修士論文

誹謗中傷検出のためのモデル学習を補助するためのサブタスクの提案

沢田 凌一

2024 年 12 月 4 日

岐阜大学大学院 自然科学技術研究科 知能理工学専攻 知能情報学領域
鈴木研究室

本論文は岐阜大学大学院自然科学技術研究科
修士（工学）授与の要件として提出した修士論文である.

沢田 凌一

指導教員：
鈴木 優 准教授

誹謗中傷検出のためのモデル学習を補助するためのサブタスクの提案*

沢田 凌一

内容梗概

本稿では，二つの研究について述べる．二つの研究は，どちらも誹謗中傷検出タスクにマルチタスク学習を採用する際に，加えるサブタスクの選定・調査を行う研究である．二つの研究に共通する目的は，誹謗中傷検出タスクにマルチタスク学習を採用する際，精度向上に有効なサブタスクを明らかにすることである．研究 A は，サブタスクとして，字面と感情の二つの観点に着目した研究である．研究 A は，サブタスクの選定に加え，重み付き損失関数を用いた学習改善手法の提案を行っている．研究 B は，サブタスクとして，主観と客観の二つの観点に着目した研究である．研究 B は，サブタスクの選定に加え，構築したモデルがどのような誹謗中傷投稿に対して，検出精度向上に有効であったか分析を行っている．我々は，二つの研究について別々に，議論を展開する．二つの研究から我々は，五つの知見を得た．(1) ランダムタスクを加えたマルチタスク学習モデルよりも，怒りの感情検出タスクを加えたマルチタスク学習モデルの方が誹謗中傷検出精度向上に有効．(2) サブタスクのラベルに約 20 倍以上の不均衡がある場合に限り，重み付き損失関数はサブタスクの検出精度向上に有効．(3) 主観サブタスクを加えたマルチタスク学習モデルは，客観サブタスクを加えたマルチタスク学習モデルに比べ，誹謗中傷検出タスクの F 値向上に有効な傾向がある．(4) 閲覧者感情分類タスクを加えたマルチタスク学習モデルは，誹謗中傷検出タスクの Precision 向上に有効．(5) 閲覧者感情分類タスクを加えたマルチタスク学習モデルはシングルタスク学習モデルに比べ，投稿者が喜びの感情を抱いた投稿に対する誹謗中傷検出タスクの Accuracy 向上に有効．

*岐阜大学大学院 自然科学技術研究科 知能理工学専攻 知能情報学領域 修士論文, 学籍番号:1224525045 1224525045, 2024 年 12 月 4 日.

キーワード

誹謗中傷，BERT，マルチタスク学習，ソーシャルメディア，機械学習

目次

図目次

表目次

第 1 章　はじめに

本稿では，二つの研究について述べる．二つの研究の目的は，どちらも誹謗中傷検出タスクにマルチタスク学習を採用した際，精度向上に有効なサブタスクを明らかにすることである．

研究 A は，サブタスクとして字面表現を検出するタスクと，投稿者感情を検出するタスクが有効ではないかと仮定し，調査を行った研究である．字面表現検出タスクは，脅迫表現検出タスク，差別表現検出タスク，容姿否定表現検出タスクの三つである．投稿者感情を検出するタスクは，正義感検出タスク，怒り感情検出タスク，失望感情検出タスクの三つである．研究 A では，サブタスクの選定に加え，サブタスクのデータ不均衡問題に対して，重み付き損失関数を用いた学習改善手法の提案を行った．

研究 B は，サブタスクとして主観サブタスクが有効ではないかと仮定し，調査を行った研究である．主観サブタスクは，投稿者感情分類タスクと閲覧者感情分類タスクの二つである．比較対象にシングルタスク学習モデル，客観サブタスクを加えたマルチタスク学習モデルを用意し，主観サブタスクが主タスクの検出精度向上に有効かどうかの検証を行った．客観サブタスクは，投稿対象分類タスクと誹謗中傷カテゴリ分類タスクの二つである．研究 B では，サブタスクの選定に加え，構築したモデルがどのような誹謗中傷投稿に対して，検出精度向上に有効であったか分析を行った．

我々は，二つの研究について別々に，議論を展開する．研究 A について 3 章，研究 B について 4 章で述べる．二つの研究から我々は，五つの知見を得た．

研究 A: ランダムタスクを加えたマルチタスク学習モデルよりも，怒りの感情検出タスクを加えたマルチタスク学習モデルの方が誹謗中傷検出精度向上に有効．

研究 A: サブタスクのラベルに約 20 倍以上の不均衡がある場合に限り，重み付き損失関数はサブタスクの検出精度向上に有効．

研究 B: 主観サブタスクを加えたマルチタスク学習モデルは，客観サブタスクを加えたマルチタスク学習モデルに比べ，誹謗中傷検出タスクの F 値向上に有効な傾向がある．

研究 B: 閲覧者感情分類タスクを加えたマルチタスク学習モデルは，誹謗中傷検出タスクの Precision 向上に有効.

研究 B: 閲覧者感情分類タスクを加えたマルチタスク学習モデルはシングルタスク学習モデルに比べ，投稿者が喜びの感情を抱いた投稿に対する誹謗中傷検出タスクの Accuracy 向上に有効. .

第 2 章　基本的事項

2.1　評価指標

評価指標とは，機械学習モデルの予測精度を評価するための指標である．評価指標は，評価の目的に応じた様々な種類が存在する．本章では，評価指標として，Accuracy，Precision，Recall，F 値について説明する．それぞれの評価指標は 2 値分類を例に説明する．評価指標を説明するために，モデルの予測と正解ラベルの関係を表した混同行列を図 2.1 に記す．図 2.1 についての説明を以下に記す．

TN: True Negative の略，モデルが Negative と予測し，実際に Negative であったテストデータの総数

FN: False Negative の略，モデルが Negative と予測し，実際には Positive であったテストデータの総数

FP: False Positive の略，モデルが Positive と予測し，実際には Negative であったテストデータの総数

TP: True Positive の略，モデルが Positive と予測し，実際には Positive であったテストデータの総数

図 2.1 の混同行列を用いて各評価指標について説明する．

2.1.1　Accuracy

Accuracy とは，モデルの予測が正解した割合を示す評価指標である．Accuracy の算出式を式 2.1.1 に示す．

$$Accuracy = \frac{TN + TP}{TN + FN + FP + TP} \tag{2.1.1}$$

2.1.2　Precision

Precision とは，モデルの適合率を示す評価指標である．適合率とは，モデルが Positive 予測したテストデータのうち，テストデータが実際に Positive であった割

	モデルの予測		
モデルの正解ラベル		Negative	Positive
	Negative	TN	FP
	Positive	FN	TP

図 2.1　提案モデル図

合である．モデルの誤検知の少なさを測る指標と考えることができる．Precision の算出式を式 2.1.2 に示す．

$$Precision = \frac{TP}{FP + TP} \tag{2.1.2}$$

2.1.3　Recall

Recall とは, モデルの再現率を示す評価指標である．再現率とは, 実際に Positive であるテストデータのうち，正しく Positive と予測できたテストデータの割合である．再現率は，True Positive Rate とも呼ぶことができる．再現率は，モデルが陽性データをどれだけ見逃すことがなく，予測することができるかを測る指標と考えることができる．Recall の算出式を式 2.1.3 に示す．

$$Recall = \frac{TP}{FN + TP} \tag{2.1.3}$$

2.1.4 F 値

F 値とは，Precision と Recall の調和平均を表した評価指標である．F 値は調和平均であるため，Precision と Recall のどちらか一方が極端に低く，もう一方が極端に高い場合よりも，Precision と Recall の両方が程よく高い方が高くなる．そのため，F 値は Precision と Recall 両方を重視したい場合に用いることが多い．F 値の算出式を式 2.1.4 に示す．

$$F = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.1.4}$$

2.2 k 分割交差検証

k 分割交差検証とは，機械学習モデルの予測精度を評価するための手法である．本研究では，10 分割交差検証を使用しているため，$k = 10$ の場合について説明する．本研究における 10 分割交差検証手順を以下に示す．

1. データセットを 10 個に分割
2. 10 個に分割したデータセットのうち，8 割を訓練データ，他の 1 割を検証データ，他の 1 割をテストデータとする
3. 訓練データで学習，検証データを用いて Early-stopping を行い，テストデータで評価指標を算出
4. 2,3 を，1 割のテストデータが 10 回異なるように繰り返す
5. 繰り返した 10 回の評価指標の平均をとり，機械学習モデルの予測精度を評価

上記の手順で検証を行うことで，正確に機械学習モデルの汎化性能を検証することができる．

2.3 対応のある 2 標本 t 検定

対応のある 2 標本 t 検定とは，対応のある 2 群の平均値に差があるかどうかの検定である．“対応のある” とは，同じ標本からデータを抽出したということである．そのため，同じデータセットを用いて，既存手法と提案手法の比較を行う際は，“対

応のある” となる．対応のある 2 標本 t 検定では，“2 群の平均値に差はない” と帰無仮説を立て，その帰無仮説が棄却された場合，“2 群の平均値に差はない” とはいえないとなる．帰無仮説が棄却できるかできないかは，検定統計量が有意水準を下回るかどうかによって決まる．有意水準とは，帰無仮説が真であるのにもかかわらず、帰無仮説を偽として棄却してしまう誤りが発生する確率のことである．有意水準に明確な指定はないが，0.01 や 0.05 を用いることが一般的である．検定統計量 t を表す式を式 2.3.1 に示す．

$$t = \frac{\bar{d} - \mu}{\sqrt{\frac{s^2}{n}}} \tag{2.3.1}$$

式 2.3.1 中の変数について説明する．$\bar{d}$ は，2 標本の差の平均，μ は差の母平均，s^2 は不偏分散，n はデータ数を表している．本研究では，帰無仮説を “2 群の平均値に差はない” としているため，μ は 0 である．s^2 の不偏分散は，母分散が不明なときに用いる．s^2 を表す式を式 2.3.2 に示す．

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.3.2}$$

式 2.3.2 中の変数について説明する．x_i は，i 番目のデータ，$\bar{x}$ はデータの平均を表す．

2.4 マルチタスク学習

マルチタスク学習とは，一つのモデルで複数のタスクを同時に解く手法である．複数のタスクを同時に解くことによって，モデルが一つのタスクに固執しない特徴を獲得し，汎化性能の向上や過学習の抑制が期待できる．複数のタスクを同時に解くためには，同時に学習する複数のタスクを同じモデルで学習し，タスクを解くときのみ分岐させる必要がある．マルチタスク学習のイメージ図を図 2.2 に示す．

図 2.2 のように，モデルを構築することによって，複数のタスクを同時に解くことができる．

マルチタスク学習を行う際には，損失を合計して逆伝播する．逆伝播する損失を

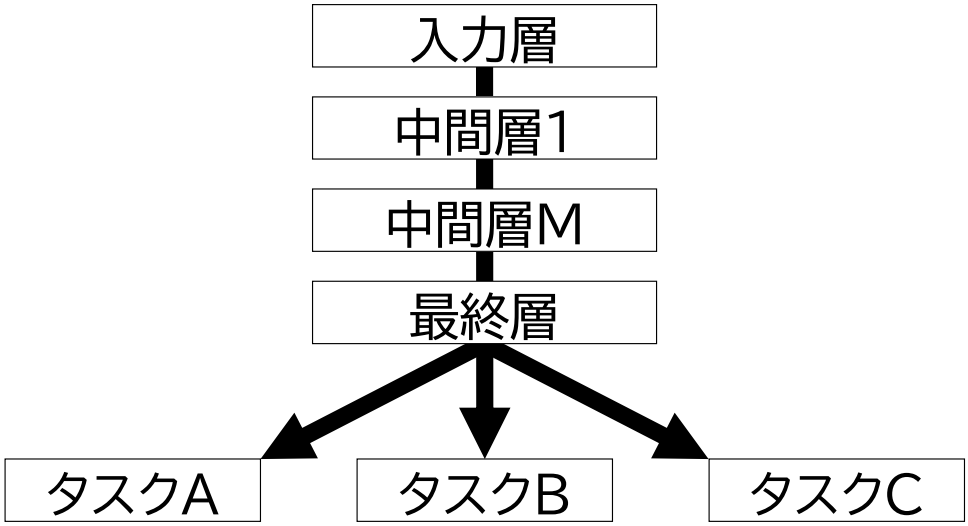

図 2.2　マルチタスク学習のイメージ図

式 2.4.1 に表す.

$$BackLoss = \sum_{i=1}^{N} Loss_i \tag{2.4.1}$$

$Backloss$ は逆伝播する損失，N は同時に解くタスクの数，$Loss_i$ は i 番目のタスクの損失である．式 2.4.1 のように，損失を合計して逆伝播することによって，複数タスクの学習を実現している．

2.5　BERT

BERT(Bidirectional Encoder Representations from Transformers) は，2018 年に登場した自然言語処理モデルである．BERT は，双方向性をもち，Transformer の Encoder を用いたモデルである．双方向性をもつとは，文の前後を理解できるということである．Transformer[1] とは，Attention 機構を用いた Encoder-Decoder モデルである．Attention とは，モデルが文を理解するためにどの単語を理解する必要があるか，単語の重要度を測るスコアである．Encoder とは，入力されたデー

タを機械学習モデルが理解できる特徴に変換するモデルである．Decoder は，機械学習モデルが理解できる特徴から出力の形に変換するモデルである．自然言語処理モデルを例として，Encoder-Decoder モデルを説明すると，入力テキストを機械学習モデルが理解できる特徴に変換するモデルが Encoder，特徴からテキストの形に変換するモデルが Decoder である．

BERT は，事前学習を行ったあとファインチューニングを行うという手順で使用されることが多い．事前学習は，自然言語を機械学習モデルが理解できるように学習することである．BERT の事前学習は，MLM(Masked Language Model) と NSP(Next Sentence Predicition) の 2 タスクに分かれて行われる．ファインチューニングは，事前学習で得たパラメータを初期値とし，実際のタスクを解くことができるようにモデルを微調整することである．事前学習の 2 タスクについて説明する．

2.5.1 MLM

MLM は，入力文中のトークンに何が入るか当てる穴埋め問題を解くタスクである．入力の全トークンのうち，トークンの 15 %をランダムに置換 (15 %のうち，80 %を [MASK]，10 %を別のトークン，10 %をそのまま) することによって，穴埋め問題タスクを作成する．MLM の例を説明する．

文 1 私は週末に友達とキャッチボールをする。(元の文)
文 2 私は週末に [MASK]] とキャッチボールをする。([MASK] への置換)

文 1 を文 2 に置き換え [MASK] に入るトークンを予測する．MLM は，置換された単語を周囲の単語から予測することによって，双方向の文脈理解を可能にしている．

2.5.2 NSP

NSP は，二つの文が連続するか否かを解くタスクである．NSP の例を説明する．

文 1 私は週末に友達とキャッチボールをする。[SEP] 私たちは毎週のようにキャッチボールしている。(この文は連続と予測)

文 2 私は週末に友達とキャッチボールをする。[SEP] 私は給食を食べる。(この文は連続ではないと予測)

NLP は，文が連続するかどうか解くことによって，文と他の文の関係理解を可能にしている．

第 3 章　研究__A(サブタスクを字面と感情の二つの観点から着目した研究)

内容梗概 A

誹謗中傷を検出する研究は行われているが，実用にはさらに精度を向上させる必要がある．機械学習の手法の一つであるマルチタスク学習は，一つのモデルで複数のタスクを解くことによって汎化性能を向上させ，精度向上が期待できる．誹謗中傷検出タスクとマルチタスク学習を組み合わせた研究は既に行われているが，どのようなサブタスクが誹謗中傷検出精度向上に有効であるかは明らかではない．そこで本研究では，誹謗中傷検出タスクの精度を向上させるサブタスクの提案を行う．また，提案したサブタスクが誹謗中傷検出タスクの精度向上に有効であるかどうかの検証を目的に実験を行った．実験の結果，怒りの感情検出タスクを誹謗中傷検出タスクと同時に解くことは，ランダムタスクを同時に解くことより有効であることが確認された．また，サブタスクのデータが不均衡であり F 値が非常に低い場合に限り，重み付き損失関数は同時に解く検出タスクの F 値向上に有効であることが確認された．

3.1　はじめに__A

SNS 上における人権侵害の増加 [2] といった，誹謗中傷が社会問題となっている．大量の投稿の中から誹謗中傷だけを人手で検出するためには，手間と費用がかかるため，誹謗中傷が投稿された後に誹謗中傷の投稿を，自動で検出する必要がある．現在でも誹謗中傷検出に関する様々な研究が行われているが [3][4][5] 現状では実用的な精度が出ておらず，さらに精度を向上させる必要がある．

本稿では，双方向エンコーダモデルである BERT にマルチタスク学習を採用する場合に有効なサブタスクの提案を行う．マルチタスク学習とは，一つの機械学習モデルで複数のタスクを解くことができる手法である．精度向上を期待する主タスクと，主タスクと関連のあるサブタスクを同時に解くことによって，モデルの汎化性能向上が期待できる．

BERT とマルチタスク学習を組み合わせた研究は行われているが，誹謗中傷検出精度向上のために有効なサブタスクを提案している研究は著者が調べた限りでは存在しない．そこで本研究では，誹謗中傷検出タスクに関連のあるサブタスクとして，字面表現を検出するタスク三つと投稿者感情を検出するタスク三つを提案する．字面表現検出するタスク三つには，脅迫表現検出タスク，差別表現検出タスク，容姿否定表現検出タスクを提案した．投稿者感情を検出するタスク三つには，正義感検出タスク，怒りの感情検出タスク，失望感情検出タスクを提案した．

また，サブタスクのデータ不均衡を改善するために重み付き損失関数を採用した．一般的にデータが不均衡である場合，データの複製を行うなどの対策が取られる．しかし，マルチタスク学習では一つのモデルで複数のタスクを解くため，一つのタスクに合わせたデータの複製はできない．そのため，サブタスクを学習する際の損失関数に重み付き損失関数を取り入れた．重み付き損失関数を取り入れることによって，データ数の少ないラベルの予測を誤った場合，大きなペナルティを科すことができ，不均衡データの学習を改善できると考えた．

提案したサブタスクとサブタスクへの重み付き損失関数の適用が，誹謗中傷検出精度向上に有効かどうか検証するために評価実験を行った．評価実験を行うために，主タスクとサブタスクに関する質問を行うことによって，アンケートをとりデータセットを用意した．主タスクである誹謗中傷かどうかの判断基準は，文章が第三者から見て故意であるかに関係なく攻撃性のある表現を含むかどうかとした．

評価実験では，シングルタスク学習モデル，提案したサブタスクを加えたマルチタスク学習モデル，提案したサブタスクに重み付き損失関数を用いて加えたマルチタスク学習モデル，ランダムタスクをサブタスクとして加えたマルチタスク学習モデルを構築し，結果を比較した．評価実験の結果から以下の知見が得られた．

1. ランダムタスクを加えたマルチタスク学習モデルよりも，怒りの感情検出タスクを加えたマルチタスク学習モデルの方が誹謗中傷検出精度向上に有効．
2. サブタスクのラベルに約 20 倍以上の不均衡がある場合に限り，重み付き損失関数はサブタスクの検出精度向上に有効．

3.2 関連研究__A

本章では，本研究の関連研究について述べる．誹謗中傷やヘイトスピーチの検出に関する研究は広く進められている．誹謗中傷検出を行う研究では，単語に着目して文章をベクトル化する手法と，分散表現を用いて文章をベクトル化する手法がある．単語に着目してベクトル化する手法として，Waseem ら [4] の研究がある．Waseem ら [4] は，ヘイトスピーチを検出するために n-gram とロジスティック回帰を用いて分類を行った．その結果，追加特徴として性別と場所に関する情報を追加することは有効であるが，その他の追加特徴は検出精度向上に不利であることを示している．分散表現を用いてベクトル化する手法として Badjatiya ら [3] の研究がある．Badjatiya ら [3] は，複数の深層学習アーキテクチャを用いてヘイトスピーチを検出しており，n-gram や TF-IDF などの既存手法よりも検出精度が向上したことを示している．本研究は，BERT で得た分散表現を用いてベクトル化を行い，誹謗中傷の検出を行う．

本研究は，BERT とマルチタスク学習を組み合わせたモデルを使用して誹謗中傷検出を行う．誹謗中傷検出に BERT とマルチタスク学習を組み合わせた研究として，Samghabadi ら [6] の研究がある．Samghabadi ら [6] は，攻撃性識別タスクと女性差別的攻撃検出タスクの二つのタスクを同時に解くマルチタスク学習モデルを提案している．この研究は，攻撃性のある文章を識別するという点が我々の研究と類似している．しかし，本研究は誹謗中傷の検出精度を向上させるためのサブタスクを提案することが目的であるため，二つのタスクを同列に重視し両方の精度を向上させようとしている Samghabadi らの研究とは目的が異なっている．また，本研究ではサブタスクのデータの不均衡を考慮し，重み付き損失関数を用いているという点でも異なっている．

不均衡データの改善の研究として Tang ら [7] の研究がある．Tang ら [7] は，不均衡データの改善手法として二つの手法を提案している．一つは，データ数が少ないラベルの文章をを英語から中国語に翻訳し英語に再翻訳することでデータを増強する手法である．もう一つは，undersampling を複数回行いサンプリングされたデータごとに分類器を構築しアンサンブル学習を行う手法である．彼らはマルチラベル感情分析において自身らの手法が既存の手法より優れていることを示している．この研究から，不均衡データの改善は分類精度を向上させることが期待でき

る．本研究では，マルチタスク学習を用いるためデータの複製やサンプリングを行うことはできない．そのため，重み付き損失関数に取り入れることで不均衡データの改善を行った．

3.3 提案手法_A

本研究では，誹謗中傷検出タスクの精度向上手法としてマルチタスク学習に着目した．シングルタスク学習は一つのモデルで一つのタスクを解く学習であるが，マルチタスク学習は，一つのモデルで複数のタスクを同時に解く学習であり，汎化性能を向上させることが期待できる．我々は，マルチタスク学習を用いて誹謗中傷検出タスクの精度向上を図る際に有効なサブタスクを提案することを目的として，サブタスクの選定を行った．また，サブタスクのデータが不均衡だった場合の改善手法として，重み付き交差エントロピー誤差を採用した．これによってモデルの学習が改善され，サブタスクと主タスクの検出精度が向上するかどうか検証した．本章では，提案するサブタスクの選定と不均衡データの学習改善手法について述べる．

3.3.1 シングルタスク学習

シングルタスク学習とは，一つのモデルで一つのタスクを解く学習である．本研究におけるシングルタスク学習は，誹謗中傷検出タスクのみを解くモデルを学習させることである．東北大学の乾・鈴木研究室が公開している事前学習済みの BERT モデル*を使用して，誹謗中傷検出タスクを解くモデルを構築した．本研究では，シングルタスク学習モデルをベースラインとして提案モデルと比較し，提案手法の有効性を検証する．

3.3.2 マルチタスク学習

マルチタスク学習 [8] とは，一つのモデルで複数のタスクを同時に解く手法である．マルチタスク学習では，主タスクと関連のあるタスクを同時に解くことによっ

*https://github.com/cl-tohoku/bert-japanese

て，モデルの汎化性能を向上させることができると言われている．そのため，マルチタスク学習においてサブタスクにどのようなタスクを選定するかは，主タスクの精度向上を図る上で非常に重要なことである．本研究で選定したサブタスクについて 3.2.1 節で述べる．

不均衡データを扱う場合，データをバランスよく学習できるようにデータの偏りをなくす手法が用いられる．一般的には，データ数の多いラベルに合わせてその他のラベルのデータを複製する oversampling や，データ数の少ないラベルに合わせてその他のラベルのデータを削減する undersampling が用いられる．しかし，マルチタスク学習では複数のタスクを同時に解くため，あるタスクのデータの偏りをなくすためにデータ数を増減させても，別のタスクのデータに偏りが生まれてしまう．そのため，マルチタスク学習ではデータ数を増減させることなくデータの不均衡を改善する手法が必要である．本研究で用いたデータの不均衡改善手法を 3.2.2 節で述べる．

サブタスクの選定

サブタスクには，字面表現を検出するタスク三つと投稿者感情を検出するタスク三つが有効であると考えた．字面表現を検出するタスクを選んだ理由として，誹謗中傷には様々な表現が存在し，この表現が誹謗中傷を検出する上で重要な特徴になることがあげられる．誹謗中傷を検出する上で，字面表現が重要な特徴になると考えた例を示す．

例 1 意外に身長低いね
例 2 身長低すぎだろ笑笑

例 1 も例 2 も “身長が低い” ということを相手に伝える投稿である．例 1 は身長が意外に低くいことに驚いている表現であり攻撃性は低いと考えられる．しかし，例 2 は身長が低いことを嘲笑する表現，つまり明らかに相手の容姿を否定するような表現であり攻撃性が高いと考えられる．この例から，表現の違いは誹謗中傷を検出する上で重要な特徴であり，サブタスクとして字面表現を検出するタスクが有効だと考えた．

投稿者感情を検出するタスクを選んだ理由として，怒りや正義感などの投稿者感

情は誹謗中傷の起因となるものであり，誹謗中傷と密接に関係していることがあげられる．実際に，炎上事例に批判的なコメントをした投稿者の多くが正義感を抱いていたという調査 [9] が存在する．誹謗中傷を検出する上で投稿者感情が重要な特徴になると考えた例を示す．怒りの感情を持った投稿者が使用する “いい加減にしろ” は，何かを全否定する攻撃性の高い言葉に感じると考えられる．しかし，怒りの感情を持っていない投稿者が使用するお笑いのツッコミのような “いい加減にしろ” は，攻撃性の低い言葉と考えられる．以上より我々は，字面表現を検出するタスク三つと投稿者感情を検出するタスク三つ計六つのサブタスクを選定した．

サブタスク 1 脅迫表現検出タスク
サブタスク 2 差別表現検出タスク
サブタスク 3 容姿否定表現検出タスク
サブタスク 4 正義感検出タスク
サブタスク 5 怒り感情検出タスク
サブタスク 6 失望感情検出タスク

全てのタスクは，検出するか検出しないかの二値分類である．これらのサブタスクと誹謗中傷検出タスクを同時に解くために，本研究では図 3.1 のモデルを用いた．図 3.1 中の FC とは全結合層のことである．それぞれのタスクの損失を合計して逆伝播することによって，全てのタスクを同時に学習させることを可能にしている．

不均衡データの学習改善手法

本研究では，不均衡データ学習の改善手法として，重み付き交差エントロピー誤差を採用した．交差エントロピー誤差 L は次の式で表される．

$$L = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{K} y_{ni} \log \hat{y}_{ni} \tag{3.3.1}$$

N，K はそれぞれデータ数とクラス数を表しており，y，$\hat{y}$ はそれぞれ実測値と予測値を表している．

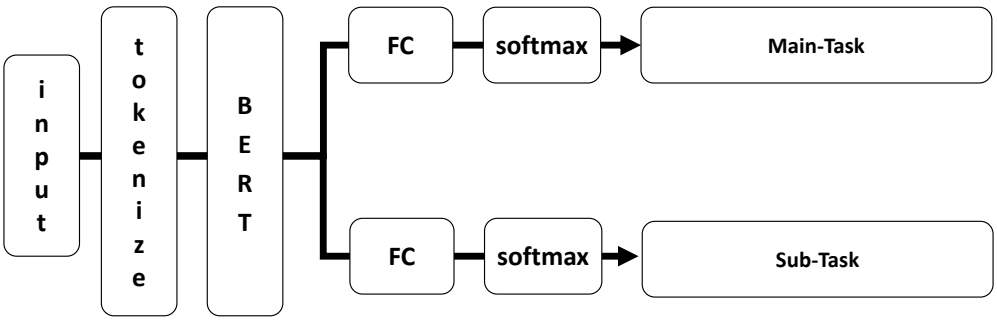

図 3.1　提案モデル図

重み付き交差エントロピー誤差 L_w は，L に重み w_i を付けた次の式で表される．

$$L_w = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{K} w_i y_{ni} \log \hat{y}_{ni} \tag{3.3.2}$$

$$w_i = \frac{N_l}{N_f} \tag{3.3.3}$$

N_l はデータ数が多い方のラベルのデータ数，N_f はデータ数が少ない方のラベルのデータ数を表している．重み付き交差エントロピー誤差を用いることによって，データ数の少ないラベルを誤って分類した際は損失が大きくなるように調整し，不均衡データをバランスよく学習させることができる．

3.4　評価実験__A

本研究では，我々の提案したサブタスクとサブタスクへの重み付き損失関数への採用が誹謗中傷検出タスクの精度向上に有効かどうか検証するために評価実験を

行った．提案したサブタスクの有効性を検証するために，シングルタスク学習モデル，提案サブタスクを加えたマルチタスク学習モデル，提案サブタスクを加えて重み付き損失関数を採用したマルチタスク学習モデル，ランダムタスクをサブタスクとして加えたマルチタスク学習モデルを構築し比較を行った．本章では，評価実験のデータセットとモデルの構築，結果・考察について述べる．

3.4.1 データセット

Twitter API を用いて収集したツイートに対して以下の質問を行った．はい・いいえ・ツイートが理解できないのうちのいずれかで答えることによってラベルを付与しデータセットを作成した．

質問 1 主タスク
　　このツイートは誹謗中傷ツイートですか？
質問 2 サブタスク 1
　　このツイートは脅迫の表現を含むツイートですか？
質問 3 サブタスク 2
　　このツイートは差別の表現を含むツイートですか？
質問 4 サブタスク 3
　　このツイートは容姿否定の表現を含むツイートですか？
質問 5 サブタスク 4
　　このツイートの投稿者は正義感を抱いてツイートしたと思いますか？
質問 6 サブタスク 5
　　このツイートの投稿者は怒りの感情を抱いてツイートしたと思いますか？
質問 7 サブタスク 6
　　このツイートの投稿者は失望の感情を抱いてツイートしたと思いますか？

ツイートは 11,200 件のツイート収集した．そのうち 622 件のツイートは理解できないと判断されたため評価実験では 10,578 件のツイートを使用した．10,578 件のツイートへの質問に対するラベル付けの結果を表 3.1 に記す．ラベル付けの結果から主タスクは，Yes/No の割合がおおよそ均衡であるが，その他のサブタスクは

Yes/No の割合が不均衡であることが確認された.

マルチタスク学習の有効性ではなく，我々の提案したサブタスクの有効性を検証するために，ランダムタスクをサブタスクとして用意した．ランダムタスクは，質問に対してランダムに半分を “はい”，半分を “いいえ” とすることによってラベルを付与したタスクである.

3.4.2 モデル構築

モデルは以下の 4 つを構築した.

モデル 1 シングルタスク学習モデル
モデル 2 各サブタスクを一つずつ加えたマルチタスク学習モデル
モデル 3 モデル 2 の学習に重み付き損失関数を採用したマルチタスク学習モデル
モデル 4 ランダムタスクをサブタスクとして加えたランダムマルチタスク学習モデル

各モデルはデータセットを Train:Validation:Test=8:1:1 に分割し学習を行った．また，全てのモデル構築においてバッチサイズを 32 に設定し，Validation の損失が 100 回最小値を下回らなかった場合，損失が最小だった際のモデルを保存した.

表 3.1　全 10,578 ツイートに対するラベル付け結果

ツイートの種類	Yes	No
誹謗中傷ツイートであるか	4,845	5,733
脅迫の字面を含むツイートであるか	505	10,073
差別の字面を含むツイートであるか	433	10,145
容姿の否定の字面を含むツイートであるか	591	9,987
正義感の感情を抱いてツイートしたか	1,165	9,413
怒りの感情を抱いてツイートしたか	1,659	8,919
失望の感情を抱いてツイートしたか	504	10,074

3.4.3 結果・考察

本研究では，F 値，Recall，Precision に対して 10 分割交差検証を行った．各評価指標に対して帰無仮説を “2 群の平均値に差はない” とし有意水準 0.05 で 2 群の対応のある t 検定を行った．10 分割交差検証は，データセットを 10 分割しその 1 割を Test，もう 1 割を Validation，その他の 8 割を Train とし，10 分割したデータそれぞれが 1 回ずつ Test になるように実験を 10 回行うものである．各モデルの評価指標の 10 回平均を表 3.4.3 に記す．表 3.4.3 について説明する．STL，MTL はそれぞれシングルタスク学習とマルチタスク学習を表している．main-F，main-R，main-P はそれぞれ主タスクの F 値，Recall，Precision を表しており，() 内の数値はシングルタスク学習モデルと比較した際の p 値を表している．太字は，シングルタスク学習モデルよりも提案モデルの方が，評価指標の値が向上している数値である．また，sub-F，sub-R，sub-P はそれぞれサブタスクの F 値，Recall，Precision を表している．各評価指標の右に記されている p-value は重みなしのマルチタスク学習と重み付きマルチタスク学習を比較した際の p 値を表している．

各モデルの評価指標

モデルの種類	評価指標											
	main-F	p-value	sub-F	p-value	main-R	p-value	sub-R	p-value	main-P	p-value	sub-P	p-value
STL	0.774	-	-	-	0.763	-	-	-	0.786	-	-	-
MTL(sub1)	0.765 (0.431)	-	0.423	-	0.738 (0.211)	-	0.373	-	**0.795** (0.322)	-	0.493	-
MTL(sub1)weight	0.771 (0.266)	0.575	0.552	0.132	0.755 (0.298)	0.324	0.741	0.001	**0.790** (0.606)	0.580	0.451	0.653
MTL(sub2)	0.766 (0.296)	-	0.110	-	0.745 (0.341)	-	0.067	-	**0.789** (0.732)	-	0.402	-
MTL(sub2)weight	0.762 (0.254)	0.678	0.269	0.495×10^{-3}	0.745 (0.365)	0.989	0.552	3.249×10^{-6}	0.783 (0.665)	0.617	0.185	0.026
MTL(sub3)	0.772 (0.777)	-	0.021	-	0.754 (0.502)	-	0.014	-	**0.791** (0.449)	-	0.047	-
MTL(sub3)weight	0.763 (0.160)	0.051	0.399	3.199×10^{-7}	0.747 (0.243)	0.450	0.608	1.220×10^{-9}	0.780 (0.477)	0.019	0.301	0.001
MTL(sub4)	0.768 (0.295)	-	0.527	-	0.746 (0.234)	-	0.435	-	**0.791** (0.585)	-	0.677	-
MTL(sub4)weight	0.764 (0.119)	0.702	0.497	7.697×10^{-10}	0.753 (0.497)	0.697	0.656	1.060×10^{-4}	0.778 (0.267)	0.142	0.400	2.372×10^{-10}
MTL(sub5)	**0.780** (0.444)	-	0.616	-	0.761 (0.891)	-	0.530	-	**0.801** **(0.032)**	-	0.744	-
MTL(sub5)weight	0.768 (0.357)	0.210	0.606	0.324	0.751 (0.290)	0.479	0.739	5.376×10^{-7}	**0.787** (0.873)	0.097	0.515	2.845×10^{-9}
MTL(sub6)	0.774 (0.960)	-	0.0	-	**0.766** (0.894)	-	0.0	-	**0.783** (0.696)	-	0.0	-
MTL(sub6)weight	0.761 (0.051)	0.067	0.226	2.201×10^{-7}	0.743 (0.217)	0.093	0.454	7.687×10^{-7}	0.783 (0.775)	0.985	0.157	1.102×10^{-6}
MTL(noise)	0.767 (0.319)	-	0.470	-	**0.766** (0.863)	-	0.463	-	0.768 (0.009)	-	0.504	-

サブタスク 1: 脅迫表現検出タスク

サブタスク 1 を加えたマルチタスク学習モデルでは，シングルタスク学習モデルと比べて主タスクの Precision のみ向上している．サブタスクの学習に重み付き損失関数を採用した場合，シングルタスク学習モデルと比べて主タスクの Precision のみ向上している．また重みなしマルチタスク学習モデルと比べて，主タスクの F 値，Recall が向上している．サブタスクの評価指標を比較すると，重みを付けた場合の方が F 値，Recall が向上しており，Recall の向上には有意な差が確認できた．

サブタスク 2: 差別表現検出タスク

サブタスク 2 を加えたマルチタスク学習モデルでは，シングルタスク学習モデルと比べて主タスクの Precision のみ向上している．サブタスクの評価指標を比較すると，重みを付けた場合の方が F 値，Recall が向上しており，どちらの評価指標にも有意な差が確認できた．

サブタスク 3: 容姿否定表現検出タスク

サブタスク 3 を加えたマルチタスク学習モデルでは，シングルタスク学習モデルと比べて主タスクの Precision が向上している．サブタスクの評価指標を比較すると，重みを付けた場合の方が F 値 Recall，Precision 全て向上しており，有意な差が確認できた．

サブタスク 4: 正義感検出タスク

サブタスク 4 を加えたマルチタスク学習モデルでは，シングルタスク学習モデルと比べて主タスクの Precision が向上している．また重みなしマルチタスク学習モデルと比べて，主タスク Recall が向上している．サブタスクの評価指標を比較すると，重みを付けた場合の方が Recall は向上している．

サブタスク 5: 怒り感情検出タスク

サブタスク 5 を加えたマルチタスク学習モデルでは，シングルタスク学習モデルと比べて主タスクの F 値，Precision が向上しており Precision の向上には有意な差が確認できた．サブタスクの学習に重み付き損失関数を採用した場合，シングル

タスク学習モデルと比べて主タスクの Precision は向上している．サブタスクの評価指標を比較すると，重みを付けた場合の方が Recall は向上しており，有意な差が確認できた．

サブタスク 6: 失望感情検出タスク

サブタスク 6 を加えたマルチタスク学習モデルでは，シングルタスク学習モデルと比べて主タスクの Precision は向上している．サブタスクの学習に重み付き損失関数を採用した場合，シングルタスク学習モデルと比べて主タスクの Recall が向上している．サブタスクの評価指標を比較すると，重みを付けた場合の方が F 値，Recall，Precision の全てが向上しており有意な差が確認できた．

ランダムタスク

ランダムタスクをサブタスクとして加えたマルチタスク学習モデルでは，シングルタスク学習モデルと比べて Recall は向上しているが F 値，Precision は低下しており Precision の低下には有意な差が確認できた．また，サブタスク 5 を加えたマルチタスク学習と比べると F 値の差が 0.13，p 値が 0.017 となり有意な差があることから，ランダムタスクを加えることに比べ提案したサブタスク 5 の怒りの感情検出タスクが有効であることを確認した．

考察

サブタスク 5 を加えたマルチタスク学習モデルに対して以下の 2 点を確認できたことからサブタスクの選定に価値はあると考えられる．しかし，シングルタスク学習モデルの F 値よりもマルチタスク学習モデルの F 値を向上させ，有意な差を確認するためには，より有効なサブタスクの選定が必要であることが考えられる．

1. シングルタスク学習モデルの F 値よりも向上．
2. ランダムタスクをサブタスクとして加えたマルチタスク学習モデルの F 値を有意に向上．

サブタスクの学習に重み付き損失関数を採用すると，サブタスク 1,4,5 の F 値の向上に有意な差が確認できないが，サブタスク 2,3,6 の F 値は向上に有意な差が確

認できる．この結果と表 3.1 のラベルの割合から，重み付き損失関数はラベルの割合に約 20 倍ほどの不均衡が確認できる場合に限り，有意な差が確認できると考えられる．しかし，サブタスクの F 値向上が必ずしも主タスクの F 値向上に繋がるわけではないことが確認された．

3.5　おわりに__A

本研究では，マルチタスク学習を用いて誹謗中傷検出精度向上を図る際に有効なサブタスクの提案し，サブタスクへの重み付き損失関数を採用した．マルチタスク学習では，主タスクと同時に主タスクと関連のあるサブタスクを解くことによって汎化性能の向上が期待できる．そのため，サブタスクに有効なタスクを選定することは非常に重要なことである．我々は，誹謗中傷検出タスクと同時に解くタスクとして，字面表現を検出するタスク三つと投稿者感情を検出するタスク三つが重要であると考えた．字面表現を検出するタスクは，誹謗中傷に様々な表現があるため，主タスクと関連が深いタスクであると考えた．字面表現を検出するタスクには，誹謗中傷と関連の深いと考えられる脅迫表現，差別表現，容姿の否定表現を検出するタスクの三つのタスクを選定した．投稿者感情を検出するタスクは，投稿者感情が誹謗中傷の原因となるため，主タスクと関連が深いタスクであると考えた．投稿者感情を検出するタスクには，誹謗中傷の動機である正義感，怒り感情，失望感情を検出するタスクの三つのサブタスクを提案した．また，マルチタスク学習では複数のタスクを同時に解くため，一つのラベルに合わせたデータの複製などによってデータの偏りを改善することはできない．そのため，重み付き損失関数を取り入れた．重み付き損失関数は，タスクごとのデータの偏りに合わせた重みを損失関数に付けることができるため，偏りのあるデータの学習改善が期待できる．提案したサブタスクの有効性を検証する評価実験を行うため，主タスク，サブタスクに関する質問を行いアンケートを取ることでデータセットを作成した．

提案したサブタスクと重み付き損失関数の有効性を検証するために，提案したサブタスクを加えたマルチタスク学習モデル，提案サブタスクを加えて重み付き損失関数を採用したマルチタスク学習モデルを構築し，シングルタスク学習モデル，ランダムタスクをサブタスクとして加えたマルチタスク学習モデルとの検出精度を比

較する実験を行った．

実験結果から，怒りの感情検出タスクを加えたマルチタスク学習モデルは，シングルタスク学習モデルと比べて有意な差にはならなかったが F 値の向上が確認できた．また，怒りの感情検出タスクを加えたマルチタスク学習モデルは，ランダムタスクを加えたマルチタスク学習モデルと比べて F 値が向上し，有意な差が確認できた．マルチタスク学習モデルでは，重み付き損失関数を採用した場合，サブタスクのデータに約 20 倍の不均衡がみられる場合に限り，サブタスクの F 値の向上に有意な差を確認することができた．脅迫表現検出タスクにおいて，重み付け損失関数を採用することによる有意な差は確認できなかったが，主タスクの F 値向上が確認できた．今後の展望として，今回のサブタスクと誹謗中傷検出精度の関係を精査することによって，より有効なサブタスクの模索，複数のサブタスクの併用を行っていきたいと考えている．また損失関数に追加する重みを学習過程で逐次変化させるなど，損失関数の設定についてもより有効な設定の模索を行いたいと考えている．

第 4 章　研究_B(サブタスクを主観と客観の二つの観点から着目した研究)

内容梗概 B

本研究の目的は二つある．一つ目は，マルチタスク学習を用いて誹謗中傷検出タスクの精度向上を図る際，我々が提案する主観サブタスクが精度向上に有効であることを明らかにすることである．二つ目は，我々が構築したモデルがどのような誹謗中傷投稿に対して，検出精度向上の傾向があるか明らかにすることである．二つの目的を果たすために，我々は主観サブタスク 2 種類と客観サブタスク 2 種類の計 4 種類のサブタスクを，追加するか追加しないかの組み合わせ，計 16 種類のモデルを構築した．構築した全モデルを，F 値，Precision，Recall で比較した．また，検出精度に違いが生じたモデルに対して，どのような投稿で検出精度に違いが出たのか分析を行った．実験の結果，以下の知見が得られた．(1) 主観サブタスクを加えた MTL(マルチタスク学習モデル) は，客観サブタスクを加えた MTL に比べ，誹謗中傷検出タスクの F 値向上に有効な傾向がある (2)MTL の F 値 (主タスク) に有意な差があった場合でも，加えるサブタスクによって，特定の投稿の検出に特化する場合がある

4.1　はじめに_B

本研究の目的は二つある．一つ目は，マルチタスク学習を用いて誹謗中傷検出タスクの精度向上を図る際，我々が提案する主観サブタスクが精度向上に有効であることを明らかにすることである．二つ目は，我々が構築したモデルがどのような誹謗中傷投稿に対して，検出精度向上の傾向があるか明らかにすることである．

一つ目の目的について述べる．マルチタスク学習 [8] とは，一つのモデルで複数のタスクを同時に解く手法である．同時に解く複数のタスクのうち，精度の向上を期待するタスクを主タスク，主タスクの学習補助の役割を果たすタスクをサブタスクと呼ぶ．マルチタスク学習モデルは，主タスクと関連のあるサブタスクを同時に解くことによって，一つのタスクにとらわれることのない特徴を獲得することがで

きる．その結果，モデルの汎化性能向上が期待できる．マルチタスク学習は，様々な自然言語処理分野の研究 [10][11][12] において，精度向上に貢献している．

我々は誹謗中傷検出タスクにマルチタスク学習を採用し，検出精度の向上を図ろうと考えた．しかし，誹謗中傷検出タスクを主タスクとした際，どのようなサブタスクを選定することが精度向上に有効であるかどうかは，著者が調べた限りでは明らかになっていない．選定したサブタスクが主タスクの精度向上に有効であるかどうかは，データセットの作成からモデルの構築，テストデータで精度検証をするまで確認できないため，膨大な時間と労力がかかるという課題がある．我々は，課題解くためには主タスクである誹謗中傷検出タスクの精度向上に，有効なサブタスクの傾向を明らかにする必要があると考えた．我々は主観サブタスクが，主タスクである誹謗中傷検出タスクの精度向上に，有効な傾向があるのではないかという仮説を立てた．

主観サブタスクとは，第 1 者または第 2 者の観点からラベルを付与することができるタスクである．第 1 者とは投稿者，第 2 者とは閲覧者のことである．主観サブタスクが，誹謗中傷検出タスクの精度向上に有効な傾向があるのではないかと考えた理由を述べる．誹謗中傷投稿には，投稿者が悪意をもって行った投稿と，投稿者に悪気はないが閲覧者との感じ方の差異によって誹謗中傷となる投稿がある．そのため，投稿者の観点と閲覧者の観点は第 3 者の観点よりも，誹謗中傷を検出するために，重要な情報だと考えた．本研究では，主観サブタスクとして投稿者感情分類タスクと閲覧者感情分類タスクを選定した．また，主観サブタスクの有効性を検証するために比較対象として，第 3 者の観点からラベルを付与することができる客観サブタスクを用意した．客観サブタスクには，投稿対象分類タスクと誹謗中傷カテゴリ分類タスクを選定した．それぞれのサブタスクについては，4.3.1 節で述べる．

二つ目の目的について述べる．誹謗中傷には，煽り，皮肉，差別など様々な種類が存在する．構築したモデルの誹謗中傷検出精度が低下したとしても，特定の誹謗中傷投稿の検出に特化したモデルである可能性がある．例えば，全体的な誹謗中傷投稿の検出精度が低下しているが，差別表現を含んだ誹謗中傷投稿の検出精度は向上した場合である．我々は，どのようなサブタスクを加えたモデルが，どのような誹謗中傷投稿を検出できたのか明らかにすることによって，状況に応じたサブタスクの選定基準を明らかにできると考えた．

我々は，二つの目的を達成するために，サブタスクの提案，データセットの作成 (4.3.1 節)，モデルの構築，評価実験 (4.4 章) を行った．データセットの作成は，収集した投稿に対してクラウドソーシングを用いてラベルを付与することによって行った．ラベルは，各投稿それぞれに対するアンケートに 5 人ずつ回答し，多数決を取ることによって決定した．作成したデータセットを用いて，主観サブタスク 2 種類と客観サブタスク 2 種類の計 4 種類のサブタスクを，追加するか追加しないかの組み合わせ，計 16 種類のモデルを構築した．構築した全モデルを，F 値，Precision，Recall で比較した．また，検出精度に違いが生じたモデルに対して，どのような投稿で検出精度に違いが出たのか分析を行った．実験の結果，以下の 3 点を明らかにした．

1. 主観サブタスクを加えた MTL(マルチタスク学習モデル) は，客観サブタスクを加えた MTL に比べ，誹謗中傷検出タスクの F 値向上に有効な傾向がある．
2. 閲覧者感情分類タスク (主観サブタスク 2) を加えた MTL は，誹謗中傷検出タスクの Precision 向上に有効．
3. 閲覧者感情分類タスク (主観サブタスク 2) を加えた MTL は STL(シングルタスク学習モデル) に比べ，投稿者が喜びの感情を抱いた投稿に対する誹謗中傷検出タスクの Accuracy 向上に有効．

4.2 関連研究_B

誹謗中傷投稿を自動検出するための研究は様々行われている．誹謗中傷投稿を検出するための研究には，誹謗中傷に関連する単語に着目し，誹謗中傷投稿を検出する手法と，文章全体に着目して誹謗中傷投稿を検出する手法がある．単語に着目した手法は，特定の単語を含んだ誹謗中傷投稿に対しての見逃しを高確率で抑えることができる．しかし，単語の登録数に限界があるため未知語を含んだ誹謗中傷投稿に対して脆弱であるというデメリットがある．単語に着目した研究として大友 [13] らの研究がある．大友らは，86,906 語の単語からなるいじめ表現辞書を作成した．彼らはいじめ表現辞書と複数の機械学習モデルを組み合わせてネットいじめ検出を

行っている．その結果，彼らは高い分類評価を記録しているが，いじめ表現辞書への登録単語の少なさを課題に挙げている．単語に着目した手法の改善を試みた研究として Waseem ら [4] の研究がある．Wassem らは，誹謗中傷単語を特定するリストベースの手法は，リストに記載されていない単語に脆弱であることを問題視し，n-gram に基づく手法を提案している．その結果，性別に関する特徴を加えることが誹謗中傷検出精度向上に有効であることを示している．文章全体に着目した誹謗中傷投稿を検出する手法として Alatawi ら [14] の研究がある．Alatawi らは，ドメイン単語埋め込みと BiLSTM を組み合わせたモデルと，文章理解に優れた機械学習モデルである BERT の二つのモデルでヘイトスピーチの検出を行い，共にベースラインより高い検出精度を記録している．BiLSTM を用いたモデルは，スラングなど特定単語を用いたヘイトスピーチに強く，BERT を用いたモデルは，全体的な検出精度が高いことを示している．BERT を誹謗中傷単語で再トレーニングした研究 [15] も存在する．Caselli ら [15] は，学習済み BERT をポリシーに違反した投稿を使用して再トレーニングすることによって誹謗中傷検出タスクに特化した BERT を構築した．その結果，攻撃性のある文章を検出するタスクにおいて一般的な BERT を大きく上回る性能を獲得している．本研究のベースラインとなるシングルタスク学習モデルも BERT[16] を使用している．

マルチタスク学習は，画像処理 [17] や強化学習 [18]，自然言語処理 [12] など様々な分野で応用されている．Lamsiyah ら [12] は，マルチタスク学習を用いて文書分類タスク，ペアワイズテキスト分類タスク，テキスト類似度タスク，関連度ランキングタスクの四つのタスクを同時に学習するモデルを提案している．提案したマルチタスク学習モデルを用いた BERT モデルで大幅な精度向上に成功している．

類似した研究としてマルチタスク学習を用いて誹謗中傷検出タスクを解くモデルを提案した研究 [6][19] がある．Samghabadi ら [6] は，攻撃性識別タスクと女性差別的攻撃性識別タスクを同列に扱いマルチタスク学習を行っている．その結果，女性差別攻撃性タスクの重み付き F 値で 15 チームのうち 3 位と高い結果を獲得している．この研究では，攻撃性識別タスクにマルチタスク学習を採用したという点で本研究と類似しているが，二つのサブタスクの精度を両方向上させようとしている．しかし，本研究では誹謗中傷検出タスクの精度向上を図るために有効性なサブタスクとして感情分類タスクを提案するため目的が違う．Aldjanabi ら [19] は，攻

撃性検出データセットと三つのヘイトスピーチデータセットをマルチタスク学習で同時に学習するモデルを提案している．その結果，四つのデータセットのうち三つのデータセットで提案手法の有効性を示している．このように，類似したタスクを同時に解くことによって全てのタスクの検出精度向上を図る研究は存在する．しかし，誹謗中傷検出タスクの精度向上を図るためにサブタスクを提案し，BERT とマルチタスク学習を用いた研究は著者が調べた限りでなはい．そこで本研究では，誹謗中傷検出精度向上を目的に，誹謗中傷検出タスクモデルの学習を補助するためのサブタスクの提案を行う．

4.3 提案手法_B

我々は，誹謗中傷検出タスクを解くモデルの学習を行う際，主観サブタスクを同時に解くことによって，モデルの学習を補助することができ，誹謗中傷検出精度が向上する傾向があるという仮説を立てた．仮説を検証するためには主観サブタスク 2 種類と客観サブタスク 2 種類の計 4 種類のサブタスクを，追加するか追加しないかの組み合わせ，計 16 種類のモデルの誹謗中傷検出精度を比較する必要がある．モデルを構築するためには，モデルの学習に必要なデータセットを作成する必要がある．そのため，我々は仮説を検証するためのデータセットを作成した．

ベースラインとなる誹謗中傷検出タスクのみを解くモデル (シングルタスク学習モデル) は，東北大学の乾・鈴木研究室が公開している事前学習済みの BERT モデル*を主タスクでファインチューニングすることによって構築した．マルチタスク学習モデルは，誹謗中傷検出タスクに主観サブタスク 2 種類と客観サブタスク 2 種類の組み合わせ全通りで加えたモデルを構築した．本章では，作成したデータセットについて 4.3.1 節，提案モデルについて 4.3.2 節で説明する．

4.3.1 データセット

データセットを構築するためには，仮説の検証に必要なタスクに応じた質問をし，アンケートを取ることによってラベルを付与する必要がある．本研究でデータセッ

*https://github.com/cl-tohoku/bert-japanese

トを構築するために行ったアンケートの質問，質問に対応するタスク，回答項目を表 4.1 に記す．データセットのそれぞれのタスクについて説明する．

主タスク

本研究における主タスクは，誹謗中傷検出タスクとする．このタスクでは投稿がどのような対象への投稿であっても誹謗中傷であれば “誹謗中傷” とする．つまり自虐のように相手を傷つけることはなくても，誹謗中傷の表現を含んでいれば “誹謗中傷” のラベルが付与される．

主観サブタスク

主観サブタスクとは，第 1 者または第 2 者の観点からラベルを付与することができるタスクである．我々は，誹謗中傷検出タスクの精度向上には主観サブタスクを同時に学習することが有効であると考えた．主観サブタスクが誹謗中傷検出精度向上に有効であると考えた背景を説明する．誹謗中傷の定義は曖昧であるため，投稿について誹謗中傷投稿であると感じる人と，別の人は正当な批判と感じる人が存在する可能性がある．この投稿から受ける感じ方の差異は，特に投稿者と閲覧者に

表 4.1　データセット構築のための質問と対応タスク

質問内容	タスク	回答項目
Q1: 投稿は誹謗中傷ですか？	誹謗中傷検出タスク (主タスク)	はい，いいえ
Q2: 投稿者の感情はどれですか？	投稿者感情分類タスク (主観サブタスク 1)	喜び，怒り，悲しみ，好き，その他
Q3: 投稿から受ける印象はどれですか？	閲覧者感情分類タスク (主観サブタスク 2)	ポジティブ, ネガティブ, ニュートラル
Q4: どれに向けての投稿ですか？	投稿対象分類タスク (客観サブタスク 1)	自分，自分のもの，他人，他人のもの，組織，その他
Q5: 投稿はどれに当てはまりますか？	誹謗中傷カテゴリ分類タスク (客観サブタスク 2)	脅迫，差別，容姿否定，その他 (複数選択)

よって生じることが多い†．そのため，我々はソーシャルメディア上の投稿が誹謗中傷かどうか判断するためには，無関係な第 3 者よりも投稿者や閲覧者の観点から考える情報が必要ではないかと考えた．本研究では，投稿者や閲覧者の観点から考えるタスクを主観サブタスクとして提案した．

投稿者感情分類タスクを主観サブタスクとして提案した理由を説明する．bot などの形式的にソーシャルメディアに投稿するものを除いたソーシャルメディアの投稿者は，何らかの感情を抱いて投稿を行う．誹謗中傷投稿者は，特に正義感や怒りの感情を抱いて誹謗中傷を投稿する傾向がある [9]．そのため我々は，投稿から読み取れる投稿者感情を明示的にモデルに伝えることができれば，誹謗中傷検出タスクの学習を補助することができるのではないかと考えた．投稿者感情が誹謗中傷検出の補助になると考えた例を以下に示す．

例 1 お前いい加減にしろ，殴るぞ笑笑
例 2 お前いい加減にしろ，殴るぞ！！

例 1 例 2 は，文末以外は “お前いい加減にしろ，殴るぞ” と相手に伝える投稿である．例 1 は，文末が “笑笑” であり，投稿相手のボケに対するツッコミのように感じる．そのため，投稿者は否定的な感情は持たず，投稿を行ったと考えることができる．例 2 は文末が “!!” であり “お前いい加減にしろ，殴るぞ” の文章を強調している．そのため，投稿者が投稿相手に対して強い怒りの感情を持っていると考えることができる．上記二つの例より，投稿者の感情が異なると投稿の伝わり方が違うことがわかる．そこで，我々は投稿者感情分類タスクを誹謗中傷検出タスクと同時に解くことによって，モデルが誹謗中傷投稿の特徴を捉える補助ができると考えた．

閲覧者感情分類タスクを主観サブタスクとして提案した理由を示す．前述したとおり，誹謗中傷の定義は曖昧であり，人によって誹謗中傷かそうでないかの判断は異なる．我々は，誹謗中傷の判断する際に重要なことは第 3 者よりも，被害を受ける閲覧者がどのように感じるかであると考えた．そのため，我々は投稿閲覧者が抱く感情は，誹謗中傷を検出する上で非常に重要な項目であると考えた．上記より，我々は，閲覧者感情分類タスクを誹謗中傷検出タスクと同時に解くことによって，モデルが誹謗中傷投稿の特徴を捉える補助ができると考えた．

†https://prtimes.jp/main/html/rd/p/000000201.000044347.html

客観サブタスク

客観サブタスクは，提案した 2 種類の主観サブタスクが有効であることを検証するための比較対象として用意した．投稿対象分類タスクとは，投稿を向けられた相手が回答項目にうちどれなのか分類するタスクである．組織とは，政党や国籍など集団を対象にした投稿である．投稿対象分類タスクをサブタスクとして提案した理由を例を用いて示す．

例 1 私は身長低いからね笑
例 2 お前は身長低いからね笑

例 1 も例 2 も文頭の “私"と “お前"以外は同じ文章である．文頭が “私"である例 1 の文は，何かに対しての言い訳をしているように感じられる．しかし，文頭が “お前"である例 2 の文は身長低い人を嘲笑しているように感じられる．このように，投稿対象が違うだけで攻撃的にも非攻撃的に感じられる．上記より我々は，投稿対象分類タスクは主タスクと関係のあるタスクであり比較対象として成立すると考えた．

誹謗中傷カテゴリ分類タスクは，誹謗中傷の種類を分割したカテゴリを分類するタスクである．誹謗中傷カテゴリは誹謗中傷に内包される．そのため，主タスクと関係のあるタスクであり比較対象として成立すると考えた．カテゴリには，誹謗中傷としてよく問題にあげられる “脅迫”，“差別”，“容姿否定” を選択した．誹謗中傷カテゴリは，そもそも誹謗中傷がある投稿にしか存在しない．そのため誹謗中傷検出タスクで “はい” と答えた場合のみ，このタスクは回答を行う．

データセット構築

データセット構築の手順を以下に記す．

1. Twitter API を用いて投稿を収集
2. クラウドソーシングを用いて収集した投稿一つずつに対して 5 人にアンケートを実施
3. アンケートに答えた 5 人のワーカの票を多数決
4. 多数決の結果最も票を集めた選択肢をラベルとして付与
5. 票割れした投稿や投稿自体が理解できないものを削除

手順 1 の Twitter API を用いて収集した投稿は，2021 年 7 月 5 日から 2022 年 7 月 5 日の投稿のうち，我々が考えた攻撃性のある単語 36 種類のいずれかを含む投稿である．36 種類の単語を次に示す．

[“ばか”,“バカ”,“馬鹿”,“あほ”,“アホ”,“ブス”,“きも”,“キモ”,“くず”,“クズ”,“しね”,“シネ”,“死ね”,“ゴミ”,“ごみ”,“雑魚”,“ざこ”,“ザコ”,“かす”,“カス”,“コロス”,“ころす”,“殺す”,“きえろ”,“キエロ”,“消えろ”,“むかつく”,“ムカつく”,“でぶ”,“デブ”,“はげ”,“ハゲ”,“無能”,“ダサ”,“いい加減にしろ”,“ウザ”]

手順 5 の票割れした投稿について説明する．票割れした投稿とは，多数決の結果最も票を集めた選択肢が 2 種類以上存在する投稿である．例えば，Q3 の 5 人へのアンケートの結果が “ポジティブ” が 2 票，“ネガティブ” が 2 票，“ニュートラル” が 1 票となった場合である．この場合，“ポジティブ” と “ネガティブ” は 2 票ずつ選択されており，どちらが正しいか判断できない．票割れした投稿は教師データとして曖昧なため削除した．Q5 は複数選択のため多数決は利用できない．そのため，3 人以上のワーカが選択した項目をラベルとして付与した．実際に作成したデータセットの各質問に対する回答の内訳は表 4.2 のようになった．

このデータセットを用いてモデルを構築する．

4.3.2 提案モデル

提案モデル図

我々が用いたモデルを図 4.1 に表す．図 4.1 中の FC は全結合層を表し，N はサ

表 4.2 各質問に対する回答の内訳

タスク	全 13,448 件の質問に対する回答のデータ数				
主タスク (Q1)	誹謗中傷	誹謗中傷ではない	-	-	-
	3,894	9,554	-	-	-
主観サブタスク 1(Q2)	喜び	怒り	悲しみ	好き	その他
	1,029	6,158	793	986	4,482
主観サブタスク 2(Q3)	ポジティブ	ネガティブ	ニュートラル	-	-
	2,676	9,457	1,315	-	-
客観サブタスク 1(Q4)	自分, 自分のもの	他人, 他人のもの	組織	その他	-
	822	10,817	885	924	-
客観サブタスク 2(Q5)	脅迫	差別	容姿の否定	その他	誹謗中傷ではない
	1,160	176	267	649	9,554

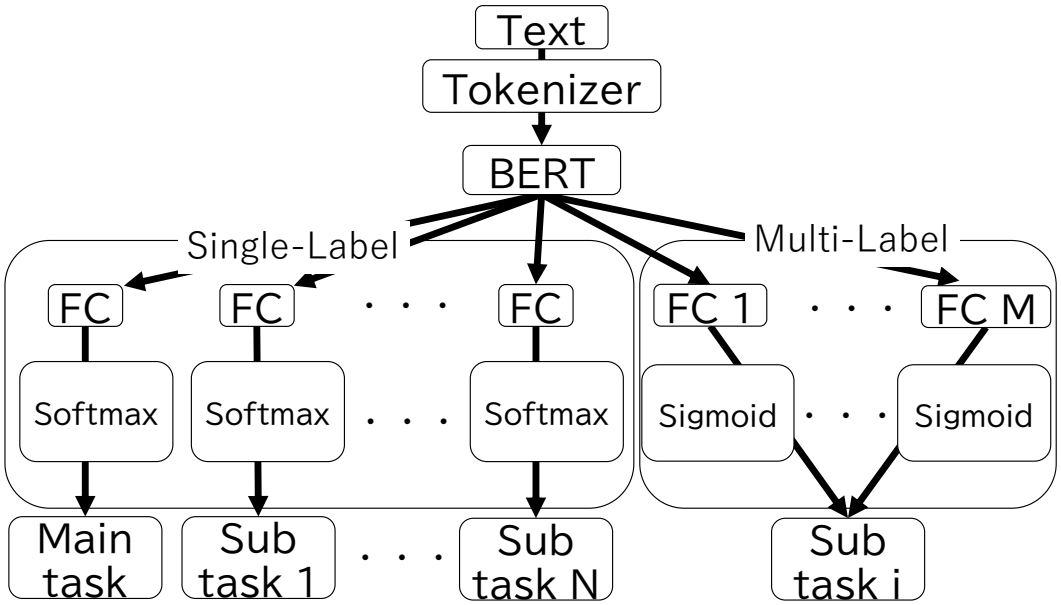

図 4.1　提案モデル図

ブタスクの総数，M はマルチラベルの選択肢の総数，$Subtaski$ は i 番目のサブタスクを表している．提案モデルは，初めに入力テキストをトークンに分割し，BERT を用いて分散表現に変換する．その後はシングルラベルかマルチラベルかによってモデルが異なる．シングルラベルは全結合層と SoftMax 関数を用いて，どの分類項目にあてはまるかどうかの確率を求める．マルチラベルは BERT から得た分散表現をマルチラベルの選択肢の数だけ，全結合層と sigmoid 関数に通す．上記のモデルを構築することによって，マルチラベルの一つ一つの分類項目が当てはまるかどうかを学習することができる．提案モデルを使用することによって主タスクとサブタスクを同時に学習する．

損失関数

モデルの学習に用いた損失関数は以下の式で表される．

$$AllLoss = MainLoss + \frac{2}{3}\sum_{i=1}^{N} SubLoss_i \tag{4.3.1}$$

$AllLoss$ は主タスクとサブタスクの損失の合計を表しており，モデルが逆伝搬する損失である．$MainLoss$ は主タスクの損失を表しておりクロスエントロピー誤差である．N は同時に学習するサブタスクの総数を表している．$SubLoss_i$ は i 番目のサブタスクの損失を表しており，サブタスクがシングルラベルの時はクロスエントロピー誤差，マルチラベルの時は，各出力に対してのバイナリークロスエントロピー誤差の合計である．本研究では，サブタスクを主タスクの学習の補助として扱い，主タスクを優先して学習させるため，サブタスクの損失の合計に 3 分の 2 をかける．

4.4 評価実験_B

評価実験の目的は二つある．一つ目は，主観サブタスクを加えたマルチタスク学習モデルが，誹謗中傷検出精度向上に有効かどうか検証することである．我々は検証のために，主観サブタスク 2 種類と客観サブタスク 2 種類の計 4 種類のサブタスクを，追加するか追加しないかの組み合わせ，計 16 種類のモデルの評価指標を比較した．二つ目は，構築したモデルが，どのような誹謗中傷投稿に対して検出精度向上したのか，または低下したのか，傾向を調査することである．我々は検証のために，全タスクの各ラベルに対応するデータごとの Accuracy を算出し，比較を行った．モデルを構築するために作成したデータセットの分析について 4.4.1 節，一つ目の検証と二つ目の検証について 4.4.2 節で述べる．

4.4.1 データセット分析

作成したデータセットについて分析する．主タスクの回答は “誹謗中傷である” が 7 割をしめる結果となった．攻撃性のある単語を含む投稿を集めたが “バカ面白い笑” など，攻撃性のない表現に用いられることも多かったため，このような結果となった．主観サブタスク 1 の回答は “怒り” が一番多くの票を集めたが総投票数の 5 割を下回り，票が割れる結果となった．上記より，攻撃性のある単語を含む投稿には様々な感情が混在していることがわかる．主観サブタスク 2 は，“ネガティブ” の投稿が 7 割をしめる結果となった．“ネガティブ” の投稿が多いのは，攻撃性

のある単語を含む投稿を集めたためと考えることができる．主観サブタスク 2 と主タスクの結果から，投稿を “ネガティブ” と感じるハードルは低いが，“誹謗中傷” と感じるハードルは高いことを確認した．客観サブタスク 1 の回答は “他人, 他人のもの” が 8 割をしめる結果となった．上記より，攻撃性のある単語を含む投稿は他人に向けた投稿が多いということを確認した．客観サブタスク 2 は主タスクで “誹謗中傷ではない” と選択された投稿が多かったため，それぞれのカテゴリへの投票は少なかった．客観サブタスク 2 の選択肢の中では “脅迫” のカテゴリが一番票を集める結果となった．

4.4.2 実験結果・考察

我々は，主観サブタスク 2 種類と客観サブタスク 2 種類の計 4 種類のサブタスクを，追加するか追加しないかの組み合わせ，計 16 種類のモデルを構築した．我々は，データセットを訓練用: 検証用: テスト用=8:1:1 の割合に分割し，訓練用データを用いてモデルの学習を行った．我々は，検証用データの損失が 100 エポック更新されなかった場合，損失が最小の時のモデルを採用した．その後，テスト用データで 10 分割交差検証を行った．

構築した 16 種類のモデルの実験結果を表 4.4.2 に記す．STL とは，誹謗中傷検出タスクのみを解くシングルタスク学習モデルである．MTL とは主タスクとサブタスクを同時に解くマルチタスク学習モデルである．括弧内の数字は，同時に学習したタスクの質問番号である．例えば，MTL(Q1,Q2) は Q1 と Q2 を同時に学習したマルチタスク学習モデルである．評価指標は，Accuracy: 正答率，F 値: 主タスクの陽性の F 値，macro-F1: 主タスクのマクロ F1 値，Precision: 主タスクの陽性の適合率，Recall: 主タスクの陽性の再現率，macro-F1(Q2): 主観サブタスク 1 のマクロ平均 F1，macro-F1(Q3): 主観サブタスク 2 のマクロ平均 F1，macro-F1(Q4): 客観サブタスク 1 のマクロ平均 F1 macro-F1(Q5): 客観サブタスク 2 のマクロ平均 F1 である．F 値，Precision，Recall は，“誹謗中傷” を陽性とした場合の評価指標である．各評価指標の下に記載されている括弧内の数値は，帰無仮説を “STL の評価指標と比較対象である MTL の評価指標の平均値に差はない” とし，2 群の対応のある t 検定を行った際の p 値である．太字は，各評価指標で最も精度が高かった

数値である．

全モデルの評価指標比較

本節では，主観サブタスク 2 種類と客観サブタスク 2 種類の計 4 種類のサブタスクを，追加するか追加しないかの組み合わせ，計 16 種類のモデルの評価指標を比較について述べる．評価指標比較の目的は，主観サブタスクを加えたマルチタスク学習モデルが，誹謗中傷検出精度向上に有効かどうか検証することである．

実験結果

	Accuracy	F 値	macro-F1	Precision	Recall	macro-F1(Q2)	macro-F1(Q3)	macro-F1(Q4)	macro-F1(Q5)
STL(Q1)	0.803	0.651	0.757	0.668	0.636	-	-	-	-
MTL(Q1,2)	0.803 (0.940)	0.657 (0.481)	0.759 (0.585)	0.664 (0.674)	0.653 (0.368)	0.469	-	-	-
MTL(Q1,2,3)	0.804 (0.739)	0.655 (0.555)	0.759 (0.576)	0.670 (0.868)	0.645 (0.635)	0.483	0.856	-	-
MTL(Q1,2,3,4)	0.802 (0.923)	0.654 (0.488)	0.758 (0.728)	0.665 (0.829)	0.647 (0.426)	**0.486**	0.847	0.343	-
MTL(Q1,2,3,4,5)	0.801 (0.624)	0.652 (0.856)	0.756 (0.954)	0.662 (0.328)	0.646 (0.457)	0.481	**0.859**	0.298	0.316
MTL(Q1,2,3,5)	0.802 (0.800)	0.647 (0.434)	0.754 (0.540)	0.670 (0.801)	0.628 (0.450)	0.482	0.857	-	0.295
MTL(Q1,2,4)	0.800 (0.613)	0.650 (0.806)	0.755 (0.686)	0.662 (0.727)	0.641 (0.623)	0.471	-	0.348	-
MTL(Q1,2,4,5)	0.799 (0.366)	0.652 (0.872)	0.755 (0.815)	0.653 (0.100)	0.653 (0.331)	0.482	-	0.358	0.353
MTL(Q1,2,5)	0.798 (0.111)	0.650 (0.827)	0.754 (0.389)	0.654 (0.070)	0.648 (0.335)	0.477	-	-	0.285
MTL(Q1,3)	0.800 (0.626)	**0.667** (0.274)	**0.762** (0.325)	**0.698** (0.033)	0.640 (0.739)	-	0.823	-	-
MTL(Q1,3,4)	0.805 (0.578)	0.650 (0.875)	0.757 (0.838)	0.683 (0.470)	0.627 (0.613)	-	0.821	0.322	-
MTL(Q1,3,4,5)	0.800 (0.330)	0.654 (0.595)	0.757 (0.997)	0.654 (0.087)	**0.656** (0.176)	-	0.823	0.33	0.253
MTL(Q1,3,5)	0.802 (0.838)	0.651 (0.968)	0.756 (0.898)	0.668 (0.973)	0.639 (0.824)	-	0.808	-	0.257
MTL(Q1,4)	**0.805** (0.434)	0.653 (0.808)	0.759 (0.650)	0.676 (0.315)	0.633 (0.796)	-	-	**0.371**	-
MTL(Q1,4,5)	0.801 (0.675)	0.648 (0.648)	0.755 (0.651)	0.665 (0.716)	0.634 (0.789)	-	-	0.333	0.310
MTL(Q1,5)	0.800 (0.446)	0.644 (0.095)	0.752 (0.147)	0.669 (0.923)	0.625 (0.288)	-	-	-	**0.544**

■F 値の比較　STL の F 値と MTL の F 値を比較する．STL の F 値と比べて MTL の F 値が向上したモデルは八つである．F 値が向上したモデルのうち，最も F 値が向上したモデルは，STL よりも 0.016 向上した MTL(Q1,3) である．MTL(Q1,3) は主観サブタスク 2 の閲覧者感情分類タスクを加えたモデルである．主観サブタスクを加えたモデルの F 値向上は，我々の仮説立証に対して望ましい結果である．しかし，表 4.4.2 括弧内に記載している MTL(Q1,3) の p 値は 0.274 である．そのため，我々は仮説の立証には更なる精度向上が必要であると考える．

構築した 16 種類のモデルの F 値を比較する．F 値上位の三つのモデルは，上から順に MTL(Q1,3)，MTL(Q1,2)，MTL(Q1,2,3) である．F 値上位の三つのモデルは，全て主観サブタスクのみを加えたモデルである．F 値下位の三つのモデルは，MTL(Q1,5)，MTL(Q1,2,3,5)，MTL(Q1,4,5) である．F 値下位の三つのモデルのうち二つは，客観サブタスクのみを加えたモデルである．16 種類のモデルの F 値を比較するために，帰無仮説を “モデルの F 値の平均値に差はない” とし，16 種類のモデル総当たりで 2 群の対応のある t 検定を行った．有意水準 0.05 を基準とすると，MTL(Q1,3) と MTL(Q1,4,5) の p 値 0.026 と，MTL(Q1,3) と MTL(Q1,5) の p 値 0.021 が有意水準を下回った．この結果から，主観サブタスク 2 の閲覧者感情分類タスクをサブタスクとして，誹謗中傷検出タスクに加えることは，(1) 客観サブタスク 4 の投稿対象分類タスクと客観サブタスク 5 の誹謗中傷カテゴリ分類タスクを加えることよりも F 値向上に有効 (2) 客観サブタスク 5 の誹謗中傷カテゴリ分類タスクを加えることよりも F 値向上に有効であることが分かった．(1)(2) はどちらも客観サブタスクよりも，主観サブタスクの方が F 値向上に有効であることを示している．そのため，マルチタスク学習を用いて誹謗中傷検出タスクの学習を行う際には，客観サブタスクよりも主観サブタスクを追加する方が F 値向上に繋がると考えることができる．

■マクロ F1 比較　STL の macro-F1 値と MTL の macro-F1 値を比較する．STL の macro-F1 値と比べて MTL の macro-F1 値が向上したモデルは七つである．macro-F1 値が向上したモデルのうち，最も macro-F1 値が向上したモデルは，STL よりも 0.005 向上した MTL(Q1,3) である．MTL(Q1,3) は主観サブタスク 2 の閲覧者感情分類タスクを加えたモデルである．主観サブタスクを加えたモデルの F 値向上は，我々の仮説立証に対して望ましい結果である．しかし，表 4.4.2 括弧内に

記載している MTL(Q1,3) の p 値 0.325 である．そのため，我々は仮説の立証には更なる精度向上が必要であると考える．

構築した 16 種類のモデルの macro-F1 値を比較する．macro-F1 値上位の三つのモデルは，上から順に MTL(Q1,3)，MTL(Q1,4)，MTL(Q1,2) である．F 値上位の三つのモデルのうち，二つが主観サブタスクのみを加えたモデルである．主観サブタスクを両方加えた MTL(Q1,2,3) の macro-F1 値は 4 位である．macro-F1 値下位の三つのモデルは，MTL(Q1,5)，MTL(Q1,2,3,5)，MTL(Q1,2,5) である．F 値下位の三つのモデルのうち最下位は，客観サブタスクのみを加えたモデルである．他の二つのモデルは主観サブタスクと客観サブタスク両方を加えたモデルである．16 種類のモデルの macro-F1 値を比較するために，帰無仮説を “モデルの F 値の平均値に差はない” とし，16 種類のモデル総当たりで 2 群の対応のある t 検定を行った．有意水準を 0.05 とすると，MTL(Q1,2,3) と MTL(Q1,2,5) の p 値 0.014 が有意水準を下回った．この結果から，主観サブタスク 1,2 の両方をサブタスクとして，誹謗中傷検出タスクに加えることは，主観サブタスク 1 の投稿者感情分類タスクと客観サブタスク 2 の誹謗中傷カテゴリ分類タスクを加えることより macro-F1 値向上に有効であることを確認した．

■Precision の比較　STL の Precision と MTL の Precision を比較する．STL の Precision と比べて MTL の Precision が向上したモデルは七つである．Precision が向上したモデルのうち，最も Precision が向上したモデルは，STL よりも 0.030 向上した MTL(Q1,3) である．有意水準 0.05 とすると，STL と MTL(Q1,3) を比較した際の p 値は 0.033 であり下回っている．この結果から，主観サブタスク 2 の閲覧者感情分類タスクを，サブタスクとして誹謗中傷検出タスクの学習に加えることは，Precision の向上に有効であるとが確認できる．

構築した 16 種類のモデルの Precision を比較する．Precision 上位の三つのモデルは，上から順に MTL(Q1,3)，MTL(Q1,3,4)，MTL(Q1,4) である．Precision 下位の三つのモデルは，MTL(Q1,2,4,5)，MTL(Q1,3,4,5)，MTL(Q1,2,5) である．比較の結果，最も精度が高い MTL(Q1,3) が他の 6 種類のモデルに対して有意水準 0.05 を下回る p 値を記録した．MTL(Q1,Q3) が他のモデルに比べ，Precision の向上に有効であることを確認した．

■Recall の比較　STL の Recall と MTL の Recall を比較する．STL の Recall と比べて MTL の Recall が向上したモデルは 10 個である．Recall が向上したモデルのうち，最も Recall が向上したモデルは，STL よりも 0.020 向上した MTL(Q1,3,4,5) である．STL の Recall と比べて MTL の Recall が低下したモデルは五つである．Recall が低下したモデルのうち三つは，客観サブタスクのみをサブタスクとして加えた MTL である．この結果から我々は，STL との有意差は小さいが，客観サブタスクが Recall の向上には繋がらない傾向を確認した．

構築した 16 種類のモデルの Recall を比較する．Recall 上位の三つのモデルは，上から順に MTL(Q1,3,4,5)，MTL(Q1,2)，MTL(Q1,2,4,5) である．Recall 下位の三つのモデルは，MTL(Q1,5)，MTL(Q1,3,4)，MTL(Q1,2,3,5) である．16 種類のモデルの Recall を比較するために，F 値の比較と同様に t 検定を行った．有意水準を 0.05 とすると，MTL(Q1,2,4,5) と MTL(Q1,3,4) の p 値 0.049 が有意水準を下回った．この結果から，主観サブタスク 2 と客観サブタスク 1,2 をサブタスクとして，誹謗中傷検出タスクに加えることは，主観サブタスク 2 と客観サブタスク 1 を加えることより Recall 向上に有効であることを確認した．

各評価指標の比較から，我々は以下の 2 点を明らかにした．

1. 主観サブタスクを加えた MTL は，客観サブタスクを加えた MTL に比べ，誹謗中傷検出タスクの F 値向上に有効な傾向がある．
2. 閲覧者感情分類タスク (主観サブタスク 2) を加えた MTL は，誹謗中傷検出タスクの Precision 向上に有効．

検出精度に違いが生じた投稿に対する考察

本節では，全タスクの各ラベルに対応するデータごとの Accuracy 比較について述べる．この比較の目的は，構築したモデルが，どのような誹謗中傷投稿に対して検出精度向上したのか，または低下したのか，傾向を調査することである．4.4.2 節で前述したとおり，F 値が最も高いモデルは MTL(Q1,3)，最も低いモデルは MTL(Q1,5) である．我々は，F 値に最も差が開いた MTL(Q1,3) と MTL(Q1,5) の比較，F 値が最も向上した MTL(Q1,3) とベースラインである STL(Q1) の比較を行った．我々は全てのタスクのどのラベルが付与されたデータの検出精度に，大

きな変化が起きたのか調査した．調査を行うことによって，我々が構築したモデルがどのような投稿の検出精度に向上に有効，または無効なのか検証することができると考えた．我々は検証のために，以下の式で全タスクの各ラベルに対応するデータごとの Accuracy を算出した．

$$A_{i,j} = \frac{N^c_{i,j}}{N_{i,j}} \tag{4.4.1}$$

i は質問番号，j は付与されるラベルである．$A_{i,j}$ は，質問 Qi に j のラベルが付与されたデータの Accuracy である．$N^c_{i,j}$ は，質問 Qi に j のラベルが付与されたデータのうち，誹謗中傷検出タスクを正解したデータ数である．$N^c_{i,j}$ は，質問 Qi に j のラベルが付与されたデータの総数である．我々は，帰無仮説を “比較する二つのモデルの $A_{i,j}$ の平均値に差はない” とし，2 群の対応のある t 検定を行った．検定は 10 分割交差検証で行い，有意水準は 0.050 とした．

■MTL(Q1,3) と STL(Q1) の比較　MTL(Q1,3) と STL(Q1) の $A_{i,j}$ を算出し，2 群の対応のある t 検定を行った．検定の結果, Q2(投稿者感情分類するタスクに対応する質問) に対して，“喜び” のラベルが付与されたデータで，我々は，MTL(Q1,3) の Accuracy:0.989 と STL(Q1) の Accuracy :0.980 の間に p 値 0.009 の有意差を確認した．この結果から，MTL(Q1,3) は STL(Q1) より喜びの投稿に対する誹謗中傷検出タスクの Accuracy 向上に，有効であることが確認できる．“喜び” のラベルが付与されたデータにおいて，STL(Q1) では，誹謗中傷検出タスクの予測を誤ったが，MTL(Q1,3) では予測が正しかった投稿の例を以下に示す．

例 1 コロス気かっ！ Thank You〜(正解ラベル: 誹謗中傷ではない)
例 2 え！？マジじゃん え！？やったーーーーーー！！！！！消した！！！！！死ね！！！！！ (該当ツイートに対して)(正解ラベル: 誹謗中傷)

例 1 は，おそらく冗談に対するツッコミで “コロス気かっ！” という表現を用いており，後半は “Thank You ” と感謝を述べているため誹謗中傷の投稿ではない．我々は，STL(Q1) は前半に着目して例 1 の投稿を “誹謗中傷” と分類したのではないかと考える．一方，MTL(Q1,3) は例 1 の投稿を “誹謗中傷ではない” と分類することができた．例 2 の前半は “え！？マジじゃん え！？やったーーーーーー！！！！！” であり，投稿者は喜びの感情を抱いていることが確認できる．例 2 の後半は “消

した！！！！！死ね！！！！！ (該当ツイートに対して)” と攻撃性の高い言葉が並んでいるため例 2 の投稿には “誹謗中傷” のラベルが付与されている．我々は，STL(Q1) は，前半に着目して “誹謗中傷ではない” と分類したのではないかと考える．一方，MTL(Q1,3) は例 1 の投稿を “誹謗中傷” と分類することができた．上記の例より我々は，投稿に “ポジティブ” な言葉と “ネガティブ” な言葉が混在している場合，STL(Q1) より MTL(Q1,3) の方が正しく分類しやすいのではないかと考える．

■MTL(Q1,3) と MTL(Q1,5) の比較　MTL(Q1,3) と MTL(Q1,5) の $A_{i,j}$ を算出し，2 群の対応のある t 検定を行った．検定の結果，Q2(投稿者感情分類タスクに対応する質問) に対して, “怒り” のラベルが付与されたデータで, 我々は, MTL(Q1,3) の Accuracy:0.717 と MTL(Q1,5) の Accuracy :0.709 の間に p 値 0.030 の有意差を確認した．この結果から，MTL(Q1,3) は MTL(Q1,5) より投稿者が怒りの感情を持った投稿に対する，誹謗中傷検出タスクの Accuracy 向上に，有効であることが確認できる．“怒り” のラベルが付与されたデータにおいて，MTL(Q1,5) では誹謗中傷検出タスクの予測を誤ったが，MTL(Q1,3) では予測が正しかった投稿の例を以下に示す．

例 1 死ね！？なんてこと言うの！ (正解ラベル: 誹謗中傷ではない)
例 2 お前に何がわかるん？そろそろいい加減にしろよまじで w(正解ラベル: 誹謗中傷)

例 1 は, 誰かから “死ね” と言われたことに対して, “なんてことを言うの！” と怒っており，誹謗中傷の投稿ではない．我々は，MTL(Q1,5) は “死ね” という単語に着目して，例 1 を “誹謗中傷” と分類したのではないかと考える．一方，MTL(Q1,3) は例 1 の投稿を “誹謗中傷ではない” と分類することができた．例 2 は，相手を煽りながら “いい加減しろよ” と怒っており，誹謗中傷の投稿である．我々は，MTL(Q1,5) は文末の “w” という表現に着目して，例 1 を “誹謗中傷ではない” と分類したのではないかと考える．一方，MTL(Q1,3) は例 2 の投稿を “誹謗中傷” と分類することができた．上記の例より，注意や相手に疑問を呼びかけるような投稿においては，MTL(Q1,5) より MTL(Q1,3) の方が正しく分類しやすいのではないかと考えることができる．

Q4(投稿者対象分類タスクに対応する質問) に対して，“他人, 他人のもの” のラベルが付与されたデータで，我々は，MTL(Q1,3) の Accuracy:0.810 と MTL(Q1,5) の Accuracy :0.805 の間に p 値 0.005 の有意差を確認した．この結果から，MTL(Q1,3) は MTL(Q1,5) より他人や他人のものに向けた投稿に対する，誹謗中傷検出タスクの Accuracy 向上に，有効であることが確認できる．“他人, 他人のもの” のラベルが付与されたデータにおいて，MTL(Q1,5) では誹謗中傷検出タスクの予測を誤ったが，MTL(Q1,3) では予測が正しかった投稿の例を以下に示す．

例 1 かわいいじゃないかー！！！死ねなんていっちゃメッ！！！ (正解ラベル: 誹謗中傷ではない)

例 2 ブスが同じダンスやってみろよ。。。ある意味流行るかも (正解ラベル: 誹謗中傷)

例 1 は，誰かが “死ね” と言ったことに対して，“死ねなんていっちゃメッ！！！” と注意しており，誹謗中傷の投稿ではない．我々は，MTL(Q1,5) は “死ね” という単語に着目して，例 1 を “誹謗中傷” と分類したのではないかと考える．一方，MTL(Q1,3) は例 1 の投稿を “誹謗中傷ではない” と分類することができた．例 2 は，“ブスが” や “ある意味” など抽象的な言葉を用いて誰かを皮肉しており，誹謗中傷の投稿である．我々は，MTL(Q1,5) は皮肉の表現を認識することができずに，例 1 を “誹謗中傷ではない” と分類したのではないかと考える．上記の例より，注意や皮肉の投稿においては，MTL(Q1,5) より MTL(Q1,3) の方が正しく分類しやすいのではないかと考えることができる．

Q5(投稿者対象分類タスクに対応する質問) に対して, “差別” のラベルが付与されたデータで，我々は，MTL(Q1,3) の Accuracy:0.415 と (Q1,5) の Accuracy :0.378 の間に p 値 0.033 の有意差を確認した．この結果から，MTL(Q1,3) は MTL(Q1,5) より差別の投稿に対する，誹謗中傷検出タスクの Accuracy 向上に，有効であることが確認できる．“差別” のラベルが付与されたデータにおいて，MTL(Q1,5) では誹謗中傷検出タスクの予測を誤ったが，MTL(Q1,3) では予測が正しかった投稿の例を以下に示す．

例 1 この国は無能な政治家によって運営されています。その多くは○○人だそうです。納得しました。(正解ラベル: 誹謗中傷)

例 2 多様性なんかいらんやろ。○○いい加減にしろ。倒産してほしい (正解ラベル: 誹謗中傷)

Q5 は誹謗中傷カテゴリ分類タスクのため，例の投稿は誹謗中傷の投稿のみである．例 1 例 2 の○○には具体的な言葉が入っていたが，本稿では○○と置き換えている．例 1 は政権批判，例 2 は会社批判であり，どちらも差別的な表現が含まれている．例より，MTL(Q1,5) の誹謗中傷カテゴリ分類のサブタスクが，差別表現の誹謗中傷検出タスクの Accuracy 向上を妨げたと考えることができる．

上記の結果より，我々は以下の 2 点を明らかにした．

1. 閲覧者感情分類タスク (主観サブタスク 2) を加えた MTL は STL に比べ，投稿者が喜びの感情を抱いた投稿に対する誹謗中傷検出タスクの Accuracy 向上に有効．
2. 加えたサブタスクによっては，加えたサブタスクの選択肢に該当する投稿の精度を低下させる可能性がある．

4.5 おわりに_B

我々は，誹謗中傷検出タスクを解くモデルの学習を行う際，主観サブタスクを同時に解くことによって，モデルの学習を補助することができ，誹謗中傷検出精度が向上するという仮説を立てた．仮説の検証を行うために，我々は主観サブタスク 2 種類と客観サブタスク 2 種類の計 4 種類のサブタスクを，追加するか追加しないかの組み合わせ，計 16 種類のモデルの誹謗中傷検出精度を比較した．比較の結果，以下の三つの知見を得た．

1. 主観サブタスクを加えた MTL は，客観サブタスクを加えた MTL に比べ，誹謗中傷検出タスクの F 値向上に有効な傾向がある．
2. 閲覧者感情分類タスクを加えた MTL は，誹謗中傷検出タスクの Precision 向上に有効．
3. 閲覧者感情分類タスクを加えた MTL は STL に比べ，投稿者が喜びの感情を抱いた投稿に対する誹謗中傷検出タスクの Accuracy 向上に有効．

今後の展望として我々は，サブタスクと実験回数を増やすことによる更なる傾向の調査や，学習時におけるサブタスクの損失への適切なバイアスを明らかにすること，選定するサブタスクと検出できる誹謗中傷の関係を明確化することが必要であると考える．

第 5 章　おわりに

本稿では，誹謗中傷検出タスクにマルチタスク学習を採用した際，検出精度向上に有効なサブタスクを明らかにすることを目的に二つの研究を行った．研究の結果，以下の五つの知見を明らかにした．

1. ランダムタスクを加えた MTL よりも，怒りの感情検出タスクを加えた MTL の方が誹謗中傷検出精度向上に有効．
2. サブタスクのラベルに約 20 倍以上の不均衡がある場合に限り，重み付き損失関数はサブタスクの検出精度向上に有効．
3. 主観サブタスクを加えた MTL は，客観サブタスクを加えた MTL に比べ，誹謗中傷検出タスクの F 値向上に有効な傾向がある．
4. 閲覧者感情分類タスクを加えた MTL は，誹謗中傷検出タスクの Precision 向上に有効．
5. 閲覧者感情分類タスクを加えた MTL は STL に比べ，投稿者が喜びの感情を抱いた投稿に対する誹謗中傷検出タスクの Accuracy 向上に有効．

今後の展望として，より有効なサブタスクの模索，損失関数の改善，選定するサブタスクと検出できる誹謗中傷の関係を明確化することが必要であると考える．

謝辞

本論文を作成するにあたり，指導教員である鈴木優准教授には，多くのご指摘と助言をいただきました．深く感謝いたします．修士から鈴木研究室に編入してきた私に，研究の楽しさと論文の難しさを教えていただいたこと忘れません．これからも後輩たちに，研究の楽しさと (助言多めで) 論文の難しさを教えてあげてください．

事務補佐員の佐野さん，井尾さんには事務的な手続きで大変お世話になりました．深く感謝いたします．研究室がにぎやかすぎて迷惑をおかけしたことが多々あるかもしれません．すいませんでした．大変だとは思いますが，これからも後輩たちの手続き作業頑張ってください．

かわいい B3 の後輩たち (尾関さん，田中くん，中村くん，まさたけ) には，後輩らしい溌剌とした元気をいただきました．深く感謝いたします．関わる期間は短かったですが，楽しくお話しできて嬉しかったです．これからも，元気を忘れず頑張ってください．

個性的な B4 の後輩たち (川上，城所，高橋，るいくん) には，いつも私のボケにツッコミを入れていただき，大変楽しい時間をいただきました．深く感謝いたします．良くも悪くも個性的なみなさんは，これからたくさん苦労すると思いますが，互いに協力しあって壁を乗り越えてください．これからも，個性的に頑張ってください．

優秀な M1 の後輩たち (エルゲン，太田さん，北村くん，桑原くん) には，優秀すぎる知識と考え方で何度も私の研究を助けていただきました．深く感謝いたします．あなたたちは，優秀すぎるので少し自重してください．後輩たちのハードルが上がりすぎてかわいそうです．これからは，後輩たちの手助けを頑張ってください．

騒がしい M2 の同期たち (小林くん，三島くん) には，いつもくだらない話に付き合っていただき，大変賑やかな時間をいただきました．深く感謝いたします．私がこの研究室で楽しく，賑やかに過ごせたのは間違いなく騒がしい M2 の同期がいたからだと思います．これからは別々の道に進みます．でもまあ，また飲みに行きましょう，絶対に．

皆様のおかげで本論文を執筆することができました．厚く御礼申し上げます．

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, Vol. 30, , 2017.

[2] 総務省. インターネット上の誹謗中傷情報の流通実態に関するアンケート調査結果, 2022. `https://www.soumu.go.jp/main_content/000813680.pdf`.

[3] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pp. 759–760, 2017.

[4] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pp. 88–93, 2016.

[5] Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, Viviana Patti, et al. Hurtbert: Incorporating lexical features with bert for the detection of abusive language. In *Proceedings of the fourth workshop on online abuse and harms*, pp. 34–43. Association for Computational Linguistics, 2020.

[6] Niloofar Safi Samghabadi, Parth Patwa, Srinivas Pykl, Prerana Mukherjee, Amitava Das, and Thamar Solorio. Aggression and misogyny detection using bert: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 126–131, 2020.

[7] Tiancheng Tang, Xinhuai Tang, and Tianyi Yuan. Fine-tuning bert for multi-label sentiment analysis in unbalanced code-switching text. *IEEE Access*, Vol. 8, pp. 193248–193256, 2020.

[8] Rich Caruana. Multitask learning. *Machine learning*, Vol. 28, No. 1, pp. 41–75, 1997.

[9] 山口真一. 炎上加担動機の実証分析. 2016 年社会情報学会 (SSI) 学会大会, 2016.

[10] Qiwei Bi, Jian Li, Lifeng Shang, Xin Jiang, Qun Liu, and Hanfang Yang.

Mtrec: Multi-task learning over bert for news recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2663–2669, 2022.

[11] Xuyang Wu, Alessandro Magnani, Suthee Chaidaroon, Ajit Puthenputhussery, Ciya Liao, and Yi Fang. A multi-task learning framework for product ranking with bert. In *Proceedings of the ACM Web Conference 2022*, pp. 493–501, 2022.

[12] Salima Lamsiyah, Abdelkader El Mahdaouy, Saïd El Alaoui Ouatik, and Bernard Espinasse. Unsupervised extractive multi-document summarization method based on transfer learning from bert multi-task fine-tuning. *Journal of Information Science*, Vol. 49, No. 1, pp. 164–182, 2023.

[13] 大友泰賀, 張建偉, 中島伸介, 李琳. いじめ表現辞書を用いた twitter 上のネットいじめの自動検出. *DEIM2020, C7-1, day2, p22*, 2020.

[14] Hind S Alatawi, Areej M Alhothali, and Kawthar M Moria. Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. *IEEE Access*, Vol. 9, pp. 106363–106374, 2021.

[15] Tommaso Caselli, Valerio Basile, Jelena Mitrovic, and Michael Granitzer. Hatebert: Retraining BERT for abusive language detection in english. *CoRR*, Vol. abs/2010.12472, , 2020.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[17] Xiangtao Zheng, Tengfei Gong, Xiaobin Li, and Xiaoqiang Lu. Generalized scene classification from small-scale datasets with multitask learning. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, pp. 1–11, 2021.

[18] Jacob Andreas, Dan Klein, and Sergey Levine. Modular multitask rein-

forcement learning with policy sketches. In *International conference on machine learning*, pp. 166–175. PMLR, 2017.

[19] Wassen Aldjanabi, Abdelghani Dahou, Mohammed AA Al-qaness, Mohamed Abd Elaziz, Ahmed Mohamed Helmi, and Robertas Damaševičius. Arabic offensive and hate speech detection using a cross-corpora multi-task learning model. In *Informatics*, Vol. 8, p. 69. MDPI, 2021.

発表リスト

[1] 沢田凌一，鈴木優『ソーシャルメディアにおける返信に着目した誹謗中傷ツイートの検出』東西関西データベースワークショップ 2022，2022
[2] 沢田凌一，鈴木優『誹謗中傷検出精度向上のためのマルチタスク学習におけるサブタスクの選定』第 15 回データ工学と情報マネジメントに関するフォーラム (DEIM2023)，2023
[3] 沢田凌一，鈴木優『重み付き損失関数を用いたマルチタスク学習における不均衡データ学習改善手法』東西関西データベースワークショップ 2023，2023
[4] 沢田凌一，鈴木優『マルチタスク学習における誹謗中傷検出精度向上のためのサブタスクの提案』WebDB 夏のワークショップ 2023，2023
[5] 沢田凌一，鈴木優『誹謗中傷検出タスクの学習を補助するためのサブタスクの提案』第 16 回データ工学と情報マネジメントに関するフォーラム (DEIM2024)，2024