自由記述からセルフ・コンパッションを推定することは可能か？ —BERT による心理学的構成概念の定量化—

岡野裕仁[1] 河原大輔[2] 野村理朗[1]
[1]京都大学 [2]早稲田大学理工学術院
{okano.hirohito.38m@st., nomura.michio.8u@}kyoto-u.ac.jp, dkw@waseda.jp

概要

苦しい状況や失敗を経験したときに自分に向けられる「自分への思いやり」を、心理学ではセルフ・コンパッション (self-compassion) と呼ぶ。従来、セルフ・コンパッション等の心理過程に関わる個人差の多くは、自己報告式の質問紙尺度によって測定・定量化されてきた。本研究は、人々から回答の自由度が高い記述式テキストを収集し、質問紙スコアを予測する BERT 回帰モデルを構築した結果、自由記述から個人のセルフ・コンパッションを高精度に推定しうることを示した。自由記述テキストによる心理の定量化は、質問紙と比べて先入観やバイアスが生じにくいという利点があり、本研究の結果から、今後の研究展開が期待される。

1 はじめに

私たちは、日々の生活において、様々な苦しみを感じたり、自分の限界にぶつかる。そのようなとき、どのように感じ考えるかには個人差がある。例えばある人は、つらい状況で、他の人々を羨んで孤独感を感じ、自分自身のふがいなさに苛立つかもしれない。その一方、そのような状況においても、弱点や欠点をもつ自分を受け入れ、自分自身に思いやりを向けることもできる。

心理学では、苦しい状況や失敗を経験したり、自分の至らなさに気づいたりしたときに自分に向けられる思いやりのことを、セルフ・コンパッション (self-compassion) と呼ぶ。例えば、セルフ・コンパッションが生じやすい個人ほど、抑うつ・ストレス等の精神病理やネガティブ感情を経験することが少なく[1]、人生満足度が高いことが示されている[2]。また、セルフ・コンパッションを育むような介入は、精神病理を有意に軽減させ、幸福感やウェルビーイングを向上させることが判明している[3]。過去 10 年ほどでセルフ・コンパッションに関する研究は指数関数的に増加しており、現在では心理学を中心に、医学・看護学・教育学などの関連分野を含む様々な領域で研究が進められている[4]。

セルフ・コンパッションのような、心理学およびその関連分野で理論的に想定される概念を、心理学的構成概念 (psychological construct) と呼ぶ[5]。心理学的構成概念は、人間の心理や行動を理解するために有用であると考えられており、従来は自己報告式の質問紙尺度によって測定・定量化されてきた。例えば、人々のセルフ・コンパッションの個人差を測定する質問紙尺度として最も広く使用されている尺度に、Self-Compassion Scale (SCS) [6] がある。SCS を用いて人々のセルフ・コンパッションを測定する際には、計 26 個の文章 (例：「苦労を経験しているとき、必要とする程度に自分自身をいたわり、やさしくする」、「自分自身の欠点と不十分なところについては、やさしい目で見るようにしている」、「感情的な苦痛を感じているとき、自分自身にやさしくする」) が呈示され、参加者は各文章に普段の自分がどの程度あてはまるかを「1. ほとんど全くない」～「5.ほとんどいつもそうだ」の選択肢から選ぶ。26 の項目に対する回答数値を平均したものが、その人のセルフ・コンパッションの高さということになる。

従来、SCS のような質問紙尺度における心理の定量化は、実施が迅速・簡便であるという理由から、心理学や精神医学等の分野で広く用いられてきた。しかし、現実社会では、人々は自分の心理状態を表現するとき、質問紙尺度にみられるように、「私の調子は5段階中4です」などと表現するのではなく、多くの場合自由度のある自然言語を使用する (例:「調子はどうですか？」と聞かれ、「忙しいけど充実していて、元気です！」と答える等) [7], [8]。実際に人々は、質問紙尺度への回答よりも、自由記述による回答のほうが、自らの心理をより明確に想起し、かつ正確に表現できると考える傾向にあり[9]、テキストデータの自然言語処理による解析は、心理学の関連分野である精神医学においても、より正確な診断やアセスメントを行う際の有望なツールであると

みなされている[10]。そこで本研究は、参加者から収集したオープンエンドな自由記述テキストから、個人のセルフ・コンパッションの高さ (すなわち、SCS 得点) を推定できるか検討した。

2 関連研究

従来、自然言語を手がかりとして、書き手の態度 (肯定的/中立的/否定的) や、基本感情 (怒り/悲しみ/喜び 等) を推定する試みは多く行われている[11], [12]。その一方で、質問紙尺度を用いて従来測定されてきたような、複雑な心理学的構成概念を推定・定量化しようとする試みは少ないものの、例えば Kjell et al. (2019) [7] は、「人生全般において、あなたは満足していますか、していませんか」というプロンプト (問い) を投げかけ、参加者から得た自由記述を潜在意味解析 (latent semantic analysis)で分析することで、参加者の人生満足度 (自分の人生にどの程度満足しているかを表す心理学的構成概念) のスコアを予測できることを示している。

近年は、Transformers アーキテクチャを使用した深層学習モデル BERT (Bidirectional Encoder Representations from Transformers) [13]を用いた研究が増加している。例えば Kjell et al. (2022) [8]は、先程と同様の人生についてのプロンプトに対する自由記述を BERT によって分析することで、従来の潜在意味解析を上回る予測精度が得られることを示した (r = .74) 。また Wang et al. (2020) [14]においては、主に中国圏で使用される SNS である Sina Weibo 上におけるテキストデータが、精神医学的抑うつ重症度を予測し、畳み込みニューラルネットワークや長・短期記憶ニューラルネットワーク等の従来の予測モデルと比べて BERT が高精度に予測することを示している。さらに、Simchon et al. (2023) [15]は、掲示板型ソーシャルサイト Reddit に投稿された内容を BERT で分析することで、投稿者の性格特性を中程度の精度で予測できることを示した (r = .33)。

なお、これまでの先行研究では、上述の「人生に満足しているか、していないか」というような Yes か No で答えられるようなプロンプトを用いた質問紙スコアの予測や [7], [8]、SNS に投稿された膨大な数 (数万~数十万) のテキストデータを用いた予測 [14], [15] が主に行われていた。こうした限界点をふまえて本研究は、より回答に自由度があるオープンエンドな自由記述データを参加者から収集し、比較的少数 (千程度) のデータを使って BERT 回帰モデルを構築することによっても、質問紙尺度 (すなわち、SCS) 得点の高精度な推定が可能であるか検証した。また、BERT による SCS の予測スコアが、他変数 (先行研究で SCS 得点との関連が示されている抑うつ・ストレス・人生満足度等) と、質問紙の真スコアと同様の関連を示すかについても併せて検討した。

3 データセット収集

調査対象者　主に日本人が登録しているクラウドソーシングプラットフォーム CrowdWorks を利用して参加者を募集し、最終的に 780 件の有効データを得た。BERT モデルを構築する際に必要となるサンプルサイズはタスクに依存するが、例えば Sun et al. (2019) [16]は、1,000 件程度のデータでも実用上十分な精度を得られる可能性を示している。

手続き　参加者は、3 つのプロンプト (表 1 参照) を画面上に呈示され、各プロンプトに対して自由記述を行った。各プロンプトが呈示される度に、参加者は 40 文字以上の自由記述文を入力し、かつ各プロンプトの呈示から 45 秒が経過しなければ次の回答に移ることができなかった。自由記述終了後、参加者は SCS [17] および抑うつ・ストレス・ネガティブ感情/ポジティブ感情・人生満足度を測定する質問紙尺度 (PHQ-9[18], PSS[19], PANAS[20], SWLS[21]) に回答した。質問紙尺度に関する詳細は、付録 1 に記載した。

表 1 呈示されたプロンプト

1. 人生のなかで、苦しい状況を経験しているとき、 自分自身についてどのように考えますか？
2. 自分自身の欠点や不十分なところに気づいたとき、 自分自身についてどのように考えますか？
3. 大きな失敗をしてしまったとき、 自分自身についてどのように考えますか？

テキスト長　参加者が記述した 3 つのテキストのテキスト長 (文字数) の中央値は 50~55 字であった。

SCS の記述統計量　SCS スコアの平均点は 2.92、標準偏差は 0.78 であった。

4 BERT モデル構築と推定精度検討

本研究では、参加者から収集した自由記述テキストを学習データとして、同じ参加者が回答した質問紙尺度得点をテキストにより予測する BERT 回帰モ

デルを構築し、その精度を評価した。

4.1 BERT モデル構築

予測に使用したテキスト　予測に使用するテキストとして、3 つのプロンプトに対する各自由記述テキストと、それらのテキストをすべて結合したテキストを用いた。テキスト結合の際は 3 つのテキストを「/」 (スラッシュ) で区切り結合した[i]。

事前学習モデルとファインチューニング　モデル構築は、東北大学が提供している BERT の事前学習モデル bert-base-japanese-v3[ii]を収集したデータによってファインチューニングすることで行った。本モデルは、日本語の大規模データセット (CC-100 と Wikipedia) によって事前学習されている最新版である。ファインチューニングのためのプログラムは Python の Transformers パッケージを用いて実装した。ファインチューニング時のオプティマイザーは AdamW を、損失関数として最小 2 乗誤差 (mean square error; MSE) 、モデル評価指標は積率相関係数 (以下単に相関係数と記す) を使用した。ファインチューニング時のハイパーパラメータとして、学習率は{2e-5, 3e-5, 4e-5, 5e-5}、バッチサイズは{8, 16}、エポック数は{1, 2, 3, 4, 5}のそれぞれを検討した (したがってハイパーパラメータの組み合わせは全部で 4×2×5=40 通りあった)。ウォームアップ率は 0.1 とし、学習率スケジューラは linear (線形減少) を使用した。全参加者のテキストの長さは BERT の処理上限であるトークン数 512 を下回ったため、テキストを切り詰める truncation は行わなかった[iii]。

Cross validation　本研究のモデル構築用データセットは比較的小さい (1,000 件以下) ため、データをランダムに分割してモデルを構築し、精度を評価する手続きを 1 回だけ行っても、推定精度に偏りが生じる可能性がある。そのため、本研究では、最適なハイパーパラメータの選択、および構築されたモデルの精度評価には nested k-fold cross validation (入れ子型 k 分割交差検証; nested CV) の手続きを使用した。Nested CV は、古典的な k-fold CV よりも計算コストがかかるが、推定精度のバイアスが少なく [22]、比較的小さなサンプルに対する機械学習で特に有効な手法である [23]。重要なことは、nested CV によって、モデル構築に使用したデータとは独立したデータに対するモデルの予測精度の推定値としてより近いものを得ることが可能となることである。本研究では k を 5 とした。Nested CV に関する詳細は、付録 2 に記載した。

4.2. 実験結果と考察

各自由記述テキストとそれらの結合テキストによる質問紙 (SCS) 得点の予測精度を示す (表 2)。

表 2 テキストによる質問紙尺度得点の予測精度

	苦しい状況	欠点に気づいた	失敗してしまった	結合テキスト
精度 (*r*)	.51	.53	.50	.67

注）予測精度は、質問紙尺度（SCS）の真のスコアと予測スコア間の積率相関係数。

3 つの単一テキストは、SCS 得点を相応の精度で予測した (*r* > .50)。3 つの各状況に関する単一テキストによる予測精度の間にほとんど差はなかった。各テキストは、最低で 40 文字、中央値で 50~55 文字程度とかなり短く、かつ自由記述を求めるプロンプトも回答に自由度がある (Yes/No で答えられない) ものだったにもかかわらず、このように高い精度を実現したことは、特筆に値するであろう。

さらに、結合テキストは、質問紙尺度得点を単一テキストよりも高い精度で予測した (*r* = .67)。単一テキストによる予測モデルのなかで最高の予測精度を実現したもの (*r* = .53) と比べて相関係数が.14 向上しており、これは質問紙尺度得点の分散説明率が 16.8% (0.67^2-0.53^2) 向上したことを示す。内容が異なる複数のテキストを結合したほうが精度が向上するのは Kjell et al. (2022) [8] と同様の結果であり、心理学的構成概念を自然言語により推定する際は、複数の自由記述を収集することが精度の向上に貢献しうるといえるだろう。表 3 には、3 名の参加者について、自由記述の結合テキストと、SCS の真スコア、および結合テキストによる BERT 予測スコアの例を示した。

[i] BERT では文の区切りに[SEP]トークンを用いることもあるが、[SEP]は主に 2 つの文章を区切るために使われるため、今回は使用しなかった。

[ii] https://huggingface.co/cl-tohoku/bert-base-japanese-v3

[iii] ただし、結合テキストのみ、1 名の参加者がトークン数 512 を上回った。そのため、結合テキストによる予測では、この参加者を除外した上でモデルを構築した。

表 3 参加者の自由記述（結合テキスト）と SCS の真スコア、および結合テキストによる SCS の予測スコアの例

結合テキスト	真	予測
周りの人達のように器用にこなすことが出来ず、不器用で要領も悪く、情けない存在だと考えます。／周りの人達と違って自慢出来るような長所が無く、価値の無い存在だと思い、自分自身を否定的に考えます。／こんな大きな失敗をしてしまう自分は本当に駄目で価値のない役立たずで情けない存在だと考えます。	1.50	1.76
自分に降りかかったこの苦しさの意味を考えます。きっと自分に起こったという事は、何か意味があるはずなので。／どうして自分はこうなんだろう…と思って落ち込むし、出来ない自分にとてもがっかりします。でも欠点はたぶん直らないので人間だから欠点があって当たり前と思うように努力します。／本当にどうしようもないダメな人間だぁ…と自分を責めると思います。 そして、こんなに落ち込んでしまう自分に嫌気がさします。	2.92	2.78
自分自身について、心身ともに「問題あるのかないのか」、「まだ耐えられるのかどうか」を逐一確認する。／自分自身について「嫌いになることは決してない」。それらを含めて「自分自身」であるから。／自分自身について「がっかり」するが、それを糧に「成長できる」ように優しく接する。	4.23	4.16

また、BERT による SCS の結合テキストによる予測スコアが、他変数 (抑うつ・ストレス・ネガティブ感情・ポジティブ感情・人生満足度) の質問紙尺度得点と間に、質問紙の真スコアと同様の関連を示すかを検討するため、相関分析を行った (表 4)。結果、予測スコアと他変数との相関パターン (すなわち、相関係数の符号や有意性) は、真の質問紙スコアの相関パターンと同様であった(ただし相関係数の絶対値は真のスコアよりも小さくなる傾向にあった)。これは BERT 予測スコアが一定の妥当性 (基準関連妥当性という) を持つことを示すものである。

表 4 質問紙（SCS）の真のスコアおよび BERT 予測スコアの各々と他変数の相関

	抑うつ	ストレス	ネガ感情	ポジ感情	人生満足度
真スコア	-.58	-.65	-.63	.50	.56
予測スコア	-.38	-.47	-.44	.38	.40

5. おわりに

本研究は、3 つの状況における思考を問うプロンプトを呈示し、収集した短い自由記述を BERT で分析することで、人々のセルフ・コンパッションを高精度に予測し得ることを示した。学習時に使用した自由記述データは、Kjell et al. (2022) 等より回答に自由度があるプロンプトによって収集したものであり、かつ学習データが 1,000 程度と少数でも、心理学的構成概念の高精度な予測が可能で、その予測スコアは一定の妥当性を持つことがわかった。

また本研究で使用したプロンプトは、特定の状況で「自分についてどう考えるか」を広く問うものであり、一切セルフ・コンパッションに相当するような概念を呈示していないことも特徴である。他方、SCS は、「自分にやさしくする」などといった、セルフ・コンパッション概念そのものを呈示し、それに自分があてはまるかを評定する手続きを経て、セルフ・コンパッションを測定する。留意すべきは、このように直接心理学的構成概念を呈示すると、参加者が何らかの先入観を持ったり、その後の回答にバイアスが生じる可能性を排除できない点である (question order bias [24] や demand characteristics [25]と呼ばれる問題)。本研究の手続きによる測定は、心理学的構成概念を直接参加者に呈示することはなく、何を測定しようとしているのか参加者にとって比較的推察しにくいと考えられるため、先入観やバイアスの問題を減じることが可能となる。また、先に述べたように、人々は、質問紙尺度への回答よりも、自由記述による回答を、自らの心理をより明確に想起し、かつ正確に表現できる方法とみなす傾向にある[9]。このような利点を持つ自然言語による心理学的構成概念の定量化手法が、心理学や関連分野研究に今後利用・応用されることが期待される。

謝辞

本研究は JSPS 科研費 22H01103 の助成を受けたものです。また本研究の BERT モデル構築には、株式会社京都テキストラボの計算環境を使用させていただきました。この場を借りて御礼申し上げます。

参考文献

[1] A. MacBeth and A. Gumley, “Exploring compassion: A meta-analysis of the association between self-compassion and psychopathology,” *Clin Psychol Rev*, vol. 32, no. 6, pp. 545–552, Aug. 2012, doi: 10.1016/j.cpr.2012.06.003.

[2] U. Zessin, O. Dickhäuser, and S. Garbade, “The Relationship Between Self-Compassion and Well-Being: A Meta-Analysis,” *Appl Psychol Health Well Being*, vol. 7, no. 3, 2015, doi: 10.1111/aphw.12051.

[3] M. Ferrari, C. Hunt, A. Harrysunker, M. J. Abbott, A. P. Beath, and D. A. Einstein, “Self-Compassion Interventions and Psychosocial Outcomes: a Meta-Analysis of RCTs,” *Mindfulness (N Y)*, vol. 10, no. 8, 2019, doi: 10.1007/s12671-019-01134-6.

[4] K. D. Neff, “Self-Compassion: Theory, Method, Research, and Intervention,” *Annu Rev Psychol*, vol. 74, no. 1, pp. 193–218, Jan. 2023, doi: 10.1146/annurev-psych-032420-031047.

[5] E. I. Fried, “What are psychological constructs? On the nature and statistical modelling of emotions, intelligence, personality traits and mental disorders,” *Health Psychol Rev*, vol. 11, no. 2, pp. 130–134, Apr. 2017, doi: 10.1080/17437199.2017.1306718.

[6] K. Neff, “The Development and Validation of a Scale to Measure Self-Compassion,” *Self and Identity*, vol. 2, no. 3, pp. 223–250, Jul. 2003, doi: 10.1080/15298860309027.

[7] O. N. E. Kjell, K. Kjell, D. Garcia, and S. Sikström, “Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs,” *Psychol Methods*, vol. 24, no. 1, 2019, doi: 10.1037/met0000191.

[8] O. N. E. Kjell, S. Sikström, K. Kjell, and H. A. Schwartz, “Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy,” *Sci Rep*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-07520-w.

[9] S. Sikström, A. Pålsson Höök, and O. Kjell, “Precise language responses versus easy rating scales—Comparing respondents’ views with clinicians’ belief of the respondent’s views,” *PLoS One*, vol. 18, no. 2, p. e0267995, Feb. 2023, doi: 10.1371/journal.pone.0267995.

[10] Z. S. Chen, P. (Param) Kulkarni, I. R. Galatzer-Levy, B. Bigio, C. Nasca, and Y. Zhang, “Modern views of machine learning for precision psychiatry,” *Patterns*, vol. 3, no. 11, p. 100602, Nov. 2022, doi: 10.1016/j.patter.2022.100602.

[11] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” *J Comput Sci*, vol. 2, no. 1, pp. 1–8, 2011, doi: 10.1016/j.jocs.2010.12.007.

[12] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment Classification using Machine Learning Techniques,” in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP ’02*, Morristown, NJ, USA: Association for Computational Linguistics, 2002, pp. 79–86. doi: 10.3115/1118693.1118704.

[13] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019.

[14] X. Wang *et al.*, “Depression risk prediction for chinese microblogs via deep-learning methods: Content analysis,” *JMIR Med Inform*, vol. 8, no. 7, 2020, doi: 10.2196/17958.

[15] A. Simchon, A. Sutton, M. Edwards, and S. Lewandowsky, “Online reading habits can reveal personality traits: towards detecting psychological microtargeting,” *PNAS Nexus*, vol. 2, no. 6, 2023, doi: 10.1093/pnasnexus/pgad191.

[16] C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to Fine-Tune BERT for Text Classification?,” in *Lecture Notes in Computer Science*, vol. 11856, 2019, pp. 194–206. doi: 10.1007/978-3-030-32381-3_16.

[17] 有光 興記, “セルフ・コンパッション尺度日本語版の作成と信頼性，妥当性の検討,” 心理学研究, vol. 85, no. 1, pp. 50–59, 2014

[18] K. Muramatsu *et al.*, “The Patient Health Questionnaire, Japanese Version: Validity According to the Mini-International Neuropsychiatric Interview–Plus,” *Psychol Rep*, vol. 101, no. 3, pp. 952–960, Dec. 2007, doi: 10.2466/pr0.101.3.952-960.

[19] 鷲見 克典, “知覚されたストレス尺度（Perceived Stress Scale）日本語版における信頼性と妥当性の検討,” 健康心理学研究, vol. 19, no. 2, pp. 44–53, 2006, doi: 10.11560/jahp.19.2_44.

[20] 佐藤 徳 and 安田 朝子, “日本語版 PANAS の作成,” 性格心理学研究, vol. 9, pp. 138–139, 2001.

[21] 角野 善司, “ ¥人生に対する満足尺度日本版作成の試み,” 日本教育心理学会総会発表論文集, vol. 36, p. 192, 1994.

[22] S. Varma and R. Simon, “Bias in error estimation when using cross-validation for model selection,” *BMC Bioinformatics*, vol. 7, 2006, doi: 10.1186/1471-2105-7-91.

[23] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, “Machine learning algorithm validation with a limited sample size,” *PLoS One*, vol. 14, no. 11, 2019, doi: 10.1371/journal.pone.0224365.

[24] M. Thau, M. F. Mikkelsen, M. Hjortskov, and M. J. Pedersen, “Question order bias revisited: A split-ballot experiment on satisfaction with public services among experienced and professional users,” *Public Adm*, vol. 99, no. 1, 2021

[25] M. T. Orne, “On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications.,” *American Psychologist*, vol. 17, no. 11, pp. 776–783, Nov. 1962, doi: 10.1037/h0043424.

付録 1（質問紙尺度の詳細)

Self-Compassion Scale (SCS) 参加者のセルフ・コンパッションを測定する質問紙尺度である。参加者は 26 の項目（例：「自分自身の欠点と不十分なところについては、やさしい目で見るようにしている」）を呈示され、各項目に「1. ほとんど全く」～「5. ほとんどいつも」の 5 件法で回答する。得点を平均して尺度得点とする（以下同様)。

Patient Health Questionnaire-9 (PHQ-9) 参加者の抑うつを測定する質問紙尺度である。参加者は 9 項目（例：「物事に対してほとんど興味がない、または楽しめない」）を呈示され、各項目に「0. 全くない」「1. 数日」「2. 半分以上」「3. ほとんど毎日」の 4 件法で回答する。

Perceived Stress Scale (PSS) 参加者の日常的ストレスを測定する質問紙尺度である。参加者は 10 の項目（例：「予想もしなかった目にあってうろたえた」）を呈示され、各項目に「0. 全くなかった」から「4. いつもあった」の 5 件法で回答する。

Positive and Negative Affect Schedule (PANAS) 参加者の普段経験するネガティブ感情およびポジティブ感情を測定する質問紙尺度である。参加者は 16 の項目（例：「びくびくした」「活気のある」）を呈示され、各項目に「1.全くない」から「6.いつも」の 6 件法で回答する。

Satisfaction with Life Scale (SWLS) 参加者の人生満足度を測定する質問紙尺度である。参加者は 5 項目（例：「大体において、私の人生は理想に近い」）を呈示され、各項目に「1. 全くそうではない」から「7. 全くそうだ」の 7 件法で回答する。

付録 2（Nested CV に関する詳細）

Nested 5-fold CV では、初めにデータセットの順番をシャッフルした後、5 つのサブデータセットに分割する。その後のプロセスは、全データセットをモデル構築（development）用データとテスト（test）データに分け、モデル構築データによって構築したモデルの精度をテストデータで推定する outer loop と、モデル構築用データを訓練（train）データと検証（validation）データに分け、最適なハイパーパラメータの組み合わせを選択し、その組み合わせを用いてモデルを構築する inner loop からなる。以下、outer loop、inner loop のそれぞれで実行されるアルゴリズムについて説明する。

Outer loop Outer loop では、5 つのサブデータセットが、1 つずつテストデータとして選ばれ、残りの 4 つのサブデータセットがモデル構築用データとなる。モデル構築用データを用いて下記で述べる inner loop の手続きに従ってモデルを構築し、構築されたモデルを使ってテストデータの質問紙スコアを予測することで、精度（質問紙スコアの真値と予測スコア間の相関係数）を得る。テストデータとなるサブデータセットは 5 つあるため、これが 5 回繰り返される。最終的に得られた 5 つの相関係数の平均値が、BERT による質問紙スコア予測の最終的な精度とみなされる。

Inner loop Inner loop では、構築用データを構成する 4 つのサブデータセットが、1 つずつ検証データとして選ばれ、残り 3 つのサブデータセットが訓練データとなる。訓練データを用いて、ハイパーパラメータの全ての組み合わせごとにモデルのファインチューニングを行った後（したがって 40 通りのモデルが出来上がる)、各ハイパーパラメータで訓練した 40 個のモデルそれぞれを用いて検証データの質問紙スコアを予測したときの相関係数を記録する。検証データとなるサブデータセットは 4 つあるため、これが 4 回繰り返される。各ハイパーパラメータの組み合わせごとに、inner loop 内で得た 4 つの相関係数を平均し、最高の平均精度を実現したハイパーパラメータの組み合わせを用いて、4 つのサブデータセット（構築用データの全体）すべてを用いてモデルを構築する。（構築されたモデルは、先述の outer loop においてテストデータによってテストされる）