卒業論文

悪意検出における有効な特徴量の検討

中村 輝

2025 年 2 月 6 日

岐阜大学 工学部 電気電子・情報工学科 情報コース
鈴木研究室

本論文は岐阜大学工学部に
学士（工学）授与の要件として提出した卒業論文である.

中村 輝

指導教員：
鈴木 優 准教授

悪意検出における有効な特徴量の検討*

中村 輝

内容梗概

本研究では，悪意検出における精度向上に寄与する特徴量の検討を行う．

SNS の普及により，誹謗中傷や炎上といったインターネット上のトラブルが深刻化しており，その一因として悪意を含むメッセージが挙げられる．本研究では，読み手が感じる書き手の悪意についての検出精度向上を目的として，BERT をベースとしたモデルに感情スコアを追加し，その効果を検証した．

5 ちゃんねるのレスデータに対し，WRIME：主観と客観の感情分析データセットを用い，8 感情スコアおよび 8 感情をポジティブ，ネガティブ，ニュートラルに変換した PNN スコアの 2 種類を感情スコアとして付与した．特徴量の追加方法として，BERT の分類層の出力である悪意を含むか含まないか，それぞれに対するスコアへの加算をする手法 1，トークン化した入力データを BERT を通すことで得られる 768 次元の実数値のベクトルである埋め込みへ結合する手法 2，両方を組み合わせる手法 3 の三つの手法を比較し，10 交差検証を行った．

実験の結果より，両方を組み合わせた手法 3 で，8 感情スコアを使用したモデルが最も高い Accuracy を記録し，他のモデルでも，追加の特徴量を加えることで，一部の評価指標において改善が見られるものも存在した．しかし，t 検定の結果，特徴量の追加による改善は有意な差とは言えなかった．

キーワード

ニューラルネットワーク, 機械学習，ソーシャルメディア，悪意検出

*岐阜大学 工学部 電気電子・情報工学科 情報コース 卒業論文, 学籍番号: 1213033104, 2025 年 2 月 6 日.

目次

図目次 iv

表目次 v

第 1 章 はじめに 1

第 2 章 基本的事項 4

2.1 BERT 4

2.2 ニューラルネットワーク 4

2.3 評価指標 4

2.3.1 Accuracy 5

2.3.2 Precision 5

2.3.3 Recall 6

2.3.4 F 値 6

2.4 対応のある 2 標本 t 検定 6

2.5 k 分割交差検証 7

第 3 章 関連研究 8

3.1 悪意の検出に関連する研究 8

3.2 特徴量を追加する研究 9

3.3 感情分析関連の研究 9

第 4 章 提案手法 11

4.1 使用データ 11

4.2 WRIME：主観と客観の感情分析データセット 12

4.3 追加の特徴量 13

4.4 モデル作成 13

4.4.1 手法 1 14

4.4.2 手法 2 15

4.4.3 手法 3 17

第 5 章　評価実験 20
5.1　実験 20
5.1.1　実験手順 20
5.1.2　実験条件 20
5.2　結果・考察 21
5.2.1　手法 1 の結果・考察 21
5.2.2　手法 2 の結果・考察 22
5.2.3　手法 3 の結果・考察 24

第 6 章　おわりに 27
6.1　まとめ 27
6.2　今後の展望 28

謝辞 29

参考文献 31

発表リスト 33

図目次

4.1　手法 1 のモデル図 15
4.2　手法 2 のモデル図 17
4.3　手法 3 のモデル図 19

表目次

2.1 予測ラベルと正解ラベルの混同行列 5

5.1 モデル一覧と使用する特徴量 21
5.2 ベースラインの混同行列 22
5.3 モデル A-1 の混同行列 22
5.4 モデル A-2 の混同行列 22
5.5 手法 1 の性能比較 22
5.6 モデル B-1 の混同行列 23
5.7 モデル B-2 の混同行列 23
5.8 手法 2 の性能比較 24
5.9 モデル C-1 の混同行列 25
5.10 モデル C-2 の混同行列 25
5.11 モデル C-2-1 の混同行列 25
5.12 モデル C-1-2 の混同行列 25
5.13 手法 3 の性能比較 26

第 1 章　はじめに

我々は，SNS のメッセージに含まれる意図的に人を傷つけたり，不快にさせるような書き手の悪意の検出において，追加の特徴量を追加することによって，検出精度が向上すると考えた．本研究では，悪意検出の精度向上に寄与する可能性のある特徴量としてメッセージの読み手が考える書き手の感情スコアを用いると，検出精度が向上するかを検証する．

SNS の普及により，多くの人々が手軽に情報を発信できるようになったが，それに伴い，誹謗中傷や炎上といったインターネット上でのトラブルが深刻化している．令和 5 年のインターネット上の人権侵害救済手続きの件数は 1,824 件と前年より 103 件増加しており，高水準で推移している* ．実際には，人権侵犯事件として扱われていない事案も存在すると考えられる．こうしたインターネット上のトラブルを未然に防ぐ方法として，SNS 上のメッセージがトラブルの原因になり得るかを判定できるようにすることが考えられる．ここで，トラブルの原因の一つとして，悪意を含んだメッセージが挙げられる．悪意を含んだメッセージとは，意図的に他人を傷つけたり，不快にさせたり，社会的な混乱を引き起こすような目的をもったメッセージである．我々は，この悪意を含んだメッセージを検出することができれば，インターネット上のトラブルを減少させることができると考えた．

悪意を検出するために，メッセージとそれが悪意を含むかのラベル付きデータを準備し，悪意検出モデルを作成することが考えられるが，この悪意検出を行うにあたっていくつかの問題がある．悪意とは明確な基準があるわけではないため，人によって悪意の基準に差があり，判断の分かれるようなメッセージも多数存在することが考えられる．また，悪意の定義が難しいことと併せて，悪意を含んだメッセージは，その範囲が広いことが考えられるため，多様な表現が出現することも考えられる．その場合，悪意を含むかのラベル付きデータを使用し，悪意検出モデルを作成しても，特徴を十分に捉えられないことが予想される．したがって，悪意を含むかのラベルの特徴だけでなく，悪意に関連性のある特徴量を追加することで悪意の検出精度向上を目指す．

*出典：法務省 https://www.moj.go.jp/content/001415625.pdf

本研究では，追加する特徴量として読み手が考える書き手の感情を実数値のベクトルとして表現した感情スコアを使用した．感情スコアに着目した理由は以下のニつである．一つ目は，悪意と感情に関連性があると考えたからである．悪意を含むと感じるメッセージは，書き手の攻撃的な姿勢や混乱を与えようといった意図を読み手が感じることで，悪意を含むと判断し，トラブルになる可能性がある．この時，読み手が感じる悪意の一つとして，メッセージの対象となる人物や集団への怒りや恨み，嫌悪などマイナスな感情を向けていることが考えられる．したがって，悪意と感情には関連があると考えられるため，感情スコアは検出精度向上に寄与する可能性がある．二つ目は，感情スコアは悪意検出の補助的役割を果たすと考えたからである．悪意検出モデルの学習をする際，特定の攻撃的な単語を含むようなわかりやすい特徴が存在した場合, 強く依存してしまい，感情的な側面の情報を見逃す可能性が考えられる．悪意は多様な表現が存在すると考えられるが，できるだけ多くの悪意を捉えられる悪意検出モデルを作成したい．ここで，感情スコアを追加することは，メッセージの感情的な側面の情報を補うことになり，感情スコアに現れる悪意の特徴があった場合，悪意検出の補助につながることが考えられる．

本研究では，5 ちゃんねるのレスデータに悪意を含むか含まないかのラベルを付与したデータを使用し，BERT[1] をベースとした悪意検出モデルを作成した．ラベルを付与したレスデータから作成した悪意検出モデルをベースラインとし，追加の特徴量として感情スコアを追加したモデルの結果と比較し，感情スコアが精度向上に有効であるか検討を行った．

追加の特徴量である感情スコアは，Plutchik の基本 8 感情 [2] に基づいており，一つ目の感情スコアとして，Plutchik の基本 8 感情である喜び，悲しみ，期待，驚き，怒り，恐れ，嫌悪，信頼，それぞれの感情強度を 8 次元のベクトルとしたものを使用した 8 感情スコア，二つ目の感情スコアとして，喜び，期待，信頼をポジティブ，悲しみ，怒り，恐れ，嫌悪をネガティブ，驚きをニュートラルとして，それぞれの感情強度を合計して，3 次元のベクトルに変換した PNN スコアを使用した．

特徴量の追加の手法として，BERT の分類層の出力である悪意を含むか含まないか，それぞれに対するスコアへの加算をする手法 1，トークン化した入力データを BERT を通すことで得られる 768 次元の実数値のベクトルである埋め込みへ結合する手法 2，手法 1 と手法 2 の両方を使用する手法 3 の三つを用いて特徴量の追加

を行い，10 交差検証を用いて検証を行った．

実験の結果，ベースラインの Accuracy0.681 に対し，追加の特徴量として 8 感情スコアを使用し，手法 3 を使用したモデルが最も高い Accuracy として 0.705 を記録し，Precision，Recall，F 値のどれも改善の傾向が見られた．しかし，t 検定を行ったところ，有意に近いながら，特徴量の追加が有意であるとは言えなかった．

本研究の貢献は以下の通りである．

- 追加の特徴量によるモデルの性能の変化を確認した．
- 感情スコアの使用によって，悪意検出の性能改善の傾向は見られたが，有意ではないことを確認した．

本論文の構成は以下の通りである．2 章では，基本的事項について述べる．3 章では，関連研究について述べる．4 章では，提案手法について述べる．5 章では，評価実験について述べる．最後に 6 章では本論文のまとめと今後の展望について述べる．

第 2 章　基本的事項

2.1　BERT

BERT（Bidirectional Encoder Representations from Transformers）は，Google が 2018 年に発表した自然言語処理のための深層学習モデルである．BERT の特徴として，双方向の文脈を考慮することで，意味をより正確にとらえることができる点が挙げられる．BERT の学習には，MLM(Masked Language Model) と NSP(Next Sentence Prediction) が使用されている．MLM は，入力文の一部をマスクし，マスクされた単語を予測するタスクであり，NSP は，2 つの文が連続しているかを判定するタスクである．この二つによって，文脈を考慮した学習が可能となっている．また，BERT はラベルなしデータを用いた事前学習により，汎用的な言語パターンを学習させ，ファインチューニングによって，様々なタスクに特化するように学習することが可能である．

2.2　ニューラルネットワーク

ニューラルネットワークは，人間の脳の神経細胞を模倣した機械学習モデルの一つであり，データのパターンを学習し，分類や予測を行う．入力層，隠れ層，出力層の 3 層から構成され，各層の間には，重みと呼ばれるパラメータが存在する．この重みを適切に更新し，目的の結果へと近づけることをニューラルネットワークの学習という．学習には誤差逆伝播法，重みの更新には勾配降下法が使用されている．また，学習の過程で活性化関数が使用されており，これによって非線形なデータにも対応できるようになる．学習の過程で手法で設定できる学習率やバッチ数といった要素を，ハイパーパラメータと呼ぶ．

2.3　評価指標

本研究では，モデルの性能評価のために，Accuracy，Precision，Recall，F 値の四つの評価指標を用いる．予測ラベルと正解ラベルの混同行列を表 2.1 に示す．

表 2.1　予測ラベルと正解ラベルの混同行列

		予測ラベル	
		Negative	Positive
正解ラベル	Negative	TN	FP
	Positive	FN	TP

表の各要素については以下の通りである．
TN：実際に負例（Negative）であり，モデルも負例と予測した数．
FP：実際には負例（Negative）だが，モデルが誤って正例と予測した数．
FN：実際には正例（Positive）だが，モデルが誤って負例と予測した数．
TP：実際に正例（Positive）であり，モデルも正例と予測した数．

2.3.1　Accuracy

Accuracy とは，全データに対して，正しく分類された割合を表す評価指標である．Accuracy は以下のように求める．

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2.3.2　Precision

Precision とは，モデルが正例と予測したデータのうち，実際に正例だった割合を表す評価指標である．Precision は以下のように求める．

$$Precision = \frac{TP}{TP + FP}$$

2.3.3 Recall

Recall とは，実際の正例のうち，モデルが正しく正例と予測できた割合を表す．Recall は以下のように求める．

$$Recall = \frac{TP}{TP + FN}$$

2.3.4 F 値

F 値とは，Precision と Recall のバランスをとるための評価指標であり，それらの調和平均で計算される．F 値は以下のように求める．

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

2.4 対応のある 2 標本 t 検定

対応のある 2 標本 t 検定とは，二つの対応のある標本の平均値に有意な差があるかの検定である．帰無仮説は，「二つの対応のある標本の平均値に差がない」とし，検定統計量 t は以下のように求められる．

$$t = \frac{(\overline{d} - \mu)}{\sqrt{\frac{s^2}{n}}}$$

2 標本の差の平均を $\overline{d}$，差の母平均を μ，不偏分散を s^2，データ数を n とする．μ は帰無仮説より，0 とし，不偏分散 s^2 は，n 個のデータ x の平均値を $\overline{x}$ としたとき，以下のように求められる．

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

検定統計量 t から p 値を求め，有意水準を満たすかどうか検定する．

2.5 k 分割交差検証

k 分割交差検証とは，機械学習モデルの汎化性能を評価するための手法の一つである．本研究では，$k = 10$ とした 10 分割交差検証を使用する．分割した十個のデータのうち，一個をテスト用データ，一個を検証用データ，残りの八個を訓練用データとし，十個のデータすべてが 1 回ずつテスト用データとして使用されるように 10 回学習を行い，各回の評価値の平均を取ることでモデルの汎化性能を検証することができる．

第 3 章　関連研究

3.1　悪意の検出に関連する研究

悪意の検出に関連する研究は現在までに複数行われている．

西原ら [3] は，電子掲示板に投稿されるコメントから誹謗中傷を含む文を抽出する手法として，辞書ベースのアプローチのバッドワードとスレッドバッドワードという 2 種類の単語リストを使用した誹謗中傷コメントの検出を行った．また，石坂ら [4] は，単語の悪口度を算出し，それを特徴量として SVM を用いた分類手法として，誹謗中傷の検出を行った．これらは，どちらも電子掲示板を対象とし，単語という特徴を用いた誹謗中傷の検出を目指したものである．単語という特徴が誹謗中傷に関与している要素の一つであり，検出に寄与している一方で，文脈を十分に考慮できない．多様な表現が存在するため，対応できるテキストに限界があるという問題がある．悪意の検出においても，単語に着目する場合，判断基準となる単語リストに依存してしまうため，どのような単語を選定するか，その単語リストに含まれないような表現が検出できるかといった問題がある．本研究では，多様な表現を含む悪意を検出することを目的とするため，単語のような辞書ベースではないアプローチによって悪意を捉え，精度向上を目指す．

山崎ら [5] は，ソーシャルメディアにおける攻撃的表現の検出を目的とし，多様な攻撃的テキストを含むデータセットの自動構築手法を提案している．ソーシャルメディア上の炎上投稿へのリプライを収集し，かつ，データの誤ラベルを自動的に修正する仕組みを導入することで精度を向上させた．炎上投稿へのリプライは攻撃的テキストの割合が高いことに注目し，炎上投稿と非炎上投稿という関係からデータセットを自動構築している．しかし，この手法は投稿とリプライという関係を利用しデータセットを構築しており，自動構築には，主となる投稿とそれに対するリプライという関係が分かっている必要がある．また，実際に悪意を判定する場合，炎上と悪意に関連性はあると考えられるが，炎上する投稿に必ず悪意が含まれているかはわからない．本研究では，投稿とそのリプライといった関係の情報は用いないこと，炎上とは異なり，メッセージの読み手が感じる悪意を検出したいことから，人手でデータにラベル付与することでデータセットを構築する．

3.2　特徴量を追加する研究

特徴量の追加をすることで精度向上を目指す研究も現在までに複数行われている．

松本ら [6] は，SNS 上でのトラブルの一要因である煽り投稿の検出を目的とし，Twitter のリプライツイートを対象に，BERT を用いた煽りツイートの分類モデルを提案している．リプライツイート単体の分類と，リプライ元ツイートとのペアでの分類の 2 種類のモデルを作成し，従来の SVM などの分類比較をし，リプライツイートとその元ツイートとのペアを活用することで，分類精度向上が確認された．また，諏訪ら [7] は，日本語における皮肉文の検出を目的とし，絵文字の特徴として考慮する手法を提案した．BERT によるテキストの特徴ベクトルに加え，絵文字の分散表現を結合することで，分類精度向上が確認された．これらの研究では，追加の特徴量を用いることで精度向上を果たしているため，特徴量を追加することで精度が向上する可能性はある．しかし，着目している特徴が，リプライとその元ツイートという対話構造や絵文字の有無であり，有効である場面が限られる．したがって，本研究では，収集したデータのテキストから抽出可能な特徴を使用し，汎用的に活用できる特徴量を利用した精度向上を目指す．

3.3　感情分析関連の研究

感情分析を検出に活用する研究も行われている．

上野ら [8] は，ソーシャルメディアにおける嫉妬感情の検出を目的とし，嫉妬に基づく皮肉や誹謗中傷表現の検出を行った．日本語評価極性辞書を用いて，嫉妬に関連する単語を抽出し，パターンマッチング手法を適用した．その後，抽出したテキストについて手動で嫉妬感情の有無を判定し，嫉妬を伴う皮肉・誹謗中傷表現を特定する辞書を構築した．また，高橋ら [9] は，ツイートの感情極性や，感情強度の変化に着目し，炎上を検出する手法を提案している．検出精度はあまり良くないが，感情と非フォロワーという属性を条件に用いることで，炎上ツイートの検出ができることが確認された．これらの研究では，感情に着目することで炎上や誹謗中傷の検出を行っている．炎上や誹謗中傷などは悪意と関連性のある可能性があるため，感情に着目することで検出ができる可能性がある．しかし，感情分析から悪意

検出を行うアプローチの場合，悪意を含むメッセージは，多様な表現が存在することが考えられ，感情分析のみから検出するのは難しいと考えられる．したがって，本研究では，感情分析によって検出するのではなく，BERT を用いたモデルに対し，追加の特徴として書き手の感情を利用することで，精度向上を目指す．

第 4 章　提案手法

本研究では，悪意検出の精度向上に寄与する可能性のある特徴量としてテキストに含まれる感情情報をベクトルで表した感情スコアに着目し，その有効性を検証することを目的とする．悪意を含むかというラベルを付与した 5 ちゃんねるレスデータから，悪意検出モデルを作成する場合，悪意というものが多様な表現を含むため検出が難しい可能性が考えられる．そこで，感情スコアを追加の特徴量とすることで，モデルが感情的な側面の悪意を捉えることができ，精度向上することが期待される．本章では，使用するデータとモデル構築の詳細について述べる．なお，感情スコアについては 4.3 節で詳しく述べる．

我々が精度向上に寄与する可能性のある特徴量として感情スコアに着目した理由は以下のニつである．

一つ目は，悪意と感情には関連性があると考えたからである．悪意を含むメッセージには，怒り，憎しみ，嫉妬といったネガティブな感情が伴うことが考えられる．そのため，メッセージに含まれる感情情報を表す感情スコアは, 悪意検出を行う上で，重要な特徴量であると考えた．

二つ目は，悪意検出を補助することができると考えたからである．悪意を含むメッセージには，攻撃的な単語が使用されることがあり，これは悪意検出において重要な特徴の一つと考えられる．しかし，モデルの学習をする際，攻撃的な単語に強く依存してしまった場合，メッセージに含まれる悪意の特徴を正しく捉えきれず，誤検出につながる可能性がある．したがって，感情スコアを使用することで，感情的な側面の情報を捉えることができるようになり，多様な表現の悪意を捉えられるようになるといった，悪意検出の補助に役立つ可能性があると考えた．

4.1　使用データ

本研究で使用するデータは，インターネット上の電子掲示板 5 ちゃんねる*から収集したレスデータである．5 ちゃんねるには様々なトピックに基づくカテゴリが

*https://5ch.net/

存在し，匿名性もあるため，悪意を含むレスが多い可能性があると考えた．

収集したレスデータは，2024 年 10 月に立てられたスレッドの中から，1000 レスに到達しているものを対象とし，5 ちゃんねるに存在するカテゴリの 1 つであるニュース速報から，悪意のある投稿がされている可能性のあるスレッドとして，3 つのスレッドを選定した．選定したスレッドに含まれるレスをスクレイピングによって計 3000 件のレスデータの収集を行った．収集したレスデータには，手作業で悪意を含むか含まないかを示す $l_{malicious}$ を付与した．悪意を含む場合は $l_{malicious} = 1$，悪意を含まない場合は $l_{malicious} = 0$ を取る．レスデータ 3000 件に $l_{malicious}$ を付与した結果，3 つのスレッドで最も悪意を含むレスが少ないものが 240 件，多いものが 339 件だった．

本研究で使用するデータとして，スレッド特有のレスの傾向に過度に依存しないように，悪意を含むレスの最小数である 240 件に合わせ，各スレッドから悪意を含むレス 240 件，悪意を含まないレス 240 件をランダムに抽出した．最終的にレスデータそれぞれに $l_{malicious}$ が付与された合計 1440 件の 5 ちゃんねるレスデータを使用データとした．

4.2 WRIME：主観と客観の感情分析データセット

テキストの感情情報を得るための感情分類モデル構築のために，WRIME：主観と客観の感情分析データセット [10] を使用した．WRIME とは，日本語テキストにおける感情分析のために作成されたデータセットである．このデータセットには，43,200 件の日本語テキストが含まれており，それぞれに Plutchik の基本 8 感情である喜び，悲しみ，期待，驚き，怒り，恐れ，嫌悪，信頼に基づいた感情ラベルが付与されている. 感情ラベルは, 主観と客観の二つの視点で付与されており，感情の強度として，無なら 0，弱なら 1，中なら 2，強なら 3 で表されている．

この WRIME を使用し，テキストに対し，Plutchik の基本 8 感情それぞれが含まれる確率を 8 次元のベクトルとして付与できる感情分類モデルを作成する．本研究では，メッセージの書き手が読み手の悪意を感じる場合，悪意を含むとするため，感情についても客観の視点の感情ラベルを使用する．また，感情強度の低いデータに関しては，感情スコアの付与において悪影響になる可能性があるため，感情強度

が 2 以上のデータを対象として感情分類モデルを作成した.

4.3 追加の特徴量

追加の特徴量である感情スコアは以下の二つを使用する.

1. 8 感情スコア
   WRIME を用いて作成した感情分類モデルの出力であり，Plutchik の基本 8 感情それぞれの感情強度を 8 次元のベクトルとして表したものを使用する. なお, 各感情の強度は［0，1］の実数値をとる.
2. PNN スコア
   WRIME を用いて作成した感情分類モデルの出力である Plutchik の基本 8 感情のそれぞれの感情強度を, 喜びは e_{joy}, 悲しみは $e_{sadness}$, 期待は $e_{anticipation}$, 驚きは $e_{surprise}$, 怒りは e_{anger}, 恐れは e_{fear}, 嫌悪は $e_{disgust}$, 信頼は e_{trust} とする．この感情強度を, ポジティブの感情強度 $e_{positive}$, ネガティブの感情強度 $e_{negative}$, ニュートラルの感情強度 $e_{neutral}$ へと変換したものを使用する.
   変換は以下の通りである.

$$e_{positive} = e_{joy} + e_{anticipation} + e_{trust}$$

$$e_{negative} = e_{sadness} + e_{anger} + e_{fear} + e_{disgust}$$

$$e_{neutral} = e_{surprise}$$

4.4 モデル作成

本研究で提案するモデルでは，自然言語処理の分野で広く利用されている BERT をベースとした悪意検出モデルを構築する．BERT の事前学習モデルは，東北大学自然言語処理研究グループが提供する日本語学習モデル[†]を使用した.

[†] https://github.com/cl-tohoku/bert-japanese

扱えるトークンの最大長は 512 とし，これに満たないものは 0 で Padding 処理を行う．バッチ数は 16 とし，検証用データの損失が 5 回連続で低下しなかった場合，Earlystopping を実施する．特徴量の追加として三つの手法を採用する．

4.4.1 手法 1

手法 1 は，5 ちゃんねるのレスデータに対して付与された感情スコア $e_{emotion}$ を全結合層により悪意を含むか含まないかそれぞれに対するスコア $logits_{emotion,2D}$ に変換し，BERT のエンコーダから得られた［CLS］トークンの出力を全結合層を通じて得られる，悪意を含むか含まないかそれぞれに対するスコア $logits_{bert_1}$ に加算する手法である．この手法では，感情スコアが分類結果に直接的な影響を与えると考えられる．手法 1 のモデル図は図 4.1 に示す．

最終的な悪意を含むか含まないかそれぞれに対するスコア $logits_1$ を得る手順は以下の通りである．

1. 感情スコア $e_{emotion}$ の変換.
   感情スコア $e_{emotion}$ を全結合層を通じて，分類するクラスに対応した次元に変換した，悪意を含むか含まないかそれぞれに対するスコア $logits_{2D}$ を得る．本研究では，悪意を含むか含まないかの 2 値分類のため，2 次元に変換する．
   変換の式は以下の通りである．

$$logits_{2D} = W_{emotion\_1} \cdot e_{emotion} + b_{emotion\_1}$$

   ここで $W_{emotion\_1}$ は重み行列，$b_{emotion\_1}$ はバイアス項である．
2. BERT のエンコーダから得られた［CLS］トークンの出力を全結合層を通じて得られるスコア $logits_{bert\_1}$ への加算.
   レスデータを入力し，BERT のエンコーダから得られた［CLS］トークンの出力を全結合層を通じて求めたスコア $logits_{bert\_1}$ に，感情スコアを変換することで得たスコア $logits_{2D}$ を加算することで，最終的な悪意を含むか含まないかそれぞれに対するスコア $logits_1$ を得る．

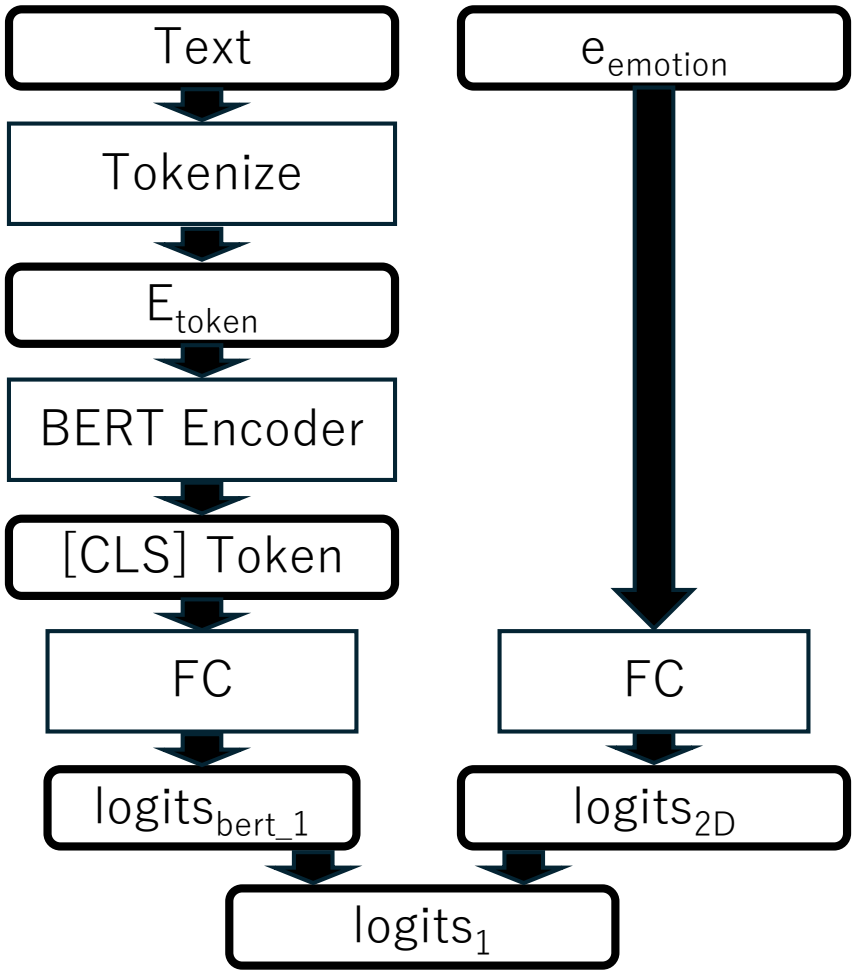

図 4.1　手法 1 のモデル図

加算の式は以下の通りである．

$$logits_1 = logits_{bert\_1} + logits_{2D}$$

4.4.2　手法 2

手法 2 は，5 ちゃんねるのレスデータに対して付与された感情スコア $e_{emotion}$ を，トークン化した入力データを BERT を通すことで得られる 768 次元の実数値のベクトルである埋め込み E_{token} へ結合する手法である．この手法では，感情ス

コアがモデル全体の学習に影響を与えると考えられる．手法 2 のモデル図は図 4.2 に示す．

最終的な悪意を含むか含まないかそれぞれに対するスコア $logits_2$ を得る手順は以下の通りである．

1. 感情スコア $e_{emotion}$ の変換．
   感情スコア $e_{emotion}$ を全結合層を通じて，BERT の埋め込み次元である 768 次元に変換した感情スコア E_{768D} を得る．
   変換の式は以下の通りである．

$$E_{768D} = W_{emotion\_2} \cdot e_{emotion} + b_{emotion\_2}$$

   ここで $W_{emotion\_2}$ は重み行列，$b_{emotion\_2}$ はバイアス項である．
2. 変換後の感情スコア E_{768D} の拡張．
   変換後の感情スコア E_{768D} を，レスデータをトークナイズした際のトークン数 seq_length に応じて拡張する．
   拡張は以下の通りである．

$$E_{768D,extended} = \begin{bmatrix} E_{768D} \\ E_{768D} \\ \vdots \\ E_{768D} \end{bmatrix} \in \mathbb{R}^{seq\_length \cdot 768}$$

3. BERT の埋め込み E_{token} への結合．
   拡張した後の感情スコア $E_{768D,extended}$ をトークン化した入力データを BERT を通すことで得られる BERT の埋め込み E_{token} と結合する．
   結合の式は以下の通りである．

$$E_{combined} = E_{token} + E_{768D,extended}$$

4. 最終的な悪意を含むか含まないか，それぞれに対するスコア $logits_2$ を得る．
   結合された埋め込み $E_{combined}$ を BERT のエンコーダに入力し，得られた［CLS］トークンを全結合層を通じて，最終的な悪意を含むか含まないか，それぞれに対するスコア $logit_2$ を得る．

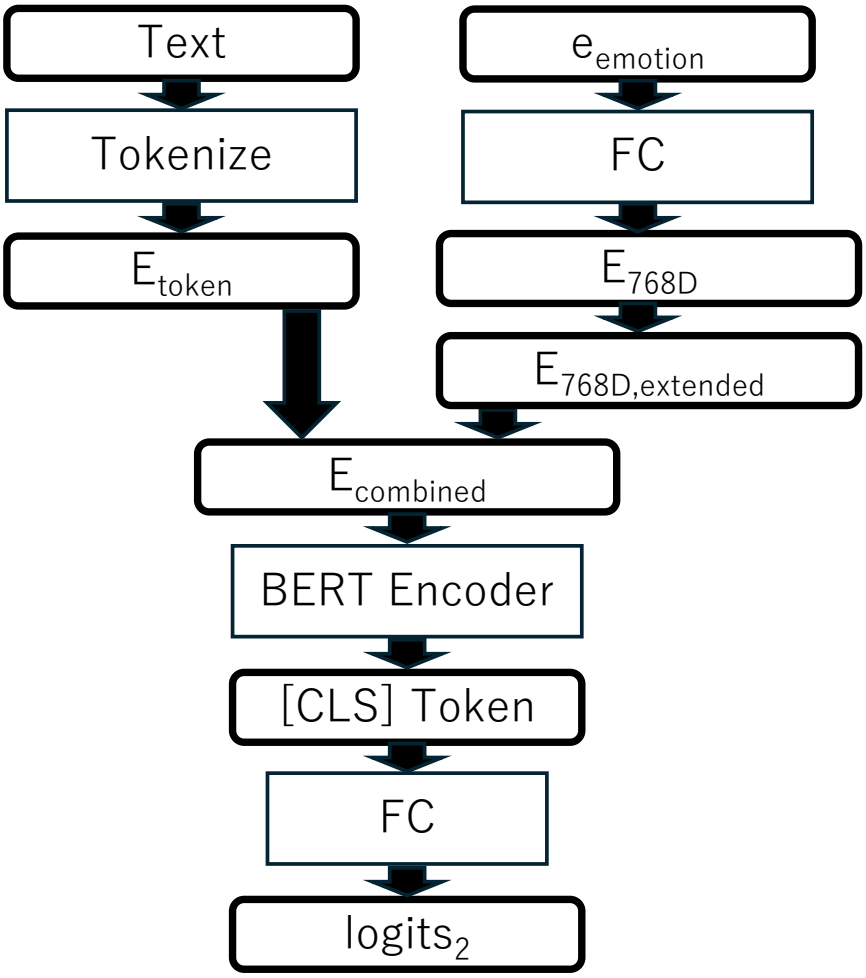

図 4.2　手法 2 のモデル図

4.4.3　手法 3

手法 3 は，手法 1 と手法 2 のどちらも使用する手法である．手法 1，手法 2 と異なり，感情スコアを二つの手法で使用するため，感情スコアの影響が大きくなる可能性がある．この影響を制御できるように，BERT の分類層の出力である悪意を含むか含まないかそれぞれに対するスコアへの加算をするためのパラメータとして α，トークン化した入力データを BERT を通すことで得られる 768 次元の実数値のベクトルである埋め込みに結合するためのパラメータとして β を導入し，寄与率を制御できるようにした．本研究では，$\alpha = \beta = 0.5$ とした．手法 3 のモデル図

は図 4.3 に示す.

最終的な悪意を含むか含まないかそれぞれに対するスコア $logits_3$ を得る手順は以下の通りである.

1. 感情スコア $e_{emotion}$ の変換
   感情スコア $e_{emotion}$ を全結合層を通じて，分類するクラスに対応した次元に変換した悪意を含むか含まないかそれぞれに対するスコア $logits_{2D}$ と BERT の埋め込み次元である 768 次元に変換し，拡張した感情スコア $E_{768D,extended}$ を得る.
2. BERT の埋め込み E_{token} への結合
   パラメータ β を活用して，トークン化した入力データを BERT を通すことで得られる BERT の埋め込み E_{token} と拡張した感情スコア $E_{768D,extended}$ を結合する.
   結合の式は以下の通りである.

$$E_{combined\_3} = E_{token} + \beta \cdot E_{768D,extended}$$

3. BERT のエンコーダから得られた［CLS］トークンの出力を全結合層を通じて得られるスコア $logits_{bert\_3}$ への加算
   パラメータ α を活用して，結合された埋め込み $E_{combined\_3}$ を BERT のエンコーダに入力し，得られた［CLS］トークンを全結合層を通じて得たスコア $logit_3$ と変換した感情スコア $logits_{2D}$ を加算し，最終的な悪意を含むか含まないかそれぞれに対するスコア $logits_3$ を得る.
   加算の式は以下の通りである.

$$logits_3 = logits_{bert\_3} + \alpha \cdot logits_{2D}$$

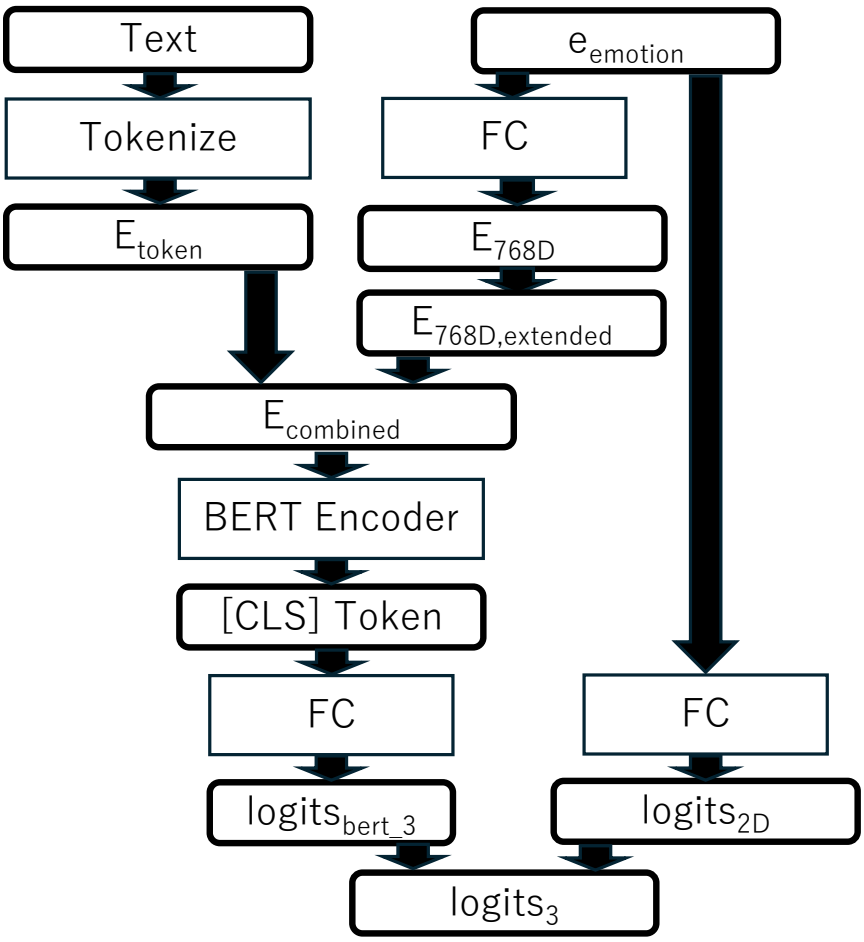

図 4.3　手法 3 のモデル図

第 5 章　評価実験

本研究での実験目的は，追加の特徴量として設定した 2 種類の感情スコアが精度向上に寄与するかを検証することである．本実験では，5 ちゃんねるのレスデータに悪意を含むか含まないかのラベルを付与したデータから作成した悪意検出モデルをベースラインとし，追加の特徴量として 2 種類の感情スコアを用いたモデルとの性能を比較する．

5.1　実験

5.1.1　実験手順

実験の手順は以下の通りである．

1. スクレイピングを用いて，5 ちゃんねるのレスデータを収集する．
2. レスデータに対して悪意のラベルを付与する．
3. BERT にレスデータを入力し，ラベルを予測し，ベースラインを作成する．
4. WRIME を用いて作成した感情分類モデルを利用し，レスデータに対し,2 種類の感情スコアを付与する．
5. 追加の特徴量を使用したモデルを作成する．
6. ベースラインと追加の特徴量を使用したモデルを比較する．

5.1.2　実験条件

本実験では，10 交差検証を用いた．データセットの割合は，訓練用：検証用：テスト用の順に 8:1:1 とし，すべてのデータがテストデータとして使用されるようにする．また，固有シードにより，データセットの分割が，どのモデルでも同じになるよう設定し，使われるデータの違いによる差が生まれないようにする．

実験で比較するモデルとそれに使用する特徴量は表 5.1 に示す．

表 5.1　モデル一覧と使用する特徴量

	手法 1 で追加した特徴量	手法 2 で追加した特徴量
ベースライン	なし	なし
モデル A-1	8 感情スコア	なし
モデル A-2	PNN スコア	なし
モデル B-1	なし	8 感情スコア
モデル B-2	なし	PNN スコア
モデル C-1	8 感情スコア	8 感情スコア
モデル C-2	PNN スコア	PNN スコア
モデル C-2-1	PNN スコア	8 感情スコア
モデル C-1-2	8 感情スコア	PNN スコア

5.2　結果・考察

5.2.1　手法 1 の結果・考察

ベースモデルの混同行列を表 5.2，モデル A-1 の混同行列を表 5.3，モデル A-2 の混同行列を表 5.4 に示す．

また，それぞれのモデルの Accuracy，Precision，Recall，F 値を表 5.5 に示す．

Accuracy を比較すると，ベースラインと比べ，モデル A-1 とモデル A-2 どちらも向上している．特に，モデル A-2 の方がより Accuracy が高くなっている．特徴量の追加が有意であるか t 検定を用いたところ，p 値は，モデル A-1 は 0.383，モデル A-2 は 0.067 だった．

結果の比較から，BERT の分類層の出力である悪意を含むか含まないか，それぞれに対するスコアへの加算による特徴量の追加において，8 感情スコアを使用するモデル A-1 は，特徴量の追加が有意であるといえるほどの改善は見られなかった．しかし，ベースラインに比べ，Recall が向上しており，8 感情スコアの使用により，悪意をより検出できるようになっている．わずかに Precision が低下はしたため，誤検出が若干増えているが，F 値の改善もみられるため，わずかに改善の可能性がある．

PNN スコアを使用するモデル A-2 は，特徴量の追加が有意であるといえるほどの改善は見られなかったが，8 感情スコアを使用したモデル A-1 に比べ，有意に近

表 5.2　ベースラインの混同行列

	pre 0	pre 1
label 0	516	204
label 1	255	465

く，性能の改善の可能性が見られた．ベースラインに比べ，Precision が向上しており，Recall もわずかではあるが向上している．よって，PNN スコアの使用により誤検出を減らすことにつながっている．本実験では，ポジティブ，ネガティブ，ニュートラルのスコアへの変換を行い，8 感情スコアに比べ，より簡易的な特徴量へと変更した．この変換が最適であるか検討する必要があるが，手法 1 の場合, 悪意に関連があるならば，よりシンプルな特徴量である方が，モデルの性能改善に影響を与える可能性があり，変換がより悪意に沿ったものにできれば，より改善する可能性もあると考えられる．

表 5.3　モデル A-1 の混同行列

	pre 0	pre 1
label 0	482	238
label 1	205	515

表 5.4　モデル A-2 の混同行列

	pre 0	pre 1
label 0	539	181
label 1	249	471

表 5.5　手法 1 の性能比較

	Accuracy	Precision	Recall	F 値
ベースライン	0.681	0.695	0.646	0.670
モデル A-1	0.692	0.684	0.715	0.699
モデル A-2	0.701	0.722	0.654	0.686

5.2.2　手法 2 の結果・考察

モデル B-1 の混同行列を表 5.6，モデル B-2 の混同行列を表 5.7 に示す．

また，それぞれのモデルの Accuracy，Precision，Recall，F 値を表 5.8 に示す．

Accuracy を比較すると，モデル B-1 は向上しているが，モデル B-2 は低下して

いる．特徴量の追加が有意であるか t 検定を用いたところ，p 値は，モデル B-1 は 0.472，モデル B-2 は 0.830 だった．

結果の比較から，トークン化した入力データを BERT を通すことで得られる 768 次元の実数値のベクトルである埋め込みへの結合による特徴量の追加において，8 感情スコアを使用したモデル B-1 は，特徴量の追加が有意であるといえるほどの改善は見られなかった．しかし，Accuracy，Precision，Recall，F 値の評価指標では，改善が見られる．特に Recall が向上しており，悪意のあるレスを見逃すことが減少している．そのため，わずかではあるが，モデルの性能改善に寄与する可能性があると考えられる．

PNN スコアを使用したモデル B-2 は，特徴量の追加が有意であるといえるほどの改善は見られなかった．また，Accuracy，Precision，Recall，F 値の評価指標を比較しても，ほぼ変化がないため，有効でないと考えられる．

8 感情スコアを使用したモデル B-1 に比べ，性能が改善しなかった原因として，二つ原因が考えられる．一つ目は，変換をする際に悪意検出において重要な感情の情報を失ってしまったことである．手法 2 の場合，感情スコアが学習にも影響を与えると考えられ，8 感情のような多くの感情の特徴の方が，情報量が多く，悪意を捉えられる可能性がある．そのため，PNN スコアへと変換するにあたって，悪意検出に有効な感情の情報を失ったことが考えられる．二つ目は，変換が悪意検出において適切でなかった可能性である．本研究では，8 感情をポジティブ, ネガティブ, ニュートラルの三つのどれに該当するかでより簡易的な感情スコアへと変換したが，手法 2 では，どの評価指標においても性能の改善は見られなかった．手法 1 でポジネガニュートラルスコアを使用したモデル A-2 では，有意であるとは言えないながらも，PNN スコアを使用することで，8 感情スコアとは違った性能改善の傾向が見られたが，手法 2 においては適切ではないと考えられる．

表 5.6　モデル B-1 の混同行列

	pre 0	pre 1
label 0	511	209
label 1	234	486

表 5.7　モデル B-2 の混同行列

	pre 0	pre 1
label 0	511	209
label 1	254	466

表 5.8　手法 2 の性能比較

	Accuracy	Precision	Recall	F 値
ベースライン	0.681	0.695	0.646	0.670
モデル B-1	0.692	0.699	0.675	0.687
モデル B-2	0.678	0.690	0.647	0.668

5.2.3　手法 3 の結果・考察

モデル C-1 の混同行列を表 5.9，モデル C-2 の混同行列を表 5.10，モデル C-2-1 の混同行列を表 5.11，モデル C-1-2 の混同行列を表 5.12 に示す．

また，それぞれのモデルの Accuracy，Precision，Recall，F 値を表 5.13 に示す．

Accuracy を比較すると，どのモデルもベースラインより Accuracy が向上していた．特徴量の追加が有意であるか t 検定を用いたところ，p 値は，モデル C-1 は 0.080，モデル C-2 は 0.508，モデル C-2-1 は 0.160，モデル C-1-2 は 0.867 であった．

結果の比較から，BERT の分類層の出力である悪意を含むか含まないかそれぞれに対するスコアへの加算と，トークン化した入力データを BERT を通すことで得られる 768 次元の実数値のベクトルである埋め込みへの結合の 2 つの手法を用いて，追加の特徴量を使用するモデルにおいても，特徴量の追加が有意であるといえるほどの改善は見られなかった．しかし，Accuracy，Precision，Recall，F 値の評価指標では改善が見られるモデルも存在しており，追加の特徴量が性能改善に寄与する可能性がある．

手法 1 と手法 2 の両方を用いて追加の特徴量を加える際，同じ特徴量を追加する場合，PNN スコアを使用したモデル C-2 より，8 感情スコアを使用したモデル C-1 の性能が高かった．また，モデル C-1 は，特徴量の追加が有意とは言えなかったが，有意に近い値を記録した．特徴量の追加として手法 1 か手法 2 のどちらか片方の追加方法を使用したモデルの結果と比較して，8 感情スコアは両方に適用することで，より高い性能のモデルを作成できた．理由としては，8 感情スコアはどちらの手法でも有効である可能性が考えられる．8 感情スコアを使用したモデル A-1 とモデル B-1 はどちらも Accuracy，F 値が向上しており，わずかに性能が改善しているため，特徴量の追加としてどちらの手段も有効であり，両方を使用すること

で，改善の効果を強めることにつながったことが考えられる．モデル C-2 は，モデル C-1 に比べ，性能が低かった．これは，PNN スコアは手法 1 で使用する場合は，モデルの性能向上の可能性があるが，手法 2 で使用する場合は，モデル性能はあまり変化がなく，効果が弱かったことが考えられる．

手法 1 と手法 2 の両方を用いて追加の特徴量を加える際，違う特徴量を使用する場合，手法 1 には PNN スコアを使用し，手法 2 には 8 感情スコアを使用したモデル C-2-1 の性能が高かった．理由としては，手法 1 に PNN スコアを使用することも，手法 2 に 8 感情スコアを使用することも，モデルの性能向上に影響する可能性があり，それぞれの効果が性能改善につながったと考えられる．

性能が比較的高かったモデル C-1 とモデル C-2-1 をベースラインと比較すると，Accuracy，Precision，Recall，F 値どれも向上しているのはモデル C-1 であり，有意とは言えないまでも，追加の特徴量として感情を用いることが有効である可能性がある．また，モデル C-2-1 に関しても，Precision は低下したが，Recall は大幅に向上していることもあり，見逃しを減らしたい場合は有効であることが考えられる．

表 5.9　モデル C-1 の混同行列

	pre 0	pre 1
label 0	517	203
label 1	222	498

表 5.10　モデル C-2 の混同行列

	pre 0	pre 1
label 0	503	217
label 1	233	487

表 5.11　モデル C-2-1 の混同行列

	pre 0	pre 1
label 0	515	205
label 1	252	468

表 5.12　モデル C-1-2 の混同行列

	pre 0	pre 1
label 0	469	251
label 1	179	541

表 5.13　手法 3 の性能比較

	Accuracy	Precision	Recall	F 値
ベースライン	0.681	0.695	0.646	0.670
モデル C-1	0.705	0.710	0.692	0.701
モデル C-2	0.688	0.692	0.676	0.684
モデル C-2-1	0.701	0.683	0.751	0.715
モデル C-1-2	0.683	0.695	0.650	0.672

第 6 章　おわりに

6.1　まとめ

本研究では，悪意検出における有効な特徴量の検討を行った．我々は，悪意というものは多様な表現を含んでいることから，テキストからの単純な検出は難しいと考えた．そこで，悪意検出の精度向上に寄与する特徴量を追加することで精度向上を目指した．

本研究では，精度向上に寄与する可能性のある特徴量として，読み手が感じる書き手の感情スコアの有効性について検討を行った．感情スコアに着目した理由は二つある．一つ目は，悪意を含む投稿には，ネガティブな感情が伴うことがあるため，感情と悪意には関連性があると考えたためである．二つ目は，悪意によって，多様な表現故に検出の難しい悪意を含む投稿に関しても検出の補助ができると考えたためである．

WRIME を用いて, 使用するデータである 5 ちゃんねるのレスデータに, Plutchik の基本 8 感情それぞれが含まれる確率を 8 次元のベクトルとしたものを付与し，8 次元のベクトルをそのまま使用する 8 感情スコアと，8 感情をポジティブ，ネガティブ，ニュートラルの 3 次元のベクトルへと変換した PNN スコアの 2 種類を感情スコアとして使用した．

実験として，BERT の分類層の出力である悪意を含むか含まないかそれぞれに対するスコアへの加算をする手法 1，トークン化した入力データを BERT を通すことで得られる 768 次元の実数値のベクトルである埋め込みへ結合する手法 2，両方を使用する手法 3 の三つを用いて，追加の特徴量の影響を検証した．

手法 1 では，ベースラインの Accuracy が 0.681 であるのに対し，8 感情スコアを使用したモデルでは 0.692，PNN スコアを使用したモデルでは 0.701 とどちらも向上がみられた．しかし，t 検定を行ったところ，PNN スコアを使用したモデルが有意に近かったものの，特徴量の追加が有意であるとは言えなかった．

手法 2 では，8 感情スコアを使用したモデルの Accuracy は 0.692 と向上したが，PNN スコアを使用したモデルの Accuracy は 0.678 と低下した．8 感情スコアは，Precision，Recal，F 値のどれも向上することにつながったが，t 検定を行ったとこ

ろ，有意であるとは言えなかった．

両方を使用する手法 3 では，8 感情スコアを追加の特徴量として使用したモデルの Accuracy が 0.705 と最も高い値を記録した．しかし，これに関しても t 検定で有意に近かったものの，有意とは言えなかった．

実験の結果，本研究で着目した感情スコアは，悪意検出の精度向上に有効とは言えなかった．

6.2 今後の展望

今後の展望を述べる．まずは，使用するデータの改善である．本研究では，5 ちゃんねるのレスデータに悪意を含むか含まないかのラベルを付与したデータを使用した．実験結果から今回着目した感情スコアは改善の可能性はあるものの，有意とは言えなかった．しかし，データ量が十分とは言えないため，より多くのデータを用いることで特徴量の有効性を判断していく必要がある．また，本研究では，ニュース速報のカテゴリから 3 つのスレッドを選定し，収集したレスデータを用いたが，カテゴリやスレッドの数についても十分とは言えないため，より多様な表現のデータを収集する必要がある．次に，特徴量の追加手法の検討である．本研究では，BERT の分類層の出力である悪意を含むか含まないかそれぞれに対するスコアへの加算をする手法 1 とトークン化した入力データを BERT を通すことで得られる 768 次元の実数値のベクトルである埋め込みへ結合する手法 2 を使用したが，手法ごとで同じ特徴量を使用していても影響が異なっていた．また，最も性能が高かったモデルは 8 感情スコアを使用し，両方の手法を適用したモデルだったが，それぞれの手法の寄与率を 0.5 として設定していたが，これを変更することも可能なため，特徴量の追加方法についても検討の必要がある．最後に，別の特徴量の検討である．本研究では，感情スコアに着目した．結果として有意とは言えなかったが，各評価指標の変化から追加の特徴量が改善につながる可能性がある．そのため，着目した感情スコアと併せて使用することで精度向上するような特徴量がないかの検討をする必要がある．

謝辞

ご指導いただいた鈴木先生には，大変お世話になりました．研究に論文執筆と慣れないことが多い中，時に厳しく，時に優しく，お忙しい中でも時間を取っていただき，深く感謝したします．ご迷惑，ご心配をおかけする機会も多々あったかと思いますが，鈴木先生のおかげで成長することができました．本当にありがとうございます．まだまだ至らぬところばかりの私ですが，大学院でも，どうぞよろしくお願いいたします．

事務補佐員の井尾さんには，事務作業や各種申請の際の書類等，大変お世話になりました．書類関連の不明点からミスや見逃した項目についても，いつも優しく対応していただき，助かりました．本当にありがとうございます．

鈴木研究室の先輩方には，研究や論文に関して，アドバイスや相談に乗っていただきました．困っている時期にも，優しい言葉や態度で支えていただき，本当に心強かったです．私の様子や状況に応じての対応，お心遣いいただき感謝申し上げます．また，先輩方の研究は，非常に良い勉強になりました．私も先輩方のようになれるようにと，今後も頑張っていきたいと思います．研究以外でも，ご飯や雑談，ゲームなどに付き合っていただき，楽しい時間も過ごすことができました．前向きに頑張れたのは先輩方のおかげです．本当にありがとうございます．

同期の田中君，将豪君，尾関さんには，いろんな面でお世話になり，本当に感謝しています．研究や大学院入試など，多くの壁に向かい合う機会がありましたが，今までやってこれたのは同期の皆さんの存在が大きかったと思います．本当にありがとうございます．時には学，時には涼，もはや鉄板ともなりつつある田中君の誤りシリーズには，本当に元気をもらいました．本当に．これからも，同期で協力して頑張っていきましょう．

新しく研究室に来た後輩の皆さん，これからよろしくお願いします．皆さんの存在が自身を先輩であると自覚させてくれ，研究にも熱が入ります．まだまだ頼りないですが，頼れる先輩になれるように頑張ります．

家族や友人には，身近であるが故に非常にお世話になりました．楽しい時間をたくさん過ごしたこともそうですが，つらいときや大変な時のさりげない気遣いや優しさが本当に嬉しかったです．

改めて，論文や研究，娯楽や雑談，その他これまでの人生でかかわっていただいたすべての皆様のおかげでここまで来れました．皆さんの支えがあってこそ，今の私がいます．心より感謝を申し上げます．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] R. PLUTCHIK. A general psychoevolutionary theory of emotion. *Emotion : Theory, research, and experience*, pp. 3–33, 1980.

[3] 西原陽子, 岩佐一樹, 福本淳一, 山西良典. 電子掲示板からの文脈を考慮した誹謗中傷コメントの抽出. 人工知能学会全国大会論文集, Vol. JSAI2014, pp. 1H4NFC01a3–1H4NFC01a3, 2014.

[4] 石坂達也, 山本和英. Web 上の誹謗中傷を表す文の自動検出. 言語処理学会 第 17 回年次大会 発表論文集 (NLP2011). 言語処理学会, 2011.

[5] 山崎慶朋. 多様な表現を含む攻撃的テキストの自動検出. 先端科学技術研究科修士論文, 03 2024. Supervisor: 白井 清昭.

[6] 松本典久, 上野史, 太田学. Bert を利用した煽りツイート検出の一手法. 第 13 回データ工学と情報マネジメントに関するフォーラム (DEIM2021) 論文集, I14-2, 3 2021.

[7] 諏訪光輔, 張建偉. Bert 及び絵文字を利用した日本語文における皮肉の検出. 第 13 回データ工学と情報マネジメントに関するフォーラム (DEIM2021). 電子情報通信学会 データ工学研究専門委員会, 2021.

[8] 貴弘上野, 慎太郎森, 正良大橋. ソーシャルメディアにおける嫉妬感情の検出の一検討. 第 81 回全国大会講演論文集, Vol. 2019, No. 1, pp. 129–130, 02 2019.

[9] 直樹高橋, 泰彦檜垣. Twitter における感情分析を用いた炎上の検出と分析. 電子情報通信学会技術研究報告 = IEICE technical report : 信学技報, Vol. 116, No. 488, pp. 135–140, 03 2017.

[10] 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, 2021.

発表リスト

[1] 中村輝，鈴木優 『悪意検出を目標とした表現分析』 東海関西データベースワークショップ 2024，2024.