JTweetRoBERTa: 大規模 SNS 投稿テキストによる事前学習と各種タスクによる性能検証

高須 遼 [1]　狩野 芳伸 [1]
[1] 静岡大学
{rtakasu,kano}@kanolab.net

概要

SNS 投稿は一般の書き言葉と異なり口語的かつ特有の傾向があり，書き言葉にもとづくモデルではうまく対応できない可能性がある．本研究では 6,000 万件のツイートデータで事前学習モデルを作成し，JGLUE，WRIME，メンタルヘルス不調の有無のツイート 2 値分類など 8 つのタスクでファインチューンを行い評価し，既存モデルとの比較を行った．すべての評価タスクで Wikipedia および CC100 で事前学習したベースラインを上回り，大規模ツイートで事前学習した提案手法の効果を示した．また，ほとんどの評価で多言語ツイート事前学習モデル TwHIN-BERT をも上回り，提案手法は日本語 SNS 投稿を対象とするタスクにおいて世界最高の性能であると考えられる．

1　はじめに

近年，SNS の利用の増加に比例して，インターネットには膨大な投稿テキストが存在する．SNS 関連のテキストの分析は様々な分野に応用が可能な一方で，あまりに膨大なデータを人手で処理することは難しい．そのため近年は，テキストに機械学習を用いる研究が頻繁に行われている．山内ら [1] は，日本語の学術文章を用いて BERT モデルを事前学習し，様々なタスクで性能の向上が見られたと報告している．また Sci-BERT[2] のようなドメイン特有のコーパスで学習されたモデルが配列タグ付け，文の分類，係り受け構文解析を含む一連のタスクで性能向上を報告している．しかし，SNS の投稿は一般の書き言葉と異なり，文が短い，文の区切りが明確でない，省略が多い，口語が多いなど話し言葉に近い傾向があり，書き言葉にもとづくモデルではうまく対応できない可能性がある．

Müller ら [3] は COVID-19 のトピックに関する大規模英語 Tweet コーパスで事前学習された事前学習モデルを公開した．Qudar ら [4] は，数百万件の英語ツイートに対して事前に訓練された，ドメインに特化した言語表現モデルである 2 つの TweetBERT モデルを公開した．また，7 つの BERT モデルを 31 の異なるデータセットで評価することで，広範な分析を提供した．Zhang ら [5] は，Twitter で作成された BERT ベースの多言語言語モデルである TwHIN-BERT を公開した．日本語だけでも 10 億ツイート程度が学習に用いられている．マスク言語モデルに加え，リプライ等で紐づいた「社会的に関連する」投稿群を Twitter 異種情報ネットワーク（TwHIN）として収集し，その予測を同時学習させている．また，ハッシュタグ予測と社会的関連性のベンチマークデータセットとして提供した[1)]．

我々は SNS 投稿データを用いた RoBERTa ベースの事前学習モデルを提案する．[2)] Twitter は SNS の中でも国内利用者が多く，データ取得 API も整っているため多くのデータを収集することが出来，多様なパターンを学習できる．我々は 6,000 万件を超える投稿データを収集し，事前学習を行った．

学習した事前学習モデルの性能評価のため，日本語理解能力測定を目的としたベンチマーク JGLUE[6] の 5 つのタスク，感情強度分析を目的とした主観と客観の感情分析データセット WRIME[7]，メンタルヘルス不調の有無でのツイート 2 値分類 [8]，診断付き被験者データセットによるツイート 2 値分類，計 8 つのタスクで性能評価を行った．

すべての評価で Wikipedia および CC100 で事前学習したベースラインを上回り，大規模ツイートで事前学習した提案手法の効果を示した．また，ほとんどの評価で多言語ツイート事前学習モデル TwHIN-BERT をも上回り，提案手法は日本語 SNS 投稿を対象とするタスクにおいて世界最高の性能であ

1）　ツイート ID から原ツイートを取得しないと実験できない．
2）　事前学習済みモデルは一般公開予定である．

ると考えられる．

2 データセット

2.1 JGLUE

日本語言語理解ベンチマーク JGLUE[6] に含まれる 5 つのタスクを説明する．それぞれに用いられたデータ数の内訳を表 1 に示す．提案手法モデルが一般的な日本語タスクでも十分な性能であることの検証として用いる．

MARC-ja 商品レビューを入力として，ポジティブかネガティブかを推定する．

JSTS 2 つの文が与えられ，文間の類似度 (0 から 5，5 が最も類似) を推定する．

JNLI 2 つの文が与えられ，含意，矛盾，中立のいずれの推論関係かを推定する．

JSQuAD 数文からなる段落とそれに関連する質問が与えられ，段落から抜き出す形で答える．

JCommonsenseQA 常識推論能力を評価するための 5 択の選択式問題を解く．

表 1 JGLUE 及び WRIME データセットの統計

データセット	train	valid	test
MARC-ja	154545	19318	19319
JSTS	11126	1390	1392
JNLI	18005	2250	2252
JSQuAD	615	76	78
JCommonsenseQA	8046	1005	1007
WRIME	40000	1200	2000

2.2 WRIME

WRIME[7] は，喜び，悲しみ，期待，驚き，怒り，恐れ，嫌悪，信頼の基本 8 感情に各 4 段階の強度ラベルを割り振っている．データ統計を表 1 に示す．

2.3 メンタルヘルス不調者データセット

我々が独自に収集した，メンタルヘルスの不調が推測される Twitter アカウントの投稿データセットである [8]．メンタルヘルス不調群の中核を占めると想定される，統合失調症圏，気分（感情）障害圏及び神経症性障害，ストレス関連障害及び身体表現性障害圏に対して，精神症状の軽減を目的に処方される薬剤（抗精神病薬，抗うつ薬，抗不安薬）の商品名（79 種類）を含む文字列をツイートしたことのあるユーザのうち，プロフィール欄に先の疾患群の訴えや症状に関連する単語を含むアカウントのツイートを抽出した．いくつかの前処理を行い，最終的に，15,266 アカウント，合計 34,505,381 件のツイート集合を正例，すなわちメンタルヘルス不調群のデータセットとした．

一般のツイート集合全体ではメンタルヘルス不調群に属する割合が十分低いと仮定し，先に述べたメンタルヘルス不調者データセット以外の全ツイート集合からランダムに取得したデータを非メンタルヘルス不調群とみなした．結果，16,519 件のアカウント，合計 28,685,425 ツイートを抽出した．取得期間は同じである．

2.4 UNDERPIN 診断付きデータセット

我々は，JST CREST 研究課題「自然言語処理により心の病の理解: 未病で精神疾患を防ぐ」や JST AIP 加速研究課題「精神医学×メディア解析技術の展開：精神疾患への介入の挑戦」の支援を受け，慶應義塾大学病院とその協力病院で募集した精神疾患患者と健常者を対象とし，許可を得た被験者の Twitter 投稿データを収集してきた．本研究ではこの収集した被験者投稿データも評価に用いる．精神疾患患者のうち CGI-S コード[3] [9] が 3 以上の方をメンタルヘルス不調群としてラベル付けした．CGI-S コードとは，疾患の重症度を 1(正常) から 7(非常に重度) の 7 段階で評価する尺度である．統計を表 2 に示す．

表 2 UNDERPIN 診断付き被験者データセットの統計

	アカウント数	ツイート数
精神疾患患者	39	230603
健常者	22	88457

3 事前学習

事前学習の訓練データとして，メンタルヘルス不調者群のツイート集合とメンタルヘルス非不調者群のツイート集合を合わせて用いた．データの統計を表 3 に示す．トークナイザには byte-pair-encoding による SentencePiece[4] を用いた．事前学習の訓練データに対し，1 ツイートを単位として改行で区切ったテキストファイルを SentencePiece の作成に用いた．語彙サイズには 50000 を設定した．

表 3 事前学習に用いたデータセットの統計

	アカウント	ツイート
メンタルヘルス不調群	15,266	34,505,381
非メンタルヘルス不調群	16,519	28,685,425

3) https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1265746

4) https://github.com/google/sentencepiece

事前学習タスクは，Masked Language Model（MLM）のみを行った．実装は huggingface[5] の提供する RoBERTa[6] を用いた．訓練データに含まれるツイートを時系列順につなげ，前述のトークナイザにより トークン数が 250 以下になるごとに分割し 1 サンプルとすることで，ツイートを途中で分割することなくできるだけ長い文脈情報を持たせるようにした．事前学習を行う際は，訓練データとテストデータが 8:2 になるようランダムシャッフルして分割した．

事前学習は 100 万ステップまで回し，5 万ステップごとに保存したモデルのうちで最も損失が小さいモデルを選んだ．事前学習の各種パラメータ設定は，付録の表 8 に記載する．

4 性能検証実験と評価結果

性能検証実験では，提案手法の事前学習モデルと，ベースラインとして学習データ違いといえる Wikipedia および CC100 で事前学習した RoBERTa モデル，およびマルチリンガルツイート事前学習モデル TwHIN-BERT を，それぞれファインチューンし評価結果を比較した．JGLUE のファインチューンに関しては，付属して公開されているコードを用いることで条件をそろえた．それ以外のファインチューンに関しては，詳細を付録 A.1 に記す．

4.1 JGLUE

MARC-ja と JSQuAD の max_seq_length に関しては，事前学習モデルの最大長に合わせて 250 とした．その他の設定に関してはデフォルトである．評価結果を表 4 に示す．公開データセットは訓練と検証データしかなく，検証とテストを同じデータで行っていたため，それらを連結し訓練，検証，テストを 8:1:1 の割合でランダムに抽出した．

5） https://huggingface.co/
6） https://huggingface.co/camembert-base

4.2 WRIME

8 つのカテゴリそれぞれに，4 段階の強度ラベルを予測する多クラス分類タスクを行った．こちらも，前節と同じように評価結果を表 5 に示す．評価指標はカッパ係数を用いた．

4.3 メンタルヘルス不調の有無 2 値分類

データセットは，事前学習と同じようにトークン数が 250 以下になるまで，ツイートを時系列順につなげて一つの入力にした．同じ文字列の 3 回以上の繰り返しがあった場合，3 回目以降を削除し 2 回分までとした．これにより，一度の入力がより多くの文脈情報を含みうるようにした．（例：あああああ→ああ，OMGOMGOMGOMG → OMGOMG）
この際，メンタルヘルス不調アカウント抽出に使用した，薬剤名とメンタルヘルス不調を直接示唆する単語を含むツイートは除去した．これら単語をツイートに含む場合は，メンタルヘルス不調の可能性が高いが，そうした直接的な単語表現以外の情報からの予測がこのデータセットの狙いだからである．訓練時間と計算機資源の限界のため，正例負例で 100 万ずつのトークンブロックを，上記のデータセットからさらにランダムにデータを抽出し，ファインチューン用データセットを作成した．

評価においては，ツイートを 250 トークンごとにまとめたブロック単位と，アカウント単位それぞれで評価を行った．ブロック単位で評価した場合，ツイート総数の多いアカウントの影響が大きくなる可能性があるためである．アカウント単位での評価では，対象アカウントのツイートブロックに対する 2 値分類推測結果がツイート単位で多い方の分類を当該アカウントの推測結果として用いた．評価結果を表 6 に示す．いずれも 5 分割交差検証における各 fold の評価値の平均を掲載した．

表 4 JGLUE の評価結果

	MARC-ja Accuracy	JSTS Pearson/Spearman	JNLI Accuracy	JSQuAD F1	JCommonsenseQA Accuracy
提案手法	**0.962**	**0.841**	**0.907**	**0.825**	0.744
ベースライン	0.953	0.790	0.881	0.809	0.685
TwHIN-BERT	**0.962**	0.836	0.869	0.819	**0.751**

表 5 WRIME の評価結果

	喜び	悲しみ	期待	驚き	怒り	恐れ	嫌悪	信頼
提案手法	**0.795**	0.652	**0.764**	**0.635**	**0.529**	0.604	**0.639**	**0.338**
ベースライン	0.770	0.566	0.713	0.586	0.390	0.502	0.563	0.276
TwHIN-BERT	0.788	**0.665**	0.756	0.631	0.520	**0.613**	0.611	0.315

4.4 UNDERPIN 診断付きデータ 2 値分類

データ数が少ないため評価のみに利用することとし，前節の 2 値分類により得られた 5fold モデルそれぞれで推論を行い平均した．結果を表 6 に示す．

4.5 評価結果

すべての評価タスクで Wikipedia および CC100 で事前学習したベースラインを上回り，大規模ツイートで事前学習した提案手法の効果を示した．また，ほとんどの評価で多言語ツイート事前学習モデル TwHIN-BERT をも上回った．

5 考察

5.1 ベースラインとの比較

メンタルヘルス不調者分類について，提案手法では正解，ベースラインで不正解だったツイート例を付録の表 9 に示す．また，UNDERPIN 診断付きデータセットについて，提案手法が正解，ベースラインが不正解だった例を付録の表 10 に示す．

いずれの例も，SNS 投稿に特有の，書き言葉には見られない口語的な語彙や表現が多数含まれており，提案手法の効果があったと考えられる．

5.2 TwHIN-BERT との比較

TwHIN-BERT は事前学習に 100 言語計 90 億ツイートを用い，うち日本語ツイートは 10 億強であったが，我々は 6000 万件を使用した．バッチサイズは 128 に対し 64，学習ステップ数は MLM のみ版で 50 万ステップ・目的関数を加えた同時学習あり版でさらに追加 50 万ステップに対し 100 万ステップであった．すなわち，全体の学習ステップ数は同等であるが，日本語は全体の 10%強程度のうえ少数言語にあわせてリサンプルしたとあり，実質的に日本語で訓練できた回数は一桁以上少なく学習が不十分な可能性がある．日本語データ量は一桁多いが，TwHIN-BERT の報告でも large モデルが base モデルに劣る評価が含まれており，これ以上訓練データを増加させても性能向上に貢献しないかもしれない．

入力最大長が 128 トークンに対し我々は 250 トークンと倍長い．また我々はツイートを連結してブロックにすることで，入力後半にあたる部分も学習が進むよう工夫しているが，TwhinBERT では特に記載がなくツイート単位で学習をしている可能性があり，そのことも性能差に表れていると考えられる．

6 おわりに

独自に収集した日本語 SNS 投稿テキスト約 6,000 万件で RoBERTa の事前学習を行った．その有用性を示すため，JGLUE，WRIME，メンタルヘルス不調の有無の 2 値分類および診断付きデータセットによる評価の 8 つのタスクでファインチューンと評価を行った．ベースラインとして RoBERTa を日本語 Wikipedia および CC100 で事前学習したモデルと，多言語ツイート事前学習モデル TwhinBERT を用いて比較し，提案手法の性能が上回ることを示した．今後の展望として，英語をはじめとした多言語に対応するモデルの学習や，アカウント間の関係性を取り入れたモデルが挙げられる．

表 6 メンタルヘルス不調者 2 値分類および UNDERPIN 診断付きデータの評価結果

	Accuracy	Recall	Precision	F1 score
メンタルヘルス不調者 2 値分類の評価結果（ツイートブロック単位）				
提案手法	**0.827** (1654244/2000000)	**0.814** (813710/1000000)	**0.836** (813710/973176)	**0.825**
ベースライン	0.702 (1404368/2000000)	0.774 (773830/1000000)	0.677 (773830/1143392)	0.722
TwHIN-BERT	0.788 (1576980/2000000)	0.793 (792624/1000000)	0.786(792624/1008268)	0.789
メンタルヘルス不調者 2 値分類の評価結果（アカウント単位）				
提案手法	**0.896** (19236/21471)	**0.904** (11784/13040)	**0.923** (11784/12763)	**0.913**
ベースライン	0.811 (17417/21471)	0.892 (11632/13040)	0.815 (11632/14278)	0.852
TwHIN-BERT	0.885(18998/21471)	0.896(11692/13040)	0.919(11692/12711)	0.907
UNDERPIN 診断付きデータの評価結果（ツイート単位）				
提案手法	**0.728** (8958/12307)	**0.757** (7102/9377)	**0.869** (7102/8176)	**0.809**
ベースライン	0.627 (7717/12307)	0.756 (5291/6996)	0.647 (5291/8176)	0.697
TwHIN-BERT	0.659 (8110/12307)	0.634 (5181/8176)	0.812 (5181/6383)	0.712
UNDERPIN 診断付きデータの評価結果（アカウント単位）				
提案手法	**0.656** (40/61)	**0.769** (30/39)	0.714 (30/42)	**0.741**
ベースライン	0.508 (31/61)	0.487 (19/39)	0.655 (19/29)	0.559
TwHIN-BERT	0.623(38/61)	0.615(24/39)	**0.750**(24/32)	0.676

謝辞

本研究は JSPS 科研費 JP22H00804, JP21K18115, JP20K20509, JST AIP 加速課題 JPMJCR22U4, およびセコム科学技術財団特定領域研究助成の支援を受けて行われた.

参考文献

[1] 山内洋輝, 梶原智之, 桂井麻里衣, 大向一輝, 二宮崇. 学術ドメインに特化した日本語事前訓練モデルの構築. 言語処理学会 第 29 回年次大会 発表論文集, 2023.

[2] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3615–3620, 11 2019.

[3] Martin Müller, Marcel Salathé and Per E. Kummervold. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. In **Frontiers in Artificial Intelligence**, 2023.

[4] Mohiuddin Md Abdul Qudar and Vijay K. Mago. Tweetbert: A pretrained language representation model for twitter text analysis. In **arXiv:2010.11091**, 2020.

[5] Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In **KDD '23: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining**, 2023.

[6] 栗原健太郎, 河原大輔, 柴田知秀. Jglue: 日本語言語理解ベンチマーク. 言語処理学会 第 28 回年次大会 発表論文集, 2022.

[7] 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.

[8] 高須 遼, 中村 啓信, 岸本 泰士郎, 狩野 芳伸. 大規模ツイートデータを用いたメンタルヘルス不調者の推測. , 人工知能学会全国大会論文集, 2022.

[9] William Guy, editor. **ECDEU Assessment Manual for Psychopharmacology, Revised**. , US Department of Health, Education, and Welfare Publication (ADM); Rockville, MD: National Institute of Mental Health, 76-338, 1976.

[10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, Vol. 86, No. 11, pp. 2278–2324, 1998.

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **arXiv:1711.05101**, 2018.

[12] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In **Advances in Neural Information Processing Systems**, Vol. 33, pp. 2958–2969. Curran Associates, Inc., 2020.

A　付録

A.1　ファインチューンの詳細設定

事前学習モデルの隠れ層の重みを凍結し，最終層を含め高次 4 層分を連結（concatenate）したものに平均値プーリング [10] を適用して次元を削減したうえで，2 値分類用に全結合層を 2 層（伝達関数は ReLU）追加し，ファインチューンを行った．[7]Roberta モデルの学習時パラメータ設定を付録表 8 に示す．学習時のパラメータの最適化手法は PyTorch が提供する AdamW[8] [11] を使用した．AdamW に関するパラメータも同じように表 8 に記載した．また，敵対的な例に対するディープニューラルネットワークの頑健性を向上させる敵対的重み摂動（Adversarial Weight Perturbation：AWP）[12] を用いた．

表 7　事前学習パラメータ

max_length	250
layer_num	12
num_hidden_layers	6
num_a-head	12
max_position_emb	514
hidden_size	768
optimizer	AdamW
learning_rate	5.00e-5
step_num	900K
batch_size	64
max_steps	1M
precision	fp16
tokenizer	sentencepiece-unigram
vocabulary	50000
char_coverage	0.9995
GPU	Quadro RTX 8000 2sheets
learning_period	2months

表 8　ファインチューン時パラメータ

hidden_act	gelu
initializer_range	0.02
layer_norm_eps	1e-12
max_position_embeddings	514
num_attention_heads	12
num_hidden_layers	6
encoder_lr	3e-5
decoder_lr	3e-4
epochs	5
max_grad_norm	1000
Dropout	0.2
lr_scheduler	linear
weight_decay(AdamW)	0.01
lr(AdamW)	1e-5
eps(AdamW)	1e-6
betas(AdamW)	(0.9, 0.999)
adv_eps(AWP)	1e-3
adv_lr(AWP)	5e-5

表 9　ツイート例 1（文脈を失わない程度に匿名化加工）

送迎の車多すぎて帰れん〜ガソリンも残ってないよやばいやばい 他部署の方がめちゃ話しかけてくれるからいつも楽しく仕事できる〜 女神ハイリアが転生したのがゼルダってことはハイリア像に話しかける時いつも話しかけてくるのは誰なんや？？ん？？久しぶりにゼルダやりたいから行ってきたけど無いやん,, 取り寄せできるか聞いてみたけど多分無理やな，この感じ 明日遊びの予定あるから残業出来ないんだけど引継ぎせんとあかんから早く行こう 朝マック食べて頃合い見て本屋でべんきょうでもするかーー，ア，あとなんか漫画でも買ってこいやー，最近ゲームばっかやってたから残業に対する耐性消えまくってるげ，でも転職する度胸もないんだよねー

表 10　ツイート例 2（文脈を失わない程度に匿名化加工）

なんて笑い上戸なんだ！！〇〇様の酔っ払い配信だ♡♡ 夕飯まで時間あるしめちゃ見れる！！ こんばんちゃんこぉぉ，今日も酔っ払い配信♡いっぱい見れるから幸せ♡甥っ子が絶賛大学留年してて，その後大学院？ってとこ行ったから金持ってるだろうし正月ねだってみよ♡あけおめ！！！♡♡ 久々に配信するよん！！ 視聴者にまだお年玉もらってるんでちゅ〜ってひといてちょーうらやましい！！あたしにもそんなときがありますた涙 マツコデラックスの物まねしてって言われたから配信でやったら同接人数ちょっと減ったんだけど！！ふざけんな！！戻ってきて〜〜〜〜♡♡ あけましておめでとう！！ 最近配信継続サボってたから頑張ります！！だからみんな見に来て♡久々に BLOG 更新したお<

7）隠れ層を凍結したのは，事前学習により得られた隠れ層の表現を破壊し，過学習を起こす可能性がある一方，我々の大規模 SNS 投稿データでの事前学習により，隠れ層は各トークンの表現を十分に獲得しており，追加的な学習の必要が低いと考えたためである．最終層を含め高次 4 層分を連結して利用したのは，Transformer モデルは層を追うごとに，表層的特徴，構文的特徴，意味的特徴というような異なる表現を学習すると考えられ，これらの特徴を低次の層のものも含め利用するためである．この連結した層の出力の次元を削減するには，[CLS] トークンに対応するベクトルを用いる手法もあるが，試したところ平均値プーリングのほうが性能が良かったためこちらを採用した．

8）https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html