方言コーパスを用いた感情分析モデルの構築と炎上・ネットいじめ検知手法の提案

加藤大造 [1]　Le-Minh NGUYEN[1]
[1] 北陸先端科学技術大学院大学
{taizo.kato, nguyenml}@jaist.ac.jp

概要

インターネット上でユーザーが自らの感情や考えを自由に，かつ簡単に投稿できるようになってから久しい．また，ネット上での会話は日常生活の一部に組み込まれている．これらの投稿は主に口語で投稿されるため，ユーザーの居住地や出身地の方言が含まれていると推察される．本研究では，方言を含む文章は書き手の感情を強く反映している，と仮定し，SNS に投稿された方言を含むテキストデータを用いて DialectBERT モデルを構築した．このモデルは，感情分析や炎上・ネットいじめの発生検知に優れた結果を示した．

1　はじめに

日本では，2004 年頃から Social Networking Service(SNS) の利用が増加した．2021 年時点で，74.2% の人々が SNS を利用している [1]．SNS 上での会話は，通常話し言葉で行われるため，ユーザーの居住地や出身地の方言が多く含まれていると推察される．廣田ら [2] は「ブログ等の CGM の普及により Web 上で方言が使用される機会が増えている．また，それに伴い，方言に対しても頑健な言語処理技術の必要性が高まっている」と述べた．また、SNS の利用が一般的になるにつれ，SNS 上での炎上やネットいじめも社会問題化している．日本では携帯電話等を使用したいじめの件数が 1 年間で 21,900 件に上り，引き続き増加傾向にある [3]．

本研究は，方言を含む文章は書き手の感情を強く反映している，と仮定し，方言データを用いた場合に，より精度の高い感情分析が可能であることを示す．まず，自然言語処理モデルの BERT を方言データで追加事前学習する．次に，感情分析のためのファインチューニングを実施し，感情分析モデルを評価する．また，構築された感情分析モデルを用いて炎上やネットいじめの発生検知の手法を検証する．この，方言で追加事前学習された感情分析モデルを用いることが，SNS 上の炎上やネットいじめ発生検知においても有効であることを示す．また，汎用性の観点から，特定のサービスに依存した特徴などは排除し，かつ投稿テキストのみを用いた分析手法の有効性を検証する．

2　関連研究

方言を用いた研究は，アラビア語圏で盛んに行われている．アルジェリア方言を利用し構築した DziriBERT [4] は，標準・古典アラビア語，その他の地域の方言を含むデータで学習したモデル [5, 6, 7, 8] と比較しても高い精度を示した．しかしながら，日本語においては，方言データを用いて感情分析を行なった研究や，方言データで事前学習された BERT モデルを感情分析にファインチューニングした研究は見当たらない．感情分析においては，主観的な感情強度の予測が，客観的な強度予測に比べて難しいことが示された [9]．また，より良い主観感情の強度推定を行うために性格情報を加えてることが有効であることも示された [10]．いじめに関する研究では，内容のポジティブ・ネガティブ予測 [11]，いじめにつながる投稿の特定 [12] といった研究が行われており，Twitter の炎上検知において，リプライの感情や非フォロワー属性が有効であることが示された [13]．

3　感情分析

3.1　データセット

まず，日本語の方言の一覧を，全国方言辞典 [1)] を用いて作成する．次に，これら方言が含まれる投稿データを Twitter(現 X) より取得する．Twitter の投稿

1） https://dictionary.goo.ne.jp/dialect/ (NTT レゾナンス社が運営)

データ，tweet(現 post) には，投稿時のユーザーの位置情報を保持しているものがある．今回，方言が話されている地域と，ユーザーが投稿した際の位置情報とを一致させ検索する．この条件にて収集した結果，約 32 万件の tweet を獲得した．tweet の特徴として，tweet 本文には，メンションと呼ばれる@の後にユーザ名が書かれたり，内容のトピック等を#(ハッシュタグ) の後に続けて記載されていたりする．また，Web ページや画像の URL がリンク等も記載されていることもある．本研究では，特定のサービスに依存せず，テキストだけに基づいて分析を行うことを目的としているため，そのような項目を前処理にて削除した．

3.2 形態素解析

形態素解析には MeCab を用いる．MeCab で利用できる辞書に，Web 上の言語資源から得た新語が追加された mecab-ipadic-NEologd[2)]がある．こちらの mecab-ipadic-NEologd をベース辞書として利用する．この辞書に加えて，新たに方言辞書を作成し，MeCab に組み込む．これを DialectMeCab とする．この辞書を用いて形態素解析した結果を表 1 に示す．「ちょす」という方言は，北海道の方言で，標準語では「触る」という意味の単語である．方言辞書が組み込まれていない MeCab では名詞の「ちょ」と動詞の「す」に分かれて解析されていが，方言辞書を組み込んだ場合では，正しく「ちょす」を 1 つの単語として，かつ方言であることが認識されている．

表 1 DialectMeCab での形態素解析結果

対象テキスト	方言辞書の有無	結果
スマホをちょす (「ちょす」は北海道の方言で「触る」)	無し	スマホ 名詞, 固有名詞, 一般, *, *, *, スマホ, スマホ, スマホ を 助詞, 格助詞, 一般, *, *, *, を, ヲ, ヲ ちょ 名詞, 動詞非自立的, *, *, *, *, ちょ, チョ, チョ す 動詞, 自立, *, *, サ変・スル, 文語基本形, する, ス, ス
	有り	スマホ 名詞, 固有名詞, 一般, *, *, *, スマホ, スマホ, スマホ を 助詞, 格助詞, 一般, *, *, *, を, ヲ, ヲ ちょす 動詞, 自立, *, *, 五段・ラ行, 基本形, ちょす, ちょす, 方言

3.3 提案手法

前処理を施したデータを方言コーパスとし，追加事前学習に使用する．追加事前学習を行うベースモデルは，東北大学が公開している BERT モデル [14] とする．追加事前学習されたモデル (DialectBERT) を基本モデルとしてし，8 つの感情の強度と極性に対してファインチューニングを行い，各感情分析に特化した BERT モデルを構築する．ファインチューニングには，感情分析のためのデータセット WRIME [9] を利用する．本手法の全プロセスを，図 1 に示す．

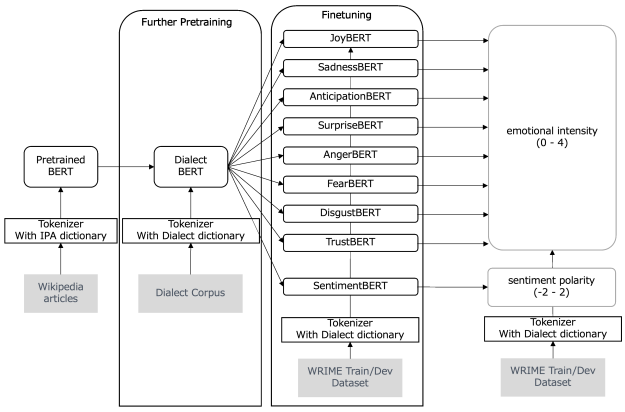

図 1 感情分析モデル構築プロセス

3.4 結果

比較のため，LSTM モデルをベースモデルとする．LSTM に加えて，DialectMeCab と DialectBERT の利用の有無を組み合わせたの 4 パターンの結果も比較する．感情強度の評価には，平均絶対誤差 (MAE)，極性の評価には，Accuracy を指標として用いる．表 2 に，8 つの感情強度の結果，表 3 に感情極性の結果を示す．8 つの感情の内，Joy，Anticipation，Surprise，Anger については DialectBERT と DialectMecab の両方を用いたパターンの精度が最も良い結果となった．また，Sadness，Disgust についても，DialectBERT を用いた場合のパターンの精度が良い結果となった．これら 6 感情については，最も精度が悪いパターンから平均して 0.163 の差があった．この結果により，感情の理解において，方言を理解したモデルを利用することが有用であることが示された．一方で，Fear，Trust については，ベースモデルの精度が最も良い結果となった．これには，Fear や Trust といった感情を持って発言する際には，方言が入りづらい状況にあり，モデルの方言の理解が結果に影響しなかったと考えられる．感情極性についても，DialectMeCab と DialectBERT の両方用いたパターンが，より正確に判定できることが示された．

次に，データ量別の精度比較を表 4，5 に示す．Joy，Sadness，Fear，Disgust，Trust においては，full の約 32 万件のデータの場合よりも精度が良い結果となった．残りの Anticipation と Surprise，Anger についても 15 万件の場合の結果と平均して 0.009 しか差がなかった．今回の実験においては，少なくとも

2) https://github.com/neologd/mecab-ipadic-neologd

10〜15 万件ほどデータがあれば感情分析において，十分な精度が達成されるということが確認された．一方で，感情極性の精度については，full が最も高かったため，さらなるデータ量を用いて検証する必要がある．

表 2　Predictions by emotion

	MAE							
	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust
Kajiwara et al,. (as reference)	0.734	0.666	0.899	0.684	0.218	0.344	0.443	0.432
LSTM	0.773	0.481	0.722	0.569	0.180	**0.246**	0.265	**0.428**
BERT + MeCab	0.693	0.442	0.699	0.578	0.179	0.275	0.264	0.468
BERT + DialectMeCab	0.691	0.440	0.700	0.578	0.178	0.268	0.258	0.476
DialectBERT + MeCab	0.658	**0.433**	0.658	0.562	**0.170**	0.260	**0.252**	0.468
DialectBERT + DialectMeCab	**0.646**	0.448	**0.652**	**0.552**	**0.170**	0.265	0.255	0.481

表 3　Predictions for emotion polarity

	Accuracy
Suzuki et al,. (as reference)	39．10%
LSTM	22.84%
BERT + MeCab	39.76%
BERT + DialectMeCab	40.88%
DialectBERT + MeCab	43.00%
DialectBERT + DialectMeCab	**43.12%**

表 4　Evaluation of sentiment analysis by data size

		MAE							
	size of data	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust
DialectBERT + DialectMecab	5k	0.666	0.453	0.684	0.571	0.173	0.273	0.249	0.476
	10k	0.658	0.442	0.667	0.561	0.174	0.263	**0.248**	**0.464**
	50k	0.659	0.431	0.656	0.562	0.171	0.263	0.251	0.475
	100k	0.654	**0.430**	0.675	0.570	0.174	0.265	0.254	0.483
	150k	**0.644**	0.434	0.665	0.564	0.173	**0.259**	0.258	0.475
	full	0.646	0.448	**0.652**	**0.552**	**0.170**	0.265	0.255	0.481

表 5　Evaluation of emotion polarity by data size

	size of data	Accuracy
DialectBERT + DialectMecab	5k	41.12%
	10k	40.84%
	50k	41.56%
	100k	40.92%
	150k	42.56%
	full	**43.12%**

4　炎上・ネットいじめ検知

4.1　データセット

炎上やネットいじめが発生している会話には，罵倒や脅しにつながる言葉が含まれていると考えられる．日本語で誹謗中傷とみなされる言葉をリストアップし，Twitter でこれらの言葉を含むデータを収集する．取得された会話の数は 7,547 件であった．これらデータをについて，炎上やネットいじめの発生有無を分類した．その結果 190 件が炎上やネットいじめを含む会話として分類された．これに，通常の会話の 190 件を加え，合計 380 件をデータセットとする．

4.2　手法

データセットと，感情分析にて獲得した 8 つの感情分析モデルを用いて，炎上・ネットいじめの発生予測を行う．これには，8 つの感情の強弱の組み合わせにより，その予測が可能であることを示す．各会話データから「感情ベクトル」という 8 次元ベクトルを作成する．その感情ベクトルを用いて，類似度を用いた手法と，機械学習アルゴリズムを用いた手法を検証する．感情ベクトルを作成する手順を図 2 に示す．8 つの感情分析モデルを用いて，各データの感情を評価する．図 2 の例では，JoyBERT を使用し，「こわっ (怖い)」，「まじきもい」などの発言のの喜びの程度を予測する．その後，各発話の評価値を平均し，その値をこの会話の Joy 値とする．他の感情についても同じ手順で感情の値を作成し，これらの値をまとめて 8 次元のベクトルを作成する．これを「感情ベクトル」と呼ぶ．

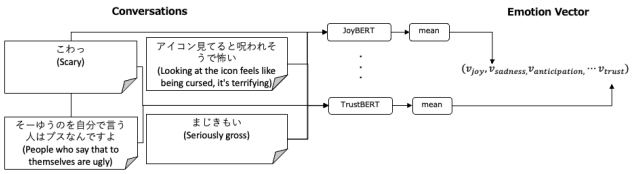

図 2　感情ベクトル作成手順

4.3　結果

4.3.1　ベクトル類似度

まず，類似度を比較する基準となる炎上・ネットいじめデータのみを利用して感情ベクトルを作成する．炎上・ネットいじめデータのみを用いているため，このベクトルを「炎上・ネットいじめベクトル」とする．この炎上・ネットいじめベクトルと評価データの感情ベクトルを cosine 類似度で比較する．80%以上の類似度で 92.1%の精度で炎上・ネットいじめと判定された．DialectBERT を用いた方が精度が高く，方言の理解が炎上・ネットいじめの検知に重要であることが明らかとなった．

表 6　Results of prediction using vector similarity.

	Accuracy of > 80% similarity	Accuracy of > 90% similarity
BERT and MeCab	90.70%	81.50%
Dialect BERT and Dialect MeCab	92.10%	86.80%

4.3.2　機械学習アルゴリズム

感情ベクトルを用いて，機械学習の分類アルゴリズムで検証を行った結果を表 7 に示す．SVM と RandomForest が最も精度が良く，共に 93.42%である．次に，感情の組み合わせによる比較を行った (表 8)．結果は，Anger，Fear と Disgust，Trust の組み合わせと，Disgust，Trust と Anticipation，Surprise の組み合わせが最も精度が良く，全アルゴリズムの平

表 7　Flaming and Cyberbullying detection accuracy

	Accuracy									
	SVM	**AdaBoost**	**Bagging**	**ExtraTrees**	**Gradient**	**RandomForest**	**KNeighbours**	**DecisionTree**	**ExtraTree**	**Average**
BERT and MeCab	90.79%	85.53%	88.16%	90.79%	89.47%	89.47%	89.47%	86.84%	85.53%	88.45%
Dialect BERT and Dialect MeCab	**93.42%**	88.16%	90.79%	92.11%	92.11%	**93.42%**	92.11%	88.16%	86.84%	**90.79%**

表 8　Result by combination of emotion pairs

	SVM	**AdaBoost**	**Bagging**	**ExtraTrees**	**Gradient**	**RandomForest**	**KNeighbours**	**DecisionTree**	**ExtraTree**	**Average**
Anger, Fear, Disgust, Trust										
BERT and MeCab	86.84%	88.16%	89.47%	90.79%	90.79%	88.16%	86.84%	86.84%	84.21%	88.01%
Dialect BERT and Dialect MeCab	**93.42%**	88.16%	92.11%	92.11%	90.79%	92.11%	92.11%	88.16%	90.79%	**91.08%**
Joy, Sadness, Anticipation, Surprise										
BERT and MeCab	82.89%	68.42%	78.95%	80.26%	77.63%	77.63%	82.89%	78.95%	75.00%	78.07%
Dialect BERT and Dialect MeCab	88.16%	81.58%	84.21%	82.89%	80.26%	80.26%	85.53%	78.95%	78.95%	82.31%
Anger, Fear, Joy, Sadness										
BERT and MeCab	88.16%	88.16%	88.16%	85.53%	86.84%	86.84%	86.84%	88.16%	82.89%	86.84%
Dialect BERT and Dialect MeCab	88.16%	88.16%	86.84%	89.47%	89.47%	89.47%	88.16%	88.16%	85.53%	88.16%
Anger, Fear, Anticipation, Surprise										
BERT and MeCab	88.16%	89.47%	90.79%	89.47%	90.79%	90.79%	90.79%	90.79%	86.84%	89.77%
Dialect BERT and Dialect MeCab	86.84%	88.16%	90.79%	88.16%	88.16%	89.47%	92.11%	88.16%	84.21%	88.45%
Disgust, Trust, Joy, Sadness										
BERT and MeCab	85.53%	88.16%	88.16%	86.84%	88.16%	88.16%	85.53%	86.84%	78.95%	86.26%
Dialect BERT and Dialect MeCab	90.79%	89.47%	86.84%	92.11%	**93.42%**	**93.42%**	88.16%	90.79%	88.16%	90.35%
Disgust, Trust, Anticipation, Surprise										
BERT and MeCab	88.16%	89.47%	88.16%	88.16%	89.47%	86.84%	86.84%	86.84%	86.84%	87.87%
Dialect BERT and Dialect MeCab	92.11%	89.47%	89.47%	**93.42%**	**93.42%**	**93.42%**	89.47%	90.79%	88.16%	**91.08%**

均は 91.08%であった．この結果は，表 7 の結果と比較しても 0.29 ポイント良い結果である．これにより，方言コーパスを用いた場合，少なくとも 4 つの感情分析モデルがあれば，精度の高い炎上・ネットいじめ検知が可能であることが確認された．これらの機械学習アルゴリズムを用いた場合の予測結果を見てみても，コサイン類似度での結果同様，DialectBERT を用いた場合の方の予測精度が高い結果であった．モデルが予測結果で誤認識したものを確認したところ，発言に攻撃的なワードが見られるものの，相手を揶揄していたり，通常面白いと感じた際に使われる表現のwや (笑) といった表現が、からかいの意図に含まれる場合に，炎上・ネットいじめが発生していない，と予測を誤った例が見られた．その逆としては，何か特定のニュース (今回確認した例では、危険運転) に，感情的に発言をしている会話の場合，炎上・ネットいじめが発生していると予測していた。

5　結論と今後の課題

本研究では，方言辞書と SNS の投稿データを用いた方言コーパスの作成と感情分析を行なった．また，そのモデルを利用した炎上・ネットいじめ検知の手法の検討し，その評価を行なった．今回評価した 8 つの感情のうち 6 感情と，感情極性において DialectBERT を利用することが有効であることを確認した．この結果により，方言を含むテキストがより書き手の感情を強く反映し，方言を利用することがモデルの感情理解に効果的であることを示した．また，学習データの量が精度に与える影響も検証した．今回の実験においては，少なくとも 10 万〜15 万件ほどデータがあれば感情分析において，十分な精度が達成されるということが示した．炎上・ネットいじめが起きている会話の判定においても，DialectBERT を用いた結果がもっと精度が高く，DialectBERT を利用することが検知に有効であることも示した．感情をより含む文として，今回の研究では方言に着目したが，同様に流行り言葉や若者言葉といったものも感情を含むと推察される．特に，若者が投稿するデータを対象にした分類においては，これらの言葉を考慮することがより精度をあげる手法であると考えられる．特に小中学生の間で発生するネットいじめという観点で考えると，それらは Twitter などの SNS 上で発生するのではなく，LINE やオンラインゲーム内のチャット機能の中の会話で発生することになる．そうした状況での検知を視野に入れると，一層，若者・流行り言葉の考慮の必要性が感じられる．

謝辞

Nguyen 教授には，本論文を執筆するにあたり研究の進め方や枠組みについて有益な助言をいただきました．この場を借りて深く御礼申し上げます．

参考文献

[1] 総務省. 令和３年通信利用動向調査報告書（世帯編）. 2021.
[2] 廣田壮一郎, 笹野遼平, 高村大也, 奥村学. 方言コーパス収集システムの構築. 2013.
[3] 総務省. 令和３年度 児童生徒の問題行動・不登校等生徒指導上の諸課題に関する調査結果について. 2022.
[4] Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. DziriBERT: a pre-trained language model for the algerian dialect. September 2021.
[5] Wissam Antoun, Fady Baly, and Hazem Hajj. AraBERT: Transformer-based model for arabic language understanding. February 2020.
[6] Muhammad Abdul-Mageed, Abdelrahim Elmadany, and El Moatez Billah Nagoudi. ARBERT & MARBERT: Deep bidirectional transformers for arabic. December 2020.
[7] Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. Pre-Training BERT on arabic tweets: Practical considerations. February 2021.
[8] Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. The interplay of variant, size, and task type in arabic pre-trained language models. March 2021.
[9] 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 **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.
[10] 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.
[11] Seiichi Ozawa, Shun Yoshida, Jun Kitazono, Takahiro Sugawara, and Tatsuya Haga. A sentiment polarity prediction model using transfer learning and its application to SNS flaming event detection. In **2016 IEEE Symposium Series on Computational Intelligence (SSCI)**, pp. 1–7, December 2016.
[12] Jianwei Zhang, Taiga Otomo, Lin Li, and Shinsuke Nakajima. Cyberbullying detection on twitter using multiple textual features. In **2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)**, pp. 1–6, October 2019.
[13] ” 高橋直樹, 檜垣泰彦”. Twitter における感情分析を用いた炎上の検出と分析. 信学技報, 2017.
[14] Tohoku NLP lab / 東北大学 乾研究室. https://www.nlp.ecei.tohoku.ac.jp/. Accessed: 2022-5-11.

A　学習パラメータ

表 9　Learning Parameters

		Further Pretraining	Fine Tuning
data size	training	258,657	30,000
	evaluation	32,332	2,500
	test	32,332	2,500
batch size			32
epoch		15	3
learning rate			2.0E-05
optimizer			AdamW
loss function			CrossEntropy
vocabulary size			32,000

B　炎上・ネットいじめベクトルのモデル別比較

図 3 は, BERT モデルと MeCab, および DialectBERT と DialectMeCab を使用した場合の，炎上・ネットいじめベクトルの値をグラフ表示したものである．どちらも怒りや嫌悪の感情が強く，恐怖，信頼，喜びなどの感情が少なくでている．特に DialectBERT と DialectMeCab を用いた場合の方が，それらの感情をより強く出ていることがわかる．

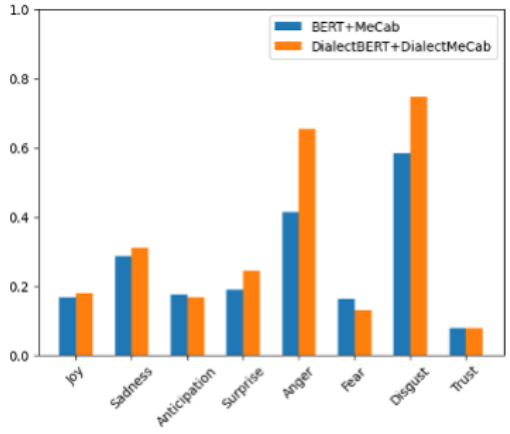

図 3　Vector values of flaming and cyberbullying conversations