D-018

ツイートを用いたユーザの持つバイアスの推定
Unconscious Bias Detection of Twitter Users

三島 惇也[1)]　鈴木 優[1)]
Junya Mishima　Yu Suzuki

1　はじめに

バイアスとは一般的に，考え方の偏りや先入観，思い込みなどのことを指す．バイアスは多種多様であり，バイアスを持った人の行動分析やバイアスに関する検証が数多く行われている．これらの行動分析，検証は，被験者を集めて実験を行い，人手による分析を行うことが主流である．本稿では以降，考え方または，立場の偏りの特徴が表れている文字列，単語のことをバイアスと呼ぶこととする．

しかし，上記の実験，分析の方法では，二つの問題がある．分析が主観になってしまう問題と，人を集めるコストがかかるという問題である．二つの問題を解決するための条件を二つ挙げる．条件 1 を，分析に人が関わらない，より客観的な分析方法に変更することとした．条件 1 を満たす方法として，BERT[1] を用いて機械学習を行い，Attention[2] によって色づけした文章を出力し，分析する方法を考えた．理由は二つある．一つ目は，人の手が加わらず，客観的な分析ができるからである．二つ目は，学習後のモデルに対し，Attention を色付き文章で確認することで特徴が表れている文字列，単語を分析する事ができると考えたからである．そして，条件 2 を，コストの問題を解決する為，被験者を集めずにバイアスの分析を行うこととした．条件 2 を満たす方法として，被験者の代わりにツイートを用いることを考えた．理由は，問題設定によって取得するツイートの条件を変更し，幅広い問題に対応できると考えたからである．

以上をまとめ，今回提案する手法では，Twitter ユーザのツイートを取集し，BERT を用いた機械学習を行う．そして，Attention からバイアスの推定を行う．

本稿では，提案手法を用いてバイアスを推定する実験を行った結果を報告する．実験の評価は人間が行う．そのため，比較的人間でも分類可能と考えられる，男性，女性のバイアス，選挙における支持する候補者の違いによるバイアスの二つの話題に対して実験を行った．選挙の話題は，ツイート数が比較的多かったアメリカ大統領選挙を選んだ．これらの話題について，実験を行った結果，バイアスの特徴を学習し，推定することができた．

2　関連研究

鈴木らの研究 [3] では，確証バイアスがウェブ検索行動へ与える影響についての調査が行われている．1 章で挙げた，被験者を集め，人手でバイアスの分析している例の一つである．それに対し，人が持つバイアスを機械学習で分析するという研究はあまり行われていない．Garg らの研究 [4] では，人種，性別に結び付きやすい職業や形容詞を定量的に分析している．性別，職業などにより，特定の単語が付随して使用されることをバイアスと定義し，バイアスを機械学習，分散表現を用いて分析している例である．本研究は，後者の機械学習が主体となり，人の持つバイアスを推定することが目的である．

1）　岐阜大学 工学部 電気電子・情報工学科

3　実験 1(男性，女性)

本章では，男性と女性という，立場の違いによるバイアスの推定の実験を行う．男性と女性では，ツイートの語尾や記号といった箇所にそれぞれの特徴が表れるのではないかと考えている．

3.1　実験手順

ツイプロ[1)]の API を使用し，Twitter のプロフィール検索を行う．取得した情報から，ユーザに男性，女性の二つのラベル付けを人手で行う．ラベルを付けたユーザのツイートを TwitterAPI を使用し，リツイートを除いたツイートを条件なしで取得する．ラベル付けは，取得したプロフィール文に男性，または女性であると明言されていた場合のみ男性，女性のラベルを付ける．

総ツイート数 2,000 件を超えるユーザを残し，取得したツイートの，リプライ情報の削除，URL の削除等の下処理を行う．無作為に選んだ男性，女性各 50 人の，無作為に選んだ 2,000 件のツイートからなる，200,000 件のデータセットを作成する．その後，データセットの 1 割をテストデータとし，BERT で学習させ，テストを行い正解率の確認を行う．得られた正解率が統計的に有意であることを示すために，検定を行う．テストデータから無作為に選んだ 100 件を学習後のモデルが推定した場合 (case1) と，無作為にラベルを付けた場合 (case2) の正解率を標本データとして用意する．標本は 10 個分用意する．得られた 2 種類の標本データを用いて，対応のある 2 標本 t 検定を行い，case1 と case2 に差がみられるのかを検定する．また，Attention によって色付けした文章を html 形式で出力し，特徴が表れている単語の確認を行う．

3.2　結果，考察

学習後，テストを行うと，正解率は 79.37 %であった．この結果が無作為なラベル付けと比べて，有意差があるか調べるため，3.1 節で述べたように，標本を取り，検定を行う．自由度は 9 で，有意確率を 0.01 とすると，t 分布の値は 3.2498 となる．このときの帰無仮説は「学習によって正解率は上がらなかった」となる．これに対し，標本データから t 値を算出すると，t 値は 26.068 であり，帰無仮説は棄却される．よって，男女のバイアスを推定することができたといえる．

また，Attention の確認をしたところ，図 1 のように特徴が表れていることがわかった．正しく分類されたデータのうち，色付けされた箇所に納得がいくもの，いかないものをそれぞれのカテゴリ毎に一つずつ例を挙げる．主観であるが，それぞれの選出理由を以下に示す．図 1 の#1 は，「ねぇねぇ！」や，「ないもん…！」など，男性が使うよりは女性が使いそうな文章があり，人間でも分類可能だと考えられるからである．図 1 の#2 は，「ジム」という単語が少し男性のイメージが強いため，分類可能だと考えられるからである．図 1 の#3 は，「LINE」という単語が女性に結びつくとは思えず，短文であるにもか

1）　https://twpro.jp

Copyright © 2021 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

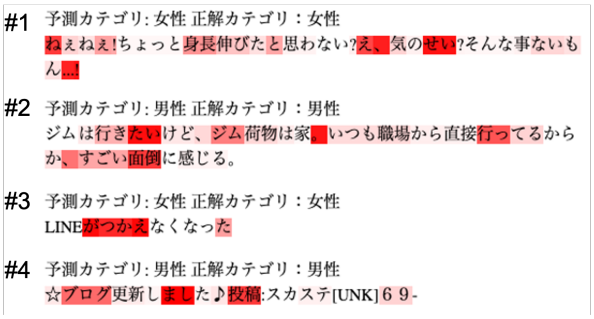

図 1　Attention の確認（男性，女性）

かわらず，女性と分類されたのは理解できないからである．図 1 の#4 は，「☆」や「♪」などの記号は女性が使用するイメージが強かったが，男性に分類されており，自分のイメージとはかけ離れていると感じたからである．

本章の実験では語尾や記号などに特徴が表れやすく，正解率も高くなると予想していたが，次章で述べる選挙の実験よりも正解率が低かった．図 1 の#4 のように，語尾や記号ではない箇所に Attention が見られる場合があり，予想よりも語尾や記号などに特徴が表れにくい，難しい問題だったことが考えられる．

4　実験 2(選挙)

本章では，選挙で支持する候補者の違いという，考え方の違いによるバイアスの推定の実験を行う．選挙に関連するツイートは数が少ないため，その他のツイートからもそれぞれの共通点を学習し，バイアスの推定ができることを期待している．

4.1　実験手順

3.1 節と同様に実験を行う．変更点は以下の 2 点である．1 点目は，ラベル付け前のデータ収集には，Twitter API を使用し，関連ツイートの取得を行ったこと，2 点目は，ラベルの種類を変更したことである．ラベルについて，人手でのラベル付けの際，トランプには賛成派，反対派が明確に表れたが，バイデンには賛成派，反対派が少なく，ラベルとして使用することが困難であった．そのため，トランプ賛成派，反対派の二つのみのラベルを使用し，2 値問題として扱うこととした．

4.2　結果，考察

学習後，テストを行うと，正解率は 82.19 %であった．この結果が無作為なラベル付けと比べて，有意差があるか調べるため，3.2 節と同様の t 検定を行った．有意確率，自由度，t 分布の値，帰無仮説は 3.2 節と同じである．t 値を求めると，t 値は 28.386 となり，帰無仮説は棄却される．よって，選挙で支持する候補者の違いによるバイアスの推定を行うことができたといえる．

また，Attention の確認をしたところ，図 1 のように特徴が表れていることがわかった．正しく分類されたデータのうち，色付けされた箇所に納得がいくもの，いかないものをそれぞれのカテゴリ毎に一つずつ例を挙げる．主観であるが，それぞれの選出理由を以下に示す．図 2 の#1 は，トランプ元大統領に関連するニュースに関するツイートで，分類は比較的簡単に見えるからである．図 2 の#2 は，安倍元総理の派閥に対しネガティブな発言であり，安倍元総理はトランプ元大統領と親しかったため，好まれていない可能性があり，カテゴリ 0 に分類されてもおかしくないからである．図 2 の#3 と，#4 は，

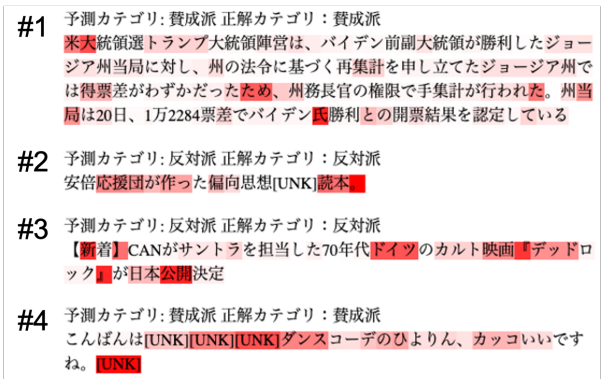

図 2　Attention の確認（選挙）

ツイート内容が選挙とは全く関係がなく，判断基準が人間には理解できないからである．

本章の実験では，長期的な話題ではなく，選挙関連のツイートは少ないため，正解率は低くなると考えていた．しかし，結果は 3 章の実験よりも高い正解率だった．それぞれのユーザが日々つぶやいていることの共通点を学習し，その特徴を得たからだと考えられる．本研究では，ツイートに表れるバイアスを推定することが目的であるため，期待した結果が得られた例だといえる．

5　おわりに

本稿は，機械学習を用いて，性別，選挙におけるバイアスをツイートから得るということを試みた．その結果，ツイートからバイアスを推定することが可能であることがわかった．バイアスを約 80 %の正解率で分類できるほど，ツイートにバイアスの特徴が表れているといえる．

今後の展望として，Attention が表れる理由が不明な単語の妥当性を検証する必要がある．同様の条件でユーザ数を増やしたデータセットまたは，異なるユーザのデータセットを作成して実験を行い．同じ単語に Attention が表れるかを見ることで検証ができると考えている．

本研究が無意識の偏見，暗黙の不平等を明らかにするきっかけとなり，偏見や差別，格差のない社会になることを期待する．

謝辞

本研究の一部は JSPS 科研費 JP19H04218, JP19H04221, JP18H03342 の助成を受けたものです．

参考文献

[1] Jacob Devlin, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc.NAACL-HLT, Volume 1*, pp. 4171–4186, 2019.

[2] Ashish Vaswani, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[3] 鈴木雅貴, 齊藤史明, 山本祐輔. 確証バイアスとウェブ検索行動の関係分析, 2020. DEIM2020 D4-3.

[4] Nikhil Garg, et al. Word embeddings quantify 100 years of gender and ethnic stereotypes. Vol. 115, pp. E3635–E3644. National Acad Sciences, 2018.

Copyright © 2021 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.