となり，正規化されなかった文は 30 文のみであった．

各感情極性ラベルが付与された投稿のうち，ある正規化分類による正規化が実施された割合を表 8 に示す．どの分類においてもラベル 0 の正規化率が最も低く，ラベル 0 以外の感情的な投稿はノイズ表現が多くなることが明らかとなった．さらに，誤字脱字・異表記・強調表現・感情表現における書き手の感情や方言・異表記・強調表現における読み手感情は，感情が強くなるほどノイズ表現が多く含まれる傾向にあった．

5. 評価実験

構築したデータセットを用いて感情分析モデルを構築し，Plutchik の 8 感情 [4] の感情強度（4 段階）と感情極性（5 段階）を推定する．加えて，SNS 特有のノイズを正規化した本データセットで感情分析モデルを構築し，テキスト正規化が感情分析モデルの性能に与える影響を調査する．

5.1 感情分析

5.1.1 実験設定

表 8　極性ごとの正規化率 (%)

	誤字脱字		方言		異表記		強調表現		平易化		感情表現		全種類の正規化	
極性	W	R	W	R	W	R	W	R	W	R	W	R	W	R
-2	93.0	90.6	94.7	96.4	73.1	78.6	48.0	50.4	48.0	48.9	45.6	47.8	99.4	100.0
-1	91.3	92.8	95.0	94.8	71.9	73.6	43.6	40.7	44.3	43.5	43.5	41.7	99.9	99.8
0	88.5	89.7	91.3	89.4	70.6	67.9	37.9	38.8	42.9	42.8	38.5	39.3	99.2	99.2
1	91.1	90.2	92.2	93.3	73.0	74.1	42.4	48.0	47.2	48.2	41.7	44.9	99.5	99.4
2	91.3	90.0	92.1	95.5	74.6	78.5	49.3	53.4	43.8	47.1	45.0	44.6	99.4	100.0

表 9　主観感情で評価した結果

	喜び	悲しみ	期待	驚き	怒り	恐れ	嫌悪	信頼	8 感情	極性
客観 (Avg. R) BoW	0.369	0.284	0.268	0.253	0.127	0.234	0.212	0.150	0.326	0.418
客観 (Avg. R) BERT	0.465	0.369	0.291	0.317	0.112	0.337	0.382	0.066	0.363	0.558
客観 (Avg. R) RoBERTa	0.497	0.421	0.345	**0.350**	0.205	**0.370**	**0.448**	0.070	0.402	**0.612**
主観 BoW	0.338	0.269	0.161	0.116	0.206	0.158	0.201	0.070	0.256	0.351
主観 BERT	0.529	0.398	0.333	0.294	0.380	0.283	0.394	**0.171**	0.417	0.540
主観 RoBERTa	**0.569**	**0.441**	**0.390**	0.338	**0.439**	0.357	0.433	0.141	**0.452**	0.605

表 10　客観感情で評価した結果

	喜び	悲しみ	期待	驚き	怒り	恐れ	嫌悪	信頼	8 感情	極性
客観 (Avg. R) BoW	0.417	0.360	0.343	0.269	0.272	0.308	0.237	0.136	0.373	0.455
客観 (Avg. R) BERT	0.636	0.547	0.622	0.536	0.298	0.510	0.483	0.310	0.603	0.707
客観 (Avg. R) RoBERTa	**0.696**	**0.634**	**0.708**	**0.580**	**0.425**	**0.622**	**0.612**	**0.379**	**0.675**	**0.778**
主観 BoW	0.273	0.269	0.147	0.116	0.086	0.194	0.191	0.042	0.210	0.371
主観 BERT	0.570	0.443	0.467	0.444	0.230	0.389	0.363	0.175	0.489	0.625
主観 RoBERTa	0.640	0.512	0.579	0.489	0.243	0.493	0.466	0.193	0.558	0.711

3 章で構築したデータセットを 30,000 件（40 人分の投稿）の訓練データ，2,500 件（10 人分の投稿）の検証データ，2,500 件（10 人分の投稿）の評価データに分割して実験する．モデルの性能は Quadratic Weighted Kappa（QWK）[21] で評価する．本稿では，主観データで評価した結果と，客観データで評価した結果の両方を報告する．

感情分析モデルとして，以下の 3 つの分類器を用いる．

- **BoW**：Bag-of-Words によって素性を抽出し，ロジスティック回帰で分類を行う．単語分割には MeCab [7] [24] を用いる．
- **BERT** [25]：事前訓練済みの日本語 BERT [8] を訓練データで fine-tuning する．
- **RoBERTa** [26]：事前訓練済みの日本語 RoBERTa [9] を訓練データで fine-tuning する．

BoW の実装には scikit-learn [10] を用い，BERT と RoBERTa の実装には Trasnformers [11] を用いた．

BERT と RoBERTa の訓練時には，最適化に AdamW [27] を使用してバッチサイズを 32 とし，検証データにおける QWK の値が 3 エポック改善されなくなるまで訓練を続けた．学習率は $\{1.0\times10^{-6}, 5.0\times10^{-6}, 1.0\times10^{-5}\}$ の中から検証データに対する最適値を選択した．異なる乱数シード値で 3 回実験した際の平均値を報告する．

7 https://taku910.github.io/mecab/
8 https://huggingface.co/tohoku-nlp/bert-base-japanese-whole-word-masking
9 https://huggingface.co/nlp-waseda/roberta-base-japanese-with-auto-jumanpp
10 https://scikit-learn.org/stable/
11 https://github.com/huggingface/transformers

5.1.2 実験結果

書き手のラベルで評価した結果を表 9，読み手のラベルで評価した結果を表 10 に示す．ここで，「8 感情」はすべての基本感情で評価した値である．

主観感情で評価した基本感情の強度推定では，信頼を除いたすべての感情で客観 RoBERTa または主観 RoBERTa が高い性能を示した．喜びや悲しみ，期待，怒りは主観 RoBERTa が高い性能を示し，驚きや恐れ，嫌悪は客観 RoBERTa の方が高い性能をした．よって，書き手の驚きや恐れ，嫌悪の感情の推定がより難しいことが示唆される．また，すべての感情で評価した「8 感情」列に注目すると，主観 RoBERTa が最も高い性能を示しており，書き手の感情を推論するには書き手の主観データで訓練した RoBERTa が適しているといえる．感情極性の実験では，客観 RoBERTa が最も高い性能を示した．4.2 節の分析より，読み手は書き手の大まか感情を読み取れることがわかっているため，これは妥当な結果であるといえる．基本感情では，書き手の感情を推定するために最適な訓練データが基本感情によって異なっていたが，感情極性においては書き手による主観データで訓練したモデルよりも読み手による客観データでモデルが一貫して高い性能を示した．

客観感情で評価した基本感情の強度推定では，読み手のデータで訓練した客観 RoBERTa がすべての感情において最も高い性能を示した．客観 BERT は喜びを除いたすべての感情で主観 RoBERTa を上回る性能であった．感情極性の実験では，基本感情の実験と同様に客観 RoBERTa が最も高い性能を示した．客観 BERT は主観 RoBERTa に僅差で劣るものの，同じ分類器から構築した主観データで訓練

表 11　主観感情で評価した結果（正規化を適用した場合）

	喜び	悲しみ	期待	驚き	怒り	恐れ	嫌悪	信頼	8 感情	極性
正規化なし	0.498	0.422	0.324	0.328	0.225	0.282	0.154	0.192	0.377	0.496
誤字脱字	0.488	0.376	0.277	**0.353**	**0.255**	0.233	**0.348**	0.152	0.376	**0.559**
方言	0.489	0.411	0.273	0.314	**0.351**	0.213	**0.329**	0.168	**0.380**	**0.577**
異表記	**0.511**	0.414	0.256	0.305	**0.297**	0.275	**0.315**	**0.208**	**0.381**	**0.532**
強調表現	0.487	**0.424**	0.309	0.275	**0.249**	0.238	**0.236**	**0.224**	0.374	**0.506**
平易化	0.497	**0.456**	0.284	0.257	**0.364**	0.260	**0.342**	**0.213**	**0.382**	**0.556**
感情表現	**0.499**	0.400	0.288	0.287	**0.332**	0.212	**0.351**	0.189	0.375	**0.539**
全種類の正規化	**0.518**	**0.423**	**0.326**	**0.329**	**0.433**	0.213	**0.339**	0.143	**0.393**	**0.603**

表 12　客観感情で評価した結果（正規化を適用した場合）

	喜び	悲しみ	期待	驚き	怒り	恐れ	嫌悪	信頼	8 感情	極性
正規化なし	0.524	0.422	0.548	0.410	0.168	0.386	0.326	0.254	0.491	0.659
誤字脱字	**0.557**	**0.545**	0.514	**0.478**	0.092	**0.393**	**0.335**	**0.258**	**0.520**	**0.679**
方言	**0.562**	**0.491**	**0.583**	**0.494**	**0.196**	**0.410**	**0.427**	**0.332**	**0.541**	**0.680**
異表記	**0.574**	**0.504**	0.521	**0.432**	**0.182**	**0.432**	**0.386**	**0.292**	**0.523**	**0.661**
強調表現	0.503	**0.488**	**0.569**	**0.438**	**0.179**	**0.431**	0.317	0.227	**0.507**	0.659
平易化	**0.538**	**0.429**	**0.590**	**0.471**	**0.254**	**0.458**	**0.343**	**0.310**	**0.527**	**0.660**
感情表現	**0.548**	**0.520**	0.544	0.365	0.143	**0.412**	**0.357**	0.247	**0.505**	0.650
全種類の正規化	**0.614**	**0.602**	**0.564**	**0.508**	**0.284**	**0.483**	**0.495**	0.191	**0.577**	**0.700**

したモデルと客観データで訓練したモデルを比較すると，客観データで訓練したモデルが一貫して高い性能を示した．

5.2 テキスト正規化

5.2.1 実験設定

60 人のユーザからそれぞれ 100 投稿ずつ抽出し，計 6,000 投稿を人手で正規化した．著者の 1 名が正規化を実施し，別の著者 1 名が正規化の可否を判断して必要に応じて修正した．また別の著者 1 名がそれぞれの正規化事例を 3.3 節の正規化分類に基づいて分類した．

モデルには BERT を用いた．正規化したデータセットを訓練データ 5,000 件，検証データ 500 件，評価データ 500 件に分割し，各分類のみを正規化したデータでモデルを fine-tuning した．5.1.1 項と同様の設定でモデルを構築し，異なる乱数シード値で 3 回実験した際の平均値を報告する．モデルの性能は QWK で評価し，主観データで訓練と評価をおこなった結果と客観データで訓練と評価をおこなった結果の両方を報告する．

5.2.2 実験結果

書き手のラベルで評価した結果を表 11，読み手のラベルで評価した結果を表 12 に示す．

主観感情で評価した基本感情の強度推定では，恐れを除くすべての感情で性能改善がみられた．特に，表 9 で客観モデルよりも低い性能を示した怒りと嫌悪では，すべての正規化分類において大幅な改善がみられた．これらの感情はノイズの影響を強く受けていると考えられる．また，ほとんどの感情は全種類の正規化を組み合わせて適用することで性能改善がみられたものの，信頼だけは逆に性能が悪化した．8 感情すべてでモデルを評価した結果，方言・異表記・平易化・全種類の正規化で性能改善がみられ，書き手特有の言い回しを統一することがモデルの性能改善に寄与することがわかった．感情極性の推定では，すべての正規化分類において性能が改善し，全種類の正規化を適用することで最高性能を達成した．すべての感情にわたり，一貫してモデル性能を改善する分類は見られず，最適な分類は感情ごとに異なることがわかった．よって，ある感情を含む投稿は特定の表現（正規化分類）がよく使用されることが示唆される．

客観感情で評価した基本感情の強度推定では，正規化によってすべての感情において性能改善がみられた．なかでも，悲しみと恐れはどの正規化を適用しても性能が改善した．信頼では，全種類の正規化を適用したときに最も低い性能を示しており，これは主観感情で評価した時と同様であった．一方で，主観感情による評価では性能が改善しなかった恐れの感情は，読み手のラベルで評価するとすべての正規化で性能が改善した．8 感情すべてでモデルを評価した結果，すべての正規化分類が客観感情の推定に役立つことが明らかとなった．感情極性の実験では，強調表現と感情表現以外の正規化分類を適用した時に性能が改善した．方言と平易化を適用することで，すべての基本感情と感情極性の性能を改善でき，客観感情を推定するときにはこれらの正規化が有効に機能することがわかった．

また，主観感情の評価結果と客観感情の評価結果を比較すると，読み手のラベルで評価した時の方が性能改善が多く見られた．よって，テキストの読み手による正規化は読み手の感情認識を補助する効果があることが示唆される．

6. おわりに

本研究では，主観と客観の観点から基本感情と感情極性をアノテーションした 35,000 件からなる日本語感情分析データセットを構築し，公開した．主観ラベルのアノテーションでは，60 人のアノテータが自身の過去の SNS 投稿テキストに対して，Plutchik の 8 感情に基づく 4 段階の感情強度と，5 段階の感情極性を付与した．客観ラベルのアノ

テーションでは，3 人のアノテータが書き手の感情を想像して感情強度と感情極性を付与した．さらに，SNS 投稿テキストに含まれるノイズ表現が感情分析モデルの性能に与える影響を調査するために，6 種類の分類を設けて人手によるテキスト正規化を実施した．

アノテーション結果を分析した結果，喜び・期待・信頼はポジティブな感情であり，驚きはニュートラル，悲しみ・怒り・恐れ・嫌悪はネガティブな感情であることが明らかとなった．さらに，読み手は書き手の感情を過小評価する傾向にあり，特に怒りの感情をうまく捉えられないことがわかった．正規化と感情極性を分析すると，書き手も読み手もニュートラルなラベルが付与された投稿の正規化率が最も低く，ポジティブまたはネガティブな感情が強くなるほど正規化率が高くなる傾向にあった．

感情分析モデルによる評価実験の結果，モデルは書き手の驚き・恐れ・嫌悪をうまく推定できないことが明らかとなった．また，テキスト正規化によって，書き手の恐れを除くすべての感情でモデルの性能を改善でき，ノイズ表現が感情分析モデルに悪影響を与えることを確認した．

今後の取り組みとして，アスペクトベースの感情分析や読み手が抱く感情の推定に取り組むことを検討している．

謝辞

本研究は、JST BOOST（課題番号：JPMJBY24036821）の支援を受けたものです．また，本稿の内容の一部は NAACL-2021 [28]，LREC-2022 [29]，WNUT-2025 [30]において報告したものです．

参考文献

[1] Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. Emotion Classification Using Massive Examples Extracted from the Web. In Proc. of COLING, pp. 881–888, 2008.
[2] Stefan Stieglitz and Linh Dang-Xuan. Emotions and Information Diffusion in Social Media ― Sentiment of Microblogs and Sharing Behavior. Journal of Management Information Systems, 29(4):217–248, 2013.
[3] Paul Ekman. A General Psychoevolutionary Theory of Emotion. Cognition and Emotion, 6(3–4):169–200, 1992.
[4] Robert Plutchik. A General Psychoevolutionary Theory of Emotion. Theories of Emotion, 1:3–31, 1980.
[5] Klaus R. Scherer and Harald G¨unter Wallbott. Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning. Journal of Personality and Social Psychology, 66(2):310–328, 1994.
[6] Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew AliAyele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection. In Proc. of SemEval, 2025.
[7] Carlo Strapparava and Rada Mihalcea. SemEval-2007 Task 14: Affective Text. In Proc. of SemEval, pp. 70–74, 2007.
[8] Anh Ngo, Agri Candri, Teddy Ferdinan, Jan Kocon, and Wojciech Korczynski. StudEmo: A Non-aggregated Review Dataset for Personalized Emotion Recognition. In Proc. of NLPerspectives, pp. 46–55, 2022.
[9] 佐々木 彬, 水野 淳太, 岡崎 直観, 乾 健太郎. 機械学習に基づくマイクロブログ上のテキストの正規化. 人工知能学会第 27 回全国大会, 4B1-4, 2013.
[10] 大崎 彩葉, 北川 善彬, 小町 守. 日本語 Twitter 文書を対象とした系列ラベリングによる表記正規化. 情報処理学会第 231 回自然言語処理研究会, 12, pp.1-6, 2017.
[11] Taishi Ikeda, Hiroyuki Shindo, and Yuji Matsumoto. Japanese Text Normalization with Encoder-Decoder Model. In Proc. of WNUT, pp. 129–137, 2016.
[12] Daniel Watson, Nasser Zalmout, and Nizar Habash. Utilizing Character and Word Embeddings for Text Normalization with Sequence-to-Sequence Models. In Proc. of EMNLP, pp. 837–843, 2018.
[13] Nobuhiro Kaji and Masaru Kitsuregawa. Accurate Word Segmentation and POS Tagging for Japanese Microblogs: Corpus Annotation and Joint Modeling with Lexical Normalization. In Proc. of EMNLP, pp. 99–109, 2014.
[14] Shohei Higashiyama, Masao Utiyama, Taro Watanabe, and Eiichiro Sumita. User-Generated Text Corpus for Evaluating Japanese Morphological Analysis and Lexical Normalization. In Proc. of NAACL, pp. 5532–5541, 2021.
[15] 斉藤 いつみ, 貞光 九月, 浅野 久子, 松尾 義博. 正規-崩れ表記のアライメントに基づく表記崩れパタンの抽出と形態素解析への導入. 情報処理学会第 214 回自然言語処理研究会, 5, pp.1-9, 2013.
[16] Saif Mohammad and Felipe Bravo-Marquez. WASSA-2017 Shared Task on Emotion Intensity. In Proc. of WASSA, Sentiment and Social Media Analysis, pp. 34–49, 2017.
[17] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. SemEval-2018 Task 1: Affect in Tweets. In Proc. of SemEval, pp. 1–17, 2018.
[18] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning Word Vectors for Sentiment Analysis. In Proc. of ACL, pp. 142–150, 2011.
[19] 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 Proc. of EMNLP, pp. 1631–1642, 2013.
[20] Yu Suzuki. Filtering Method for Twitter Streaming Data Using Human-in-the-Loop Machine Learning. Journal of Information Processing, 27:404–410, 2019.
[21] Jacob Cohen. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. Psychological Bulletin, 70(4):213–220, 1968.
[22] 笹野 遼平, 黒橋 禎夫, 奥村 学. 日本語形態素解析における未知語処理の一手法―既知語から派生した表記と未知オノマトペの処理―. 自然言語処理, 21(6):1183‐1205, 2014.
[23] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzm´an, Armand Joulin, and Edouard Grave. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In Proc. of LREC, pp. 4003–4012, 2020.
[24] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In Proc. of EMNLP, pp. 230–237, 2004.
[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proc. of NAACL, pp. 4171–4186, 2019.
[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Joshi. Mandar, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692, 2019.
[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In Proc. of ICLR, 2019
[28] 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 Proc. of NAACL, pp. 2095–2104, 2021.
[29] Haruya Suzuki, Yuto Miyauchi, Kazuki Akiyama, Tomoyuki Kajiwara, Takashi Ninomiya, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. A Japanese Dataset for Subjective and Objective Sentiment Polarity Classification in Micro Blog Domain. In Proc. of LREC, pp. 7022–7028, 2022.
[30] Risa Kondo, Ayu Teramen, Reon Kajikawa, Koki Horiguchi, Tomoyuki Kajiwara, Takashi Ninomiya, Hideaki Hayashi, Yuta Nakashima, and Hajime Nagahara. Text Normalization for Japanese Sentiment Analysis. In Proc. of WNUT, pp. 149–157, 2025.