CE-007

感情分類と攻撃性推定を統合した特徴量による SNS 上の攻撃的投稿検知モデル

A Model for Detecting Offensive Posts on SNS Using Features Integrated with Sentiment Classification and Offensive Estimation

草野 雅也[1)]　佐久間 拓人[1)]　加藤 昇平[1)]
Masaya Kusano　Takuto Sakuma　Shohei Kato

1　はじめに

1.1　研究背景

近年，日本におけるソーシャルネットワーキング・サービス (以下，SNS と呼ぶ) の利用者数は年代問わず増加している．しかし，SNS の利用者の増加に伴い，インターネットの匿名性を悪用したヘイトスピーチや誹謗中傷などの不特定多数に対して攻撃的な内容の投稿 (以下，攻撃的投稿と呼ぶ) も増加しており，社会問題としてテレビや新聞などで取り上げられている．

この攻撃的投稿の検知に関する研究は，国内外で数多く報告されている．Mnassri ら [1] は，Twitter(現 X) 上で英語のヘイトスピーチや攻撃的な言葉を検知するために，感情情報を補助タスクとして導入した BERT 系ベースのマルチタスク型のモデルを提案し，ヘイトスピーチ検出タスクでは，単一データのみを学習したモデルよりも高い性能を示したが，攻撃的な言葉のタスクでは大きな性能向上は見られなかった．Plaza-Del-Arco ら [2] は，スペイン語のヘイトスピーチを対象として，感情情報と極性値を補助タスクとして取り入れたマルチタスク学習手法を用いることで，検出精度の向上を達成している．さらに，攻撃性の定義に着目した藤原ら [3] は，ソーシャルメディア上での円滑なコミュニケーションを促進するため，読み手や文脈によって攻撃的とも非攻撃的とも受け取れる発言（グレーゾーンの発言）を自動で検出する攻撃性検知モデルを構築している．

1.2　先行研究の課題点

Mnassri らの研究 [1] には，感情情報が性能向上にどの程度寄与しているかが不明確であるという課題が存在する．同研究では，モデルの共有エンコーダ部分で感情分類と攻撃性検出を同時に学習させた結果，感情情報の追加によって両タスクにおいて性能向上が見られたと報告されている．しかし，感情極性値やコサイン類似度などのテキストから抽出可能な他の特徴量，あるいは対象や皮肉といった人手で付与された特徴量との比較検証も必要であると述べられている．また，感情情報の活用によってヘイトスピーチおよび攻撃的言語の検出性能が向上することは示されているが，どの感情が特に性能向上に寄与したのかについての詳細な分析は実施されていない．これは，マルチタスク学習における共有エンコーダの構造上，個別の感情がどの程度モデル出力に影響を与えたかを明示的に評価することが困難であるためと考えられる．

1.3　本研究の概要

以上の課題を踏まえ，本研究では SNS 上の攻撃的投稿と特定の感情との間に一定の相関が存在するという仮説を立て，攻撃性推定モデルと感情分類モデルの出力を組み合わせた特徴量を用いてアンサンブル学習した攻撃的投稿検知モデルを提案する．提案モデルでは，SNS 上の投稿に対して，先行研究と同じ事前学習済みモデルから構築した攻撃性推定モデルから出力された攻撃性の推定確率と，感情分類モデルから出力された各感情の推定確率を結合することで，投稿の感情情報と攻撃性の関連を捉える特徴量を生成する．これらの統合した特徴量を LightGBM に入力し，最終的な攻撃性を判定する．

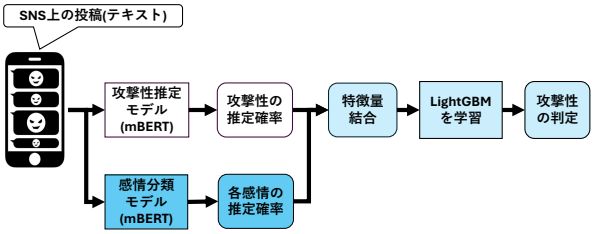


図 1: 提案モデル概観

提案モデルの性能評価は，Mnassri らの研究 [1] で構築されたモデルを比較対象とし，正答率，F1 スコア，および混同行列を用いて実施する．また，提案モデルでは，攻撃性と感情情報の特徴量がそれぞれ別のモデルによって出力されるため，感情と攻撃性との相関を定量的に評価できる．そのため，性能評価に加えて，攻撃性推定結果と感情分類結果を用いて混同行列を生成し，攻撃的投稿と特定の感情との相関について定量的に調査した．

2　提案手法

提案モデルを図 1 に示す．本研究で提案するモデルでは，事前学習済みの攻撃性推定モデルと感情分類モデルとして構築し，それらの出力を結合した特徴量をもとに攻撃性を判別する検知モデルを構築した．攻撃性推定モデルおよび感情分類モデルでは，BERT（Bidirectional Encoder Representations from Transformers）[4] をベースに，複数の言語の大規模なコーパスを用いて訓練された多言語対応の事前学習済み言語モデルである Multilingual BERT(mBERT) をファインチューニングすること構築した．各モデルにおいては，第 3 節で説明するデータセットを用いて，学習率 1×10^{-5}，エポック数 3，バッチサイズ 8 に設定して構築し，攻撃性推定モデルでは「攻撃的」または「非攻撃的」である 2 次元の推定確率を出力する．一方，感情分類モデルでは，喜び，悲しみ，怒り，驚き，恐れ，嫌悪，中立の 7 次元の感情における推定確率を出力する．これら 2 つのモデルの出力結果を統合し，全体で 9 次元の特徴量を作成する．統合後の特徴量は攻撃性と感情の特徴量のバランスを取る

1)　名古屋工業大学 大学院工学研究科 工学専攻 情報工学系プログラム
Computer Science Program, Dept. of Engineering, Graduate School of Engineering, Nagoya Institute of Technology

表 1: ハイパーパラメータの探索範囲

	探索範囲
learning_rate	[0.001, 0.1]
num_leaves	[8, 64]
max_depth	[5, 20]
min_data_in_leaf	[10, 50]
feature_fraction	[0.5, 1.0]

表 2: 再構成後の Goemotion のデータ分布

感情ラベル	件数
怒り (anger)	5199
嫌悪 (disgust)	601
恐怖 (fear)	651
喜び (joy)	17746
悲しみ (sadness)	2753
驚き (surprise)	4693
中立 (neutral)	14428
合計	48834

ために Softmax 関数を用いて正規化した．この特徴量を LightGBM に入力することで，投稿が「攻撃的」か「非攻撃的」かを二値分類する攻撃性検知モデルを構築した．また，提案モデルにおける LightGBM のパラメータは，Python ライブラリの最適化フレームワーク Optuna を用いてマクロ平均 F1 スコアが最良になるように 500 回探索して決定した．提案モデルにおける Optuna のパラメータの探索範囲を表 1 に示す．

3 データセット

3.1 Davidson データセット

提案モデルの性能評価と内部の攻撃性推定モデルの構築には，Davidson データセット [5] を使用した．このデータは，Twitter(現 X) からヘイトスピーチに該当するワードの辞書を用いて英語の投稿を収集し，3 人のアノテータによってヘイトスピーチ，攻撃的，どちらにも該当しないの 3 クラスに分類された約 24000 件のラベルが付与されている．

本研究では，攻撃性を検知するモデルの構築のために，「攻撃的（offensive)」，「どちらにも該当しない（normal)」のデータ約 23,000 件を使用した．本研究では，「どちらにも該当しない」を「非攻撃的」とし，このデータを各ラベルの比率を保ったまま，提案モデルの構築用データと内部の攻撃性推定モデル構築用データの 2 種類に分割して使用した．分割後の各データにおけるラベル分布は，攻撃的は約 9600 件，非攻撃的は約 2000 件である．

3.2 Goemotion データセット

感情分類モデルの構築には，Goemotion データセット [6] を用いた．このデータは，英語圏の掲示板サイト Reddit の投稿を約 58000 件収集し，「中立」を含めた計 28 種類の感情ラベルが手動で付与されている．

本研究では，先行研究と同様に，Ekman の基本 6 感情に中立の感情を加えた「怒り (anger)」「嫌悪 (disgust)」「恐怖 (fear)」「喜び (joy)」「悲しみ (sadness)」「驚き (surprise)」「中立 (neutral)」の計 7 種類の感情に再構成して使用した．再構成後のデータ分布を表 2 に示す．

4 性能評価実験

4.1 データ前処理

本研究では，Davidson データセットに対して Mnassri らの研究 [1] と同様に，Ekphrasis ライブラリ [7] を用いてデータの前処理を実施した．最初にすべてのテキストを小文字に変換し，URL やメールアドレス，ユーザ名・メンションを削除した．次に，「yeeessss」のような引き伸ばされた文字列を「yes」に正規化し，ストップワードは削除せずに保持した．そして，句読点や不要な区切り文字を除去し，ハッシュタグ記号（#）を削除したうえで，テキスト部分を補正（例：「#notracism」→「not racism」）した．最後に，2 単語未満のツイートは除去し，絵文字についてもすべて削除した．

4.2 実験設定

提案モデルの性能を評価するため以下の 3 種類のモデルを比較することで性能評価を実施した．

Proposed: 攻撃性タスクと感情タスクのそれぞれ独立のモデルを構築した提案モデル．

BERT MTL[1]: BERT の共有エンコーダ部分で両タスクも共有しながら同時に学習したモデル．

mBERT MTL[1]: mBERT の共有エンコーダ部分で両タスクも共有しながら同時に学習したモデル．

評価指標には，正答率（Acc.)，適合率（Pr.)，再現率（Recall)，マクロ平均 F1 スコア（F1-m)，重み平均 F1 スコア (F1-w) を用いた．また，BERT MTL および mBERT MTL では，訓練データの全体を用いた単一評価している一方で，Proposed モデルは，感情分類モデルと攻撃性分類モデルを個別に構築し，それぞれの出力を統合しているため，使用可能なデータ量が半分であることを考慮し，Proposed モデルについては層化 5 分割交差検証を用いて，性能評価実験を実施した．

4.3 実験結果

表 3 に性能評価の結果を示す．BERT MTL および mBERT MTL は先行研究で示された結果，Proposed は交差検証に用いた評価のため，全フォールドの平均値で結果を示している．全ての評価指標において，Proposed モデルは先行研究のモデルと同等の精度，もしくはわずかながらの精度向上が見られた．また，Proposed における正解ラベルと予測ラベルの混同行列を作成し，先行研究で比較対象とされていた BERT MTL との比較を実施した．mBERT MTL については，先行研究において混同行列による比較が行われていなかったため，本研究では対象外にした．その混同行列を図 2 に示す．「非攻撃的」の正答率は BERT STL では 88.41%だったものが，Proposed では 94.44%まで向上した．さらに，Proposed モデルでは「非攻撃的」の投稿を「攻撃的」と誤検出した割合が 11.59%から 5.56%へと減少しており，「非攻撃的」の検知精度が先行研究のモデルよりも向上したことが確認された．

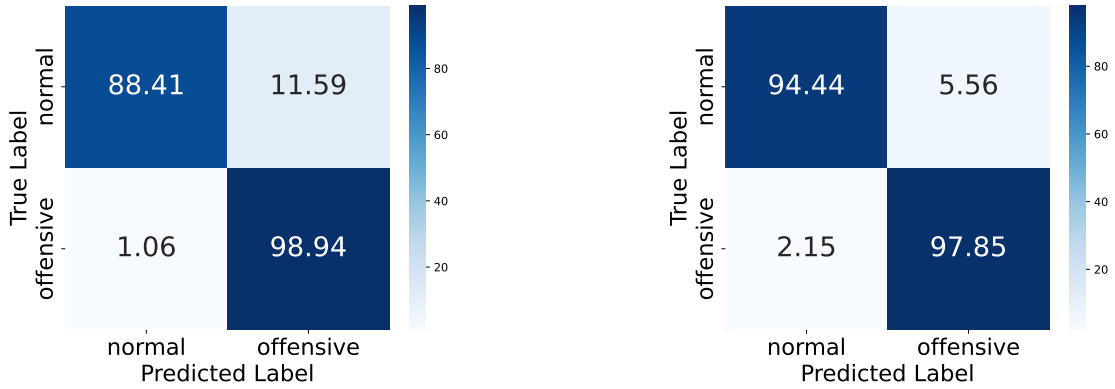
以上の実験結果より，Proposed モデルは，特に非攻撃的投稿の検知精度が向上したことで，先行研究モデルより高い性能であることが示された．

4.4 考察

今回の実験結果より，アンサンブル学習においても検知精度が向上したことから，感情情報が攻撃的投稿の検知において有効な特徴量であることが，異なるモデル構

表 3: 各モデルの性能評価結果

	Acc.	Pr.	Recall	F1-m	F1-w
BERT MTL[1]	0.9691	0.9367	0.9625	0.9489	0.9695
mBERT MTL[1]	0.9674	0.9346	0.9558	0.9459	0.9678
Proposed	0.9724	0.9878	0.9785	0.9538	0.9727



（a）BERT MTL [1]　　（b）Proposed

図 2: 混同行列 (性能評価実験)

造においても一貫して確認された．特に「非攻撃的」投稿の検知精度の向上が大きく，非攻撃的な投稿が誤って削除されるリスクを低減できることが期待される．非攻撃的な投稿の誤検出が減少することは，ユーザからの信頼獲得に繋がるので，実際の SNS プラットフォームにおいて重要であると考えられる．このことから，今回比較した 3 つのモデルの中では，実際の SNS 上で常時稼働させる運用を想定した場合において，Proposed モデルが他の 2 モデルよりも実用性が高いと考えられる．一方で，多少の誤検知を許容してでも攻撃的投稿の見逃しを最小限に抑え，SNS 全体の安全性を優先する運用方針をとる場合には，他の 2 モデルの方が Proposed モデルよりも適している可能性が考えられる．

また，Proposed が他の 2 つのモデルよりもわずかに高い性能を示した要因の一つとして，パラメータの最適化の有無が考えられる．先行研究では，学習率やバッチサイズといったハイパーパラメータを固定値で設定していたのに対し，本研究では LightGBM に対して Optuna によるパラメータ最適化を実施しており，これが Proposed モデルの性能，特に再現率の向上に寄与したと考えられる．ただし，今回比較対象とした先行研究のモデルにおいても同様に Optuna による最適化を適用すれば，Proposed と同等の精度が得られる可能性も十分に考えられる．

以上の結果から，Proposed モデルにおいて，感情情報は特に非攻撃的な投稿の検知に有効な特徴量であることが示唆された．特に非攻撃的な投稿の検知において精度が向上した点については，誤って投稿が削除されないことによるユーザの信頼や，SNS における表現の自由の確保という観点では非常に重要であると考えられる．

5　感情相関調査

5.1　調査方法

性能評価実験で，攻撃性検知において感情情報の有効性が示唆されたことで，投稿の攻撃性と各感情との相関を混同行列を用いた調査も実施した．この調査では，正解ラベルと感情ラベル，予測ラベルと感情ラベルの 2 種類の混同行列を使用した．この混同行列は各攻撃性ラベルごとに，感情分類モデルから出力された各感情の推定確率を合計し，正規化することで作成した．この混同行列の各要素は，以下の式により算出される．ここで，$C_{a,e}$ は攻撃性ラベル a に対する感情ラベル e の混同行列の要素，I_a は攻撃性ラベルが a に分類される投稿の集合，$N_a = |P_a|$ は攻撃性ラベル a に属する投稿の総数，$q_{i,e}$ は投稿 i に対する感情分類モデルによる感情 e の出力確率を示している．

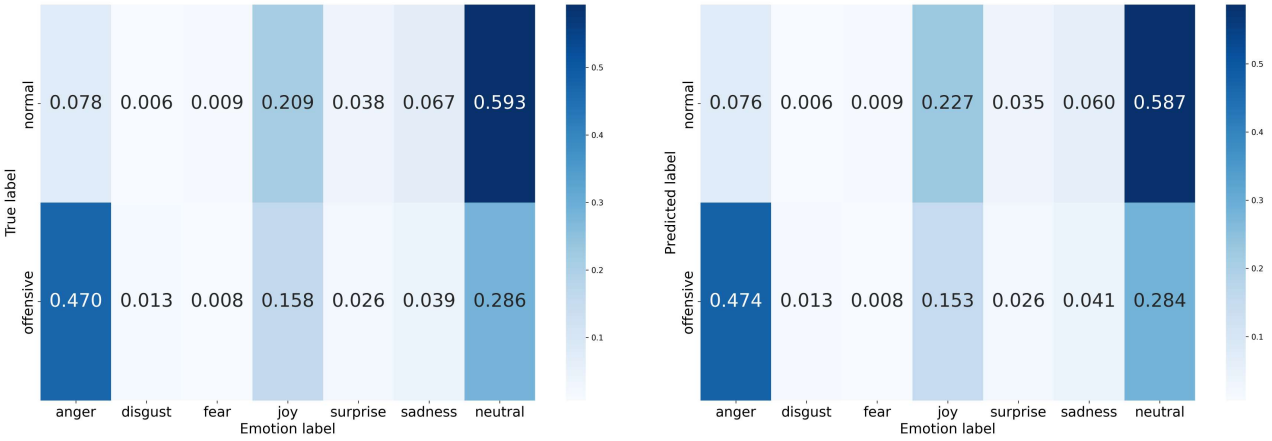
$$C_{a,e} = \frac{1}{N_a} \sum_{i \in P_a} q_{i,e} \tag{1}$$

5.2　結果と考察

攻撃性と感情の混同行列を図 3 を示す．正解ラベルと予測ラベルどちらに混同行列においても，「攻撃的」においては「怒り (anger)」の割合がともに約 50%を占めており，「非攻撃的」においては「中立 (neutral)」が約 60%を占めていた．他の感情においても差異がないことから，感情情報がモデルにとって攻撃性の判定をするうえで有効な情報として機能している可能性が示唆される．

一方で，感情情報が攻撃性の識別に有効であるが，単一の感情カテゴリが一意に攻撃性を決定づけるわけではないことも明らかとなった．特に，「喜び」や「中立」の感情がどちらの攻撃性ラベルにも一定の割合で含まれており，これは投稿に含まれる皮肉や冗談が感情分類モデルの出力に影響を与えている可能性が考えられる．

以上の結果から，感情情報は攻撃性の判定に一定の寄与を示すが，感情と攻撃性の相関は単純ではなく，文脈や表現の多様性も影響を及ぼしている可能性が考えられる．本研究では英語を対象にしたが，日本語やスペイン語などの多言語の場合は，もっとこの影響が大きく，別の相関が現れる可能性が十分に考えられる．このことから，感情や文脈的情報の統合によって明示的な攻撃性の



（a）正解ラベル　　（b）予測ラベル

図 3: 混同行列 (感情相関調査)

推定確率に依存せずに攻撃性を捉えることが可能ではないかと考えられる．

6　今後の展望

今後の展望として，他特徴量を用いたときの検証と，多言語データでの検証の 2 つが挙げられる．

他特徴量を用いたときの検証については，本研究では感情情報が性能向上にどの程度寄与しているかが不明確であるという課題点に着目し，攻撃性と感情情報の特徴量を抽出しやすい Proposed モデルを構築した．しかし，今回の性能評価実験では感情情報以外の特徴量との比較検証は十分に実施されていない．なので，今後は感情極性値や攻撃的言語の出現頻度など，テキストから自動で抽出可能な特徴量や，攻撃対象や皮肉性といったラベリング時に付与される特徴量と比較することで，今回使用した感情情報が他の特徴量と比べてどの程度有効であるかを明らかにする必要がある．

また，多言語データでの検証については，多言語でも本研究と同様の手順で，性能評価実験と感情相関調査を実施し，Proposed モデルの有効性と多言語における投稿の攻撃性と感情情報の相関を調査する必要がある．この多言語での検証において，攻撃性と感情情報との相関が本研究の結果と類似している場合，多言語に応用可能な攻撃的投稿検知モデルの提案が可能になるだろうと考えられる．さらに，時代とともに表現や言語使用が変化する SNS の特性を踏まえると，多言語かつ感情情報を考慮した攻撃的投稿検知モデルは，表現の多様性に柔軟に対応可能な実用的なモデルになると期待できる．

7　おわりに

本研究では，SNS 上の攻撃的投稿と個別の感情には何かしらの相関があるという仮説より，感情分類モデルと攻撃性推定モデルの出力を統合した特徴量を用いた攻撃的投稿検知モデルを提案した．事前学習済みの mBERT を用いて構築した 2 つのモデルから得られた推定確率を統合し，LightGBM による最終的な判定をすることで，攻撃性と感情の両方の情報を考慮した特徴量を用いて判定する Proposed モデルを開発した．Proposed モデルを先行研究のマルチタスク学習型の攻撃性検知モデルと比較した結果，すべての評価指標において同等またはわずかに上回る性能を示した．特に，「非攻撃的」の投稿の誤検出が減少し，誤検出されるリスクが軽減されることが確認された．また，感情情報と攻撃性ラベルの混同行列を用いた分析により，特定の感情と攻撃的投稿との相関が定量的に示され，感情情報が攻撃性判定において有効な特徴量であることが示唆された．今後は，感情以外の特徴量との比較検証や，多言語データへの適用を実施することで，より汎用性の高い攻撃的投稿検知モデルの構築を目指す．

謝辞

本研究は，一部，文部科学省科学研究費補助金（課題番号 JP24H00741），ならびに，国立研究開発法人情報通信研究機構委託研究の助成により行われた．

参考文献

[1] Mnassri, K., Rajapaksha, P., Farahbakhsh, R. and Crespi, N.: Hate Speech and Offensive Language Detection Using an Emotion-Aware Shared Encoder, pp. 2852–2857 (2023).

[2] Plaza-Del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A. and Martín-Valdivia, M. T.: A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis, *IEEE Access*, Vol. 9, pp. 112478–112489 (2021).

[3] 藤原知樹, 伊藤彰則, 能勢隆：ソーシャルメディア上の発話の攻撃性推定と会話補助, 言語処理学会第 30 回年次大会発表論文集, pp. 505–510 (2024).

[4] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: 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 (2019).

[5] Davidson, T., Warmsley, D., Macy, M. and Weber, I.: Automated hate speech detection and the problem of offensive language, Vol. 11, No. 1, pp. 512–515 (2017).

[6] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G. and Ravi, S.: GoEmotions: A dataset of fine-grained emotions, *arXiv preprint arXiv:2005.00547* (2020).

[7] Baziotis, C., Pelekis, N. and Doulkeridis, C.: Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis, in *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pp. 747–754 (2017).