日本語レビューに対するレーティング予測の精度比較

森廣 勇樹 [1]　南條 浩輝 [2]　馬 青 [1]
[1] 龍谷大学理工学研究科　[2] 滋賀大学データサイエンス学部
[1]t22m005@mail.ryukoku.ac.jp
[2]hiroaki-nanjo@biwako.shiga-u.ac.jp
[1]qma@math.ryukoku.ac.jp

概要

本研究では多言語 amazon レビューコーパスを用い，日本語のレビュー文から評価点（レーティング）がどの程度正確に予測できるかを調査した．BERT と RoBERTa に対し合計 360 通りのハイパーパラメータの組み合わせでグリッドサーチを行った．得られたハイパーパラメータで予測を行うと，BERT より RoBERTa の方が精度がよいことがわかった．英語レビューへのレーティング予測の関連研究の結果に比べ，精度があまり高くはないがレーティング予測の傾向自体の学習はできていると考える．

1　はじめに

本稿では，日本語の商品レビュー文に対し星数のような評価点（レーティング）がどの程度正確に予測できるかの調査について述べる．

商品レビューは消費者の正直な感想であるため広告より信頼性が高いが，ヤラセ・サクラなど，レビュー の悪用や利用者増加に伴う役に立たないレビューの増加で，真に有用なレビューが埋もれている．推薦・推薦文においてこのようなレビュー等が主に使われており，それらの結果に疑問が残るのではないかと考える．それに比べ，Twitter のツイートなどの生の声は，レビュー文よりも実際の気持ちを表していると予想される．

本研究では Twitter のツイートデータを用いた推薦文（及びその根拠文）の生成することを目標としている．第一歩として，Twitter のツイートなどレビューそのものを目的としていない生のユーザの声（テキスト）が，商品に対してどの程度のお勧め度を表しているか (レーティング) を予測する必要がある．しかしながら，ツイートデータには正解が付いていないため学習に用いるのが困難である．多言語 amazon レビューコーパスに正解データが付いており，最終的にツイートデータに対するレーティング予測を行うとしても，その予測に用いる機械学習の学習データとしては使えると考える．そこで本研究ではまず，多言語 amazon レビューコーパスを用い，日本語でのレビューのレーティング予測を調査した．

2　関連研究

Liu [1] は英語レビューのレーティング予測を機械学習と深層学習を用いて行っている．コーパスはレビュー文と星数のラベルのペアとなっており，レビュー文に対してのレーティング（整数値）を予測するタスクとなっている．この研究では，Naive Bayes，Logistic Regression，Random Forest，Linear Support Vector Machine の 4 つの機械学習モデルと BERT，DistilBERT，RoBERTa，XLNet の 4 つの transformer ベースの深層学習モデルを使用し，モデル間での精度の比較やモデルの学習時間の比較も行っている．一方，日本語レビューのレーティング予測についての研究は見当たらない．

3　手法

Liu の研究では，4 つの機械学習モデルと 4 つの transformer ベースの深層学習モデルを使用している．その実験結果から機械学習モデルの予測精度が低いことがわかったため，本研究では 4 つの深層学習モデルのうち，まず BERT [2] と RoBERTa [3] を用いることにした．

BERT のモデルとして，東北大学乾・鈴木研究室の Wikipedia で訓練済み日本語 BERT モデル (cl-tohoku/bert-base-japanese-v2) [4] を使用した．BERT による分類モデルの入出力例を図 1 に示す．

図 1 のように文（文章）を入力し，出力 C トークンから全結合層の分類器 1 層（classifier）を通して 5 クラス分類を行う．

表 1　データセットの星評価の内訳

星評価＼データ	学習データ	検証データ	テストデータ
★	37,000	4,000	1,000
★★	37,000	4,000	1,000
★★★	37,000	4,000	1,000
★★★★	37,000	4,000	1,000
★★★★★	37,000	4,000	1,000

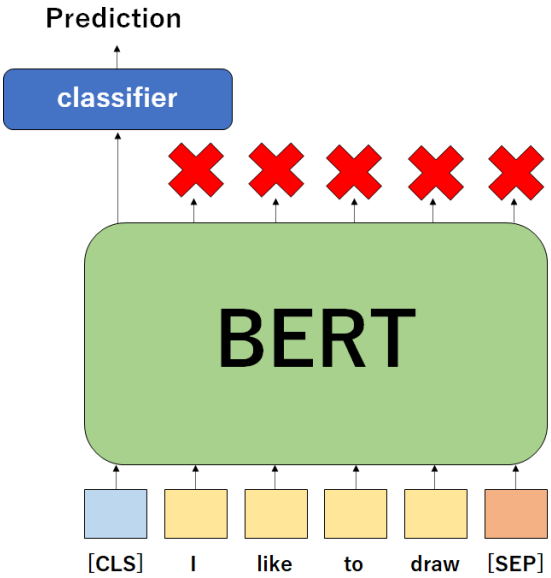

図 1　BERT による分類モデルの入出力例

RoBERTa のモデルは，xlm-roberta-base を Language Identification データセットに基づいて fine-tuning した RoBERTa モデル（papluca/xlm-roberta-base-language-detection）[5] を使用した．

4　実験

4.1　データセット

多言語 amazon レビューコーパス [6] と呼ばれるデータセットを用いて実験を行った．データセットには，英語，日本語，ドイツ語，フランス語，中国語，スペイン語の合計 6 言語のレビューが含まれておりその中で日本語のみのレビューを使用した．データセットの各レコードには，レビューテキスト，レビュータイトル，星評価，匿名のレビュアーが含まれている．各言語ごとに合計 210,000 のレコードがあり，185,000 を学習データ，20,000 を検証データ，5,000 をテストデータとした．すべてのレビューは 2,000 文字を超えると切り捨てられ，すべてのレビューは少なくとも 20 文字の長さである．本研究で使用するデータセットの星評価の内訳を表 1 に示す．

表 1 よりコーパスは星評価（星 1〜星 5）でバランスが取れているため，各星評価は各言語のレビューの 20 % で構成されている．

表 2　モデルのハイパーパラメータ

パラメータ＼モデル名	BERT	RoBERTa
最適化アルゴリズム	RMSProp	RMSProp
学習率	1e-05	5e-06
バッチサイズ	8	16
エポック数	2	3

4.2　ハイパーパラメータの決定

各モデルに対して適切なハイパーパラメータを決定するためにグリッドサーチを行った．ハイパーパラメータとして最適化アルゴリズム，学習率，バッチサイズ，エポック数の 4 つのパラメータを可変としグリッドサーチを行った．最適化アルゴリズムは，AdamW，SGD，RMSProp の 3 種類，学習率は各モデル，各最適化アルゴリズムによって変わるが 3 種類ずつ行い，バッチサイズは 32, 16, 8, 4 の 4 種類，エポック数は 1 から 10 の 10 種類，合計 360 通りのハイパーパラメータの組み合わせでグリッドサーチを行った．検証データはハイパーパラメータを決定するために使用し，決定後は学習データと合わせて実験の学習に使用する．この組み合わせの中で最も検証データでの正答率が高いパラメータを最適なものとしハイパーパラメータを決定した．得られたハイパーパラメータと RoBERTa でのグリッドサーチの結果の一部を表 2，表 3 に示す．

表 2 の条件の通り BERT の最適化アルゴリズムは RMSProp，学習率は 1e-05，バッチサイズは 8，エポック数は 2 としたとき検証データでの精度

表 3 RoBERTa のグリッドサーチの結果（一部）

エポック数	学習率		
	1e-06	5e-06	1e-05
1	0.5386	0.5550	0.5589
2	0.5573	0.5783	0.5800
3	0.5641	**0.5852**	0.5835
4	0.5641	0.5825	0.5825
5	0.5657	0.5820	0.5787
6	0.5659	0.5830	0.5811
7	0.5681	0.5797	0.5793
8	0.5704	0.5825	0.5742
9	0.5725	0.5805	0.5729
10	0.5719	0.5804	0.5730

（Accuracy）が 0.5531 と 360 通りのパラメータの中で一番高い結果であったためこのハイパーパラメータで決定し，RoBERTa の最適化アルゴリズムは RMSProp，学習率は 5e-06，バッチサイズは 16，エポック数は 3 としたとき検証データでの精度（Accuracy）が 0.5852 と一番高い結果であったためこのハイパーパラメータで決定した．

表 3 は RoBERTa の最適化アルゴリズムが RMSProp，バッチサイズが 16 のときのグリッドサーチの結果である．グリッドサーチを行って気づいたこととして，よりよいパラメータは学習率の値が小さいときエポック数が大きくなり，学習率の値が大きいときエポック数が小さくなる傾向にあった．

4.3 実験結果

4.2 節で可変にしたパラメータ以外を固定にして精度比較を行った．データセットは多言語 amazon レビューコーパスを使用し，205,000 を学習データ，5,000 をテストデータとし精度の算出を行った．その条件をもとに学習を行った際のテストデータでの精度を表 4 に示す．

表 4 の各モデルの精度を比較すると BERT では Accuracy が 0.5624，F score が 0.5612 に対し RoBERTa では Accuracy が 0.5956，F score が 0.5893 となり RoBERTa の方がより精度がよいことがわかる．各モデルの Confusion Matrix（混同行列）を表 5 と表 6 に示す．

表 4 の精度は Liu の英語に関する先行研究の 0.6-0.7 より低かったが，レーティング予測は人手評価からそれほどかけ離れていないことが表 5 と表 6 の Confusion Matrix からわかる．すなわち，レーティング予測の傾向自体の学習はある程度できていると考える．

表 4 各モデルの Accuracy と F score

モデル名	BERT	RoBERTa
Accuracy	0.5624	**0.5956**
F score	0.5612	**0.5893**

表 5 BERT の Confusion Matrix

人手評価クラス ＼ レーティング予測クラス	1	2	3	4	5
1	**0.65**	0.29	0.04	0.01	0.01
2	0.20	**0.53**	0.21	0.04	0.01
3	0.07	0.24	**0.44**	0.20	0.06
4	0.01	0.05	0.21	**0.46**	0.28
5	0.01	0.02	0.05	0.20	**0.73**

表 6 RoBERTa の Confusion Matrix

人手評価クラス ＼ レーティング予測クラス	1	2	3	4	5
1	**0.75**	0.20	0.04	0.01	0.01
2	0.25	**0.51**	0.20	0.03	0.01
3	0.07	0.22	**0.45**	0.19	0.07
4	0.01	0.04	0.15	**0.48**	0.32
5	0.01	0.01	0.03	0.17	**0.79**

5 おわりに

本研究では多言語 amazon レビューコーパスを用い，日本語のレビュー文に対しレーティングがどの程度正確に予測できるかを調査した．BERT と RoBERTa に対し合計 360 通りのハイパーパラメータの組み合わせでグリッドサーチを行った．グリッドサーチを行った結果，BERT のハイパーパラメータは，最適化アルゴリズムが RMSProp，学習率が 1e-05，バッチサイズが 8，エポック数が 2 で決定し，RoBERTa のハイパーパラメータは，最適化アルゴリズムが RMSProp，学習率が 5e-06，バッチサイズが 16，エポック数が 3 で決定した．これらのハイパーパラメータでレーティング予測を行った結果，BERT では Accuracy が 0.5624，F score が 0.5612 に対し，RoBERTa では Accuracy が 0.5956，F score が 0.5893 となり，RoBERTa の方がより精度がよいことがわかった．精度は Liu の英語レビューに関する先行研究の 0.6-0.7 より低かったが，レーティング予測が人手評価からそれほどかけ離れていないことか

ら，レーティング予測の傾向自体の学習はある程度できていると考える．

次のステップとして，amazon の日本語レビューコーパスで学習した BERT と RoBERTa を用い，ツイートデータのレーティング予測を行う予定である．

謝辞

本研究は JSPS 科研費 19K12241 の助成を受けたものです．

参考文献

[1] Zefang Liu. Yelp review rating prediction: Machine learning and deep learning models. **arXiv**, 2020. https://arxiv.org/abs/2012.06690.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. **arXiv**, 2018. http://arxiv.org/abs/1810.04805.

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. **arXiv**, 2019. http://arxiv.org/abs/1907.11692.

[4] 東北大学乾・鈴木研究室 bert モデル. cl-tohoku/bert-base-japanese-v2. https://huggingface.co/cl-tohoku/bert-base-japanese-v2.

[5] RoBERTa モデル. papluca/xlm-roberta-base-language-detection. https://huggingface.co/papluca/xlm-roberta-base-language-detection.

[6] Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. The multilingual amazon reviews corpus. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**, 2020.