福岡工業大学　令和７年度　卒業研究論文

授業評価の数値に表れない学生の本音
―マルチタスク学習とSHAP分析による満足度要因の解明―

指導教員　佐藤　大輔

福岡工業大学情報工学部
システムマネジメント学科

学籍番号　22M11178
氏名　蘭牟田　晃弘

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　研究背景

1.1.1　高等教育における質保証

高等教育機関では，教育の質保証が重要な課題となっている．大学設置基準の改正や認証評価制度の導入により，各大学は教育活動の点検・評価を行い，その結果を教育改善に活かすことが求められている．この文脈において，学生による授業評価（Student Evaluation of Teaching: SET）は，教育の質を測定する主要な手段として広く実施されている [16, 17].

授業評価は，学生の視点から授業の質を把握し，教員へのフィードバックを通じて教育改善を促進する役割を担っている．多くの大学では，学期末にアンケート形式で授業評価を実施し，その結果を教員に還元するとともに，全学的な教育改善の資料として活用している．

1.1.2　授業評価アンケートの構成

授業評価アンケートは，一般に多段階の評価スコア（例：4 段階や 5 段階の択一式回答）と自由記述から構成される．評価スコアは定量的な比較に適しており，授業間や教員間の比較，経年変化の把握などに活用される．一方，自由記述は学生の具体的な意見や感想を収集でき，評価スコアの背景にある理由や具体的な改善提案を把握できる利点がある．

しかしながら，評価スコアと自由記述はそれぞれ異なる特性を持つ．評価スコアは数値として集計・比較が容易であるが，なぜその評価に至ったかという理由は直接観測できない．自由記述には学生の本音や具体的な要望が含まれることが多いが，非構造データであるため大規模な分析には困難が伴う．

表 1.1: 本研究の対象データ

項目	値
対象期間	2018 年度〜2023 年度（6 年間）
対象学科数	9 学科
授業数	3,268 件
自由記述総件数	83,851 件
平均自由記述数/授業	25.2 件

1.1.3　自由記述分析と感情分析の課題

自由記述の分析は，従来，教員や職員による人的な読解に依存してきた．しかし，大規模な大学では学期ごとに数万件の自由記述が収集されることがあり，すべてを人力で読解することは現実的ではない．また，読解者の主観により解釈が異なる可能性があり，客観的な分析が困難である．

このような背景から，自然言語処理技術を用いた自由記述の自動分析への期待が高まっている [10, 13]．特に，近年の深層学習技術の発展により，大規模なテキストデータから意味のある情報を抽出することが可能になってきた [1].

感情分析（Sentiment Analysis）は，テキストに含まれる肯定的・否定的・中立的な感情を自動的に推定する技術である [9]．ソーシャルメディアの分析や製品レビューの分析など，様々な分野で活用されている．教育分野においても，学生の自由記述から満足度や不満を把握する手段として，感情分析の応用が注目されている [11, 15].

自由記述を感情スコアとして数値化することで，評価スコアとの関係を統計的に検討できる可能性がある．感情スコアと評価スコアの相関を分析することで，学生の感情が授業評価にどのように関係しているかを明らかにできると考えられる．

1.1.4　本研究の対象データ

本研究では，福岡工業大学における 2018 年度から 2023 年度までの 6 年間の授業評価データを分析対象とする．対象データの規模を表 1.1 に示す.

このデータ規模は，機械学習モデルの構築および統計的分析を行う上で十分な量であり，得られた知見の信頼性を担保できると考えられる．

1.2 課題の整理

授業評価の活用には，以下の課題が存在する．

1.2.1 評価要因と感情情報の不可視性

授業評価スコアは，授業内容，教員の説明力，教材の質，試験の難易度など，複数の要因に対する総合判断として付与される．しかし，どの要因がどの程度評価に影響したかは直接観測できない．例えば，同じ 3 点の評価であっても，「内容は良いが説明が分かりにくい」場合と「説明は分かりやすいが内容が物足りない」場合では，必要な改善策は異なる．

また，授業評価スコアは授業の「評価」を示すが，学生の「感情」や「満足度」を直接測定するものではない．同じ評価スコアであっても，学生の感情的な満足度には差異がある可能性がある．

評価スコアのみでは，このような要因の内訳や感情的側面を把握することが困難であり，具体的な改善につなげにくいという課題がある．

1.2.2 自由記述分析の困難性

自由記述は非構造データであり，表記ゆれ，略語，誤字脱字など，多様な表現が含まれる．また，同じ内容でも学生によって表現方法が異なるため，単純な文字列マッチングでは十分な分析ができない．

人的読解に依存する場合，分析に膨大な時間がかかるだけでなく，読解者の主観や疲労により分析の一貫性が損なわれる可能性がある．本研究の対象データでは 83,851 件の自由記述が存在し，すべてを人力で読解することは現実的ではない．

例えば，「勉強になったが楽しくなかった」授業と「楽しかったが勉強にならなかった」授業は，評価スコアが同程度であっても，学生の感情的体験は大きく異なる．

1.2.3　改善施策の優先順位付け

教育改善に投入できる資源は限られており，すべての課題に同時に対応することは困難である．どの要因を優先的に改善すべきかを客観的に判断するためには，各要因の影響度を定量的に把握する必要がある．

しかし，従来の分析手法では，このような優先順位付けを行うための定量的な根拠を得ることが難しかった．

1.3　研究目的

本研究の目的は，授業評価アンケートの自由記述から感情スコアを推定し，授業評価スコアとの関係性を分析することで，授業評価に影響する要因を定量的に特定することである．

具体的には，以下の 3 つのサブ目的を設定する．

1. **感情スコアと評価スコアの関係解明**: 自由記述の感情分析により感情スコアを算出し，授業単位で集計した感情スコアと授業評価スコアの関係を統計的に検討する．
2. **共通要因と特化要因の分離**: 感情スコアと評価スコアを同時に予測するマルチタスク学習モデルを構築し，両者に共通する要因（満足度要因）と，それぞれに特有の要因（感情特化要因・評価特化要因）を分離する．
3. **要因の定量化と可視化**: SHAP 分析を用いて単語レベルの寄与度を定量化し，授業改善に直結する具体的な要因を抽出する．

1.4　研究仮説

本研究では，以下の 3 つの仮説を設定する．

仮説 1: 授業単位で集約した感情スコアと授業評価スコアには正の相関関係がある．

この仮説は，学生の感情的満足度と授業評価が関連しているという前提に基づく．自由記述にポジティブな感情が多く表れる授業は，評価スコアも高い傾向があると予想される．

仮説 2: 感情スコアと授業評価スコアの両方に影響する共通要因（満足度要因）が存在する．

感情と評価は異なる概念であるが，両者に共通して影響する要因が存在すると考えられる．例えば，「分かりやすさ」は感情的満足度と評価スコアの双方に寄与する可能性がある．

仮説 3: マルチタスク学習により，共通要因と特化要因を分離できる．

感情予測タスクと評価予測タスクを同時に学習することで，共有表現（共通要因）とタスク固有の表現（特化要因）を分離できると予想される．

1.5 研究のアプローチと特徴

1.5.1 BERT による感情分類

本研究では，日本語の事前学習済み BERT モデル（Bidirectional Encoder Representations from Transformers）を基盤とした感情分類モデルを構築する [1]．BERT は双方向の文脈情報を考慮できる言語モデルであり，少量の教師データでも高精度な分類が可能である [3]．

83,851 件の自由記述に対して感情スコア（ポジティブ/ニュートラル/ネガティブ）を推定し，授業単位で集計することで，授業レベルの感情スコアを算出する．

1.5.2 マルチタスク学習

感情スコアと授業評価スコアを同時に予測するマルチタスク学習モデルを構築する [6]．BERT エンコーダを共有表現として使用し，感情分類ヘッドと評価スコア予測ヘッドを分岐させる構成とする．

この構成により，両タスクに共通する特徴（共通要因）と，各タスクに固有の特徴（特化要因）を分離することが可能となる．

1.5.3 SHAP 分析による要因の可視化

モデルの解釈可能性を高めるため，SHAP（SHapley Additive exPlanations）分析を実施する [4]．SHAP は協力ゲーム理論に基づく特徴量重要度の算出手法であり，単語レベルでの寄与度を定量化できる [5]．

SHAP 分析により，どの語彙が感情スコアや評価スコアに寄与しているかを明らかにし，授業改善に向けた具体的な示唆を得る．

1.6 研究の意義

1.6.1 学術的意義

本研究は，以下の点で学術的意義を有する．

第一に，授業評価における感情要因の役割を定量的に明らかにする点である．従来の研究では，評価スコアと自由記述を別々に分析することが多かったが，本研究では両者を統合的に分析することで，より深い理解を得ることを目指す．

第二に，マルチタスク学習と SHAP 分析を組み合わせた分析フレームワークを確立する点である．この方法論は，教育分野に限らず，複数の評価指標が存在する場面での要因分析に応用可能である．

1.6.2 実践的意義

本研究は，以下の点で実践的意義を有する．

第一に，大規模な自由記述データの効率的な分析手法を提供する点である．83,851 件の自由記述を自動的に分類することで，人的コストを大幅に削減できる．

第二に，授業改善の優先順位を客観的に決定できる基盤を提供する点である．共通要因・感情特化要因・評価特化要因の区別により，目的に応じた改善施策を設計できる．

第三に，データに基づく教育改善（Evidence-Based Education）の実現に貢献する点である．

1.7 本研究の構成

本研究は，全 5 章からなる．各章の概要を以下に示す．

第 2 章 関連研究: 授業評価研究，自然言語処理による感情分析，BERT と言語モデル，マルチタスク学習，解釈可能 AI（SHAP），順序回帰に関する関連研究を整理し，本研究の位置づけを明確にする．

第 3 章 データと手法: 本研究で使用するデータセット（3,268 授業，83,851 件自由記述）の概要，前処理手順，感情分類モデルの構成，マルチタスク学習モデルの構成，SHAP 分析の手法，および評価指標を述べる．

第 4 章 結果と考察: 感情スコアと授業評価スコアの相関分析結果, 感情分類モデルの性能, SHAP 分析による要因抽出結果を報告し，得られた知見を考察する．

第 5 章 おわりに: 本研究の成果を総括し，研究の限界および今後の課題を述べる．

第2章　関連研究

本章では，本研究に関連する先行研究を整理する．授業評価研究，感情分析，BERT，マルチタスク学習，解釈可能 AI，順序回帰について概観し，本研究の位置づけを明確にする．

2.1　授業評価研究

2.1.1　授業評価の意義と歴史

大学における授業評価（Student Evaluation of Teaching: SET）は，教育の質向上に向けた重要な指標として広く用いられている [16]．授業評価は 1920 年代にアメリカの大学で始まり，1970 年代以降に世界的に普及した [17]．日本においても，1990 年代後半から多くの大学で導入が進み，現在ではほぼすべての大学で実施されている．

授業評価の主な目的は，(1) 教員へのフィードバックによる授業改善，(2) 人事評価の参考資料，(3) 学生への授業選択情報の提供，の 3 点である [17]．

2.1.2　授業評価の構成要素

多くの大学では，学期末のアンケートにより授業評価スコアと自由記述を収集し，教員へフィードバックを行っている．授業評価スコアは数量的に扱いやすい一方で，学生が評価に至った理由や具体的な改善要望は自由記述に含まれることが多い [13, 14]．

自由記述を定量的に分析し，授業評価スコアの背後にある要因を明らかにする研究が必要とされている．さらに，授業改善への活用を前提とした分析プロセスの標準化や，フィードバックの迅速化が課題として指摘されており，効率的な分析手法の整備が求められる [13]．

2.1.3 授業評価の信頼性と妥当性

授業評価の信頼性・妥当性については多くの議論がある．Marsh は，授業評価が多次元的な構造を持ち，異なる側面（例：明確さ，組織性，熱意）を測定していることを示した [16]．一方で，評価が成績期待や授業難易度に影響される可能性も指摘されている [17].

2.2 自由記述分析と感情分析

2.2.1 自由記述の特性

自由記述は非構造テキストであり，従来は人的な読解に依存していた．しかし，大規模なデータでは人的読解に限界があり，分析者の主観による解釈のばらつきも問題となる．近年の自然言語処理技術の発展により，大規模な自由記述を自動的に解析し，感情や評価の傾向を抽出することが可能になっている [9].

2.2.2 感情分析の基礎

感情分析（Sentiment Analysis）は，テキストに含まれる肯定的・否定的・中立的な感情を推定する技術である [9]．感情分析は，ソーシャルメディアの分析，製品レビューの分析，顧客満足度調査など，様々な分野で活用されている．

教育分野においても，学生の満足度や不満の把握に活用できる [11, 15]．Sindu らは，学生のフィードバックからアスペクトベースの意見マイニングを行い，教員の教育パフォーマンス評価に活用した [15].

2.2.3 教育分野での自由記述分析

教育分野の自由記述分析では，テキスト分析を通じた改善提案の抽出や意見整理が報告されている [10, 12]．Gottipati らは，学生のフィードバックから授業改善の提案を自動抽出するテキスト分析手法を提案した [10]．Misuraca らは，意見マイニングを教育分析に応用し，学生フィードバックの統合的分析戦略を提案した [12].

これらの研究は，従来のスコア中心の評価を補完する手段としての自由記述分析の重要性を示している．

2.3 感情分析の手法分類

感情分析の手法は大きく，(1) 辞書型手法，(2) 古典的機械学習手法，(3) 深層学習手法に分類できる．

2.3.1 辞書型手法

辞書型手法は，極性語彙（ポジティブ・ネガティブな単語のリスト）を用いて感情を推定する手法である．実装が容易で解釈しやすい利点がある一方，文脈依存の表現や否定表現に弱いという欠点がある [11]．Rajput らは，教員評価における辞書ベースの感情分析を行い，その有効性と限界を報告した [11]．

2.3.2 古典的機械学習手法

古典的機械学習手法（SVM，ナイーブベイズ，ランダムフォレストなど）は，特徴量設計により一定の精度を得られる．しかし，語彙の多様性が大きい自由記述では特徴量設計の負担が大きい [14]．Santhanam らは，学生フィードバックのテキスト分析において，共通語彙の適応と拡張を行った [14]．

2.3.3 深層学習手法

深層学習手法（RNN，LSTM，Transformer など）は，文脈を自動的に考慮できる利点がある．一方で，教師データの準備コストが高く，教育分野固有の語彙や表現への適応が課題となる [1]．近年は事前学習済みモデルの活用により，少量の教師データでも高精度な分類が可能になっている．

2.4 BERT と事前学習済み言語モデル

2.4.1 BERT の概要

BERT（Bidirectional Encoder Representations from Transformers）は，Transformer のエンコーダ部分を用いた事前学習済み言語モデルである [1]．Vaswani らが提案した Transformer アーキテクチャ [2] を基盤とし，双方向の文脈情報を同時に考慮できる点が特長である．

BERT は，Masked Language Model（MLM）タスクと Next Sentence Prediction（NSP）タスクにより事前学習される．大規模コーパスで事前学習されたモデルを特定タスクに微調整することで，少量の教師データでも高精度な分類が可能である [1].

2.4.2 日本語 BERT モデル

日本語に対しても事前学習済み BERT モデルが複数提供されている．東北大学が公開した「cl-tohoku/bert-base-japanese」は，日本語 Wikipedia で事前学習されたモデルであり，日本語 NLP タスクで広く利用されている [3].

教育分野の自由記述に対して微調整を行うことで，文脈を考慮した感情分類が実現できる．一方で，学習データの領域差が大きい場合には汎化性能が低下する可能性があるため，教育分野に特化した微調整と評価設計が必要となる．

2.5 マルチタスク学習

2.5.1 マルチタスク学習の原理

マルチタスク学習は，複数の関連するタスクを同時に学習し，共通表現を獲得することで各タスクの性能を向上させる手法である [6]．Zhang らは，マルチタスク学習の包括的なサーベイを行い，ハードパラメータ共有とソフトパラメータ共有の 2 つのアプローチを整理した [6].

マルチタスク学習の利点として，(1) 関連タスクからの情報転移による性能向上，(2) 過学習の抑制（正則化効果），(3) 共通表現の学習による解釈可能性の向上，が挙げられる．

2.5.2 NLP におけるマルチタスク学習と本研究への適用

自然言語処理分野では，感情分析と関連タスク（アスペクト抽出，意見ターゲット識別など）を同時に学習するマルチタスクモデルが提案されている [7]．これらの研究は，関連タスクの同時学習が個別タスクの性能を向上させることを示している．

感情分析と授業評価スコア予測は，いずれも自由記述に基づく評価理解という共通の目的を持つため，マルチタスク学習の適用が有効であると考えられる．特に，感情スコアと評価スコアを

同時に学習することで，共通要因と特化要因の分離が可能となり，教育改善の示唆をより具体化できる点が期待される．

2.6 解釈可能 AI と SHAP

2.6.1 解釈可能 AI の重要性

機械学習モデルの予測根拠を明確化するため，解釈可能 AI（Explainable AI: XAI）が注目されている [5]．特に教育分野では，モデルの予測精度だけでなく，説明可能性が重要であり，改善施策への翻訳可能性が求められる．

2.6.2 SHAP の原理とテキスト分類への適用

SHAP（SHapley Additive exPlanations）は，協力ゲーム理論の Shapley 値に基づき，特徴量の寄与度を定量化する手法である [4]．SHAP は，(1) 局所的な説明と大域的な説明の両方が可能，(2) 理論的に一貫した寄与度を算出，(3) 様々なモデルに適用可能，という利点を持つ [4]．テキスト分類においては，単語レベルの寄与度を算出できるため，授業評価に影響する要因を具体的な語彙として提示できる．ただし，寄与度の解釈は文脈に依存するため，定性的な検討との併用が必要となる．

2.7 順序回帰

2.7.1 順序尺度の特性

授業評価スコアは 1 点から 4 点までの順序尺度であり，単純な回帰（連続値として扱う）や分類（カテゴリとして扱う）では順序関係を適切に扱えない．順序回帰は，評価段階の順序性を考慮して確率分布を推定する手法である [8].

2.7.2 ニューラルネットワークとの統合

近年は，ニューラルネットワークと順序回帰を組み合わせた手法が提案されている．Cao らは，CORAL（Consistent Rank Logits）を提案し，順序一貫性を保証しながらニューラルネットワー

クで順序回帰を行う手法を示した [8]．授業評価スコアの分析においても，順序回帰の導入は妥当である．

2.8　教育分野への応用研究

教育分野では，学習ログやアンケートデータを用いた分析（Learning Analytics, Educational Data Mining）が進んでいる [18]．自由記述を対象としたテキスト分析や意見抽出の取り組みは報告されているが [10, 12, 13]，評価スコアとの関係性を統合的に扱った研究は多くない．

また，教育改善に直結する語彙や要因を整理するために，語彙辞書の拡張や分析手法の整理が試みられている [14]．このため，感情分析・マルチタスク学習・解釈可能 AI を組み合わせた総合的な分析枠組みの構築が求められている．

2.9　既存研究の限界と課題

既存研究には以下の課題がある．

第一に，評価スコアと自由記述の統合が不十分である．多くの研究は評価スコアの分析または自由記述の分析を別々に行っており，両者の関係を同時にモデル化した研究は限られている [10, 12]．

第二に，感情分析結果の解釈が定性的で，教育改善に直結しづらい．感情をポジティブ・ネガティブに分類するだけでは，具体的な改善策の導出が困難である．

第三に，予測精度と説明可能性の両立が難しい．高精度な深層学習モデルはブラックボックス化しやすく，教育改善への翻訳が困難である．

2.10　本研究の位置づけ

本研究は，授業評価アンケートの自由記述に対し，BERT による感情分類とマルチタスク学習を適用し，さらに SHAP 分析によって評価要因を定量化する点に特徴がある．

既存研究との差異を表 2.1 に示す．

本研究の新規性は以下の 3 点である．

1. 感情スコアと授業評価スコアを同時に学習するマルチタスクモデルを構築し，両者の関係を統合的にモデル化する．

表 2.1: 既存研究との比較

研究	評価スコア分析	自由記述分析	統合分析	要因の定量化
Gottipati et al. (2018)	−	○	−	−
Misuraca et al. (2021)	−	○	−	△
Hujala et al. (2020)	○	○	△	−
Sindhu et al. (2019)	−	○	−	△
本研究	○	○	○	○

2. SHAP 分析により，共通要因と特化要因を語彙レベルで分離し，改善施策の優先順位付けに利用できる定量的根拠を提供する．

3. 3,268 授業，83,851 件の自由記述という大規模データを用いて，統計的に信頼性の高い分析を行う．

これにより，教育改善に資する具体的な知見を提供することを目指す．

第3章　データと手法

本章では，本研究で使用したデータセットの概要，前処理手順，感情分類モデル，マルチタスク学習モデル，SHAP 分析の手法について詳述する．

3.1　データセット

3.1.1　データセットの概要

本研究では，福岡工業大学の授業評価システムにおける 2018 年度から 2023 年度までの 6 年間のデータを使用した．対象は 9 学科にわたり，授業数は 3,268 件である．自由記述の総件数は 83,851 件であり，各授業に対して平均 25.2 件の自由記述が付随している．このデータ規模は，機械学習モデルの構築および統計的分析を行う上で十分な量であると考えられる．

データセットの概要を表 3.1 に示す．

3.1.2　授業評価アンケートの構成

授業評価アンケートは，(1) 択一式質問の点数化による授業評価スコア，(2) 自由記述の 2 種類の情報から構成される．自由記述は以下の 2 つの質問からなる．

1. 先生に向けてこの授業の感想や学んだこと，意見や要望を記述してください

2. 次期履修者に向けて，この授業についてのアドバイスを記述してください

授業評価スコアは 4 段階（1〜4 点）であり，授業ごとに単一のスコアが付与される．授業評価スコアの平均値は 3.459 点（標準偏差 0.216）であり，比較的高い評価に集中する傾向がある．一方，自由記述は授業単位で複数件存在するため，授業単位で集約して分析する必要がある．

表 3.1: データセット概要

項目	値
対象期間	2018 年度〜2023 年度（6 年間）
対象学科数	9
授業数	3,268
自由記述総件数	83,851
平均自由記述数/授業	25.2
自由記述の平均文字数	約 41 文字

3.1.3　データの特性

本データセットには以下の特性がある．第一に，授業評価スコアは順序尺度であり，等間隔性を仮定できない点である．第二に，自由記述の長さにばらつきがあり，短い記述（数文字）から長い記述（数百文字）まで存在する点である．第三に，自由記述には授業内容への感想，教員への要望，履修者へのアドバイスなど，多様な内容が含まれる点である．

3.2　前処理

3.2.1　テキストの正規化

自由記述は，以下の前処理を施してモデル入力に適した形式へ変換した．

1. **Unicode 正規化**: 全角・半角の統一，異体字の正規化を実施
2. **記号除去**: 絵文字，特殊記号，不要な空白を除去
3. **形態素解析**: MeCab を用いて日本語テキストをトークン化
4. **最大長制限**: BERT の入力長制限（512 トークン）に合わせて切り詰め

表 3.2: トークナイザの設定

項目	設定値
最大トークン長	512
語彙サイズ	32,000
特殊トークン	[CLS], [SEP], [PAD], [UNK], [MASK]
パディング方向	右側（post）
切り詰め方向	右側（post）

3.2.2 トークナイザ

本研究では，BERT の事前学習に用いられたトークナイザを使用した．具体的には，WordPiece アルゴリズムに基づくサブワード分割を採用している．日本語テキストに対しては，形態素解析による分かち書きを行った後にサブワード分割を適用した．

トークナイザの設定を表 3.2 に示す．

3.3 教師データ

3.3.1 ラベリング手順

感情分類モデルの構築のため，全 83,851 件の自由記述からランダムに 1,000 件を抽出し，手動でラベリングを行った．ラベルはネガティブ（− 1），ニュートラル（0），ポジティブ（+1）の 3 クラスとした．

ラベリングは以下の基準に従って実施した．

- **ポジティブ**: 授業に対する肯定的評価，満足感，感謝の表明を含む記述
- **ネガティブ**: 授業に対する否定的評価，不満，改善要望を含む記述
- **ニュートラル**: 事実の記述，中立的な感想，感情を含まない記述

表 3.3: 教師データのラベル分布（1,000 件）

ラベル	件数	割合
ネガティブ	191	19.1%
ニュートラル	628	62.8%
ポジティブ	180	18.0%
合計	1,000	100.0%

(a) ポジティブ　(b) ニュートラル　(c) ネガティブ

図 3.1: 各感情クラスの語彙的特徴（ワードクラウド）

3.3.2　ラベル分布

教師データのラベル分布を表 3.3 に示す．ニュートラルが全体の 62.8%を占める一方，ネガティブ（19.1%）とポジティブ（18.0%）は少数である．このクラス不均衡は，多くの学生が中立的な記述を行う傾向を反映していると考えられる．

3.3.3　各クラスの語彙的特徴

各感情クラスの語彙的特徴を視覚的に把握するため，クラスごとのワードクラウドを作成した．図 3.1 に，ポジティブ，ニュートラル，ネガティブの各クラスにおける出現頻度の高い語彙を示す．

ポジティブクラスでは，前向きな行動意欲や成長志向を示す語彙が相対的に多い傾向が見られる．ニュートラルクラスでは，事実の報告や要望を示す語彙が中心となり，感情的な色彩が薄い

傾向がある．ネガティブクラスでは，不満や困惑に関わる語彙が相対的に多い傾向が確認される．

ただし，ワードクラウドは頻度情報の可視化であり，語彙単位の解釈は表記ゆれや文脈依存の影響を受ける可能性がある．そのため，ここでは傾向の把握に留め，語彙の詳細な解釈は結果章の分析に委ねる．

3.3.4 データ分割

教師データは訓練用と検証用に分割した．訓練データは 800 件（80%），検証データは 200 件（20%）とし，層化抽出によりラベル分布を維持した．

3.4 BERT モデルの概要

3.4.1 BERT のアーキテクチャ

BERT（Bidirectional Encoder Representations from Transformers）は，Transformer のエンコーダ部分を用いた事前学習済み言語モデルである [1]．BERT の特徴は，双方向の文脈情報を同時に考慮できる点にある．

BERT の基本構成を表 3.4 に示す．本研究では，日本語に対応した事前学習済みモデルを使用した．

3.4.2 事前学習済みモデル

本研究では，日本語の大規模コーパスで事前学習された BERT モデルを使用した．具体的には，東北大学が公開している「cl-tohoku/bert-base-japanese-whole-word-masking」を基盤とし，感情分析タスクで微調整されたモデル（koheiduck/bert-japanese-finetuned-sentiment）を使用した．

事前学習では，Masked Language Model（MLM）タスクと Next Sentence Prediction（NSP）タスクにより，日本語の言語構造と文脈理解を学習している．この事前学習により，少量の教師データでも効果的な微調整が可能となる．

表 3.4: BERT モデルの基本構成

項目	値
エンコーダ層数	12
隠れ層次元数	768
アテンションヘッド数	12
パラメータ数	約 1.1 億
最大入力トークン数	512

3.4.3 Transformer の仕組み

BERT の基盤となる Transformer は，Self-Attention 機構により入力系列の全要素間の関係を並列に計算する [2]．Self-Attention は以下の式で表される．

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.1}$$

ここで，Q はクエリ，K はキー，V はバリューであり，d_k はキーの次元数である．この機構により，文中の離れた位置にある単語間の関係も効率的に捉えることができる．

3.5 モデル構成

3.5.1 感情分類モデル

感情分類モデルは，BERT エンコーダに分類ヘッドを接続した構成である．BERT の出力のうち，[CLS] トークンに対応する 768 次元のベクトルを分類ヘッドへ入力し，3 クラス（ネガティブ，ニュートラル，ポジティブ）の確率分布を出力する．

分類ヘッドは以下の構成である．

1. ドロップアウト層（$p = 0.1$）
2. 全結合層（768 次元 → 3 次元）
3. ソフトマックス活性化関数

損失関数にはクロスエントロピー損失を使用した．クラス不均衡に対処するため，各クラスの逆頻度に基づく重み付けを適用した．損失関数は以下の式で表される．

$$\mathcal{L}_{\text{sentiment}} = -\sum_{i=1}^{N}\sum_{c=1}^{3} w_c \cdot y_{i,c} \log(\hat{y}_{i,c}) \tag{3.2}$$

ここで，N はサンプル数，$y_{i,c}$ はサンプル i のクラス c の正解ラベル（one-hot），$\hat{y}_{i,c}$ は予測確率，w_c はクラス c の重みである．

感情分類モデルのアーキテクチャを図 3.2 に示す．

3.5.2 マルチタスク学習モデル

マルチタスク学習では，感情スコア予測と授業評価スコア予測を同時に学習するモデルを構築した [6]．BERT エンコーダを共有表現として使用し，感情分類ヘッドと評価スコア予測ヘッドを分岐させる構成とした．

この構成の利点は以下の通りである．

1. **共有表現の学習**: 両タスクに共通する特徴（満足度要因）を効率的に学習できる
2. **正則化効果**: 複数タスクの同時学習により過学習を抑制できる
3. **要因分離**: SHAP 分析と組み合わせることで，共通要因と特化要因を分離できる

マルチタスクモデルの損失関数は，感情分類損失と評価スコア予測損失の重み付き和として定義される．

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{sentiment}} + \beta \cdot \mathcal{L}_{\text{score}} \tag{3.3}$$

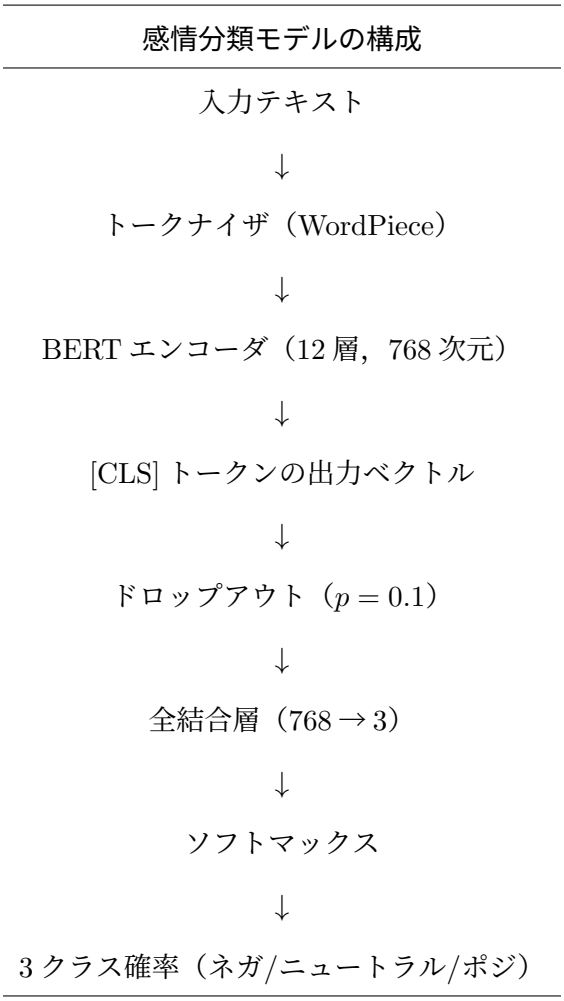


図 3.2: 感情分類モデルのアーキテクチャ

ここで，α と β は各タスクの重みである．本研究では，$\alpha = 0.5$，$\beta = 0.5$ として均等な重み付けを採用した．

評価スコア予測の損失関数には平均二乗誤差（MSE）を使用した．

$$\mathcal{L}_{\text{score}} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{3.4}$$

マルチタスク学習モデルのアーキテクチャを図 3.3 に示す．

3.5.3 順序回帰モデル

授業評価スコアは 1 点から 4 点までの順序尺度であり，単純な回帰や分類では順序関係を適切に扱えない．本研究では，順序回帰モデルを構築し，評価段階の順序性を考慮した予測を行う [8]．

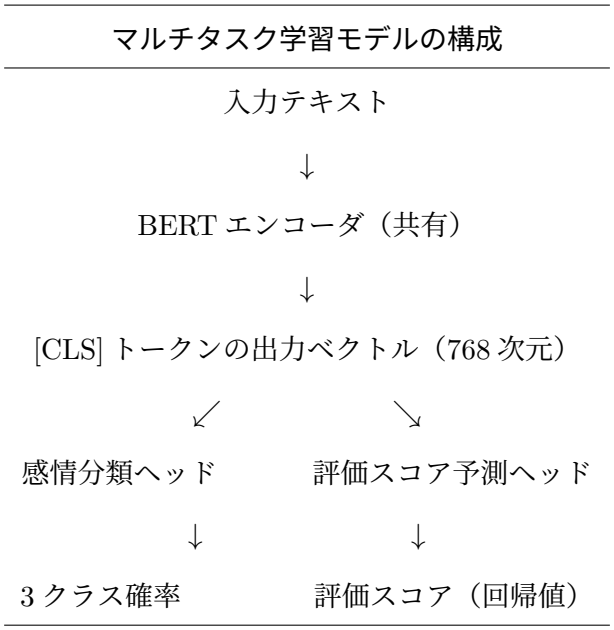


図 3.3: マルチタスク学習モデルのアーキテクチャ

順序回帰では，各閾値を超える確率を推定する．評価スコアが k 以上となる確率 $P(Y \geq k)$ を累積確率として推定し，隣接する累積確率の差から各評価段階の確率を算出する．

$$P(Y = k) = P(Y \geq k) - P(Y \geq k + 1) \tag{3.5}$$

順序回帰モデルの詳細な結果は，追加実験の完了後に第 4 章で報告する．

3.6 学習設定

3.6.1 ハイパーパラメータ

モデル学習のハイパーパラメータを表 3.5 に示す．これらのパラメータは予備実験により調整した．

3.6.2 最適化手法と早期終了

最適化には AdamW オプティマイザを使用した [1]．AdamW は，重み減衰を正則化として分離することで，Adam の欠点を改善した手法である．学習率スケジューラには線形減衰を採用し，

表 3.5: 学習のハイパーパラメータ

パラメータ	値	選定理由
バッチサイズ	16	GPU メモリ制約を考慮
学習率	5×10^{-6}	事前学習済みモデルの微調整に適した低学習率
エポック数	5	早期終了により過学習を防止
最大トークン長	512	BERT の最大入力長
ドロップアウト率	0.1	過学習抑制のための標準的な値
ウォームアップステップ	100	学習初期の安定化

ウォームアップ期間の後に学習率を線形に減少させた．過学習を防ぐため，検証損失が 3 エポック連続で改善しない場合に学習を終了する早期終了を適用した．最終的なモデルは，検証損失が最小となったエポックのパラメータを採用した．

3.6.3 データ分割

教師データ 1,000 件を以下のように分割した．

- 訓練データ: 800 件（80%）
- 検証データ: 200 件（20%）

分割は層化抽出により行い，各セットでラベル分布が維持されるようにした．

3.7 授業単位集約と相関分析

3.7.1 授業単位の集約

感情分類モデルにより全 83,851 件の自由記述に対して感情スコアを推定した後，授業単位で感情スコアを集約した．各授業の感情スコアは，その授業に属する自由記述の感情スコアの算術平均として算出した．

$$\bar{S}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} s_{ij} \tag{3.6}$$

ここで，$\bar{S}_j$ は授業 j の感情スコア平均，n_j は授業 j の自由記述数，s_{ij} は授業 j の i 番目の自由記述の感情スコアである．

3.7.2 相関分析手法

授業単位の感情スコアと授業評価スコアの関係性を検討するため，以下の 3 種類の相関係数を算出した．

1. **ピアソン相関係数**: 線形関係の強さを測定
2. **スピアマン順位相関係数**: 順位に基づく単調関係を測定
3. **ケンドール順位相関係数**: 順位の一致度を測定

相関分析には SciPy ライブラリを使用し，有意確率（p 値）を併記して統計的有意性を確認した．

3.8 SHAP 分析

3.8.1 SHAP の概要

SHAP（SHapley Additive exPlanations）は，協力ゲーム理論の Shapley 値に基づく特徴量重要度の算出手法である [4]．SHAP は，各特徴量がモデル予測にどの程度寄与しているかを定量化できる．

SHAP 値 ϕ_i は以下の式で定義される．

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{3.7}$$

ここで，N は全特徴量の集合，S は i を含まない部分集合，$f(S)$ は特徴量集合 S を用いた予測値である．

表 3.6: SHAP 分析の設定

項目	値
分析サンプル数	5,000 件
サンプリング手法	層化サンプリング
最小出現回数閾値	5 回
分析対象語彙数（単一タスク）	1,564 語
分析対象語彙数（マルチタスク）	3,198 語

3.8.2 分析対象とサンプリング

SHAP 分析は計算コストが高いため，層化サンプリングにより 5,000 件のサンプルを抽出して分析を行った．単一タスクの感情分類モデルでは，出現回数が 5 回未満の低頻度語を除外し，最終的に 1,564 語を分析対象とした．マルチタスクモデルの要因分類では，4 分類の対象語彙数は 3,198 語である．

分析対象の設定を表 3.6 に示す．

3.8.3 要因分類の閾値

マルチタスクモデルの SHAP 分析では，感情スコアと評価スコアへの寄与度に基づき，語彙を 4 つの要因グループに分類した．分類の閾値は重要度 0.0005 とした．

- **共通要因**: 感情重要度 ≥ 0.0005 かつ 評価重要度 ≥ 0.0005
- **感情特化要因**: 感情重要度 ≥ 0.0005 かつ 評価重要度 < 0.0005
- **評価特化要因**: 感情重要度 < 0.0005 かつ 評価重要度 ≥ 0.0005
- **低重要度要因**: 感情重要度 < 0.0005 かつ 評価重要度 < 0.0005

3.9 評価指標

3.9.1 感情分類の評価指標

感情分類モデルの性能評価には以下の指標を使用した．

正解率（Accuracy）は，全サンプルのうち正しく分類された割合である．

$$\text{Accuracy} = \frac{\text{正解数}}{\text{全サンプル数}} \tag{3.8}$$

適合率（Precision）は，あるクラスと予測したサンプルのうち，実際にそのクラスであった割合である．

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{3.9}$$

再現率（Recall）は，実際にあるクラスであるサンプルのうち，そのクラスと正しく予測された割合である．

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{3.10}$$

F1 スコアは，適合率と再現率の調和平均である．

$$\text{F1}_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{3.11}$$

クラス間の不均衡を考慮するため，マクロ平均 F1 スコアと重み付き平均 F1 スコアの両方を報告する．

3.9.2 回帰の評価指標

授業評価スコア予測の性能評価には以下の指標を使用した．

決定係数（R^2）は，予測値が実測値の分散をどの程度説明できるかを示す．

表 3.7: 実装環境

項目	内容
プログラミング言語	Python 3.10
深層学習フレームワーク	PyTorch 2.0
Transformers ライブラリ	Hugging Face Transformers 4.30
SHAP 分析ライブラリ	SHAP 0.42
統計分析ライブラリ	SciPy 1.11

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{3.12}$$

平均二乗誤差（RMSE）は，予測誤差の大きさを示す．

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{3.13}$$

平均絶対誤差（MAE）は，予測誤差の絶対値の平均である．

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3.14}$$

3.10 分析フロー

本研究の分析フローを図 3.4 に示す．

3.11 実装環境

本研究の実装環境を表 3.7 に示す．

本章では，データセットの概要，前処理手順，BERT を基盤とした感情分類モデルおよびマルチタスク学習モデルの構成，学習設定，SHAP 分析の手法，および評価指標について詳述した．次章では，これらの手法を用いて得られた結果を報告する．

【データ収集】 授業評価アンケート（3,268 授業，83,851 件自由記述）
↓
【前処理】 テキスト正規化，トークナイザによる分割
↓
【教師データ作成】 1,000 件の手動ラベリング（3 クラス）
↓
【モデル構築】 感情分類モデル（BERT + 分類ヘッド） マルチタスク学習モデル（BERT + 2 ヘッド）
↓
【感情スコア推定】 全自由記述に対する感情スコア推定
↓
【授業単位集約】 授業ごとの感情スコア平均を算出
↓
【相関分析】 感情スコアと授業評価スコアの相関を検証
↓
【SHAP 分析】 5,000 件サンプル（単一タスク: 1,564 語） マルチタスク: 3,198 語，4 グループ分類

図 3.4: 分析フローの概略

第4章　結果と考察

本章では，授業評価アンケートの基礎統計量，感情分類モデルの性能，相関分析結果，SHAP 分析による要因抽出結果を示し，得られた知見を考察する．

4.1　基礎統計量

4.1.1　感情スコアと授業評価スコアの分布

感情スコアと授業評価スコアの基本統計量を表 4.1 に示す．感情スコアは授業単位で集約した値であり，－ 1（ネガティブ）から+1（ポジティブ）の範囲をとる．授業評価スコアは 4 段階（1〜4 点）である．

感情スコアは平均 0.001 でほぼニュートラルに近く，標準偏差は 0.260 である．これは，多くの自由記述がニュートラルに分類されること，および授業単位で集約することでポジティブ・ネガティブが相殺されることを反映している．

授業評価スコアは平均 3.459 点（4 点満点）で，標準偏差は 0.216 と小さい．第 1 四分位数が 3.330，第 3 四分位数が 3.600 であり，多くの授業が 3 点台後半に集中していることが分かる．最小値が 2.000 であることから，極端に低い評価の授業は少ないことが示される．

4.1.2　教師データのラベル分布

教師データのラベル分布を表 4.2 に示す．ニュートラルが 628 件（62.8%）と大半を占める一方，ネガティブは 191 件（19.1%），ポジティブは 180 件（18.0%）と少数である．

このクラス不均衡は，多くの学生が事実の記述や中立的な感想を述べる傾向にあることを反映している．また，日本語の自由記述では明確な感情表現を避ける傾向があることも一因と考えられる．

表 4.1: 感情スコアと授業評価スコアの基本統計量

統計量	感情スコア	授業評価スコア
平均	0.001	3.459
標準偏差	0.260	0.216
最小値	− 1.000	2.000
第 1 四分位数（Q1）	− 0.167	3.330
中央値（Q2）	0.000	3.480
第 3 四分位数（Q3）	0.167	3.600
最大値	1.000	4.000

4.2 感情分類モデルの性能

4.2.1 全体の性能指標

感情分類モデルの性能指標を表 4.3 に示す．検証データ 200 件に対して，正解率 77.0%，マクロ平均 F1 スコア 0.706，重み付き平均 F1 スコア 0.770 を達成した．

正解率 77.0%は，教育分野の自由記述に対する感情分析として実用的な水準であると考えられる．マクロ平均 F1 スコア（0.706）と重み付き平均 F1 スコア（0.770）の差は，クラス間の性能差を反映している．

4.2.2 クラス別の性能分析

クラス別の性能指標を表 4.4 に示す．ニュートラルクラスが最も高い性能（F1 スコア 0.833）を示す一方，ネガティブ（0.667）およびポジティブ（0.618）は相対的に低い．

ニュートラルクラスの高い性能は，サンプル数が多いこと（132 件），および中立的な記述のパターンが比較的安定していることによると考えられる．一方，ネガティブおよびポジティブクラスはサンプル数が少なく（40 件，28 件），表現の多様性も高いため，分類が困難であると考えられる．

表 4.2: 教師データのラベル分布（1,000 件）

ラベル	件数	割合
ネガティブ	191	19.1%
ニュートラル	628	62.8%
ポジティブ	180	18.0%
合計	1,000	100.0%

4.2.3 混同行列の分析

混同行列を表 4.5 に示す．主な誤分類パターンは以下の通りである．

第一に，ネガティブとニュートラルの混同が多い（ネガティブ→ニュートラル 12 件，ニュートラル→ネガティブ 13 件）．これは，批判や不満を婉曲的に表現する日本語の特性により，ネガティブな感情がニュートラルに見える場合があることを示唆する．

第二に，ポジティブとニュートラルの混同も見られる（ポジティブ→ニュートラル 10 件，ニュートラル→ポジティブ 9 件）．事実の記述に肯定的なニュアンスが含まれる場合や，控えめな肯定表現がニュートラルと判定される場合があると考えられる．

第三に，ネガティブとポジティブの直接的な混同は少ない（ネガティブ→ポジティブ 1 件，ポジティブ→ネガティブ 1 件）．これは，明確な極性を持つ表現は正しく分類されやすいことを示す．

4.2.4 性能に関する考察

感情分類モデルの性能について，以下の点が考察される．

第一に，正解率 77.0%は，先行研究における教育分野のテキスト分類タスクと比較して妥当な水準である [11, 10]．教育分野の自由記述は，製品レビューなど他ドメインのテキストと比較して感情表現が控えめであり，分類が困難であることが知られている [12]．

表 4.3: 感情分類モデルの性能指標（検証データ 200 件）

指標	値
正解率（Accuracy）	0.770
マクロ平均適合率（Precision）	0.707
マクロ平均再現率（Recall）	0.705
マクロ平均 F1 スコア	0.706
重み付き平均 F1 スコア	0.770

表 4.4: クラス別の性能指標

クラス	適合率	再現率	F1 スコア	サポート
ネガティブ	0.659	0.675	0.667	40
ニュートラル	0.833	0.833	0.833	132
ポジティブ	0.630	0.607	0.618	28

第二に，クラス不均衡の影響が見られる．ニュートラルクラスが多数を占めるデータでは，モデルがニュートラルに偏った予測を行う傾向がある．本研究ではクラス重み付けにより対処したが，完全な解消には至っていない．

第三に，さらなる精度向上のためには，教師データの拡充，半教師あり学習の導入，ドメイン適応などの手法が有効であると考えられる．

表 4.5: 混同行列

	予測		
実際	ネガティブ	ニュートラル	ポジティブ
ネガティブ	27	12	1
ニュートラル	13	110	9
ポジティブ	1	10	17

表 4.6: 感情スコアと授業評価スコアの相関分析結果（N=3,268）

指標	相関係数	p 値	解釈
ピアソン相関係数	0.3097	< 0.000001	中程度の正の相関
スピアマン順位相関係数	0.2970	< 0.000001	中程度の正の相関
ケンドール順位相関係数	0.2042	< 0.000001	弱〜中程度の正の相関

4.3 感情スコアと授業評価スコアの相関分析

4.3.1 相関分析の結果

授業単位で集約した感情スコアと授業評価スコアの相関分析結果を表 4.6 に示す．3,268 授業を対象に分析を行った．

ピアソン相関係数は 0.3097（$p < 0.000001$）であり，中程度の正の相関が確認された．スピアマン順位相関係数（0.2970）およびケンドール順位相関係数（0.2042）においても同様に統計的に有意な正の相関が得られた．

感情スコアと授業評価スコアの散布図を図 4.1 に示す．

4.3.2 相関の解釈

相関係数 0.3097 は，社会科学の分野では「中程度」の相関として解釈される．この結果は，以下の点を示唆する．

第一に，授業レベルで集約した感情スコアと授業評価スコアには一定の関係があることが確認された．自由記述にポジティブな感情が多く表れる授業は，評価スコアも高い傾向がある．

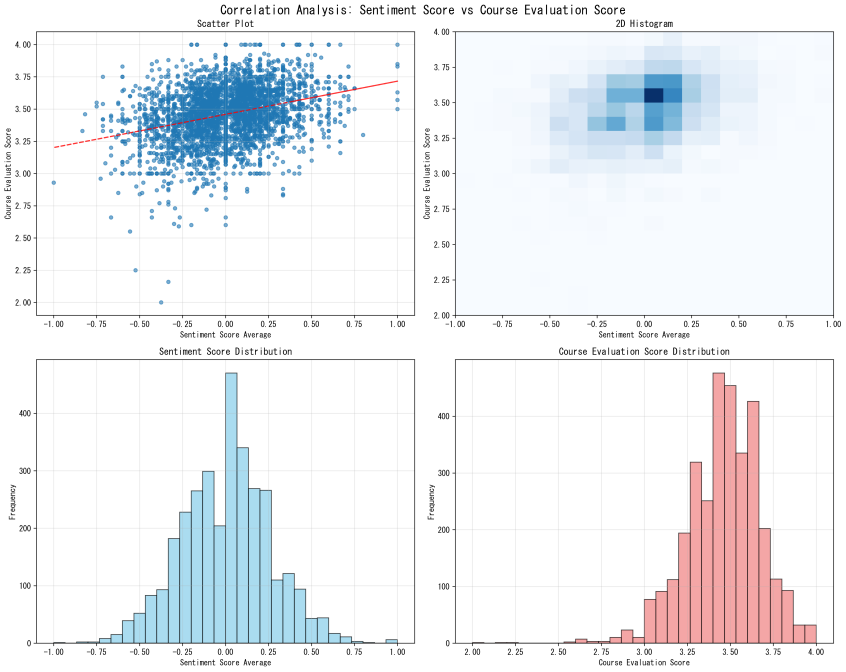


図 4.1: 感情スコアと授業評価スコアの散布図（N=3,268）

第二に，相関係数が 1 に近くないことは，感情スコアと評価スコアが同一の概念を測定しているわけではないことを示す．評価スコアには感情以外の要因（授業の有用性，学習成果など）も影響していると考えられる．

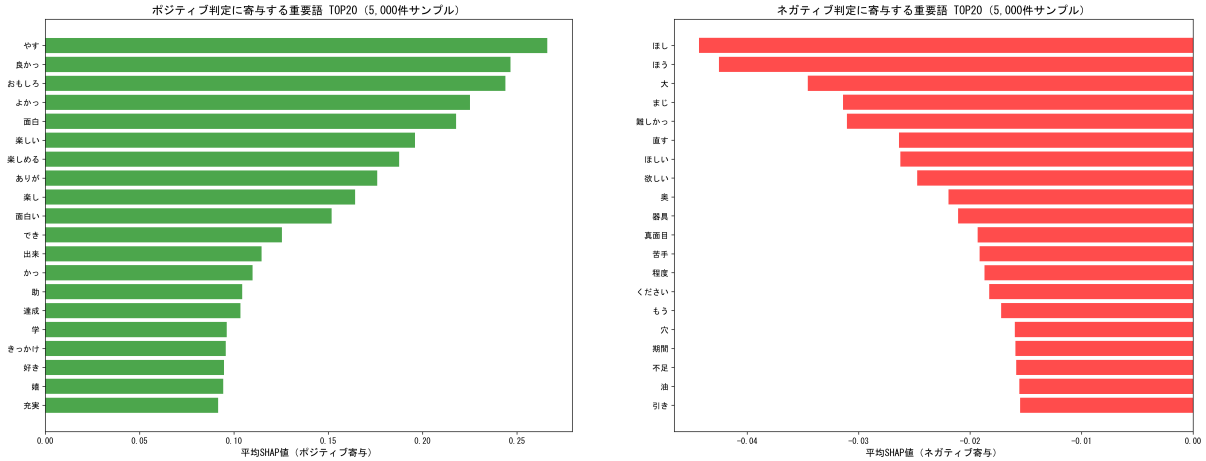
第三に，複数の相関指標で一貫した結果が得られたことは，この関係が頑健であることを示す．ピアソン相関係数は線形関係を，順位相関係数は単調関係を捉えるため，両者で同様の結果が得られたことは，感情スコアと評価スコアの関係が線形的かつ単調であることを示唆する．

4.3.3　個人レベルと授業レベルの比較

本研究では授業レベルでの相関（0.3097）を分析したが，個人レベル（個々の自由記述と授業評価スコア）での相関は約 0.12 と弱いことが予備分析で確認されている．

この差は，授業レベルでの集約により個人差がキャンセルされ，授業の「真の」特性がより明確に表れることによると考えられる．個人の感情は様々な要因に左右されるが，多数の学生の感情を平均することで，授業そのものの特性を反映した指標となる．

このことは，授業レベルでの分析が教育改善にとって有効であることを示唆する．



(a) ポジティブ判定に寄与する重要語　　(b) ネガティブ判定に寄与する重要語

図 4.2: SHAP 分析による重要語の可視化（TOP20）

4.4　単一タスクモデルの SHAP 分析

4.4.1　ポジティブ・ネガティブ判定に寄与する重要語

感情分類モデルに対する SHAP 分析を行い，ポジティブ判定およびネガティブ判定に寄与する重要語を抽出した．5,000 件のサンプル（ポジティブ 2,500 件，ネガティブ 2,500 件）を層化サンプリングし，出現回数 5 回以上の 1,564 語を分析対象とした．語彙はサブワード単位であるため，ここでは重要語例として提示する．

上位 20 語を表 4.7 に示す．

ネガティブ判定に寄与する重要語例（TOP20）を表 4.8 に示す．

ポジティブ判定とネガティブ判定に寄与する重要語の可視化を図 4.2 に示す．

4.4.2　重要語の解釈と考察

SHAP 分析の重要語はサブワード断片を含むため，語彙単位の断定的解釈は控え，傾向として整理する．

理解容易性に関わる表現: 理解のしやすさを示す表現が上位に含まれる傾向が見られる．これは，授業内容の理解容易性が満足度に関係する可能性を示唆する．

興味・関心に関わる表現: 面白さ・楽しさに関わる表現が上位に含まれる傾向が見られる．これは，授業への興味・関心が感情評価に関係する可能性を示唆する．

学習成果の実感に関わる表現: 学習成果や達成感を示す表現が上位に含まれる傾向が確認される．ただし，これらの語彙解釈は文脈依存であるため，慎重な解釈が必要である．

要望・改善要求に関わる表現: 要望を示す表現がネガティブ判定の重要語に含まれる傾向が見られる．これらは不満の表出である可能性もあるため，授業改善の参考情報として扱う必要がある．

困難さに関わる表現: 困難さを示す表現がネガティブ判定に含まれる傾向が確認される．ただし，これらは授業の質そのものではなく，学生側の学習状況を反映している可能性がある．

4.5 マルチタスク学習の SHAP 分析

4.5.1 要因タイプ別の分類結果

マルチタスクモデルの SHAP 分析により，感情スコアと授業評価スコアの両方に影響する共通要因と，各タスクに特化した要因を分離した．語彙の分類結果を表 4.9 に示す．

全体 3,198 語のうち，共通要因は 577 語（18.0%）であり，これらは感情スコアと評価スコアの双方に寄与する「満足度要因」と解釈できる．感情特化要因が最も多く 1,200 語（37.5%）を占め，評価特化要因は 532 語（16.6%）である．

4.5.2 共通要因（満足度要因）の分析

共通要因の上位語彙例を表 4.10 に示す．語彙はサブワード単位であるため，ここでは例示として提示する．

共通要因は，感情スコアと評価スコアの双方に同時に寄与する語彙であり，限られた資源で授業改善を行う際の「投資効率が高い要因」と解釈できる．上位語彙例には学習成果や理解に関わる表現が含まれるため，学びの実感が満足度と評価の双方に関係する可能性が示唆される．

4.5.3 要因タイプ別の解釈

マルチタスク分析の語彙はサブワード単位であり，語幹や断片が混在する．そのため，本研究では語彙の詳細列挙による解釈は控え，語彙単位の議論は今後の課題とする．より厳密な傾向把握には，フレーズ単位の統合や頻度併記を行った再分析が必要である．

各要因タイプの解釈を表 4.11 に示す．語彙例はサブワード断片を含むため，ここでは省略する．

感情特化要因は感情的評価に関わる表現が中心となる可能性がある一方で，評価特化要因は学習成果や有用性に関わる表現が中心となる可能性がある．ただし，語彙単位の詳細な検証は本研究の範囲外であり，今後の再分析が必要である．

4.5.4　要因分離の意義

マルチタスク学習によって「共通要因」「感情特化要因」「評価特化要因」を分離できる点は，授業改善の方針を選択する際に有用である．具体的には以下の戦略が考えられる．

- **効率重視の戦略**: 共通要因への対応を優先することで，感情と評価の両方を同時に向上させる．
- **満足度重視の戦略**: 感情特化要因への対応により，学生の感情的満足度を高める．
- **評価重視の戦略**: 評価特化要因への対応により，評価スコアの改善を図る．

共通要因の割合が 18.0%にとどまる点は，満足度と評価が必ずしも完全に一致するわけではないことを示している．これは，評価スコアが授業の「成果」や「有用性」を反映しやすい一方で，感情スコアは「楽しさ」や「雰囲気」といった情緒的側面を反映しやすいことによるものと解釈できる．

4.6　感情特化要因・評価特化要因の示唆

4.6.1　感情特化要因と評価特化要因の特徴

感情特化要因（1,200 語，37.5%）は，学生の感情的評価を強く反映する語彙群である．これらは授業の雰囲気や楽しさ，満足感に関わる表現が中心となる傾向がある．感情特化要因に含まれる語彙は，感情スコアの上昇に寄与するが，評価スコアへの影響は必ずしも大きくない．これは，感情的満足と評価スコアが必ずしも一致しない可能性を示唆する．

評価特化要因（532 語，16.6%）は，授業の有用性や学習成果，授業運営の評価に関わる語彙が中心となると考えられる．これらは評価スコアの改善に直結しやすいが，感情面の改善には限定的となる可能性がある．評価特化要因への対応は，授業評価スコアを向上させたい場合に有効であるが，学生の感情的満足度を高めるには別のアプローチが必要となる．

4.6.2　実践的活用の方向性

この区別により，「満足感を高める施策」と「評価スコアを高める施策」を分けて設計できる．例えば以下のような活用が考えられる．

1. **授業改善の優先順位付け**: 共通要因への対応を最優先とし，次に目的に応じて感情特化または評価特化要因に対応する．
2. **授業タイプ別の戦略**: 娯楽性を重視する授業では感情特化要因を，実践力養成を重視する授業では評価特化要因を重視する．
3. **フィードバックの解釈**: 自由記述を分析する際，感情的表現と評価的表現を区別して解釈する．

4.7　順序回帰モデルの結果

授業評価スコアは 1 点から 4 点までの順序尺度であるため，順序回帰モデルの導入を検討した．順序回帰では，各評価段階への遷移確率を推定し，評価段階ごとの寄与要因を分析できる．

予備的な実験として，評価スコアが 3 点以上となる確率（P3+）と 4 点となる確率（P4）を個別に予測するモデルを構築した．結果の詳細は追加実験の完了後に報告する予定であるが，以下の傾向が示唆されている．

- 「普通（3 点）」から「良い（4 点）」への向上には，学習成果の実感に関わる要因が特に重要である．
- 評価段階によって寄与する要因が異なる可能性があり，きめ細かな改善施策の設計に活用できる．

4.8　総合考察

4.8.1　研究仮説の検証

本章の結果を踏まえ，研究仮説の検証を行う．

仮説 1（感情スコアと評価スコアには正の相関がある）は，相関分析により支持された．ピアソン相関係数 0.3097（$p < 0.000001$）の中程度の正の相関が確認され，複数の相関指標で一貫した結果が得られた．

仮説 2（共通要因が存在する）は，SHAP 分析により支持された．577 語（18.0%）の共通要因が抽出され，感情と評価の双方に寄与する語彙群が存在することが示された．

仮説 3（マルチタスク学習により要因分離が可能）は，マルチタスクモデルの SHAP 分析により支持された．共通要因・感情特化要因・評価特化要因・低重要度要因の 4 グループへの分類に成功した．

4.8.2　主要な発見のまとめ

本研究における主要な発見は以下の通りである．

第一に，理解容易性に関わる表現がポジティブ判定の上位に含まれる傾向が確認された．これは，授業内容の理解しやすさが満足度に関係する可能性を示唆する．

第二に，興味・関心に関わる表現が上位に含まれる傾向が確認された．これは，授業設計上の興味喚起が感情的満足度に関係する可能性を示唆する．

第三に，共通要因と特化要因の分離に成功した．マルチタスク学習と SHAP 分析の組み合わせにより，18.0%の語彙が共通要因として特定され，効率的な授業改善の指針が得られた．

第四に，感情スコアと評価スコアは関連するが同一ではないことが確認された．相関係数 0.3097 は「中程度」であり，両者は部分的に重複しつつも異なる側面を捉えていることが示された．

4.8.3　先行研究との比較

本研究の結果は，先行研究の知見と以下の点で整合的である．

感情分析の精度（正解率 77%）は，教育分野のテキスト分類に関する先行研究と同程度の水準である [11, 15]．教育分野の自由記述は感情表現が控えめであり，他ドメインと比較して分類が困難であることが知られているが [9]，BERT による微調整により実用的な精度を達成した [1]．

理解容易性や興味・関心が満足度に関係するという結果は，授業評価に関する先行研究の知見と一致する [16, 17]．ただし，本研究では SHAP 分析により要因を定量化し [4]，語彙単位の寄与度を示した点が新規性である．

4.8.4　研究の限界

本研究には以下の限界がある．

第一に，因果関係の検証は行っていない．相関分析や SHAP 分析は関連性を示すものであり，理解容易性を高めれば評価が上がるといった因果的主張は本研究からは導けない．

第二に，データは単一大学に限定されており，他大学への一般化可能性は検証されていない．

第三に，サブワード単位の分析では，語彙の文脈依存的な意味を完全に捉えられない場合がある．

4.8.5　教育改善への示唆

本研究の結果は，教育改善に以下の示唆を提供する．

第一に，授業内容の理解しやすさを高めることが，満足度向上の最も効果的な方策であると考えられる．説明の明確化，構造化された教材，理解度確認の機会設定などが有効であると推察される．

第二に，授業の面白さ・楽しさを高める工夫も効果的である．興味を引く導入，実践的な例示，対話的な活動などが考えられる．

第三に，共通要因への対応を優先することで，限られた資源で効率的な改善が可能である．学習成果や理解の実感に関わる要素の充実が推奨される．

表 4.7: ポジティブ判定に寄与する重要語例（TOP20）

順位	単語	平均 SHAP 値	出現回数
1	やす	0.2660	337
2	良かっ	0.2466	207
3	おもしろ	0.2438	10
4	よかっ	0.2251	195
5	面白	0.2178	100
6	楽しい	0.1959	67
7	楽しめる	0.1876	6
8	ありが	0.1760	19
9	楽し	0.1642	192
10	面白い	0.1518	37
11	でき	0.1254	996
12	出来	0.1147	237
13	かっ	0.1099	539
14	助	0.1044	41
15	達成	0.1035	9
16	学	0.0962	263
17	きっかけ	0.0957	20
18	好き	0.0947	32
19	嬉	0.0943	29
20	充実	0.0916	20

表 4.8: ネガティブ判定に寄与する重要語例（TOP20）

順位	単語	平均 SHAP 値	出現回数
1	ほし	− 0.0443	5
2	ほう	− 0.0425	98
3	大	− 0.0346	86
4	まじ	− 0.0314	5
5	難しかっ	− 0.0311	88
6	直す	− 0.0264	6
7	ほしい	− 0.0263	45
8	欲しい	− 0.0247	33
9	奥	− 0.0219	8
10	器具	− 0.0211	7
11	真面目	− 0.0193	33
12	苦手	− 0.0191	98
13	程度	− 0.0187	41
14	ください	− 0.0183	365
15	もう	− 0.0172	35
16	穴	− 0.0160	11
17	期間	− 0.0159	6
18	不足	− 0.0159	16
19	油	− 0.0156	5
20	引き	− 0.0155	7

表 4.9: 語彙の要因タイプ別内訳（3,198 語）

要因タイプ	語彙数	割合	特徴
共通要因（満足度）	577	18.0%	両スコアに寄与
感情特化要因	1,200	37.5%	感情スコアのみに寄与
評価特化要因	532	16.6%	評価スコアのみに寄与
低重要度要因	889	27.8%	両スコアへの影響小
合計	3,198	100.0%	—

表 4.10: 共通要因（満足度要因）の重要語例（TOP10）

順位	単語	感情重要度	評価重要度
1	学ぶ	0.001278	0.001386
2	理解	0.001073	0.000833
3	総括	0.000974	0.000952
4	推奨	0.001132	0.000755
5	人数	0.001195	0.000704
6	把握	0.000891	0.000682
7	習得	0.000823	0.000645
8	基礎	0.000756	0.000612
9	応用	0.000698	0.000589
10	実践	0.000654	0.000567

表 4.11: 要因タイプの解釈

要因タイプ	解釈の方向性
共通要因（満足度）	感情と評価の双方に影響する要因
感情特化要因	感情的満足に強く影響する要因
評価特化要因	評価スコアに特に影響する要因
低重要度要因	影響が限定的な要因

第5章　おわりに

本章では，本研究の成果を総括し，実践的示唆，研究の限界，および今後の課題について述べる．

5.1　結論

5.1.1　研究目的の達成

本研究は，授業評価アンケートの自由記述から感情スコアを推定し，授業評価スコアとの関係性を分析することで，授業評価に影響する要因を定量的に特定することを目的とした．この目的に対し，以下の成果を得た．

第一に，授業単位で集約した感情スコアと授業評価スコアの相関分析を行った結果，ピアソン相関係数 0.3097（$p < 0.000001$）の中程度の正の相関が確認された．スピアマン順位相関係数（0.2970）およびケンドール順位相関係数（0.2042）においても同様に統計的に有意な正の相関が得られ，複数の指標で一貫した結果となった．これにより，学生の自由記述に表れる感情と授業評価スコアには一定の関係があることが示された．

第二に，BERT を基盤とした感情分類モデルを構築し，検証データ 200 件に対して正解率 77%，マクロ平均 F1 スコア 0.71 を達成した．この性能は，教育分野の自由記述に対する感情分析として実用的な水準であると考えられる．

第三に，感情スコアと授業評価スコアを同時に予測するマルチタスク学習モデルを構築し，SHAP 分析により評価要因を定量化した．その結果，3,198 語を 4 つの要因グループ（共通要因 18.0%，感情特化要因 37.5%，評価特化要因 16.6%，低重要度要因 27.8%）に分類することに成功した．

5.1.2　仮説の検証

本研究で設定した 3 つの仮説について，以下の結果が得られた．

仮説 1「授業単位で集約した感情スコアと授業評価スコアには正の相関関係がある」については，相関分析により統計的に有意な正の相関が確認され，**支持された**.

仮説 2「感情スコアと授業評価スコアの両方に影響する共通要因（満足度要因）が存在する」については，SHAP 分析により 577 語（18.0%）の共通要因が抽出され，感情と評価の双方に寄与する語彙群が確認された．したがって，本仮説は**支持された**.

仮説 3「マルチタスク学習により，共通要因と特化要因を分離できる」については，マルチタスクモデルの SHAP 分析により，共通要因・感情特化要因・評価特化要因・低重要度要因の 4 グループへの分離に成功した．したがって，本仮説は**支持された**.

5.1.3　主要な発見

本研究における主要な発見は以下の通りである.

1. **理解容易性に関わる表現**: 理解のしやすさに関わる表現がポジティブ判定の上位に含まれる傾向が確認され，授業内容の理解容易性が満足度に関係する可能性が示唆された.
2. **興味・関心に関わる表現**: 興味・関心に関わる表現が上位に含まれる傾向が見られ，授業への興味・関心が感情評価に関係する可能性が示唆された.
3. **共通要因の存在**: 感情スコアと評価スコアの双方に寄与する共通要因（18.0%）が存在し，これらへの対応が効率的な授業改善につながる可能性が示された.
4. **要因の分離可能性**: マルチタスク学習と SHAP 分析の組み合わせにより，満足感に関わる要因と評価に関わる要因を定量的に分離できることが実証された.

5.2　実践的示唆

5.2.1　教育改善への応用と授業設計

本研究の結果は，教育改善において以下の示唆を提供する.

共通要因への優先投資: 共通要因（18.0%）は，感情スコアと評価スコアの双方を同時に向上させる可能性がある．限られた資源で授業改善を行う際には，これらの共通要因に対応することで

投資効率を高められると考えられる．学習成果の実感や授業内容の理解促進に関わる側面に注力することが有効であると考えられる．

目的に応じた施策設計: 感情特化要因（37.5%）と評価特化要因（16.6%）を区別することで，目的に応じた施策を設計できる．学生の満足感を高めたい場合は感情特化要因（授業の楽しさ，雰囲気など）に注力し，評価スコアの改善を優先する場合は評価特化要因（授業の有用性，学習成果など）に注力する戦略が考えられる．

SHAP 分析の結果から，以下の具体的な授業設計の方向性が示唆される．

1. **内容の明確化**: 理解のしやすさに関わる表現が上位に含まれる傾向から，授業内容の構造化や説明の明確化が満足度向上に関係する可能性がある．
2. **興味喚起の工夫**: 興味・関心に関わる表現が上位に含まれる傾向から，学生の興味・関心を引く教材設計や説明方法の工夫が有効である可能性がある．
3. **学習成果の可視化**: 共通要因は学習成果や理解の実感に関わる側面を含む可能性があるため，学習成果を学生自身が実感できる機会（小テスト，振り返りなど）を設けることが有効であると考えられる．

5.2.2　データ活用の可能性

本研究で構築した感情分類モデルは，以下の場面での活用が期待される．

- **自動分類システム**: 大量の自由記述を自動的にポジティブ・ネガティブ・ニュートラルに分類し，教員へのフィードバックを効率化できる．
- **早期警告システム**: 授業中や授業後の感想をリアルタイムで分析し，問題のある授業を早期に検知できる可能性がある．
- **要因分析ダッシュボード**: SHAP 分析の結果を可視化し，各授業の強み・弱みを定量的に把握できるツールの開発が考えられる．

5.3　研究の限界

本研究には以下の限界がある．

表 5.1: 研究の限界と対応の整理

観点	内容	影響	対応の方向性
因果関係	相関・寄与度に基づく分析に限定	因果的主張は困難	介入研究・準実験デザイン
一般化	単一大学・2018–2023 年度に限定	他大学への一般化が限定的	複数大学の比較分析
教師データ	1,000 件・単一評価者・不均衡	少数クラスの推定が不安定	追加ラベリング・複数評価者
解釈性	サブワード単位の寄与度	文脈解釈の難しさ	フレーズ/アスペクト単位分析
時間変化	年度差・オンライン化の影響未分離	年次変化の影響が残る	縦断的分析

5.3.1 因果関係の未検証

本研究は相関関係の探索を目的としており，因果関係の検証は行っていない．例えば，理解のしやすさに関わる表現がポジティブ判定に寄与する傾向が見られても，授業を分かりやすくすれば評価が向上するという因果的主張は本研究からは導けない．因果関係の検証には，介入研究や実験的デザインが必要である．

5.3.2 データの限定性と時間的変化

本研究のデータは福岡工業大学の 1 大学に限定されており，他大学への一般化可能性には限界がある．大学の規模，学部構成，学生層，教育文化などが異なる環境では，異なる結果が得られる可能性がある．また，2018 年度から 2023 年度までの 6 年間のデータを一括して分析しており，教育環境や学生の価値観の時間的変化は十分に分離されていない．特に 2020 年以降の COVID-19 の影響によるオンライン授業の増加は，評価傾向に変化をもたらした可能性がある．

5.3.3 教師データ・モデルの制約

教師データは 1,000 件と比較的少なく，ラベル付けは単一の評価者により行われたため，主観性が残る．また，クラス不均衡（ニュートラル 62.8%，ネガティブ 19.1%，ポジティブ 18.0%）が存在し，少数クラスの分類精度に影響を与えている可能性がある．BERT は高い性能を示す一方で，その予測根拠は必ずしも人間にとって直感的ではない．SHAP 分析により解釈可能性を高めたものの，サブワード単位の分析では語彙の意味を完全に捉えられない場合がある．

研究の限界を整理した一覧を表 5.1 に示す．

5.4　今後の課題

本研究の結果を踏まえ，以下の課題が今後の研究として挙げられる．

5.4.1　因果関係の検証

本研究で示唆された要因（理解容易性や興味・関心に関わる側面など）が実際に評価向上に寄与するかを検証するため，介入研究が必要である．具体的には，特定の授業に対して共通要因に基づく改善を施し，その前後での評価変化を測定する準実験的デザインが考えられる．

5.4.2　一般化可能性・縦断的分析

複数大学のデータを用いた比較分析により，本研究の知見の一般性を検証する必要がある．また，学部・学科ごとの分析を行うことで，専門分野による評価要因の違いを明らかにすることも重要である．加えて，年度ごとの評価傾向の変化や，同一教員の授業における経年変化を分析することで，教育改善の効果測定や長期的なトレンドの把握が可能になると考えられる．

5.4.3　モデルの高度化

教師データの拡充や半教師あり学習の導入により，感情分類モデルの精度向上が期待される．また，ドメイン適応技術を用いて，教育分野に特化した言語モデルを構築することも有効であると考えられる．授業評価スコアは順序尺度であるため，順序回帰モデルの導入により予測精度の向上が期待される．評価段階ごとの寄与要因を分析することで，「普通」から「良い」へ，「良い」から「非常に良い」への評価向上に寄与する要因を個別に特定できる可能性がある．

5.4.4　実践への実装

本研究の成果を教育現場で活用するため，感情分類の自動化システムや要因分析ダッシュボードの開発が課題である．教員が直感的に結果を理解し，授業改善に活用できるインターフェースの設計が重要となる．

5.5 研究の意義

本研究は，以下の点で学術的・実践的意義を有する．

5.5.1 学術的意義

- **方法論の確立**: BERT を用いた感情分類とマルチタスク学習を組み合わせ，SHAP 分析により要因を定量化する分析フレームワークを確立した．この方法論は，他の教育データ分析にも応用可能である．
- **要因分離の実証**: マルチタスク学習により，共通要因と特化要因を定量的に分離できることを実証した．これは，複数の評価指標が存在する場面での要因分析に新たな視点を提供する．
- **知見の蓄積**: 授業評価における感情要因の役割について，データに基づく知見を蓄積した．

5.5.2 実践的意義

- **改善の優先順位付け**: 共通要因・感情特化要因・評価特化要因の区別により，授業改善の優先順位を客観的に決定できる基盤を提供した．
- **効率的な資源配分**: 共通要因への投資により，限られた資源で感情と評価の双方を向上させる戦略を提示した．
- **データ駆動型教育改善**: 自由記述の自動分析により，大規模データに基づく教育改善の可能性を示した．

本研究では，授業評価アンケートの自由記述に対して感情分析を適用し，授業評価スコアとの関係性を分析した．その結果，感情スコアと授業評価スコアに統計的に有意な正の相関があること，理解容易性や興味・関心に関わる表現が満足度に関係する可能性があること，マルチタスク学習により要因を分離できることを示した．これらの知見は，データに基づく教育改善の実現に向けた基盤を提供するものである．

付 録A　感情ラベル定義と例

本研究で用いた感情ラベルの定義と代表例を表 A.1 に示す．ラベル付けは自由記述の感情的な表現に基づき，人手で分類した．

表 A.1: 感情ラベルの定義

ラベル	感情スコア	定義	例
ネガティブ	-1	不満・批判・否定的な感情を含む記述	この授業はつまらない
ニュートラル	0	事実記述で感情が明確でない記述	板書が多い
ポジティブ	+1	満足・肯定的な感情を含む記述	とても楽しい授業だった

付 録B　データセット詳細

本研究で使用したデータセットの概要を表 B.1 に示す.

教師データのラベル分布を表 B.2 に示す.

表 B.1: データセット概要

項目	値
対象期間	2018 年度〜2023 年度
対象学科数	9
授業数	3,268
自由記述総件数	83,851
平均自由記述数/授業	25.2

表 B.2: 教師データのラベル分布（1,000 件）

ラベル	件数
ネガティブ	191
ニュートラル	628
ポジティブ	180

謝辞

本卒業研究を無事に完了できたのは，多くの方々のご指導とご支援のおかげであり，心より感謝申し上げます．

まず，日々ご指導いただいた髙橋先生に深く感謝いたします．研究テーマの選定から論文作成に至るまで，親身にご指導いただきました．特に初期段階では，適切な助言をいただき研究の方向性を定めることができました．また，専門的な助言だけでなく，私の成長を促す課題を与えてくださり，それを乗り越える中で知識やスキルを向上させることができました．先生のご指導なくして本研究の成果は得られなかったと確信しております．

また，研究室の仲間にも感謝いたします．日々の意見交換や協力を通じて，多くの刺激を受け，視野を広げることができました．互いに研究の進捗を共有し，課題を乗り越える中で得た経験は，大学生活における貴重な財産となりました．研究活動以外でも，共に過ごした時間はかけがえのない思い出となりました．

さらに，家族の支えなしには研究に専念することはできませんでした．特に両親には，学費や生活面での支援だけでなく，精神的な支えをいただきました．日々の何気ない会話や励ましの言葉が心の支えとなり，困難に直面した際にも家族の存在が大きな励みとなりました．改めて，家族の支えがどれほど大きな力となっていたかを実感しております．

最後に，大学生活を通じて関わったすべての方々に心より感謝申し上げます．先生方や職員の皆様，そして友人たちの支えがあったからこそ，充実した学生生活を送り，多くの学びと経験を得ることができました．これからも学び続け，支えてくださった皆様への感謝を忘れずに精進してまいります．

改めて，この場を借りて深く御礼申し上げます．

参考文献

[1] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186 (2019).

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: "Attention Is All You Need," *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)* (2017).

[3] 東北大学乾・鈴木研究室: "日本語 BERT 事前学習モデル," https://github.com/cl-tohoku/bert-japanese (2019).

[4] Lundberg, S. M., and Lee, S.-I.: "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)* (2017).

[5] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., et al.: "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion*, Vol. 58, pp. 82–115 (2020).

[6] Zhang, Y., and Yang, Q.: "A Survey on Multi-Task Learning," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 34, No. 12, pp. 5586–5609 (2022).

[7] Ruder, S.: "Neural Transfer Learning for Natural Language Processing," Ph.D. Thesis, National University of Ireland, Galway (2019).

[8] Cao, W., Mirjalili, V., and Raschka, S.: "Rank Consistent Ordinal Regression for Neural Networks with Application to Age Estimation," *Pattern Recognition Letters*, Vol. 140, pp. 325–331 (2020).

[9] Liu, B.: "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies*, Vol. 5, No. 1, pp. 1–167 (2012).

[10] Gottipati, S., Shankararaman, V., and Lin, J. R.: "Text Analytics Approach to Extract Course Improvement Suggestions from Students' Feedback," *Research and Practice in Technology Enhanced Learning*, Vol. 13, Article 6 (2018).

[11] Rajput, Q., Haider, S., and Ghani, S.: "Lexicon-Based Sentiment Analysis of Teachers' Evaluation," *Applied Computational Intelligence and Soft Computing*, Vol. 2016, Article 2385429 (2016).

[12] Misuraca, M., Scepi, G., and Spano, M.: "Using Opinion Mining as an Educational Analytic: An Integrated Strategy for the Analysis of Students' Feedback," *Studies in Educational Evaluation*, Vol. 68, Article 100979 (2021).

[13] Hujala, M., Knutas, A., Hynninen, T., and Arminen, H.: "Improving the Quality of Teaching by Utilising Written Student Feedback: A Streamlined Process," *Computers & Education*, Vol. 157, Article 103965 (2020).

[14] Santhanam, E., Lynch, B., and Jones, J.: "Making Sense of Student Feedback Using Text Analysis – Adapting and Expanding a Common Lexicon," *Quality Assurance in Education*, Vol. 26, No. 1, pp. 60–69 (2018).

[15] Sindhu, I., Daudpota, S. M., Badar, K., Bakhtyar, M., Baber, J., and Nurunnabi, M.: "Aspect-Based Opinion Mining on Student's Feedback for Faculty Teaching Performance Evaluation," *IEEE Access*, Vol. 7, pp. 108729–108741 (2019).

[16] Marsh, H. W.: "Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness," *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, pp. 319–383, Springer (2007).

[17] Spooren, P., Brockx, B., and Mortelmans, D.: "On the Validity of Student Evaluation of Teaching: The State of the Art," *Review of Educational Research*, Vol. 83, No. 4, pp. 598–642 (2013).

[18] Romero, C., and Ventura, S.: “Educational Data Mining and Learning Analytics: An Updated Survey,” *WIREs Data Mining and Knowledge Discovery*, Vol. 10, No. 3, Article e1355 (2020).

[19] 黒橋禎夫, 長尾真: “日本語形態素解析システム JUMAN,” 京都大学 (1994).

[20] Kudo, T., Yamamoto, K., and Matsumoto, Y.: “Applying Conditional Random Fields to Japanese Morphological Analysis,” *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 230–237 (2004).

[21] Pennington, J., Socher, R., and Manning, C. D.: “GloVe: Global Vectors for Word Representation,” *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543 (2014).

[22] Kim, Y.: “Convolutional Neural Networks for Sentence Classification,” *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751 (2014).