

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

第1章 はじめに	1
1.1 研究背景	1
1.2 課題の整理	1
1.3 研究目的と仮説	1
1.4 研究のアプローチ	2
1.5 本研究の構成	2
第2章 関連研究	3
2.1 授業評価研究	3
2.2 自由記述分析と感情分析	3
2.3 BERTと言語モデル	3
2.4 マルチタスク学習	3
2.5 解釈可能 AI と SHAP	4
2.6 本研究の位置づけ	4
第3章 データと手法	5
3.1 データセット	5
3.1.1 データセットの概要	5
3.1.2 授業評価アンケートの構成	5
3.1.3 データの特性	5
3.2 前処理	6
3.2.1 テキストの簡易整形	6
3.3 教師データ	6
3.3.1 ラベリング手順	6

3.3.2	ラベル分布	7
3.3.3	各クラスの語彙的特徴	7
3.3.4	データ分割	8
3.4	モデル構成	8
3.4.1	BERT ベース感情分類モデル	8
3.4.2	マルチタスク学習モデル	8
3.5	授業単位集約と相関分析	8
3.6	SHAP 分析	9
3.7	評価指標	10
3.8	分析フロー	10
第 4 章	結果と考察	12
4.1	基礎統計量	12
4.1.1	感情スコアと授業評価スコアの分布	12
4.1.2	教師データのラベル分布	12
4.2	感情スコアと授業評価スコアの相関分析	12
4.2.1	相関分析の結果	12
4.2.2	相関の解釈	13
4.3	単一タスクモデルの SHAP 分析	13
4.3.1	ポジティブ判定に寄与する重要語	13
4.3.2	ネガティブ判定に寄与する重要語	14
4.3.3	重要語の解釈と考察	15
4.4	マルチタスク学習の SHAP 分析	15
4.4.1	要因タイプ別の分類結果	15
4.4.2	共通要因（満足度要因）の分析	16
4.4.3	要因分離の意義	16
4.5	感情特化要因・評価特化要因の示唆	17
4.5.1	実践的活用の方角性	17
4.6	総合考察	18

	iii
4.6.1 研究仮説の検証	18
4.6.2 主要な発見のまとめ	19
4.6.3 先行研究との比較	19
4.6.4 研究の限界	20
4.6.5 教育改善への示唆	20
第 5 章 おわりに	21
5.1 まとめ	21
5.2 研究の限界	21
5.3 今後の課題	21
付 録 A 感情ラベル定義と例	23
付 録 B データセット詳細	24
参考文献	26

目 次

3.1 各感情クラスの語彙的特徴（ワードクラウド）	7
3.2 マルチタスク学習モデルのアーキテクチャ	9
3.3 分析フローの概略	11
4.1 感情スコアと授業評価スコアの散布図（N=3,268）	15
4.2 SHAP 分析による重要語の可視化（TOP10）	18

表 目 次

1.1	本研究の対象データ	2
2.1	既存研究との比較	4
3.1	データセット概要	6
3.2	教師データのラベル分布 (1,000 件)	7
3.3	SHAP 分析の設定	10
4.1	感情スコアと授業評価スコアの基本統計量	13
4.2	教師データのラベル分布 (1,000 件)	14
4.3	感情スコアと授業評価スコアの相関分析結果 (N=3,268)	14
4.4	ポジティブ判定に寄与する重要語 TOP10	16
4.5	ネガティブ判定に寄与する重要語 TOP10	17
4.6	語彙の要因タイプ別内訳 (3,198 語)	18
4.7	共通要因 (満足度要因) TOP10	19
4.8	感情特化要因・評価特化要因の代表語彙 (抜粋)	20
5.1	主要結果の要約	22

第1章 はじめに

1.1 研究背景

高等教育機関では、教育の質保証が重要な課題であり、授業評価は教育改善のための主要な手段として広く実施されている [9, 10]. 授業評価アンケートは一般に、評価スコアと自由記述から構成される. 評価スコアは集計・比較が容易である一方、評価の理由や具体的な改善点は直接観測できない.

自由記述には授業への具体的な意見が含まれるが、非構造テキストであるため大規模分析には困難が伴う. そのため、近年は自然言語処理技術を用いた自動分析が進められている [7, 8, 1]. また、感情分析は学生の満足度や不満を把握する方法として活用されている [6].

本研究の対象データは、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である. 対象データの規模を表 1.1 に示す.

1.2 課題の整理

授業評価の活用には、(1) 評価スコアから評価要因の内訳が直接把握できない、(2) 自由記述は大規模かつ表現が多様であり手作業の分析が困難である、(3) 感情的側面と評価スコアの関係が明確でない、(4) 改善資源の優先順位付けに客観的根拠が不足する、という課題がある. これらは授業改善に必要な要因の抽出と整理を難しくしている.

1.3 研究目的と仮説

本研究の目的は、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を分析した上で、共通要因と特化要因を定量的に整理することである. これにより、授業評価スコアの高さと関連する要因を明確化し、改善施策の検討に資する知見を得ることを目指す.

本研究では、以下の 2 点を仮説として設定する.

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. 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がある.
2. 仮説 2: 感情スコアと授業評価スコアの双方に関連する共通要因（満足度要因）が存在する.

1.4 研究のアプローチ

本研究では、BERT を基盤とした感情分類により自由記述の感情スコアを推定し [1, 3]，感情スコアと授業評価スコアを同時に扱うマルチタスク学習モデルを構築する [5]．さらに，SHAP 分析により語彙単位の寄与度を算出し [4]，共通要因と特化要因を整理する．

1.5 本研究の構成

本研究は全 5 章からなる．第 2 章では関連研究を整理し，第 3 章でデータと手法を述べる．第 4 章で結果を示し，第 5 章でまとめと今後の課題を述べる．

第2章 関連研究

本章では、授業評価研究、自由記述分析、感情分析、BERT、マルチタスク学習、SHAP に関する先行研究を概観し、本研究の位置づけを示す。

2.1 授業評価研究

授業評価は教育改善のための重要な指標であり、評価の信頼性・妥当性やバイアスに関する議論がある [9, 10]。評価スコアは比較が容易である一方、評価の理由を説明できない点が課題とされる。

2.2 自由記述分析と感情分析

自由記述は授業評価の理由や具体的な改善要望を含むが、大規模分析には自然言語処理が必要である [7, 8]。感情分析はテキストから肯定・否定・中立を推定する手法であり [6]、評価スコアとの統合的な分析は限定的である。

2.3 BERTと言語モデル

BERT は双方向の文脈情報を考慮できる事前学習済み言語モデルであり [1, 2]、日本語モデルも公開されている [3]。自由記述のような多様な表現を含むテキストに対して、高い表現力を持つ点が特長である。

2.4 マルチタスク学習

マルチタスク学習は関連タスクを同時に学習し、共通表現を共有する枠組みである [5]。評価スコアと感情スコアの同時学習に適用可能である。

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)	–	○	–	–
Hujala et al. (2020)	○	○	△	–
本研究	○	○	○	○

2.5 解釈可能 AI と SHAP

予測根拠を説明する解釈可能 AI が注目されており, SHAP は協力ゲーム理論に基づく特徴量重要度の算出手法として広く利用されている [4]. テキストでは語彙単位の寄与度を提示できる点が利点である.

2.6 本研究の位置づけ

既存研究では, 自由記述分析や感情分析は行われているが, 評価スコアとの統合的分析や要因の定量化は十分ではない [7, 8]. 本研究は, BERT による感情分類とマルチタスク学習を組み合わせ, SHAP 分析により共通要因・特化要因を語彙レベルで整理する点に特徴がある.

既存研究との差異を表 2.1 に示す.

この位置づけに基づき, 評価スコアの高さと関連する要因を定量的に示すことを目的とする.

第3章 データと手法

本章では、本研究で利用したデータセットの概要、前処理手順、モデル構成、SHAP 分析の設定、評価指標を述べる。

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。

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) 授業評価スコア (1~4 点)、(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。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。

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る。第二に、自由記述の長さにはばらつきがあり、短い記述から長い記述まで存在する。第三に、自由記述には授

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

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．

3.2 前処理

3.2.1 テキストの簡易整形

自由記述の前処理は最小限に留め，以下のみ実施した．

1. 空白・改行の整理: 連続する空白や改行を整形し，空欄を除外
2. 最大長制限: モデル入力長の上限（512 トークン）に合わせて切り詰め

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．

- 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

表 3.2: 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%



図 3.1: 各感情クラスの語彙的特徴 (ワードクラウド)

- 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

3.3.2 ラベル分布

教師データのラベル分布を表 3.2 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.

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.

語彙傾向はクラス間で差があり、感情分類モデルが語彙パターンを学習できる可能性を示唆する。

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 モデル構成

3.4.1 BERT ベース感情分類モデル

感情分類には日本語 BERT を基盤とするモデルを用いた [1, 3]。BERT エンコーダの出力から [CLS] トークンのベクトルを取得し、3 クラスの確率分布を出力する分類ヘッドを接続した。

3.4.2 マルチタスク学習モデル

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マルチタスク学習モデルを構築した [5]。BERT エンコーダを共有表現として使い、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。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、 $\alpha = \beta = 0.5$ とした。

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{sentiment}} + \beta \cdot \mathcal{L}_{\text{score}} \quad (3.1)$$

マルチタスク学習モデルの構成を図 3.2 に示す。

3.5 授業単位集約と相関分析

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。

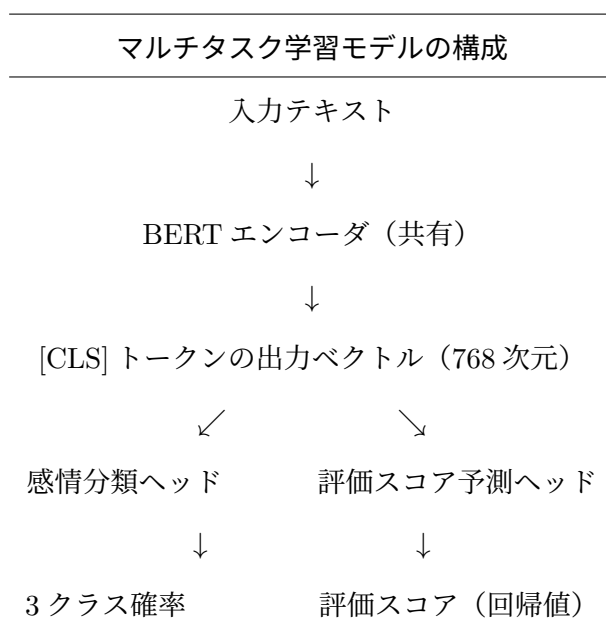


図 3.2: マルチタスク学習モデルのアーキテクチャ

$$\bar{S}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} s_{ij} \quad (3.2)$$

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、ピアソン相関係数、スピアマン順位相関係数、ケンドール順位相関係数を算出した。

3.6 SHAP 分析

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。また、出現回数が 5 回未満の低頻度語は除外し、最終的に 3,198 語を分析対象とした。

分析対象の設定を表 3.3 に示す。

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 グループに分類した。

表 3.3: 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	3,198 語

3.7 評価指標

感情分類モデルの評価には Accuracy と F1 スコア（マクロ平均・重み付き平均）を用いた。授業評価スコア予測の評価には R^2 , RMSE, MAE を用いた。

3.8 分析フロー

本研究の分析フローを図 3.3 に示す。

本章では、データセットの概要、前処理手順、BERT を基盤としたモデル構成、SHAP 分析の設定、および評価指標について述べた。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】 簡易整形， 入力長の調整
↓
【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， 3,198 語を対象に要因分析 要因を 4 グループに分類

図 3.3: 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，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 から $+1$ 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（標準偏差 0.260）であり，授業単位集約によりポジティブ・ネガティブが相殺されやすい傾向が見られる．授業評価スコアは平均 3.459（標準偏差 0.216）で，分布は 3 点台後半に集中している．

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，ネガティブ 191 件（19.1%），ポジティブ 180 件（18.0%）である．

4.2 感情スコアと授業評価スコアの相関分析

4.2.1 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 4.3 に示す．3,268 授業を対象に分析を行った．

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

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。

4.2.2 相関の解釈

相関係数 0.31 は中程度の正の相関と解釈され、感情スコアと評価スコアには一定の関係があることが示唆される。一方で、相関係数が 1 に近くないことは、評価スコアが感情以外の要因も反映している可能性を示す。複数の相関指標で一貫した結果が得られたことから、関係性は一定の頑健性を持つと考えられる。

4.3 単一タスクモデルの SHAP 分析

4.3.1 ポジ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

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%

表 4.3: 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

以上の 3,198 語を分析対象とした。

上位 10 語を表 4.4 に示す。「やす」「良かつ」「おもしろ」「面白」など、理解のしやすさや授業の面白さを示す語彙が上位に現れた。

4.3.2 ネガティブ判定に寄与する重要語

ネガティブ判定に寄与する重要語 TOP10 を表 4.5 に示す。「ほしい」「難しかつ」「苦手」など、要望や困難さを示す語彙が上位に現れた。

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。

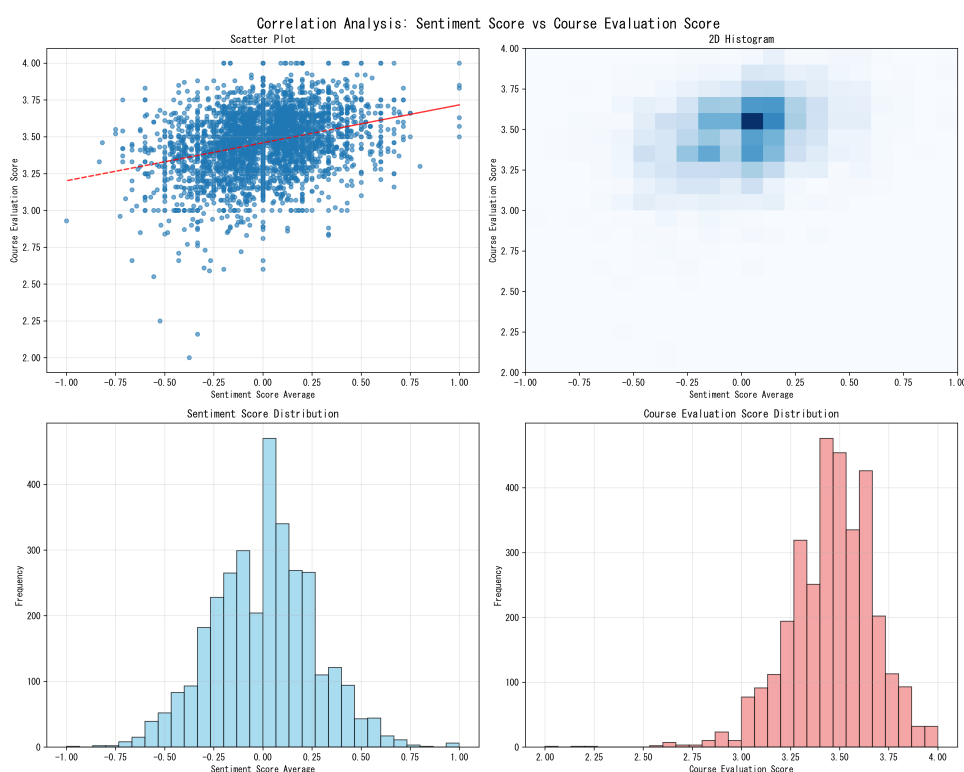


図 4.1: 感情スコアと授業評価スコアの散布図 (N=3,268)

4.3.3 重要語の解釈と考察

ポジティブ判定では理解のしやすさや面白さを示す語彙が多く、感情評価に関連する可能性が示唆される。一方、ネガティブ判定では「ほしい」「難しかった」など要望・困難さを示す語彙が現れ、改善要望として解釈できる。

4.4 マルチタスク学習のSHAP 分析

4.4.1 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 4.6 に示す。

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。

表 4.4: ポジティブ判定に寄与する重要語 TOP10

順位	単語	平均 SHAP 値	出現回数	解釈
1	やす	0.2660	337	分かりやすい, 理解しやすい
2	良かつ	0.2466	207	良かった (過去形)
3	おもしろ	0.2438	10	面白い
4	よかつ	0.2251	195	良かった (ひらがな)
5	面白	0.2178	100	面白い
6	楽しい	0.1959	67	楽しい
7	楽しめる	0.1876	6	楽しむことができる
8	ありが	0.1760	19	ありがとう
9	楽し	0.1642	192	楽しい (語幹)
10	面白い	0.1518	37	面白い

4.4.2 共通要因 (満足度要因) の分析

共通要因の上位語彙を表 4.7 に示す。「学ぶ」「理解」「総括」「推奨」「人数」など、学習成果に関わる語彙が上位に現れている。

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。例えば「学ぶ」「理解」「総括」といった語彙は、学習成果の実感に関わる可能性がある。また、感情特化要因には情緒的な語彙が、評価特化要因には有用性や運営に関わる語彙が含まれる可能性がある。

4.4.3 要因分離の意義

マルチタスク学習によって共通要因と特化要因を分離できる点は、授業改善の方針選択に有用である。共通要因の割合が 18.0%にとどまることは、満足度と評価が必ずしも一致しない可能性を示し、評価スコアと感情スコアが異なる側面を反映することを示唆する。

表 4.5: ネガティブ判定に寄与する重要語 TOP10

順位	単語	平均 SHAP 値	出現回数	解釈
1	ほし	- 0.0443	5	～してほしい
2	ほう	- 0.0425	98	～した方がいい
3	大	- 0.0346	86	大変, 大きい
4	まじ	- 0.0314	5	まじめ, まじで
5	難しかっ	- 0.0311	88	難しかった
6	直す	- 0.0264	6	直してほしい
7	ほしい	- 0.0263	45	～してほしい
8	欲しい	- 0.0247	33	～が欲しい
9	奥	- 0.0219	8	奥が深い等
10	器具	- 0.0211	7	実験器具等

4.5 感情特化要因・評価特化要因の示唆

感情特化要因（1,200 語, 37.5%）は楽しさや満足感に関わる語彙が中心であり, 評価特化要因（532 語, 16.6%）は有用性や授業運営に関わる語彙が中心となる可能性がある. 両者は役割が異なるため, 改善目的に応じた解釈が必要である.

感情特化要因・評価特化要因の代表的な語彙例（意味のある単語のみ）を表 4.8 に示す.

4.5.1 実践的活用の方向性

この区別により, 改善の優先順位付けや授業タイプ別の施策設計が可能となる. 共通要因への対応を優先しつつ, 目的に応じて感情特化要因または評価特化要因を重視する戦略が考えられる.

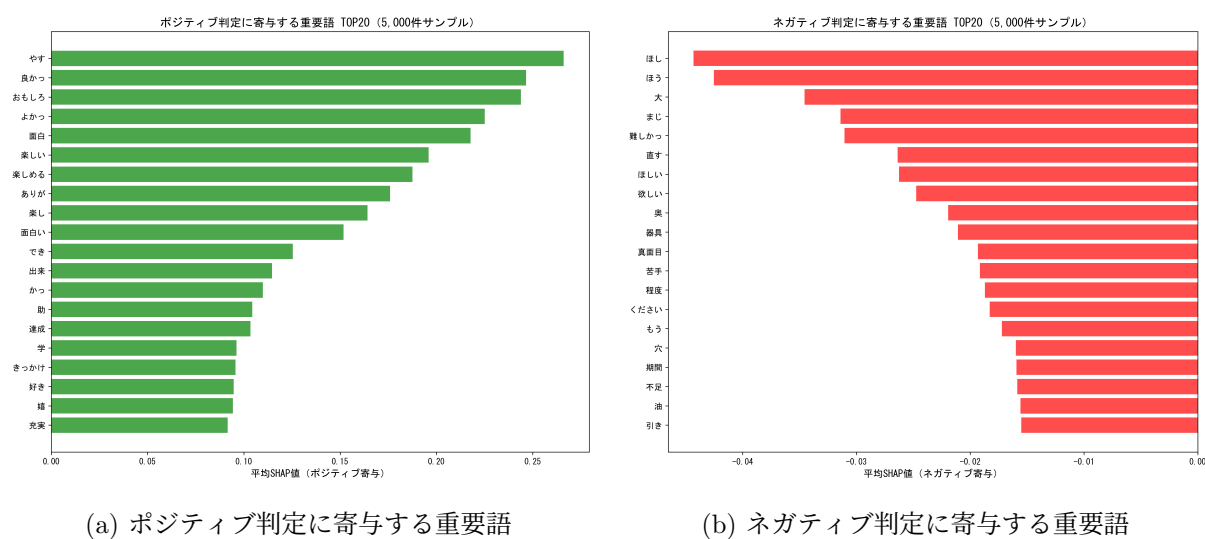


図 4.2: SHAP 分析による重要語の可視化 (TOP10)

表 4.6: 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%	—

4.6 総合考察

4.6.1 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 は、相関分析により中程度の正の相関が確認された。仮説 2 は、SHAP 分析により共通要因が抽出されたことから支持された。

表 4.7: 共通要因（満足度要因）TOP10

順位	単語	感情重要度	評価重要度	解釈
1	学ぶ	0.001278	0.001386	学びがあった
2	理解	0.001073	0.000833	理解できた
3	総括	0.000974	0.000952	総括・まとめ
4	推奨	0.001132	0.000755	推奨したい
5	人数	0.001195	0.000704	人数が適切
6	把握	0.000891	0.000682	内容を把握
7	習得	0.000823	0.000645	スキル習得
8	基礎	0.000756	0.000612	基礎が身についた
9	応用	0.000698	0.000589	応用力
10	実践	0.000654	0.000567	実践的

4.6.2 主要な発見のまとめ

本研究では、理解しやすさや面白さに関わる語彙が感情評価に強く関連すること、感情スコアと評価スコアが中程度の相関を持つこと、共通要因と特化要因を語彙レベルで分離できることが示された。

4.6.3 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。

理解しやすさや面白さが満足度に寄与するという結果は、授業評価に関する先行研究の知見と一致する [9, 10]。ただし、本研究では SHAP 分析により要因を定量化し [4]、語彙レベルでの寄与度を明らかにした点が新規性である。

表 4.8: 感情特化要因・評価特化要因の代表語彙（抜粋）

要因タイプ	代表語彙（抜粋）
感情特化要因	感謝，ありがとう，面白い，楽しい，嬉しい，対話
評価特化要因	設計，回路，関数，人材，調整，符号

4.6.4 研究の限界

本研究には以下の限界がある．

第一に，相関分析や SHAP 分析は関連性を示すものであり，因果関係は検証していない．第二に，データは単一大学に限定されているため，一般化可能性には限界がある．

4.6.5 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．

理解しやすさや学習成果の実感に関わる要因への対応は，満足度と評価の双方に寄与する可能性が高い．また，授業の面白さや雰囲気に関わる要因は感情的満足度に影響し得るため，改善目的に応じた施策設計が有効と考えられる．

第5章 おわりに

本章では、本研究の成果を簡潔にまとめ、研究の限界と今後の課題を述べる。

5.1 まとめ

本研究では、2018年度～2023年度の授業評価データ（3,268授業，83,851件自由記述）を対象に、自由記述の感情スコア推定と授業評価スコアとの関係分析を行った。感情スコアと授業評価スコアには正の相関（ピアソン 0.3097／スピアマン 0.2970／ケンドール 0.2042）が確認され、両者が一定の関係を持つ可能性が示唆された。

さらに、BERTを基盤としたマルチタスク学習とSHAP分析により、3,198語を共通要因（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%）に分類した。これにより、授業評価スコアと感情スコアの双方に関わる語彙を定量的に整理し、授業改善の優先順位付けに資する基盤を提示した。

5.2 研究の限界

第一に、本研究は相関関係と寄与度の分析に基づくため、因果関係は検証していない。第二に、対象が単一大学に限定されており、一般化可能性には限界がある。第三に、教師データは1,000件であり、クラス不均衡の影響が残る可能性がある。また、2018年度～2023年度の6年間を一括して分析しているため、時間的変化を詳細には捉えていない。

5.3 今後の課題

今後は、介入研究や準実験的デザインにより因果関係を検証することが課題である。また、他大学データを用いた比較や学科別分析により、一般化可能性を評価する必要がある。さらに、教師データ拡充を通じて、評価段階ごとの要因分析を精緻化することが求められる。

表 5.1: 主要結果の要約

項目	値
対象期間	2018 年度～2023 年度（6 年間）
授業数 / 自由記述数	3,268 / 83,851
相関係数（ピアソン/スピアマン/ケンドール）	0.3097 / 0.2970 / 0.2042
共通要因の割合	18.0%（577 語）

付 録 A 感情ラベル定義と例

感情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 値とし，自由記述の表現に基づいて人手で分類した．

付 録 B データセット詳細

データセットの規模やラベル分布は第 3 章の表で示しているため，本付録では重複掲載を省略する．必要な数値は本文の該当箇所を参照する．

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高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] 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).
- [6] Liu, B.: “Sentiment Analysis and Opinion Mining,” *Synthesis Lectures on Human Language Technologies*, Vol. 5, No. 1, pp. 1–167 (2012).
- [7] 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).
- [8] 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).
- [9] 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).
- [10] 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).