

Twitter データを用いた新型コロナワクチン 賛成派と反対派の傾向分析 Tendency Analysis of Proponents and Opponents of COVID-19 Vaccines Using Twitter Data

田中 創太

Sota Tanaka

法政大学情報科学部デジタルメディア学科

E-mail: sota.tanaka.6a@stu.hosei.ac.jp

Abstract

In Japan, there have been many occurrences of mass train rides with no masks and the spread of unsubstantiated information by opponents of COVID-19 vaccines. These nuisances may increase the damage of disasters and should be addressed. To curb these nuisances, it is necessary to prevent the increase in the number of the opponents. This paper proposes how to analyze the tendencies of the proponents and the opponents of COVID-19 vaccines, with the aims of providing help in formulating strategies to prevent the increase in the number of the opponents, and of providing information on vaccines that will satisfy the opponents. We begin by collecting Japanese tweets from Twitter, and next form communities based on retweet relationships among Twitter users. We analyze user biases by comparing the distribution of the users in these communities who tweeted positively and negatively about the vaccines. For this purpose, we propose a method for distinguishing between proponents and opponents by using the Twitter data. We also perform sentiment analysis and topic analysis based on the full text of positive and negative tweets about the vaccines. Although the distinction is not highly precise, we were able to discover some differences of formed communities, emotions that can be read from tweets, and topics of attention between the proponents and the opponents.

1. はじめに

日本国内で初めて新型コロナウイルスへの感染が確認されてから現在まで 3 年近く、日本国内では感染者数のデータ、政府の対応、ワクチンの安全性など様々な話題で議論が繰り広げられてきた。この議論の多くは、新型コロナウイルスへの感染を避けワクチンを自ら進んで接種する賛成派と、新型コロナウイルス自体の存在を否定しワクチンは有害だと主張する反対派の間で行われてきた。反対派は少数ではあるが SNS 等で進んで情報発信する人が多く、反対派の存在が国民の不安を煽り集団免疫の形成を遅らせている要因であると考えられる。今後同

じような災害が発生したときのためにも、賛成派、反対派になる人はどういったコミュニティに属しているのか、どういった話題に反応するのか、また、反対派に影響されやすい人はどういった人なのかといった傾向を分析する必要があると考える。

本研究では、鳥海ら [1]の研究にワクチンの賛成派と反対派の概念を加えることで、さらに実用的なデータを得て反対派の増加を抑えるための戦略を考える助けとすることを目的とする。具体的には、(1)コロナワクチンに関するツイートをしたユーザー全体のネットワーク内に存在する賛成派と反対派について、人数分布の差と時系列による変化を見るユーザーの偏り分析、(2)賛成派と反対派のツイート全文から感情を抽出し、その種類と数の差とその時系列による変化を見る感情分析、(3)賛成派と反対派それぞれが注目している話題の差とその時系列による変化を見るトピック分析の三つを行う。

2. 関連研究

新型コロナに関するツイートを収集し分析した研究がある。鳥海ら [1]は、コロナ禍が始まってからの Twitter ユーザーの感情の変化を 10 種類の感情を用いて示した。集めたデータが実験に用いるデータとしてふさわしいことを示すために、ユーザーの偏り度合いを Kullback-Leibler (KL) 情報量を用いて示している。Baj-Rogowska [2]は、ツイート全文をトピックモデリングにより分析し、独自の分類手法を用いることで人々の新型コロナウイルスワクチン接種の意思決定要因を抽出し分類した。これらはいずれもコロナに関するツイートをしたユーザーを対象にした分析であり、賛成派と反対派の区別はされていない。

3. 準備

3.1. ユーザーの偏り分析

本研究のユーザーの偏り分析は、鳥海ら [1]の研究で行われた実験を参考にする。その流れは、(1)ツイートデータの収集、(2)リツイートデータの収集、(3)ネットワークの構築、(4) Louvain 法の適用によるコミュニティ抽出、(5) Fruchterman-Reingold 法を用いたグラフの描画である。

ここで用いられる Louvain 法はグラフデータから密に結合しているノードの集合（コミュニティ）を抽出する手法であり、アルゴリズムは以下の通りである。

1. 各ノードに個別のコミュニティを設定する。
2. ノードの移動がなくなるまで以下を繰り返す。
 - (a) ノードをランダムに選択し、このノードを隣接するコミュニティに移動させた場合のモジュラリティ（コミュニティの質を表す数値）の変化を隣接するすべてのコミュニティに対して計算する。
 - (b) 計算したモジュラリティの変化量の最大値が0を超える場合、ノードをモジュラリティの変化量が最大値を取るよう移動させる。

3.2. 感情分析

本研究の感情分析は、鳥海ら [1]の研究で行われた感情分析を参考に、Python のライブラリーである ML-Ask を用いて行う。ML-Ask は、感情表現辞典に記載された表現をテキストデータから探索し、マッチした表現の数から感情を推定するツールである。感情は哀、恥、怒、厭、怖、驚、好、昂、安、喜の10種類に分類される。本研究では賛成派のツイートと反対派のツイートに対して ML-Ask で処理し、結果の差と時系列による変化を分析する。

3.3. トピック分析

Baj-Rogowska [2]の研究でトピック分析はソフトウェアを用いてトピックモデリングを行っていたが、本研究では LDA を用いてトピック分析を行う。概要は以下の通りである。最初に、賛成派、反対派のツイートそれぞれに対して学習を行う。次に、トピックごとの単語分布を出力し、各トピックについてラベリングを手動で行う。このとき出力するトピック数を五つとする。最後に、その結果について、賛成派と反対派の差と時系列による変化を見比べる。

4. 提案手法

4.1. 傾向分析

新型コロナウイルスワクチンの賛成派と反対派の傾向分析を行うための方法を提案する。その流れは、(1)ツイートデータの収集、(2)リツイートデータの収集、(3)賛成派と反対派の区別、(4)ユーザーの偏り分析、(5)感情分析、(6)トピック分析である。

データの収集と賛成派と反対派の区別を行った後、三つの分析を行う。まずユーザーの偏り分析でそれぞれの人数やコミュニティの形成状況を大まかに見る。その後、感情分析とトピック分析で細かい傾向を見る。これらの結果を併せて見ることでより詳細で正確な分析を行う。

4.2. 賛成派と反対派の区別

賛成派と反対派はツイート本文に含まれる単語によって区別する。例を挙げると、ツイートの中に「ワクチン」と「有害」を同時に含んだ場合はこのユーザーを反対派であると区別する。本研究では国民の不安を煽るような人を反対派とするため、ワクチンを接種するメリットを

理解しつつも接種しない選択をした人は反対派とせず賛成派とする。この単語による区別に加え、反対派のツイートに対してリツイートを行ったユーザーも反対派に賛同していると捉え、反対派に加える。賛成派の範囲はかなり広く単語で絞り込むことが難しいうえ、本研究では反対派の抽出を分類の目的とするため、指定した単語を含まないツイートはすべて賛成派であるとする。

本研究で指定した単語は、ワクチン信者、ワク信、コロナ脳、兵器、茶番、有害、ワクチン臭、人殺し、殺人、毒（消毒、毒性、弱毒、強毒、気の毒を除く）、マスクはずそう、マスクを外そう、コロナ劇場、奴隷、ワクチン中止、薬害、普通の風邪、ただの風邪、杵珍の19個である。

4.3. ユーザーの偏り分析

ユーザーの偏り分析のうち、ネットワークの構築、コミュニティの抽出までは鳥海ら [1]の研究を参考に行う。結果の示し方について、鳥海らは KL 情報量を用いていたが、本研究では偏りだけでなく賛成派と反対派の人数の差や偏り方も結果として示すため、Fruchterman-Reingold 法を用いたグラフ描画で結果を視覚的に示す。これによって人数の差や偏り方を直観的に把握できるようにする。

5. 実装

5.1. データ収集

データ収集は2段階に分けて行う。ツイートデータの収集とリツイートデータの収集である。ツイートデータの収集では、Twitter API を用いて「コロナ」か「ワクチン」を含むツイートを、投稿日時を指定して収集し、これを CSV ファイルに出力する。次に、収集したツイートに対してリツイートを行ったユーザーの情報を Python ライブラリーである tweepy を用いて収集する。収集したデータをもとに、リツイート関係のデータをテキストファイルに保存する。

5.2. 賛成派と反対派の区別

収集したツイートの文章の中に指定した単語が含まれていた場合はそのツイートをしたユーザーの ID をリストに追加するようにして反対派のユーザーを集めたリストを作る。それ以降、分析をする際にユーザーがこのリストに含まれているかどうかによって賛成派と反対派を区別する。

5.3. 偏り分析

Python ライブラリーの NetworkX を用いてリツイート関係を保存したテキストファイルからネットワークを作成し、同じく Python ライブラリーの community によって Louvain 法によるクラスタリングを行う。完成したコミュニティ全体のグラフを作成してから、賛成派だけ、反対派だけの2種類のグラフを描画する。

5.4. 感情分析

Python ライブラリーの ML-Ask を用いてツイートの文章に対して10種類の感情を割り当てる。どの感情にも当

てはまらない場合は「none」を割り当てる。賛成派、反対派のそれぞれで 10 種類の感情が割り当てられたツイートの数をカウントし、積み上げグラフで描画する。

5.5. トピック分析

Python ライブラリーの scikit-learn にある LatentDirichletAllocation を用いて、賛成派のツイート群、反対派のツイート群をもとに LDA のトピックモデルを学習する。その後トピックごとの単語分布を出力し、手動でトピックにまとめる。

6. 実験

6.1. データの概要

本研究の実験に用いたツイートデータ、リツイートデータの概要を以下に示す。

6.1.1. ツイートデータ

収集したデータの条件は「コロナ」「ワクチン」のいずれかを含むツイートである事で、時系列による変化を見るため投稿日時を 2021 年 8 月, 12 月, 2022 年 4 月, 8 月の四つの月に絞った。また、平日と祝日で国民の生活が変化し、Twitter 上での振る舞いも変化する可能性を考慮し、収集したツイートの投稿日時は上記の各月の第 1 日曜日と第 1 水曜日の二つの日で合計八つの日を指定している。データ数は各日約 24,000 件である。収集するデータの内容は、ツイート ID、ツイートをしたユーザーの ID、ツイート本文の三つである。

6.1.2. リツイートデータ

収集したツイートに対してリツイートしたユーザーの ID を収集し、ツイートをしたユーザーの ID とともに保存する。リツイートデータの件数は、最も少ない 2022 年 8 月 7 日で 12,307 件で、最も多い 2021 年 12 月 1 日で 18,620 件である。

6.2. ユーザーの偏り分析

収集したデータからネットワークの作成、コミュニティの抽出、グラフの描画を行った。結果を図 1 に示す。図中の点はユーザーを表し、点の色は所属するコミュニティを表す。時系列順にみると、賛成派のグラフは大きな変化が見られないが、反対派のグラフからはその勢力は時間が経つにつれて拡大し、密なコミュニティを形成し、賛成派のグラフの形に近づいていく傾向にあることが見て取れる。

6.3. 感情分析

収集したツイートの文章に対して ML-Ask で感情分析を行い、10 種類の感情がどのくらい割り当てられたのかを積み上げグラフとして図 2 に示す。この図から、賛成派、反対派ともに嫌の感情が最も多くの割合を占めていることが分かる。賛成派と反対派を比較すると、反対派は安と喜の感情が少なく、怖と嫌の感情が多くなっていることが分かる。

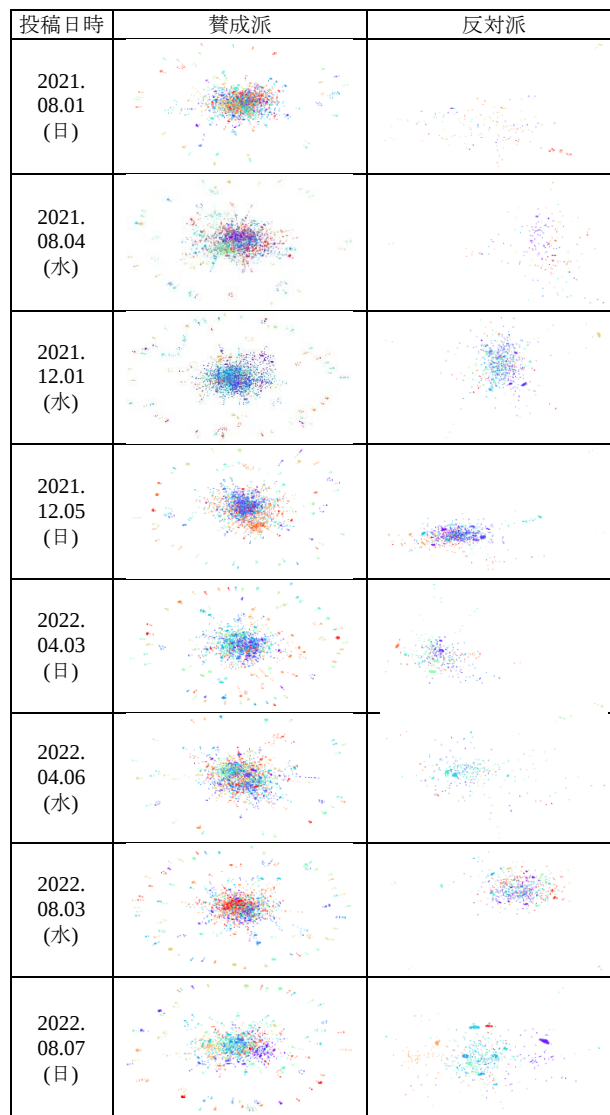


図 1 ユーザーの偏り分析

6.4. トピック分析

ツイートデータを用いて LDA で学習を行い、賛成派と反対派それぞれのトピックを抽出した。結果を表 1 に示す。賛成派ではすべての時系列において、自分や周りの状況、ニュースに対する感想が大半を占め、コロナ禍やワクチンに対する個人的な意見はほとんどないことがわかった。反対派では当初、ワクチンに関する話題は少ないが、時間が経つにつれてワクチンは危険だと発信するようになることが見て取れる。

6.5. 評価

賛成派と反対派を区別する手法の評価をするために以下の実験を行った。

1. ツイートデータを賛成派と反対派、各投稿日時から均等に 240 件取り出す。

- 2. 協力者 3 名にツイート本文だけ見てもらい、そのツイートをしたユーザーが賛成派か反対派か判断してもらう。
- 3. その結果を正解データとして、提案手法による結果と見比べて正解率を導出する。

結果を表 2 に示す。賛成派に関してはかなり高い正解率だが、反対派に関してはあまり高くない正解率だった。また、リツイート(RT)関係を除く反対派、つまり単語による区別のみ行った場合の反対派に関する正解率の方が反対派より少し高かった。

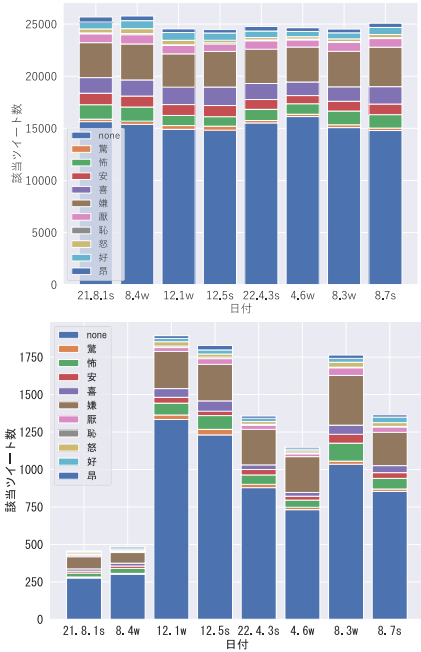


図 2 感情分析結果（上：賛成派，下：反対派）

表 1 トピック一覧

投稿日時	賛成 1	賛成 2	反対 1	反対 2
2021.08.01 (日)	イベント	緊急事態宣言	政治	コロナは政府の嘘
2021.08.04 (水)	オリンピック	ワクチン効果とリスク	データの信憑性	医療崩壊原因、責任は政府
2021.12.01 (水)	イベント	感染状況	ワクチンは危険	コロナ禍の子供
2021.12.05 (日)	感染状況	変異	ワクチンは危険	ワクチンは妊婦に悪影響
2022.04.03 (日)	ワクチン副作用	自宅療養リモート	政治	情報の信憑性
2022.04.06 (水)	新生活	ニュース	情報の信憑性	接種会場への侵入による逮捕報道
2022.08.03 (水)	イベント	マスク効果とリスク	ワクチンマスクは危険	感染症学会による発表
2022.08.07 (日)	ニュース	旅行、検査、隔離	ワクチンは危険	公文書の不開示

表 2 提案手法の評価

	賛成派	反対派	全体	反対派(RT関係除く)
正解率	96.6%	61.3%	79.1%	64.1%

7. 議論

提案手法の精度は賛成派に対しては良かったが反対派に対しては高くななかった。最大の原因として、反対派に対して否定的な意見を述べているツイートも指定した単語を含んでいることが多いことが考えられる。リツイート関係を含まない方が高い正解率だった原因も、単語による分類の正解率が低いためだと考えられる。この問題の解決策として、賛成派と反対派による議論がすべて反対派に分類されることを防ぐためにリプライの除外や、指定した単語に対して「なんか」「なんて」「じゃない」といった否定的な単語を並べているツイートの除外が挙げられる。

偏り分析から、反対派は時間が経つにつれて勢力を拡大させ、密なコミュニティを作ることが分かる。トピック分析の 2021 年 8 月の部分を見ると、当初から反対派である人は政治への関心が高く、政府への不満も高いことが分かる。つまり、政府への不信感により、政府が出した情報は信頼できないという感情からコロナは存在しない、ワクチンは危険だという発想が生まれ、それが改善されないコロナ禍や、効果が不確実なワクチンに対して不満を持つ人に広がって、反対派の勢力を拡大することにつながっていると考えられる。

感情分析から反対派は否定的な感情が占める割合が多いことが読み取れるが、それ以外で賛成派と反対派の差や時系列による変化はあまり読み取れず、感情を割り当てられなかったツイートが大半を占めた。結果として、本研究の手法で感情分析から望ましい結果を得ることはできなかった。

8. おわりに

ツイートデータから新型コロナウイルスワクチンの賛成派と反対派を分類し、それぞれの傾向について三つの分析を行った。偏り分析とトピック分析では賛成派と反対派の間に明確な差が見受けられたが、感情分析においては大きな差が見られなかった。

反対派に対する分類の精度の低さが大きな問題である。指定する単語の見直しやリプライを除外するなど、指定した単語を使用した賛成派を反対派に分類しなくするなどの改善が必要である。

文 献

[1] 鳥海不二夫, 榊剛史, 吉田光男, "ソーシャルメディアを用いた新型コロナ禍における感情変化の分析," 人工知能学会論文誌, vol. 35, no. 4, pp. F-K45_1-7, 2020.

[2] A. Baj-Rogowska, "Mapping of the Covid-19 Vaccine Uptake Determinants From Mining Twitter Data," IEEE Access, vol. 9, pp. 134929-134944, 2021.