D-001

新型コロナ関連トピックと非関連トピック間のコミュニティ類似性の分析
Analysis of community similarity between COVID-related topics and non-related topics

成田 省吾* 横山 昌平* † ‡
Shogo Narita Shohei Yokoyama

1 はじめに

SNS 上では日々多様な情報が拡散されている．その中には社会的な問題に対する意見の相違を浮き彫りにするものも多くある．このような立場が分かれるようなトピックに関して，特に SNS 上で情報の拡散や議論が行われる時は，異なるスタンスを持つユーザ同士が分極化していく傾向にある．[1]

新型コロナ禍では，SNS 上で新型コロナワクチンの賛否や感染症予防対策の議論などが特に活発に行われてきた．それに伴い，新型コロナに関連する情報の拡散の性質や，分極化に関する先行研究が存在する．[2][3] しかし，いずれも新型コロナ関連トピックの視点のみからのユーザの性質や分極化を分析したものである．

本研究の目的は，立場が分かれるような新型コロナ関連トピックに対して，異なるスタンスを持つユーザ群同士の違いを明らかにすることである．そこで，新型コロナワクチンをテーマとして，新型コロナワクチン関連トピックと非関連トピックの関係を分析する．具体的にはまず，SNS の一つである Twitter を取り上げ，潜在的ディリクレ配分法（LDA: Latent Dirichlet Allocation）[4] を用いて，ツイート文書からトピックを抽出する．その後，通常 SNS のユーザは複数のトピックに関心を持つため，各トピックに興味を持っているユーザ群の重複度を類似度として，コミュニティが類似しているトピックペアを見つけ出し分析を行った．

結果として，新型コロナワクチンに否定的なユーザ群は，他の一般的なトピックとのコミュニティ類似性が比較的低いことが明らかになった．これは新型コロナワクチンに否定的なユーザが関心を持つトピックは多様である，もしくは新型コロナワクチン関連のみに興味がある可能性が示唆されるものである．

2 分析手法

2.1 データセット

本研究では株式会社 NTT データ「なずきのおと」を用いて，2021 年 6 月 1 日から 7 月 1 日までに投稿された全ての日本語ツイートを扱った．その中から，「ワクチン」の語を含むツイートと，全てのツイートからランダムにサンプリングしたものの 2 つをデータセットとした．

2.2 LDA によるトピックの抽出

日本語のツイート文書を扱っているため，まず形態素解析を行い単語に分割する．形態素解析には MeCab を使用し，一般的なストップワードを除去した．また，記号や絵文字等も除去した．その後文書ごとに，それに含まれる単語とその出現頻度を表す BoW ベクトルを作成した．

続いて BoW ベクトルを入力とし，LDA を用いてトピックを抽出する．今回は，新型コロナワクチンツイート文書とランダム文書で独立して LDA モデルを作成した．LDA は文書の生成過程を確率的にモデル化したトピックモデルのひとつであり，一つの文書が複数の潜在的なトピックから生成されると仮定したモデルである．

LDA では，パラメータとしてトピック数を指定する必要があるが，Coherence スコアが最も大きいトピック数を採用した．Coherence スコアはトピックの中の頻出単語が共起しているほど高くなるものであり，Coherence スコアが高いとトピックの品質が高く人間が解釈しやすいものであると言われている．その結果，新型コロナワクチン関連トピック数は 19，全体のトピック数は 95 であった．

2.3 トピック間のコミュニティ類似度計算

まずトピック間のユーザ重複度を求める．LDA では各文書がどのトピックに含まれるかを得ることができる．そこで，あるツイート X がトピック A，B に含まれるとしたとき，ツイート X を行ったユーザとツイート X に反応 (リツイートや言及) したユーザをトピック A，B に興味を持っているユーザとする．このようにして各トピックに対し興味を持っているユーザ集合を構築していく．新型コロナワクチン関連トピックと全体トピックの全ての組み合わせに対して，ユーザ集合のシンプソン係数を計算し，トピック間のユーザ重複度とする．

次に，トピック間の意味的類似度を求める．LDA では出力としてトピックごとの単語分布ベクトルを得ることができる．この単語分布は，各トピックにおいてその単語の出現確率が高いほど高い確率を取るものであり，トピックの意味的特徴を表している．したがって，二つのトピックの単語分布ベクトルのコサイン類似度をトピック間の意味的類似度とみなし，新型コロナワクチン関連トピックと全体トピックの全ての組み合わせに対し計算をする．

最後にトピック間のコミュニティ類似度を求める．コミュニティが類似しているトピックペアを探し出すために，手順 3 で求めたユーザ重複度を手順 2 で求めた類似度で割った値が高い順にトピックペアを抽出する．単にユーザ重複度をコミュニティ類似度とすることも考えられるが，例えば新型コロナワクチンの自治体対応トピックと地方政治トピックのように，意味的に近いトピックペアばかりが出てくる．そこで，ユーザ重複度をトピック間の意味的類似度で割った値で並び替えることで，意味的には遠いがコミュニティが類似しているトピックペアを抽出する．

3 結果

表 1 は，新型コロナワクチン関連トピックの一覧である．ただし，各トピックからランダムに抽出したツイートを確認し，意味解釈を行ったうえトピックテーマを人手でつけた．その後，ワクチンに対する賛否という観点から，グループ分けを行なった．親ワクチングループ

* 東京都立大学 Tokyo Metropolitan University.
† 国立情報学研究所 National Institute of Informatics.
‡ 東京大学 University of Tokyo.

は，新型コロナワクチン接種を促すもの，新型コロナワクチンを接種もしくは接種予定を読み取れるものである．中立は，新型コロナワクチン接種に対して直接の言及がないか，もしくは客観的な情報をシェアしているものとしている．反ワクチンは，新型コロナワクチンを接種したくない意思や，新型コロナワクチンの危険性等を強調しているトピックである．

ワクチン関連トピックテーマ
1 親ワクチン
1.1 ワクチン接種後経過報告
1.2 ワクチン接種者の優先順位
1.3 ワクチン接種意思表明
1.4 ワクチン二回目接種報告
1.5 地方自治体のワクチン接種対応
1.6 自粛
1.7 陰謀論否定
2 中立
2.1 変異株
2.2 世界のワクチン接種状況
2.3 国内のワクチン接種状況
2.4 ワクチン接種準備状況
2.5 政治的な影響，マスク
2.6 オリンピック
2.7 ワクチン予約，廃棄
3 反ワクチン
3.1 ワクチンへの恐怖
3.2 反与党，反政府
3.3 ワクチンの効果
3.4 ワクチンの身体への害
3.5 人口抑制

表 1　ワクチン関連テーマ 19 トピック

	全体トピック	ワクチン関連トピック
1	テレビ，ラジオ	世界のワクチン接種状況 (2.2)
2	テレビ，ラジオ	ワクチン接種経過報告 (1.1)
3	個人的事柄	国内のワクチン接種状況 (2.3)
4	テレビ，ラジオ	ワクチン接種準備状況 (2.4)
5	天気，雑談	世界のワクチン接種状況 (2.2)
6	テレビ，ラジオ	ワクチンの副反応 (2.7)
7	ゲーム，Vtuber	ワクチン二回目接種報告 (1.4)
8	音楽，コンサート	国内のワクチン接種状況 (2.3)
9	天気，雑談	ワクチン接種者の優先順位 (1.2)
10	ゲーム，Vtuber	ワクチン接種意思表明 (1.2)

表 2　コミュニティ類似度上位 10 件トピックペア

また，表 2 にはコミュニティが類似しているトピックペアの上位 10 件を記載している．ワクチン関連トピックには表 1 との対応番号を振っている．全体トピックに関しては，ワクチン関連トピック同様，ランダムに抽出したツイートを確認し，トピック名をつけている．ただし，本研究においては全体トピックの解像度がワクチン関連の解像度より低くなっており，全体トピックにおけるトピック一つあたりの範囲が大きくなっている．

表 2 から読み取れるように，全体のトピックとしてはテレビラジオやゲーム等の一般的なエンタメや娯楽のトピックが上位に出現している．しかし，上位のワクチン関連テーマには，親ワクチンや中立にあたるトピックが並んでいるものの，反ワクチンテーマのトピックが少ない．これは，新型コロナワクチンに否定的な感情を持っているユーザ群は，他の意味的に遠いトピックに対して反応しない傾向，もしくはバラバラの他のトピックに興味を持っており全体として別の単一のトピックに反応することはないという傾向を示すものである．または，一般的なエンタメ等のトピックに関心は持たずとも，より小さな範囲のトピックに関心を持っており，先述のトピック分類の解像度の影響により抽出できていない可能性もある．もし，他の意味的に遠いトピックに対して反応しない傾向にあるのであれば，単一のトピックとその周辺トピックにしか興味がなく，いわゆるエコーチェンバー状態になっているといえる．

4　おわりに

本研究の目的は，立場が分かれる話題が多かった新型コロナ禍において，スタンスの異なるユーザ間の違いを分析することであった．その目的のもと新型コロナワクチントピックを取り上げることとし，SNS のユーザは通常複数のトピックに興味を持っていることから，コミュニティが類似している新型コロナワクチン関連トピックと非関連トピックの関係を分析した．

結果として，新型コロナワクチンに否定的でないユーザ群は，一般的な娯楽等のトピックとのコミュニティ類似性が確認できた．しかし，新型コロナワクチンに否定的なユーザ群はそうでないユーザ群と比較して，他の意味的に遠いトピックに対して反応しない傾向，もしくは否定的なユーザ群の中でバラバラなトピックに興味を持っている傾向があることが明らかになった．

今後の課題として以下の二つが挙げられる．まず，新型コロナワクチンに否定的なユーザ群に関して，コミュニティが類似した他のトピックが少ない傾向にあるのかの要因の分析をより細かく行うことである．次に，トピックの解像度をより上げることである．今回は全体のトピックの解像度がワクチン関連ツイートのみでのトピックの解像度と比べて低くなっている．トピックの解像度を揃えることで，さらに細かいレベルでの分析が期待できる．

参考文献

[1] K. Asatani, H. Yamano, T. Sakaki, and I. Sakata. Dense and influential core promotion of daily viral information spread in political echo chambers. Scientific Reports, Vol. 11, No. 7491, 4 2021.

[2] D. Lim, F. Toriumi and M. Yoshida. Do you trust experts on Twitter?. IEEE/WIC/ACM International Conference on Web Intelligence. 2021.

[3] S. Hisamitsu, S. Cho, H. Jin, M. Toyoda, and N. Yoshinaga. Diachronic analysis of users' stances on covid-19 vaccination in japan using twitter. In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2022.

[4] Blei DM, Ng AY, Jordan MI., (2003). "Latent dirichlet allocation.", J Mach Learn Res 3: pp. 993- 1022.