新型コロナウイルス感染症に関連する偏見・差別ツイート判別手法の提案

髙山 勝彦 † 熊本 忠彦 ‡

†, ‡ 千葉工業大学 情報科学部 情報ネットワーク学科

1. はじめに

法務省・人権擁護局の人権啓発サイト[1]では，新型コロナウイルス感染症の蔓延に伴い，感染者や医療従事者，エッセンシャルワーカーへの偏見や差別が多く見られることや，こういった偏見や差別が人々の思い込みや不安により生じ，無自覚になされうることが指摘されている．このような偏見や差別は「コロナ差別」と呼ばれており，医療従事者やエッセンシャルワーカーの離職に加え，PCR 検査を避けたり感染を隠したりする人の増加を引き起こす可能性も指摘されている．

また，こういったコロナ差別は，Twitter のような匿名性の高い SNS 上で起こりやすく，社会問題にもなっているが，未だ有効な解決策は政策的にも技術的にも見出されていない．

そこで本稿では，Twitterに投稿されたツイートが新型コロナウイルス感染症に関連する偏見や差別を含むか否かを判別する手法を提案することで，偏見ツイートや差別ツイートをフィルタリングしたり，可視化したりすることの実現可能性を示す．具体的には，自然言語処理に適した深層学習手法である BERT[2]とその文書分類モデルである BertForSequence Classification[3]（本稿では BERTFSC と略す）を用いて 2 つの手法を実装し，その精度を比較することで，提案手法の有効性を検証する．なお，BERT も BERTFSC も実装には Hugging Face から提供されている Transformers ライブラリを用いる．

2. 偏見・差別ツイート判別手法の提案

2.1 深層学習に用いる訓練/テストデータの準備

本稿では，深層学習のための訓練/テストデータとして新型コロナウイルス感染症に関連する偏見や差別を含んだツイートと含んでいないツイートを必要とする．そこで，まず，「医療従事者　近寄りたくない」や「ワクチン未接種者　迷惑」といったキーワードを含むツイートを収集し，合計で 2,010 件のツイートを収集した．

次に，男女 15 人（男性 7 人，女性 7 人，不明 1 人）が参加するアンケート調査を行い，この 2,010 件のツイートのそれぞれに対し新型コロナウイルス感染症に関する偏見や差別を感じるかどうかを「感じる(3)」，「少し感じる(2)」，「あまり感じない(1)」，「全く感じない(0)」の 4 段階で評価してもらった．その結果得られたデータの平均をツイート毎に求め，各ツイートの偏見度・差別度とした．

また，偏見度・差別度が中央値以上のツイートを正例，中央値未満のツイートを負例として用いることとした．偏見に関するデータの中央値は 1.53 であり，正例が 1,021 件，負例が 989 件となった．一方，差別に関するデータの中央値は 1.13 であり，正例が 1,084 件，負例が 926 件であった．ここで，偏見度・差別度が最も高かったツイートと最も低かったツイートを抽出し，表 1 に示す．なお，同率があったため，表 1 には複数のツイートが表示されている．

表 1. 偏見度・差別度が最も高かった/低かったツイート

偏見	差別	ツイート
2.80	0.93	茶番コロナ以前より、軍産複合体西側＆米ディープステート達に依る処の ポイズン迷惑ワクチンビジネスは、迷惑度を益々深化させ、人間を家畜やモルモット扱いし、自分達の利益優先に世界的テロを展開し現在に至る。(ｰ`дｰ´)ｷﾘｯ
0.20	0.20	このご時世ですから、禁止されてる出待ち入り待ちは、本当にみんな守って欲しいです 無症状の感染者が、知らないうちに選手や関係者に近寄らないようにして欲しいですね
0.20	0.13	コロナで本当に苦しんでいる人に政府は支援を差し伸べて欲しい
1.73	2.73	おたくらのボスが県外から来るなとほざいているよ。島根県なんて日本には要らないね
0.33	0.07	コロナで国民がみんながんばってます。一人一人がどーあがいても結局声は届かないよねだから国がなんとか頑張ってくれないともうみんないっぱいいっぱいですよ。
1.27	0.07	コロナはこれからも続く病気だろう。そのうち、風邪やインフルエンザのようなありふれた病気になるだろう。

偏見に関するデータと差別に関するデータのそれぞれにおいて，正例から 300 件，負例から 300 件の計 600 件を無作為に抽出し，テストデータとした．残りの 1,410 件が訓練データとなる．

2.2 BERT と BERTFSC の実装

BERT と BERTFSC の実装には，Python 向けの深層学習ライブラリである PyTorch[4]を用いた．このとき，BERT と BERTFSC に共通の設定として，トークナイザには Bert JapaneseTokenizer（MeCab と WordPiece を併用）を用い，事前学習済みの日本語 BERT モデルには東北大学の乾研究室から提供されている cl-tohoku/bert-base-japanese を用いた．

Proposal of a Method for Detecting Prejudice and Discrimination Related to COVID-19 on Twitter

† Katsuhiko Takayama, Chiba Institute of Technology

‡ Tadahiko Kumamoto, Chiba Institute of Technology

なお，BERT/BERTFSC では 512 トークンを超える長さの入力文字列は扱えないため，512 トークンを超える部分は削除した．また，入力文字列に対する前処理として，絵文字やハッシュタグは削除した．これは絵文字やハッシュタグがノイズになると考えたためである．加えて，1ツイート＝1行になるようにツイート内の余分な改行を削除した．

2.3 BERT/BERTFSC モデルのファインチューニング

BERT モデルと BERTFSC モデルのファインチューニングを 3.1 節で準備した訓練データ（1,410 ツイート分）を用いて行った．このとき，バッチサイズは 4（BERT）と 3（BERTFSC）とし，損失関数にはクロスエントロピー誤差を，最適化関数には確率的勾配降下法を採用し，学習率は 0.0005 とした．また，エポック数は 50 エポックとした．

3. 精度評価

3.1 節で得たテストデータ（600 ツイート分）を用いてファインチューニングした各モデルの精度（正解率）をエポック数毎に求めた．結果を図1に示す．なお，図中，黒線が BERT モデルを用いて差別データに関する正解率を求めたもの，青線が BERTFSC モデルを用いて差別データに関する正解率を求めたもの，赤線が BERT モデルを用いて偏見データに関する正解率を求めたもの，オレンジ線が BERTFSC モデルを用いて偏見データに関する正解率を求めたものとなっている．

図 1 によれば，それぞれのモデルにおいて最も高い正解率は，偏見データに対しては，〔BERT モデル〕エポック数が 41 のとき，70.5%，〔BERTFSC モデル〕エポック数が 34 のとき 70.7%であり，差別データに対しては，〔BERT モデル〕エポック数が 30 と 35 のとき，73.2%，〔BERTFSC モデル〕エポック数が 22 のとき，74.2%であった．偏見データと差別データのいずれに対しても BERTFSC の方が正解率は高い結果となったが，モデル間の差は小さかった．

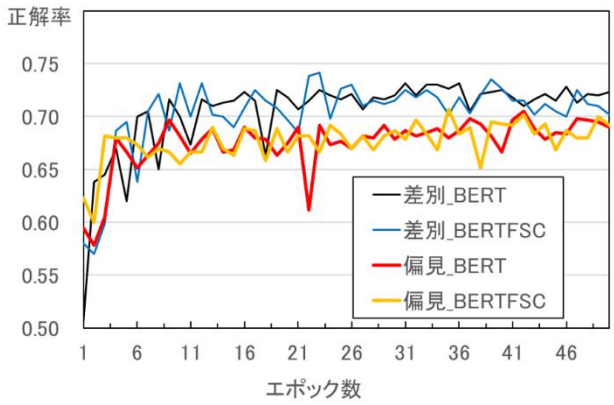

図 1 偏見・差別モデルの精度比較

一方，図1において偏見データと差別データに対する正解率をそれぞれのモデルにおいて比較してみると，どちらのモデルにおいても偏見データの方が低い結果となっている．この原因としては差別を生じさせるツイートより偏見を含んでいるツイートの方が多様な表現であり，内容にあまり一貫性がないことが考えられる．

ここで，偏見データと差別データのそれぞれから正例 50 件，負例 50 件を無作為に抽出し，BERT モデルと BERTFSC モデルとで出力結果が異なるものを調べてみた．結果，偏見データに対しては 15 件，差別データに対しても 15 件が抽出された．

偏見データに対しては，15 件中 7 件が BERT モデルの不正解であり，うち 6 件が負例を正例と誤判断したもの（False Positive）であった．一方，15 件中残りの 8 件は BERTFSC モデルの不正解であり，うち 5 件が正例を負例と誤判断したもの（False Negative）であった．また，差別データに対しては，15 件中 9 件が BERT モデルの不正解であり，うち 7 件が負例を正例と誤判断したもの（False Positive）であった．一方，9 件中残りの 6 件が BERTFSC の不正解であり，うち5件が負例を正例と誤判断したもの（False Positive）であった．したがって，偏見データに対しては，モデルによって傾向が異なったが，差別データに対しては，いずれのモデルでも差別ありと誤判断したもの（False Positive）が多かったと言える．

なお，本研究では偏見や差別を感じそうなツイートを広く収集し，アンケート調査の結果に基づいて正例と負例に分けたが，絶対数が少ないうえ，正例と負例の数がアンバランスになってしまった点も正解率が低い原因の一つと考えられる．

4. まとめ

本稿では，新型コロナウイルス感染症に関連する偏見や差別がツイートに含まれているか否かを判別する手法を提案した．精度評価により，偏見に関しては 65%以上の精度で，差別に関しては 70%以上の精度で判別できることを示した．

提案手法の精度はまだ実用レベルとは言えないが，訓練データを増やしたり，正例数と負例数のバランスをとったり，対象に適した深層学習手法を導入したりすることで，実用レベルでの実装も十分可能と考えられる．今後の課題としたい．また，そのような手法を取り入れることで，偏見ツイートや差別ツイートをフィルタリングしたり，可視化したりすることも可能と考えられる．今後，そのようなシステムの開発を目指していきたい．

謝辞

本研究の一部は，JSPS 科研費 20K12085 により実施されている．ここに記して深く感謝の意を表す．

参考文献

[1] 法務省・人権擁護局，新型コロナウイルス感染症に関連して－差別や偏見をなくしましょう，https://www.moj.go.jp/JINKEN/jinken02_00022.html（2023 年 1 月 8 日閲覧）．

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv:1810.04805v2, 2018.

[3] Hugging Face, https://huggingface.co/docs/transformers/model_doc/bert#bertforsequenceclassification（2023 年 1 月 8 日閲覧）．

[4] 新納浩幸，PyTorch 自然言語処理プログラミング，インプレス，東京，2021.