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

沢田　凌一†　　鈴木　　優††

† 岐阜大学大学院自然科学技術研究科知能理工学専攻　〒 501–1193 岐阜県岐阜市柳戸 1 番 1
†† 岐阜大学工学部電気電子・情報工学科　〒 501–1193 岐阜県岐阜市柳戸 1 番 1
E-mail: †b4525045@edu.gifu-u.ac.jp, ††ysuzuki@gifu-u.ac.jp

あらまし　我々は誹謗中傷の検出精度向上のためにマルチタスク学習に着目した．マルチタスク学習とは，1 つのモデルで複数のタスクを行うことができる学習方法である．複数のタスク間で損失の重みを共有することでタスクの精度向上が期待できる．誹謗中傷の検出精度向上に効果があるサブタスクを選定するため，本研究では，2 種類のサブタスクに着目した．1 つ目は字面からわかる誹謗中傷の種類を分類するタスクである．誹謗中傷の種類が多いことによる定義の難しさが検出の難易度を上げているという問題に対して，誹謗中傷の種類を分類するタスクをサブタスクに選定することで検出精度の向上につながると考えた．2 つ目は誹謗中傷の原因となる投稿者の感情を分類するタスクである．誹謗中傷投稿者の多くは正義感や怒りの感情など特定の感情を抱き，自身の投稿を誹謗中傷だと自覚していないという特徴がある．そのため投稿者の感情を分類するタスクをサブタスクに選定することで検出精度の向上につながると考えた．シングルタスクモデル，サブタスクとして，字面からわかる誹謗中傷の種類を分類するタスクを加えたマルチタスクモデル，誹謗中傷の原因となる投稿者の感情を分類するタスクを加えたマルチタスクモデル，字面からわかる誹謗中傷の種類を分類するタスクと誹謗中傷の原因となる投稿者の感情を分類するタスクの両方を加えたマルチタスクモデルの検出精度を比較した．その結果，2 種類のサブタスクを別々に加えたモデル 2 つでは評価指標の向上が見られたが，両方を加えたモデルでは評価指標の向上が見られなかった．そのため，誹謗中傷検出タスクにおいてサブタスクは増やすほど検出精度が向上するわけではなく，サブタスクの種類や組み合わせによって検出精度が向上するか決まることを確認した．

キーワード　BERT，SNS，Twitter，誹謗中傷検出，マルチタスク学習

1　はじめに

SNS の普及に伴って誹謗中傷の増加が問題になっている．人手で誹謗中傷の投稿を削除できれば問題ないが，全ての投稿を一つ一つ確認して誹謗中傷を検出するにはあまりに時間と労力がかかりすぎる．そのため，誹謗中傷の自動検出に期待が寄せられている．誹謗中傷に関する禁止語句を設定し，禁止語句を含む投稿全てを自動検出してしまえば誹謗中傷の投稿はなくなるだろう．しかし，それでは表現の自由を損害してしまう．そのため，単語ではなく文章から誹謗中傷かどうか判断することが必要である．文章から誹謗中傷かどうか判断する方法として機械学習を用いた手法が考えられる．しかし，機械学習を用いた手法においてすべての誹謗中傷を検出するのは難しく，検出精度の向上が必要である．

検出精度向上には様々な手法が考えられるが，その 1 つにマルチタスク学習 [1] がある．マルチタスク学習は最も解きたいタスクである主タスクに加えて補助の役割を果たすサブタスクを同時に解くことで，共有された表現を獲得することができる．サブタスクの選定方法によって検出精度に大きな効果をもたらすため，有効なサブタスクを適切な数選定することが重要である．

本研究では，マルチタスク学習を用いた誹謗中傷検出タスクの精度を向上させるために最適なサブタスクの選定を目指す．我々は，誹謗中傷検出タスクにサブタスクとして選定できる関連の深いタスクが存在すると考えマルチタスク学習に着目した．誹謗中傷タスクと関連が深いと考えたタスクは 2 つある．1 つ目は，誹謗中傷の表現を分類するタスクである．誹謗中傷には表現が多いことによる定義の難しさが検出の難易度を上げているという問題がある．実際，誹謗中傷で罪に問われる際には名誉棄損罪，侮辱罪，信用毀損・業務妨害罪，脅迫罪の 4 つの罪がある．この誹謗中傷の表現は誹謗中傷を細分化したものであり誹謗中傷の検出と関連の深いタスクになると考えた．2 つ目は，誹謗中傷投稿者の感情を分類するタスクである．誹謗中傷の投稿者は，正義感や怒りの感情など自身を正当化する感情を抱き，自分の投稿を誹謗中傷だと自覚していないことが多いという特徴がある．実際に炎上事例に批判的なコメントをした投稿者のもつ動機を調査した結果，正義感が 6 割を超えたという調査 [2] もある．この誹謗中傷投稿者の感情は誹謗中傷の原因となるものであり誹謗中傷の検出と関連の深いタスクになると考えた．以上より，サブタスクの候補として字面からわかる誹謗中傷の種類を分類するタスクと誹謗中傷の原因となる投稿者の感情を分類するタスクの 2 つに着目した．

データの収集には Twitter API を用いた．収集したデータに

対して誹謗中傷かどうかの質問と着目したサブタスク 2 つについてそれぞれ 3 つずつ，計 7 つの質問を行いラベルをつけることでデータセットを用意した．誹謗中傷の定義は投稿相手を傷つけるものとした．投稿相手とはツイートを向けられた人，集団，モノのことである．“傷つける” の基準は客観的に見て，投稿相手を傷つける要素を含むかどうかとした．ツイートから取得できない前提を加えた場合のみ傷つくツイートは除いた．誹謗中傷とはならないツイートの例を記す．

例 1　身長低いね
例 2　○○さん魚嫌いらしいよ

例 1 は，投稿の受け手が低い身長にコンプレックスを抱いていた場合誹謗中傷となる恐れがあるが，抱いていなければ誹謗中傷とはならない．このような字面からわからない前提を踏まえなければ誹謗中傷とはならないツイートは本研究では検出しない．例 2 はこの情報が真実かどうか定かではない噂話である．もしこのツイートが真実ならば投稿の受け手は傷つかないだろうが，真実でなければ嘘を広められており傷つくだろう．本研究では，ツイートが真実かどうか判断できないものは検出はしない．ただし，“○○さん△△さんのこと嫌いらしいよ” のツイートは真実でもそうでもなくとも，受け手は人に知られたくない情報を広められているとして傷つく可能性が高いためこのようなツイートは検出することとする．

誹謗中傷ツイートの例を記す．

例 1　お前バカだな，ふざけんな！いい加減にしろ (直接的な表現)
例 2　○○さんってゴキブリみたいですよね (遠回しな表現)
例 3　タヒね，まじきえ ro(遠回しな語句)

例 1 は，直接的な表現であり明らかな誹謗中傷である．例 2 は直接的な表現は使用していないが文章全体を読むと明らかな誹謗中傷である．例 3 は語句として成立していないが，意味が通じる文章で明らかな誹謗中傷である．

用意したデータセットは訓練データ，検証データ，試験データに分割した．訓練データを用いてシングルタスクモデル，字面からわかる誹謗中傷の種類を分類するタスクをサブタスクとして加えたマルチタスクモデル，誹謗中傷の原因となる投稿者の感情を分類するタスクをサブタスクとして加えたマルチタスクモデル，字面からわかる誹謗中傷の種類を分類するタスクと誹謗中傷の原因となる投稿者の感情を分類するタスクの両方のサブタスクを加えたマルチタスクモデルの 4 つを構築した．学習には BERT [3] を使用した．BERT は双方向 Transformer によって文章理解を実現し，高い汎用性を持つ事前学習用の機械学習の手法である. 構築した分類器で試験データを用いて検出精度を比較した．

シングルタスクモデルと比べて字面からわかる誹謗中傷の種類を分類するタスクをサブタスクとして加えたマルチタスクモデルは，全ての評価指標で精度が向上する結果となった．誹謗中傷の原因となる投稿者の感情を分類するタスクを加えたマルチタスクモデルは，Accuracy が 4 つの分類器の中で最も高い数値を記録したが，重要な項目である Recall の数値がシングルタスクモデルと変わらない結果となった．字面からわかる誹謗中傷の種類を分類するタスクと誹謗中傷の原因となる投稿者の感情を分類するタスクの両方のサブタスクを加えたマルチタスクモデルは，全ての評価指標で精度がシングルタスクモデルと比べて変わらない結果となった．

この研究の貢献を以下に示す．

（1） 誹謗中傷検出タスクにおけるサブタスク選定において字面情報から予想しないとわからない感情より，字面をみて瞬間的にわかる情報の方が適していることを示した．
（2） マルチタスク学習においてサブタスクは増やすほど検出精度が向上するわけではなく，サブタスクの種類，数もしくは組み合わせによって検出精度が向上するか決まることを確認した．

2 関 連 研 究

本章では，本研究の関連研究を紹介する．誹謗中傷の検出を行う研究は多数行われている．誹謗中傷を検出するには，誹謗中傷文の特徴量を得る必要がある．特徴量を得る方法には，人手で誹謗中傷語を収集した辞書を作成し，文章に含まれる単語との関連性から特徴量を得る方法 [4] [5] と，BERT を用いて文章から特徴量を得る方法 [6] [7] がある．また，タスクの精度を向上させる手法の 1 つにマルチタスク学習がある. 誹謗中傷タスクに適応された事例は見当たらないが，様々なタスク [8] [9] において精度の向上が確認されている．BERT とマルチタスク学習を組み合わせることで，文章分類の精度が向上した研究 [10] もある．これらの関連研究から，BERT を用いた誹謗中傷検出タスクにマルチタスク学習を使用することで精度の向上が期待できると考えた．

独自で誹謗中傷に関係の深い特徴語を収集することで辞書を作成し、辞書に含まれる特徴語と文章の関連性から誹謗中傷を検出する関連研究を紹介する．磯野ら [4] は，不適切な投稿を検出し，書き改めるように勧めることを目的に誹謗中傷の検出を行っている. 誹謗中傷は表層的な語句を手がかりに検出できると考え，誹謗中傷の投稿に出現するワードの数から独自に特徴語を集め辞書を作成し SVM を用いて検出している．また，大友ら [5] は，ネットいじめの自動検出を目的に研究を行っている．いじめ文の投稿から単語ごとの TF-IDF を求めることで独自のいじめ表現辞書を作成し，6 種類の機械学習モデルで検出精度を比較している．このように人手で特徴語を収集する研究は多数行われている．

2 つの関連研究はどちらも課題として特徴量の算出方法や辞書に登録する単語の改善を挙げている．本研究では，遠回しな表現に対応するため独自の辞書を作成するわけではなく，事前学習済みの BERT を使用している．BERT を使用することで単語ではなく文章全体に着目して誹謗中傷の検出を行うことができ，単語に依存する問題を解決することができる．

機械学習を用いて特徴量を取り出す手法として，BERT を用いた表現抽出を利用し、分類タスクを行った関連研究を紹介す

る．Lee ら [6] は，BERT と自身らの提案した新たなデータ補強技術を組み合わせ，Twitter データセットを用いた皮肉検出で高い検出精度を達成している．このように BERT は言語表現の検出において高い検出精度を達成した事例が報告されている．また，Caselli ら [7] は，Reddit のコメントを集めたデータセット RAL-E を用いて BERT に英語の罵倒表現を学習させている．Reddit とは，アメリカ合衆国で広く使用されている掲示板型のソーシャルサイトである．その結果，罵倒表現を学習させる前の BERT より高い性能を示すことに成功している．本研究では，データ補強技術や事前学習させるわけではなくマルチタスク学習に着目して実験を行っている．

マルチタスク学習を用いることで精度向上を試みた関連研究を紹介する．Collobert ら [8] は，品詞タグ付け，チャンキング，固有表現認識などの 6 つの代表的な自然言語処理タスクにおいてマルチタスク学習を適応し，様々なタスクに適応可能であることを示している．特に SRL に関しては最先端の性能を達成したことを示している．また，Luong ら [9] は，Seq2Seq を用いた英語ドイツ語間の翻訳においてマルチタスク学習を用いることで大幅に精度向上を果たしたことを報告している．

誹謗中傷に関係のない分類タスクに BERT とマルチタスク学習の両方を利用した関連研究を紹介する．Samghabadi ら [10] は，非攻撃性のテキスト，間接的な攻撃性のテキスト，直接的な攻撃性のテキストの 3 クラス分類を行うタスクと，ジェンダー問題に対して肯定的か否定的かの 2 クラス分類を行うタスクの 2 つを同じ分類器で解くマルチタスク学習を行っている．分類器を構築する際の文脈の情報抽出に BERT を用いている．本研究では，同じく BERT とマルチタスク学習を用いているが，1 つのタスクを解くために複数のサブタスクを選定するという点で関連研究とは目的が異なっている．

以上の関連研究より単語辞書生成や BERT による誹謗中傷の自動検出，自然言語処理分野にマルチタスク学習を用いた研究は行われている．しかし，BERT とマルチタスク学習を組み合わせた誹謗中傷の自動検出，及び有効なサブタスクの選定は見当たらない．本研究では，マルチタスク学習の適応による精度向上が誹謗中傷検出タスクにも期待できるのではないかと考え，提案手法として BERT とマルチタスク学習を利用し誹謗中傷の自動検出，及び有効なサブタスクの選定を行う．

3 提案手法

本研究では，誹謗中傷検出タスクの精度向上を目的としてマルチタスク学習におけるサブタスクの選定を行う．サブタスクには主タスクと関連の深いタスクを選択するのが一般的である．なぜなら，関連の深いタスクの方が主タスクと並列に学習させる際に共通の表現を獲得しやすく，主タスクを解く助けになるからである．誹謗中傷と関連の深いタスクとして，字面からわかる誹謗中傷の種類を分類するタスクと誹謗中傷の原因となる投稿者の感情を分類するタスクの 2 つに着目した．2 つのタスクをサブタスクとして主タスクの学習に加えることで検出精度の向上を図る．以降，誹謗中傷かそうではないかを分類するタスクを主タスク，字面からわかる誹謗中傷の種類を分類するタスクをサブタスク A，誹謗中傷の原因となる投稿者の感情を分類するタスクをサブタスク B とする．提案手法の流れ，マルチタスク学習，サブタスクの定義と選定，提案モデルについて説明する．

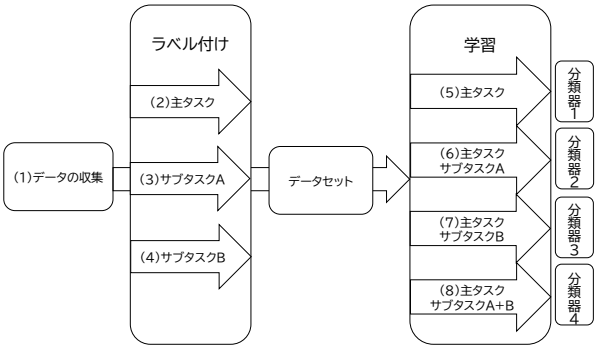

図 1　提案手法の概要図

提案手法の流れを以下に，概要図を図 1 示す．

（1） Twitter API を用いてツイートを収集する．
（2） 収集したツイートに対して主タスクについての質問をしラベル付けを行う．
（3） 収集したツイートに対してサブタスク A についての質問を 3 つしラベル付けを行う．
（4） 収集したツイートに対してサブタスク B についての質問を 3 つしラベル付けを行う．
（5） ツイートのテキスト情報を用いて主タスクを解く分類器を構築する．
（6） ツイートのテキスト情報を用いて主タスクとサブタスク A を解く分類器を構築する．
（7） ツイートのテキスト情報を用いて主タスクとサブタスク B を解く分類器を構築する．
（8） ツイートのテキスト情報を用いて主タスクとサブタスク A，サブタスク B を解く分類器を構築する．

3.1 マルチタスク学習

マルチタスク学習 [1] とは，1 つのモデルで複数のタスクを解くことができる学習方法である．タスクの名称は様々だが本稿では最も解きたいタスクを主タスク，主タスクを補助する役割のタスクをサブタスクと呼ぶこととする．タスクを並列処理することで，主タスクとサブタスクの共有された表現を獲得し主タスクの学習を助けることができる．この性質から関連の深いタスクをサブタスクに選定することで主タスクの精度向上が期待できる．

本研究では，誹謗中傷検出を主タスクとした際，関連の深いタスクとしてサブタスクに選定できるタスクが存在すると考えマルチタスク学習に着目した．このように考えた理由は 2 つある．1 つ目は，誹謗中傷の種類が多いことによる定義の難しさが検出の難易度を上げているという問題点に対して，サブタスクに誹謗中傷の種類を分類するタスクを選定することで検出精

度の向上につながると考えたからである．2 つ目は，誹謗中傷の投稿者は，正義感や怒りの感情など自身を正当化する感情を持っていることが多いため，自分の投稿を誹謗中傷だと自覚していないことが多いという特徴から，サブタスクに誹謗中傷の原因となる投稿者の感情を分類するタスクを選定することで検出精度の向上につながると考えたからである．

3.2 サブタスクの定義と選定

本研究では，字面からわかる誹謗中傷の種類と誹謗中傷の原因となる投稿者の感情の 2 つに着目した．字面からわかる誹謗中傷の種類に着目した理由について説明する．誹謗中傷の定義をすることは難しい．なぜなら，誹謗中傷は種類が多いからである．例えば誹謗中傷が罪に問われる際，名誉棄損罪，侮辱罪，信用毀損・業務妨害罪，脅迫罪の 4 つがある．また，1 章の誹謗中傷ツイートの例に示した通り，直接的な表現や遠回しな表現，遠回しな語句を用いた誹謗中傷が存在する．誹謗中傷を定義する際はその全てを網羅するようにしないといけない．そこで本研究では，字面からわかる誹謗中傷の種類を分類するタスクが誹謗中傷と関連の深いサブタスクとして機能し，検出精度向上に繋がるのではないかと考えた．実際のサブタスクには，罪名を問うとわかりにくいため，代表的な例の質問を行った．

誹謗中傷の原因となる投稿者の感情に着目した理由について説明する．誹謗中傷をやめるように呼び掛けても無くならない原因の 1 つに投稿者が自身の投稿を誹謗中傷と自覚していないことがあげられる．なぜなら，投稿者は悪気を持って誹謗中傷の投稿をする訳ではなく，集団心理における正義感や怒りの感情など特定の感情を抱いて誹謗中傷をすることが多いからである．そのため，投稿者の感情は誹謗中傷の原因となるものと言える．そこで本研究では，誹謗中傷の原因となる投稿者の感情を分類するタスクが誹謗中傷と関連の深いサブタスクとして機能し，検出精度向上に繋がるのではないかと考えたこの 2 種類のサブタスクの大きな違いは，字面からわかる情報か字面から予測しないとわからない情報かという点である．誹謗中傷の種類を分類するタスクには明確な答えがあり，字面をみて判断することができる．しかし，投稿者の感情は実際は投稿者しかわからず第三者はあくまで予測でしか判断することができない．そのため，ラベル付けを行う際にはラベル付けを行う作業者によってばらつきが生じる可能性がある．

サブタスクの質問と例を表 1 に示す．サブタスクは表 1 の質問に対して Yes か No を分類するタスクとした．サブタスク A-1 から A-3 までは字面からわかる誹謗中傷の種類を分類するサブタスク，サブタスク B-1 から B-3 までは誹謗中傷の原因となる投稿者の感情を分類するサブタスクである．

サブタスク A-1 を選定した理由を説明する．脅迫のツイートは誹謗中傷の中でも危険性が高い，直接的な表現が多い，出現回数が多いという考えから選定した．サブタスク A-2 を選定した理由を説明する．差別表現は誹謗中傷問題としてよく取り上げられている．言葉単体では誹謗中傷と判断しにくく，文章全体を理解しないと検出が難しいという点で他のタスクとは違った点を持つという考えから選定した．サブタスク A-3 を選定した理由を説明する．容姿の否定は字面としてわかりやすい，直接的な表現と間接的な表現を両方含む，受け手が傷つきやすい表現であるという考えから選定した．

サブタスク B-1 を選定した理由を説明する．正義感は感情としてわかりにくいという課題があるが，誹謗中傷の投稿者が抱く感情として多いものの 1 つであるという考えから選定した．サブタスク B-2 を選定した理由を説明する．怒りの感情は正義感同様誹謗中傷の投稿者が抱く感情として多いものの 1 つである，他人の感情であっても字面から想像しやすい感情であるという考えから選定した．サブタスク B-3 を選定した理由を説明する．失望の感情は相手を責めるのではなく投稿者の感情を投稿することで間接的に相手を傷つけるという点で他のサブタスクと違った点を持つという考えから選定した．

3.3 提案モデル

本研究では，BERT [3] とマルチタスク学習を組み合わせたモデルを使用している．BERT(Bidirectional Encoder Representations from Transformers) [3] とは，自然言語処理における事前学習用の機械学習モデルである．双方向の Transforemer によって，ラベルのないデータから高い汎用性を持つ事前学習モデルを構築できることや，単語ごとではなく文章全体を理解できるという特徴を持つ．実際にタスクを解く際には，最終層をタスクにあった形に微調整するだけで利用できるため様々な研究に使用されている．これらの特徴から BERT は単語ごとではなく，文章全体をみて誹謗中傷かどうか判断したい今回のタスクに適していると考えた．また本研究では，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 [1]を，誹謗中傷検出タスクにあった形に微調整して使用している．

BERT とマルチタスク学習は，図 2 のように組み合わせている．テキストを前処理したのち input として BERT に入力する．マルチタスク学習を用いて主タスクとサブタスクを BERT で並列に学習する．学習したそれぞれの出力の合計を softmax 関数を用いて 1 になるように調整する．その後 output で主タスクとサブタスクの損失の重みを統合することで，共通の要因を獲得することができる．また，主タスクとサブタスクの損失の重みの割合を変更することで，主タスクを解く際にどれほどサブタスクからの影響を加えるかを調整することができる．

4 評価実験

マルチタスク学習における誹謗中傷検出タスクの精度向上に適したサブタスクを選定するため実験を行った．実験は，シングルタスクモデルと次に説明するマルチタスク学習を用いた 3 種類のモデルを構築し比較を行った．1 つ目は，字面からわかる誹謗中傷の種類を分類するサブタスク A を加えたマルチタスクモデルである．2 つ目は，誹謗中傷の原因となる投稿者の感情を分類するサブタスク B を加えたマルチタスクモデルである．2 つのモデルは，着目した 2 つのサブタスクの効果に有用性があるかどうか調べるために構築した．3 つ目は，サブタス

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

表 1　サブタスクの質問

サブタスク	サブタスクの質問	例
A-1	このツイートは投稿相手を脅迫しているツイートですか? (脅迫とは投稿者が自らの行動によって投稿相手を危険にさらそうとすることの告知とする．)	殺すぞ，殴るぞ，監禁するぞ，秘密をばらしてやる，家を燃やすぞ
A-2	このツイートは投稿相手を差別しているツイートですか? (差別とは投稿相手の特定の思想や所属団体，属性に関して否定的な考えを押し付ける行為とする。)	女はわがままだな，男は馬鹿だな，○○人はきえろ
A-3	このツイートは投稿相手の容姿を否定しているツイートですか?	ブサイクだな，顔面終わってるな，服装ダサすぎ
B-1	このツイートの投稿者は正義感を持ってツイートしたと思いますか? (正義感を持ってとは自分が正しいもしくは投稿相手が間違っていると考え，考え押し通そうとする感情とする．)	○○党の方が優れているに決まっているだろ，誤字ってるよ馬鹿なの?
B-2	このツイートの投稿者は怒りの感情を持ってツイートしたと思いますか?	ふざけんな，お前むかつくなぁ
B-3	このツイートの投稿者は失望の感情を持ってツイートしたと思いますか?	なんかがっかりだわ、期待してたのにクソゲーやん

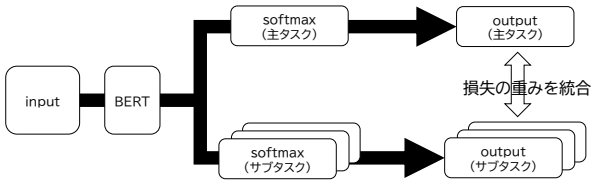

図 2　提案手法のモデル図

ク A，サブタスク B の両方を加えたマルチタスクモデルである．このモデルは，両サブタスクを同時に学習させることが有効である可能性があるため構築し比較を行った．本章ではデータセット，実験手順，結果，考察について記す．

4.1　データセット

データの収集方法には Twitter API を用いた．Twitter は，誹謗中傷の投稿が目撃される数が多い SNS の 1 つである [11]．そのため，本研究で扱うデータの収集に適していると考えた．対象ツイートは以下の条件で絞り込みを行った．

（1） 明確に投稿相手が定まっているツイートに絞るためメンション付きツイートのみを対象とする．
（2） リツイートや引用ツイートは本人の意思で述べておらず，本研究で検出したいツイートとは異なるため省くものとする．
（3） 本研究の手法では，テキスト情報のみを用いており，画像や動画を用いたツイートは考慮するとができないため省くものとする．

取得したツイートに対して主タスク，サブタスクの問いを行いラベル付けを行うことでデータセットを作成した．ラベル付けは，1 人 800 件を 14 人で行った．その結果，14 人の作業結果を合計し 11,200 ツイートのデータセットが作成されたが，その内 692 ツイートは作業者に内容が理解できないと判断され本研究では使用しなかった．その結果，データセットは 10,508 ツイートとなった．全ツイートのラベル付け結果を表 2 に示す．各タスクのラベル付けは全て独立に行った．つまり，1 つのツイートに対して主タスク，サブタスクを合わせた 7 つのラベルがついている．

表 2　全 10,508 ツイートに対するラベル付け結果

ツイートの種類	Yes	No
誹謗中傷であるか	2,862	7,646
脅迫の字面を含むか	494	10,014
差別の字面を含むか	409	10,099
容姿の否定の字面を含むか	382	10,126
正義感の感情を含むか	769	9,739
怒りの感情を含むか	1,238	9,270
失望の感情を含むか	399	10,109

4.2　実 験 手 順

データセットは訓練データ，検証データ，試験データを順に 8:1:1 に分割した．学習は (1) テキストのみを用いたシングルタスクモデル，(2) サブタスク A を加えたマルチタスクモデル，(3) サブタスク B を加えたマルチタスクモデル，(4) サブタスク A，サブタスク B の両方を加えたマルチタスクモデルの 4 つを行った．(2)，(3)，(4) においてはあくまでも主タスクの精度向上が目的なので，損失の重みを統合する際の割合を主タスクと全サブタスク比率が 6:4 となるように設定した．この設定によってサブタスクよりも主タスクを重視することができる．また学習の際には，訓練データの誹謗中傷ツイートの方が少ない問題を解消するため，誹謗中傷ツイートを複製し誹謗中傷でないツイートとデータ数を揃えた．実験環境を表 3 に記す．

また，得られたシングルタスクモデルの実験結果とマルチタスク学習を加えた提案手法 3 種類それぞれの実験結果に有意な差があるかどうか調べた．有意水準 0.05 で帰無仮説を “2 群の平均値に差はない” とし，対応のある 2 群の t 検定を行った．

4.3 結　　果

本研究では，評価指標に Accuracy，Precision，Recall，F-measure を用いた．いずれも誹謗中傷ツイートを正例としている．実験結果の評価指標を表 4 に混同行列を表 5，6，7，8 に示す．

サブタスク A を加えたマルチタスクモデルはシングルタスクモデルと比べ，Accuracy が 0.28，Precision が 0.54，Recall が 0.56，F-measure が 0.55 向上し，全ての評価指標でシングルタスクモデルを上回る結果となった．サブタスク B を加えたマルチタスクモデルはシングルタスクモデルと比べ Accuracy が 0.37，Precision が 0.41，F-measure が 0.19 向上したが，Recall はシングルタスクモデルと同じ数値となった．サブタスク A，サブタスク B を加えたマルチタスクモデルはシングルタスクモデルと比べ Precision が 0.08，Recall が 0.35，F-measure が 0.21 向上したが，Accuracy は 0.09 減少している．また，サブタスク A を加えたマルチタスクモデルとサブタスク B を加えたマルチタスクモデルよりもシングルタスクモデルとの差がない結果となっている．評価指標ごとに精度を比べると Accuracy はサブタスク B を加えたマルチタスクモデルが一番よく，Precision，Recall，F-measure はサブタスク A を加えたモデルが一番よい結果となった．よってサブタスク A を加えたマルチタスクモデルが一番誹謗中傷の見逃しも誤検知も少ないという結果となった．Accuracy のみサブタスク B を加えたマルチタスクモデルが上回っているのは，試験データの誹謗中傷と誹謗中傷でないもののデータ数にばらつきがあるためだと考えられる．実際に，混同行列を見ると数の多い誹謗中傷ではないツイートを正解している数は，サブタスク B を加えたマルチタスクモデルの方が多いが，数の少ない誹謗中傷のツイートを正解している数は，サブタスク A を加えたマルチタスクモデルの方が多くなっている．

また，シングルタスクモデルと提案手法それぞれの有意水準 0.05 における対応のある 2 群の t 検定の結果は，サブタスク A を加えたマルチタスクモデルの p 値が 0.882，サブタスク B を加えたマルチタスクモデルの p 値が 0.047，サブタスク A，サブタスク B を加えたマルチタスクモデルが p 値が 0.147 となった．この結果からサブタスク B を加えたマルチタスクモデルのみ p 値が 0.05 を下回り，帰無仮説が棄却され 2 群の平均値に差はないとは言えない結果となった．

以上より，提案手法とシングルタスクモデルと比べると，サブタスク A を加えたマルチタスクモデルは Precision，Recall は向上したが，t 検定の結果より有意差があるとは言えなかった．サブタスク B を加えたマルチタスクモデルは t 検定の結果より有意差があると言えたが，Precision，Recall は向上しなかった．サブタスク A，サブタスク B を加えたマルチタスクモデルは t 検定の結果より有意差があるとは言えず，Precision，Recall も向上しなかった．この結果よりサブタスクを増やすほど精度が向上するわけではない，字面からわかる誹謗中傷の種類を分類するタスクのみを加えたマルチタスクモデルのほうが，誹謗中傷の原因となる投稿者の感情を分類するタスクを加えたマルチタスクモデルよりも Presicion，Recall が向上する，検出精度に有意な差が認められるのは字面からわかる誹謗中傷の種類を分類するタスクのみを加えたマルチタスクモデルであることが確認された．

表 3　実験環境

バッチサイズ	32
エポック数	10,000
Earlystopping	100
損失関数	交差エントロピー誤差
最適化アルゴリズム	Adam

表 4　実験結果の評価指標

タスクの種類	Accuracy	Precision	Recall	F-measure
シングル	0.595	0.250	0.244	0.247
A	0.623	0.304	0.300	0.302
B	0.632	0.291	0.244	0.266
A+B	0.586	0.258	0.279	0.268

4.4 考　　察

サブタスク A を加えたマルチタスクモデルが全ての評価指標でシングルタスクを上回ったことは，本研究にとって望ましい結果となっている．しかし，サブタスクの選定には未だ改善の余地があり更なる改善が期待できる．サブタスク B を加えたマルチタスクモデルにおいて，Accuracy の評価指標は，全てのタスクの中で 1 番高い数値を記録し，t 検定の結果からシングルタスクモデルと有意な差があったが，Recall がシングルタスクモデルと比べて変わらない結果となっている．Recall は誹謗中傷の見逃しをはかる数値として重要視される項目である．そのため，有効なサブタスクとは言えない結果となった．この原因として，投稿者の感情は字面から予想しなければわからないもためあり，ラベル付けを行う人によってばらつきがあることが考えられる．これを解決するためには，定義を明確にすること，ラベル付けを行う際に多数決による投票方式を採用することが考えられる．サブタスク A，サブタスク B を加えたマルチタスクは，シングルタスクモデルと比べて精度は向上しなかった．この原因として，サブタスクを加えすぎた結果タスク同士の共通する要因が定まらなかったことが考えられる．

シングルタスクモデルと比べ Precision，Recall が上がったサブタスク A を加えたマルチタスクモデルのみに検出できたテキスト例を以下に示す．

例 1　うるせえよぶち殺すぞクソが
例 2　あの顔なんかムカつくからさ...
例 3　底意地が悪い。消えろ。

例より，投稿相手を脅迫をしているかの問いであるサブタスク A-1 の影響で例 1,3 が検出できたのではないかと考えられる．また，投稿相手の容姿の否定をしているかの問いであるサブタスク A-3 の影響で例 2 が検出できたのではないかと考えられる．

シングルタスクモデルと比べ検出精度に有意な差が認められたサブタスク B を加えたマルチタスクモデルのみに検出できた

混同行列	誹謗中傷と予測	誹謗中傷ではないと予測
実際に誹謗中傷	70	216
実際は誹謗中傷ではない	209	556

表 5　シングルタスクモデルの混同行列

混同行列	誹謗中傷と予測	誹謗中傷ではないと予測
実際に誹謗中傷	86	200
実際は誹謗中傷ではない	196	569

表 6　サブタスク A を加えたマルチタスクモデルの混同行列

混同行列	誹謗中傷と予測	誹謗中傷ではないと予測
実際に誹謗中傷	70	216
実際は誹謗中傷ではない	170	595

表 7　サブタスク B を加えたマルチタスクモデルの混同行列

混同行列	誹謗中傷と予測	誹謗中傷ではないと予測
実際に誹謗中傷	80	206
実際は誹謗中傷ではない	229	536

表 8　サブタスク A，サブタスク B を加えたマルチタスクモデルの混同行列

テキスト例を以下に記す．

例 1　ごみくずかす 4 ね

例 2　誰がハゲや 56 すぞ

例 3　もっと国民の底を引き上げて下さい。今の○○党では格差社会は広がっていくばかり。寄付に対する減税の幅を増やす事や弱者の負担が大きくなる消費税特に食料品や生活必需品の撤廃など弱者はエンゲル係数高い。経済成長していた昭和時代を見習ってほしい。国家の経済を家計簿同様と○○省もアホか

例より，感情を考慮したサブタスクを加えると直接的な関係はわからないが，例 1,2 のように数字を使った遠回しな表現に対応できた．また，投稿者が正義感を持っているかの問いであるサブタスク B-1 の影響で例 3 が検出できたのではないかと考えられる．

今後の改善点としてデータ数の不足やばらつき，ラベル付けの観点のばらつきが考えられる．特にデータ数のばらつきは誹謗中傷でないツイートが誹謗中傷ツイートの約 2.6 倍となり改善の余地があると考えられる．この問題はサブタスクに更に見られる．サブタスク A-1 は約 20.2 倍，サブタスク A-2 は約 24.6 倍，サブタスク A-3 は約 26.5 倍，サブタスク B-1 は 12.6 倍，サブタスク B-2 は 7.4 倍，サブタスク B-3 は 25.3 倍の差がある．このばらつきが原因でサブタスクとしての機能が完全ではなかった可能性がある．今後は，データのばらつきを解消するために誹謗中傷のデータを更に収集する，本研究で使用したサブタスクより誹謗中傷に関連のあるサブタスクを選定する，サブタスクは二値分類ではなく分類結果が均一になるような多クラス分類にすることを考えている．

5　おわりに

本研究では，SNS の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を行った．誹謗中傷の量は膨大で全てを人手で検出するのは不可能なため，誹謗中傷を減らしていくには自動検出精度の向上は必須である．そこで，自動検出精度の手法としてマルチタスク学習に着目した．なぜなら，誹謗中傷にはマルチタスク学習のサブタスクとして選定できる関連の深いタスクが存在すると考えたからである．関連するサブタスクを選定する際には，誹謗中傷の種類が多いことで定義が難しいこと，誹謗中傷の投稿者は正義感や怒りなどの特定の感情を抱いていることが多いことの 2 つの特徴に着目した．それぞれの特徴から字面からわかる誹謗中傷の種類を分類するタスクをサブタスク A-1 から A-3 に，誹謗中傷の原因となる投稿者の感情を分類するタスクをサブタスク B-1 から B-3 に選定した．

サブタスクは以下の質問に対して Yes か No を分類するタスクとした．

A-1　投稿相手を脅迫しているツイートですか?
A-2　投稿相手を差別しているツイートですか?
A-3　投稿相手の容姿を否定しているツイートですか?

B-1　投稿者は正義感を持ってツイートしたと思いますか?
B-2　投稿者は怒りの感情を持ってツイートしたと思いますか?
B-3　投稿者は失望の感情を持ってツイートしたと思いますか?

提案手法として，BERT とマルチタスク学習を組み合わせることで，シングルタスクモデル，サブタスク A を加えたマルチタスクモデル，サブタスク B を加えたマルチタスクモデル，サブタスク A，サブタスク B の両方を加えたマルチタスクモデルをそれぞれ構築し，検出精度の向上を図ることを提案した．

評価実験を行った結果，サブタスク A を加えたマルチタスクモデルは，シングルタスクモデルと比べて全ての評価指標で精度が向上する結果となった．しかし，対応のある 2 群の t 検定においてシングルタスクモデルと有意な差があるとは言えなかった．サブタスク B を加えたマルチタスクモデルは Accuracy が 4 つの分類器の中で最も高い数値を記録し，対応のある 2 群の t 検定においてシングルタスクモデルと有意な差が認められた．しかし，誹謗中傷の見逃しを判定する重要な項目である Recall の数値がシングルタスクモデルと変わらない結果となった．サブタスク A，サブタスク B の両方を加えたマルチタスクモデルは，全ての評価指標で精度がシングルタスクモデルと比べて変わらない結果となった．

この結果から，サブタスクの数は多ければ良いというものではなく，加えるサブタスクの内容，サブタスクの数もしくは組み合わせによって検出精度が向上するか決まることを確認した．

また，加えるサブタスクの内容によって検出できる誹謗中傷の種類が変わることを確認した．本研究の結果は暫定的なものであり，誹謗中傷検出タスクにおけるマルチタスク学習，サブタスクの有用性については今後更に追及する必要がある．各サブタスクと検出精度の関係性を明確にすることでより検出精度を向上させることのできるサブタスクの選定を行いたいと考えている．

謝辞　本研究の一部は JSPS 科研費 19H04218 および越山科学技術振興財団の助成を受けたものです．

文　献

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

[2] 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.

[3] 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.

[4] 磯野史弥, 松吉俊, 福本文代ほか. Web 掲示板における皮肉の分類および自動検出. 研究報告自然言語処理 (NL), Vol. 2013, No. 7, pp. 1–8, 2013.

[5] 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.

[6] Hankyol Lee, Youngjae Yu, and Gunhee Kim. Augmenting data for sarcasm detection with unlabeled conversation context. *CoRR*, Vol. abs/2006.06259, , 2020.

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

[8] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 160–167, 2008.

[9] Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[10] 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.

[11] 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`.