

議論分析のための大規模ツイートデータセットの構築

張 翔[†] 豊田 正史^{††} 吉永 直樹^{††}

[†] 東京大学大学院情報理工学系研究科 〒113-8656 東京都文京区本郷 7-3-1

^{††} 東京大学生産技術研究所 〒153-8505 東京都目黒区駒場 4-6-1

E-mail: [†]{cs,toyoda,ynaga}@tkl.iis.u-tokyo.ac.jp

あらまし 実社会では多様な観点で意見が交わされる話題があるが、人は自分と異なる意見への抵抗感を抱いたり、自分と同じ意見を持つ人間とのみ交流したりしがちであるため、実際に接触する意見には偏りが生じる。こうした現象は意思決定の質を損ね、社会的分断を助長しうするため、意見交換・議論を対象とした様々な研究が行われている。しかし、既存研究のほとんどは少数の話題を個別に分析するに留まっており、多様な話題を対象とした包括的な分析はまだなされていない。そこで本研究では、Twitter を対象として、議論の対象となっている多様な事物・イベントを自動で収集する手法を提案し、既存の議論対象収集手法により収集した結果と質・量を比較することで、評価を行った。

キーワード Twitter, 分極化, 議論, データセット

1 はじめに

現代では情報技術の発展により、多様な話題について、人は時間的・空間的制約に縛られず多様な意見に触れることができる。特に、COVID-19 のパンデミック以降「アベノマスク」や「給付金」など、社会全体に大きな影響を及ぼすイベントが連日のように発生し、その是非について多くの議論がなされている。こうした多様な意見が存在する状況下においては、自分とは異なる意見に接触することにより多様な立場の意見を盛り込んだ意思決定が可能になるとされている [1]。

しかしながら、人は自分と異なる意見を拒絶したり、自分と同じ意見を有する人間のみと交流関係を持ったりする傾向にあることが指摘されている。例えば、Festinger は人が自分の信念・思想に矛盾する情報に触れたときに感じる不快感を認知的不協和と呼び、この認知的不協和を解消する方法の一つとして人は矛盾した情報を拒絶しがちである¹と述べている [2]。このような人間の心理や交流の性質などを考慮すると、人々が必ずしも多様な意見に触れているかという点には疑問が残る。

こういった社会心理学的な知見を踏まえて、ソーシャルメディア上における議論や意見交換データを用いて人が自分の信念や思想に矛盾する情報を拒絶するメカニズムについて分析した研究が数多く存在する。しかしながらそういった研究のほとんどが人手で収集した高々数十件程度の話の議論を分析対象としており、多様な話題に関する分析はほぼ行われていない。自動で多様な話題の議論を収集する研究もいくつか存在はするが、主に著名人の名前 [3]、や投稿 [4]、重大なイベントを指すハッシュタグ [5, 6] をベースにソーシャルメディア上の投稿を集めるといった方式であり、著名人が中心に位置したり、ハッシュタグが存在したりするイベントしか取り扱えないという制限が存在する。

そこで本論文では、多様な話題の議論が公開で行われるソー

シャルメディアの 1 つである Twitter から議論分析のためのデータセットを自動で構築する手法を提案する。具体的には、議論対象を人々の賛否の対象となっている事物で、「現実世界の事物を指している」かあるいは「何らかの動作・行動を指している」ものとして、まず人々の賛否の対象となっている事物を獲得する既存手法 [5] を用いて獲得する。次に、現実世界の事物は固有名詞で、何らかの動作・行動はサ変活用名詞で表されることが多いというヒューリスティクスから、獲得した事物の中から、固有名詞かサ変活用名詞を含むものだけを抽出する。これにより、多様な議論対象を自動で獲得可能である。

評価実験として、2017 年初めから 20 か月間のツイートを用いた議論対象の収集を既存手法 [5, 6] と提案手法とで行い、獲得できる議論対象の数と質とを比較した。このツイートデータは著者の所属研究室で 2011 年より継続的に収集しているもので、2020 年 8 月時点で訳 150 万人ほどのユーザの投稿が収集対象となっている。

最後に、ケーススタディとして、提案手法を用いて本年度社会に大きな影響を与えた COVID-19 パンデミックに関する議論の分析を行う。COVID-19 のように大きな社会的インパクトを持つ話題に関しては、多様な議論が発生するため、主要な議論の対象を網羅的に抽出し、議論の推移を追跡調査した。抽出元のツイートデータとしては 2020 年の 2 月以降の全量ツイートから COVID-19 関係のキーワードマッチにより抽出したものをを用いた。本手法を用いて、各月ごとに注目度の高い議論対象をランキングにしたところ、確かに当時世間からの注目度の高いイベントを指す議論対象が確認された。

本研究の主な貢献は以下の 2 点である。

- 議論分析のための議論対象の自動収集手法を提案し、大規模ツイートデータを用いて収集した
- 既存手法と獲得できる議論対象の質・量を比較し、提案手法の評価を行った

1: 他の方法として自分の思想・信念を変えるといた例も挙げられている。

2 関連研究

2.1 人手による議論対象収集

ソーシャルメディア上での議論分析を目的とした研究の多くは、先述したようにイベントを指すキーワードのリストを手で作成し、それをもとに投稿を収集しデータセットを構築している。例えば、Kiran ら [7] や Demszky ら [8] は Twitter 上において、国内外の政治家の罷免や任免、賛否両論の法案、航空事故、巨大な音楽・映画の催事、あるいは銃乱射事件など数十件の主要なイベントのリストを作成し、各イベントを指すようなキーワードを含む投稿を集めることで、各イベントに言及したツイートを収集し、それによりデータセットを構築した。これはソーシャルメディアデータを使用して議論分析を行う際の典型的なデータセットの構築方法である。

このように、既存研究では高々数十程度の議論を手で収集する研究が多い。その結果、個別のケースへの深い洞察は得られているものの、時系列に沿った議論対象の変遷のような、より大規模なデータセットを必要とする調査はほとんど行われてこなかった。

2.2 自動的な議論対象収集

本節では議論をウェブ上から自動的に収集した研究を紹介する。議論を自動的に収集した研究もいくつか存在するが、著名人や特定のツイート、あるいは特定のハッシュタグに関連する議論しか収集できないなど制限も多い。

Popescu らは議論を引き起こすようなイベントには中心人物が存在すると考え、Wikipedia に記事の存在するような著名人の名前を用いてツイートを抽出し、各ツイート集合に対して議論が発生しているかどうかをラベル付けすることで Twitter における議論データセットを構築した。これはツイートデータを自動で収集するという点で先駆的な研究ではあるが、著名人が中心に位置するイベントしか取り扱えないという制限や、著名人の名前ごとにまとめたツイートデータのうち、どのデータで議論が起きているのかを結局ラベル付けを手で行う必要があるという制限が存在する [3]。

Colleto らは著名人のリストからトランプ元大統領のような、投稿の 9 割以上が議論を引き起こす controversial なユーザーを特定し、そのユーザーの全ツイートを根とするリプライのツリーを一つの議論データとして扱うという枠組みを考えた [4]。この枠組みは一度特定ユーザーが controversial であるとラベル付けさえすれば、自動でデータセットが拡張されていくために Popescu らの手法で必要であったラベリングのコストがいらぬというアドバンテージを有する。しかしながら、中心的な人物の存在を前提としているという問題点は依然として存在している。

Sasaki らは「#XX 賛成」あるいは「#XX 反対」というハッシュタグを集めて XX の部分にくる文字列を議論対象の事物を指すキーワードとして獲得する手法を提案した [5, 6]。彼らは「#XX 賛成 (反対)」というハッシュタグを集め XX の部分に位置するキーワードを抽出することにより、Kiran らや Demszky

Table 1 パタンにマッチしたツイートの例。上段が賛成パタン、下段が反対パタン。太字は抽出される対象。

パターン	マッチしたツイート文
賛成	回答者数は少ないですが、 ミサイル防衛 は80%以上の人が賛成。
してくれ	都営地下鉄って言ってんだから、全部地下に してくれよ
の危険性	コラム:世界で高まる「 ドル不足 」の危険性
許すな	アベ改憲 を許すな!

らが人手で集めたような、議論を引き起こす人物・イベントを指すキーワードを自動で集めた。この手法も低コストで議論データを集められるものの、「#XX 賛成 (反対)」というハッシュタグ内に出現した事物しか集められないという制限がある。

また Sasaki らは、人が賛成あるいは反対を表明するときの言語パターンを用いて議論対象を収集する手法を提案した。彼らは「#XX 賛成 (反対)」というハッシュタグを使用するユーザは XX に賛成 (反対) 的であり、同ユーザが XX に言及していればそれは賛成 (反対) 的意見であるはずという仮定に基づき、先ほど集めた「#XX 賛成 (反対)」ハッシュタグを投稿したユーザの XX に対する他の言及を集めることで、人が何かに対して賛否を述べる時の言語パターンリストを作成し (以降、賛否パターン)、この賛否パターンを含む投稿のうち、賛否パターンの直前に先ほど集めた XX が位置している場合にその XX を議論対象として獲得することで議論対象を収集した。例えばあるユーザが「#特定秘密保護法反対」というハッシュタグを用いており、同じユーザが「特定秘密保護法やめろ」という投稿をしていた場合、「XX やめろ」というパターンが獲得され、他のユーザが「YY やめろ」という投稿をしていた場合、YY が新たな議論対象として抽出される。この手法は先ほどの単純なハッシュタグによるものと比べてより多様な議論対象が獲得されることが予想されるが、やはり最終的に「#XX 賛成 (反対)」の形で出現したものしか得られないという制限が残っている。

本研究では基本的に Sasaki らの手法に沿って議論を収集するため、我々の提案手法は Sasaki らの手法を発展させたものと捉えることが可能である。具体的には、Sasaki らの手法が課している「賛否パターンの直前に位置する文字列で、#XX 賛成 (反対) というハッシュタグの形で事前に出現していたもののみを議論対象として得る」という制限を緩和することで、獲得できる議論対象の数を増やす。ただし、当然ながら制限を緩和すると議論対象ではないキーワードも混入してしまうため、それらのノイズを取り除くために、賛否パターンの選別や、ヒューリスティックスによる議論対象の新たなフィルタリングを行う。

3 提案手法

本節では議論分析のための意見表明を有する議論対象の手法を提案する。なお、先述したように、議論対象とは人々の賛否の対象となっている事物で、「現実世界の事物を指している」かあるいは「何らかの動作・行動を指している」ものとする。提

Table 2 議論対象を指す対象に対する各パタンのマッチ回数 (2018年9月に投稿された日本語ツイート 178,687,801 件を使用)。

使用賛否パターン	上位 N 件中の議論対象を指す対象数			対象数	マッチツイート数	パターン数
	$N = 10$	$N = 50$	$N = 100$			
Sasaki, 2017	4	21	48	278,854	500,052	200
- 賛のみ	1	12	55	203,318	371,301	100
- 否のみ	7	35	78	91,654	146,682	100
選別パターン	10	46	91	10,115	24,865	124
- 賛のみ	9	43	80	2,998	5,509	60
- 否のみ	9	47	84	7,673	19,816	64

案手法は以下のように行われる。

- (1) 賛否パタンの選別
- (2) 選別パターンを用いた賛否対象を含むツイートの獲得
- (3) 前処理
- (4) 係り受け解析に基づく賛否対象の抽出
- (5) 品詞フィルタリングによる議論対象の抽出

まず, Sasaki らの作成した賛否表明パターン²を用いて人々の賛否の対象となっている事物を獲得する [5, 6]. 彼らは Twitter ユーザの賛否の対象となっている事物を収集するためにこのパターンを作成した. Table 1 に賛否表明パタンの例とマッチするツイートの例を示す. 以降, 賛否表明パターンを単に賛否パターンと呼ぶ.

しかし, 単に賛否パターンを用いて賛否の対象となっている事物を集めようとする, 賛否以外の意見の対象となっている事物も集まるという問題点がある. 例えば「～お願いします」というパターンは「熱中症対策 (お願いします)」や「カープ優勝 (お願いします)」のような, 単なる注意喚起や願望の対象にもマッチしてしまう³.

そこで, 賛否の対象にのみマッチするパターンを手で選別する. 具体的には, オリジナルの賛否パターンを用いて収集した議論対象の候補からパターンにマッチした回数の多い上位 50 件に対して, 各議論対象候補を含むツイートを何件か見て, (注意喚起や願望の対象ではなく) 賛否の対象であるかどうかのラベル付けを行う. 同時に, 各議論対象候補にどのパターンが何回マッチしたのかを確認する. この 2 つの結果を合わせて, マッチする議論対象候補の 8 割以上が賛否の対象であるようなパターンを選別パターンとして獲得し, このパターンを用いて確かに賛否の対象となっているような事物を含むツイートを獲得する.

次に, 獲得したツイートに前処理を行う. 具体的には, 選別パターンにマッチしたツイートを句読点や括弧記号などで文分割し, 再度選別パターンにマッチする文のみを抽出した. その後, 全角の数字・アルファベットを半角に直したり空白文字を削除したりといった前処理を行い正規化した⁴.

2: http://www.cl.ecei.tohoku.ac.jp/ja_stance

3: 「カープ」はプロ野球チームの名称. 支持するプロチームが違う場合も賛成的・反対的意见は発生するのである種の議論と捉えることは可能だが本研究では思想的な対立を対象とするため, 議論ではないものとして扱う.

4: この処理は形態素解析器 MeCab 用の辞書である NEologd 辞書の使用時に推奨されている前処理を参考にした. <https://github.com/neologd/mecab->

更に, ツイートに係り受け解析を行い, 選別パターンを含む文節が修飾している文節内に存在する名詞を議論対象として抽出する. 係り受け解析には J.DepP [9] を用いる. このとき, 複数の名詞が連続している場合にはひとまとめにして扱い, これを名詞連続と呼ぶ. ただし以降では, 便宜上単一の名詞も名詞連続と呼ぶ.

最後に, 獲得した名詞連続のなかから, 現実世界の事物を指しているものか, 何らかの動作・行動を指しているものを抽出する. ここで, 現実世界の事物を指している場合は固有名詞が, 何らかの動作・行動を指している場合はサ変活用名詞が含まれていることが多いと考え, そのどちらかが含まれている名詞連続のみを抽出する.

4 評価

本節では議論対象収集というタスクにおける提案手法の収集能力を既存手法と比較することで評価する. ここで使用するデータは筆者の所属研究室で 2020 年 8 月時点で 150 万人程度のユーザの投稿を継続的に収集している Twitter データセットである. このデータセットは著名人など, 影響力の大きいと考えられる 30 程度のアカウントを投稿のシードとなる収集対象として設定して, それらのアカウントの投稿を獲得し, 更に投稿の収集対象のアカウントがリツイートやメンションなどを行ったユーザを再帰的に投稿の収集対象とすることで構築したものである.

4.1 賛否パタンの選別

Sasaki らの賛否パターンリストから, 実際には賛否対象になっていない事物にマッチするパターンを取り除いた結果を示す. 今回は上述のツイートデータのうち 2018 年の 9 月一ヶ月間のツイート 959,162,996 件 (うち日本語ツイートは 178,687,801 件) を使用してこの処理を行い, 結果賛成パターン 60 個, 反対パターン 64 個が選別された. この一覧を付録 2 に示す.

また, 選別したパターンによって獲得される賛否対象が確かに賛否の対象であるか確認するために, 各パタンの賛否対象抽出性能を調査する. 具体的には, 元々の Sasaki らのパターンと選別パターンそれぞれを用いて賛否対象を収集したときに, 各パタンのマッチ頻度上位 N 件中に賛否対象が含まれている件数によ

Table 3 使用データ.

抽出期間	2017/1/1 – 2018/8/31
全ツイート	18,439,627,692
日本語ツイート	3,370,564,751
ハッシュタグ#XX 賛成 の数	142
ハッシュタグ#XX 反対 の数	1,488
賛否対象 XX の数	1,594

Table 4 賛成ハッシュタグの例 (期間内の使用ユーザ数の上位 5 件).

タグ	使用ユーザ数
特定秘密保護法賛成	64
受動喫煙防止法に賛成	63
憲法改正賛成	41
安保法案賛成	38
テロ等準備罪賛成	24

Table 5 反対ハッシュタグの例 (期間内の使用ユーザ数の上位 5 件).

タグ	使用ユーザ数
共謀罪反対	12,428
築地市場の豊洲移転反対	5,634
高江ヘリパッド反対	5,611
TPP/FTA 反対	5,605
移民反対	4,282

Table 6 各手法で獲得できる議論対象の数.

使用賛否パターン	Sasaki	提案手法
Sasaki, 2017	898,072 (1,194)	2,580,342 (1,152)
選別	47,237 (1,124)	153,021 (1,113)

て評価する。ただし、このとき各パターンにマッチした賛否対象が願望や注意喚起の対象ではなく、確かに賛否の対象であるかどうかは、人手で判断する。本稿では著者が賛否対象とマッチしたパターンを見て判断した。この結果を Table 2 に示す。この表より確かに選別パターン (表中「選別パターン」) の方がより高精度に賛否対象を収集することが確認できた。

4.2 議論対象の収集能力

提案手法によって議論対象をどの程度収集できたのか、その結果を Sasaki らの既存手法からハッシュタグの制限を無くしたものをういた場合の結果とともに示す。なお、今回は 2017 年 1 月 1 日から 2018 年 8 月 31 日までの日本語のツイート、メンション、および引用リツイートを使用する⁵。このデータセットの統計値と Sasaki らの手法を再現するにあたって抽出した賛否ハッシュタグ・賛否対象の数を Table 3 に、収集した賛否それぞれのハッシュタグの例を Table 4, Table 5 に示す。

Table 6 に Sasaki らの手法および、提案手法で得られる議論対象の数を比較した表を示す。括弧内は Sasaki らの手法に沿った場合の数 (既に「#XXX 賛成 (反対)」の形で得られていた 1,664 件の賛否対象と重複している議論対象の数) である。この

Table 7 各手法で獲得された議論対象からランダムに抽出した 100 件のうち、実際に議論が含まれている議論対象の割合.

手法	割合
Sasaki	0.13
提案手法	0.15

表を見ると、Sasaki らの手法のとおりハッシュタグ内の議論対象のみ抽出した場合に獲得できる議論対象数が 1,194 個であるのに対し、提案手法では 1,113 個と、元の手法の 94% の数の議論対象を獲得できている。本文中の対象をすべて抽出したときには、オリジナルパターンと選別パターンのどちらの場合も Sasaki らの手法の 3 倍程度の議論対象を獲得できている。このように、提案手法を用いることで既存手法と同程度の精度を保ちつつより多くの議論対象を収集できることがわかる。

更に、実際に議論時に言及されている議論対象をどの程度獲得できているかを既存手法と比較して確認した。具体的には、「ある議論対象に対して肯定的意見を有する人物と否定的意見を有する人物が同時に存在する」場合をその議論対象について議論が起きているものとし、Sasaki らの手法と提案手法で獲得された議論対象を含むツイート中で実際に議論が起きているかどうかを割合を調べた。このアノテーションは第一著者が人手で行った。具体的には、議論対象の文字列だけを見て即座に議論が起こる・起こらないと判断がつく場合はすぐアノテーションを行い、議論対象だけでは判断が難しい場合はアノテーションする議論対象を含む投稿をツイートデータセットから抽出し、その中から 100 件のツイートをランダムに抽出した投稿リストを見て確認した。ただし、ユーザの発信する意見に着目するために、リツイート・メンション・引用リツイートは除去し、また URL を含んでいる投稿も除去したうえでランダムサンプリングを行った。ただし、「自民党」や「改革」などそもそも話題が一意に特定されない議論対象も存在しており、今回はそれらも「議論が起きていない」ものとして扱った。この結果を Table 7 に示す。この結果から、提案手法で獲得できる議論対象の質が確かに既存手法と同程度であることが確認された。

5 ケーススタディ

本手法は、話題を事前に指定せずに使用可能ではあるが、COVID-19 のように大きな社会的インパクトを持つ話題に関しては、多様な議論が発生するため、主要な議論の対象を網羅的に抽出し、議論の推移を追跡調査するような用途を考えている。COVID-19 は我々の社会を大きく変容させ、「GoTo トラベル」や「3 密回避」など、感染への対応策や新しい生活様式について多くの議論が発生した。今後これらの議論を収集し分析対象とする研究は増加すると考えられるが、議論の分析を行う研究に適用可能かどうかを確認するために、本手法によってどのような議論対象が獲得できるのかを示す。

使用データは 2020 年 2 月から 2021 年の 2 月までの 1 年間

5: 単純なリツイートは使用せず、引用リツイートの引用部分も除去する。

Table 8 COVID-19 データから得た議論対象の言及したユーザ数でのランキング.

年/月	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
20/2	コロナウイルス感染	緊急事態宣言	新型コロナウイルス感染	PCR 検査	感染症対策	新型コロナウイルス対策	エアロゾル感染	東京オリンピック開催	専門家会議	東京オリンピック中止
20/3	緊急事態宣言	専門家会議	PCR 検査	コロナウイルス感染	新型コロナウイルス感染	感染症対策	クラスター感染	東京オリンピック延期	東京封鎖	クラスター発生
20/4	緊急事態宣言	PCR 検査	コロナウイルス感染	新型コロナウイルス感染	緊急事態宣言発令	感染症対策	緊急事態宣言延長	専門家会議	緊急事態宣言解除	クラスター感染
20/5	緊急事態宣言	緊急事態宣言解除	検察庁法改正	PCR 検査	持続化給付金	甲子園中止	緊急事態宣言延長	専門家会議	コロナウイルス感染	新型コロナウイルス感染
20/6	緊急事態宣言	持続化給付金	PCR 検査	緊急事態宣言解除	コロナウイルス感染	新型コロナウイルス感染	専門家会議	感染症対策	クラスター発生	新型コロナウイルス感染拡大
20/7	緊急事態宣言	PCR 検査	コロナウイルス感染	感染症対策	持続化給付金	新型コロナウイルス感染	クラスター発生	緊急事態宣言解除	専門家会議	GoTo 延期
20/8	緊急事態宣言	PCR 検査	コロナウイルス感染	感染症対策	新型コロナウイルス感染	クラスター発生	持続化給付金	家庭内感染	緊急事態宣言解除	営業時間短縮
20/9	PCR 検査	緊急事態宣言	コロナウイルス感染	感染症対策	新型コロナウイルス感染	持続化給付金	クラスター発生	普通に生活	クラスター感染	消費税増税
20/10	PCR 検査	緊急事態宣言	持続化給付金	コロナウイルス感染	感染症対策	新型コロナウイルス感染	クラスター発生	新型コロナウイルス感染拡大	普通に生活	大阪市廃止
20/11	緊急事態宣言	PCR 検査	Goto 中止	感染症対策	コロナウイルス感染	新型コロナウイルス感染	持続化給付金	国民投票法改正	クラスター発生	大阪市廃止
20/12	緊急事態宣言	PCR 検査	感染症対策	コロナウイルス感染	新型コロナウイルス感染	GoTo 停止	成人式中止	指定感染症延長	クラスター発生	持続化給付金
21/1	緊急事態宣言	PCR 検査	成人式中止	緊急事態宣言発令	感染症対策	コロナウイルス感染	持続化給付金	新型コロナウイルス感染	緊急事態宣言解除	営業時間短縮
21/2	緊急事態宣言	緊急事態宣言延長	PCR 検査	組織委員会	緊急事態宣言解除	感染症対策	持続化給付金	コロナウイルス感染	新型コロナウイルス感染	クラスター発生

にわたる日本語の全ツイートから⁶から、COVID-19 に関するキーワードを本文に含むものを抽出して構築した 1,168,489,934 ものツイートからなるデータセットである。このキーワードリストは「COVID」や「コロナ」のような総称的なものや、「GOTO」や「10 万円給付金」などの当時の政府施策に関するもの、「リモートワーク」や「オンライン会議」など新しい生活様式に関するもの、WHO 事務局長の「テドロス」氏や対策分科会会長の「尾身茂」氏などの人名などを、NHK の COVID-19 関係の主要ニュースアーカイブサイトをもとに筆者らが人手で集めて作成した。抽出に作成したキーワード 156 個の一覧を付録に示す。このキーワードを用いて抽出したツイートデータに本手法を適用し、COVID-19 という話題における議論対象を 206,356 件得た。

次に、獲得した議論対象の時系列に沿った変遷を確認する。Table 8 に獲得した議論対象を、投稿したユニークユーザの人数で月ごとにランキングにした表を示す。事例に着目してみると、例えば 2020 年の 3 月に東京オリンピックの延期が安倍首相と IOC のバッハ会長との間で正式に合意されたが、2020 年の 3 月の 8 番目の議論対象として「東京オリンピック延期」という議論対象が出現している。ほかにも、2020 年の 4 月と 5 月にはそれぞれ緊急事態宣言の発令と解除が行われたがそれも「緊急事態宣言発令 (2020/4, 5th)」と「緊急事態宣言解除 (2020/5, 2nd)」という議論対象として表れている。また、「持続化給付金

(2020/5, 5th)」や「GoTo 延期 (2020/7, 10th)」, 「営業時間短縮 (2020/8, 10th)」や「Goto 中止 (2020/11, 3rd)」など、提案手法によって当時の世間からの関心の大きいと考えられる話題が確かに獲得されていることが確認された。

6 おわりに

議論分析研究は人文科学的にも注目されておりソーシャルメディア上のデータを使用した実証研究が近年盛んになってきている。その一方で個々の研究者による独自の条件下での研究が多く、統一的な見解を欠いているのが現状である。その要因の一つとして間違いなくデータセットの作成手法が研究ごとにまちまちであることが考えられる。本研究で提案した手法がこの現状を改善させる一善となることを期待する。

本稿では議論や意見交換の包括的な分析の一助として、Twitter を用いた議論分析のためのデータセット自動構築手法について提案した。また、実際に獲得できる議論対象の質や量の比較や、実世界データに本手法を適用した場合の結果の例示などにより、提案手法の有効性も評価した。

本稿で収集した議論対象リストおよび、選別した賛否パターンは著者の所属研究室のページにて公開している⁷。

6: Twitter データは Twitter, NTT データより提供を受けたものである。

7: www.tkl.iis.u-tokyo.ac.jp/~cs/DEIM2021/index.html

謝 辞

本研究は、JST、CREST、JPMJCR19A4の支援を受けたものです。

文 献

- [1] J. Habermas, *The structural transformation of the public sphere*. MIT press, 1991.
- [2] L. Festinger, *A theory of cognitive dissonance*, vol. 2. Stanford university press, 1957.
- [3] A. M. Popescu and M. Pennacchiotti, "Detecting Controversial Events from Twitter," *International Conference on Information and Knowledge Management, Proceedings*, pp. 1873–1876, 2010.
- [4] M. Coletto, K. Garimella, A. Gionis, and C. Lucchese, "Automatic controversy detection in social media: A content-independent motif-based approach," *Online Social Networks and Media*, vol. 3-4, pp. 22–31, oct 2017.
- [5] A. Sasaki, K. Hanawa, N. Okazaki, and K. Inui, "Other Topics You May Also Agree or Disagree: Modeling Inter-Topic Preferences using Tweets and Matrix Factorization," in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, vol. 1, pp. 398–408, apr 2017.
- [6] A. Sasaki, K. Hanawa, N. Okazaki, and K. Inui, "Predicting Stances from Social Media Posts using Factorization Machines," *Coling-2018*, pp. 3381–3390, 2018.
- [7] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, "Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship," in *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*, (New York, New York, USA), pp. 913–922, ACM Press, 2018.
- [8] D. Demszky, N. Garg, R. Voigt, J. Zou, M. Gentskow, J. Shapiro, and D. Jurafsky, "Analyzing polarization in social media: Method and application to tweets on 21 mass shootings," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 2970–3005, 2019.
- [9] N. Yoshinaga and M. Kitsuregawa, "A self-adaptive classifier for efficient text-stream processing," in *Proceedings of the 25th International Conference on Computational Linguistics*, pp. 1091–1102, 2014.

付 録

1 COVID-19 ツイート抽出に作成したキーワード一覧

コロナ, COVID, COVID, 肺炎, インフルエンザ, 風邪, 感染, パンデミック, エピデミック, オーバーシュート, 世界的流行, 病原性, 無症状, PCR, PCR, 抗体検査, 抗原検査, BCG, BCG, アビガン, レムデシビル, ワクチン, イソジン, ヨード, うがい薬, 再検査, 濃厚接触, 医療崩壊, 臨床試験, 治験, WHO, WHO, 専門家会議, 分科会, 医師会, マスク, トイレトペーパー, ソーシャルディスタンス, テレワーク, テレカン, テレ会議, テレ講義, テレ授業, テレ診察, リモートワーク, リモート会議, リモート講義, リモート授業, リモート診察, オンライン会議, オンライン講義, オンライン授業, オンライン診察, オンライン飲, オンライン呑, オンライン面接, オンライン収録, オンラインレッスン, オンライン

ミーティング, オンライン研修, オンライン面談, オンライン説明会, 緊急事態宣言, ロックダウン, 都市封鎖, 自粛, 給付金, 助成金, 補助金, 設備導入, 協力金, データダイエット, 国内パスポート, 東京アラート, 倒産, 閉店, 赤字, 解雇, 転職, 自殺, 中止, 再開, 延期, 買い占め, 買占, 売り切れ, 売切, 接触確認, 接触追跡, COCOA, COCOA, ココア, 生物兵器, GO TO, Go To, GO TO, GoTo, GoTo, 10万円, 10万円, K1, K-1, K1, K-1, 要請無視, クルーズ, 観光客, インバウンド, チャーター機, 自己責任, オリンピック, パラリンピック, k値, K値, K値, k値, エピセンター, ワークーション, 組織委員会, テドロス, シルビーブリアン, シルビー・ブリアン, 尾身茂, 尾身会長, 尾身先生, 尾身教授, 尾身氏, 押谷仁, 押谷教授, 押谷先生, 押谷氏, 西浦教授, 西浦先生, 西浦氏, 山中教授, 山中先生, 脇田隆字, 脇田教授, 脇田先生, 岡部信彦, 岡部教授, 岡部先生, 中野貴志, 中野教授, 中野先生, 児玉龍彦, 児玉教授, 児玉先生, 岡田晴恵, 岡田教授, 岡田先生, 上昌広, 上教授, 上先生, 孫正義, 孫社長.

2 選別パターン

賛成パターン 賛成, 賛成です, に賛成, に賛成です, に賛成します, 賛成します, 大賛成, 成立させるぞ, 成立おめでとうございます, 可決おめでとうございます, は合憲, を支持します, を支持, 賛成, 賛成派です, には賛成,」に賛成, には賛成です, 賛成だ, を支持する, に賛成しています, 法制賛成, 成立おめでとう, を実現させてください, 賛成!! , に賛成する, 賛成」, に賛成だ, 可決, おめでとうございます, を実現させましょう, 可決おめでとう, 賛成です, を実現しましょう, は絶対に必要, 賛成ですが, は賛成派です, 賛成の声を挙げよう, を支持しています, で連携強化, を支持している, の実現を, に賛成しよう, 反対って誰が得するの, を成立させましょう, なんか問題あるの, に賛成している, うちにも来るかなあ・, を望みます, 大賛成です, 賛成ですよ, 法制に賛成, 以外にない, が正解, にすべきだ, には賛成してます, 実現しましょう, 賛成でした, 賛成してる.

反対パターン 反対, 絶対反対, に反対, 断固反対, 反対です, に反対します, を許すな, を許さない, に反対です, は反対, を廃案に, には反対です,, 絶対反対, は反対です, に反対する, 絶対廃案, 許さない, は廃案に, に抗議, 建設反対, に反対しています, には反対, を廃案へ, 許すな, 絶対阻止, は許さない, に断固反対します, 再稼働反対, は絶対反対, 大反対, は憲法違反, に反対している, などしている場合か, はんたーい, に断固反対, 断固阻止, という暴挙をとめる, したくなくてふるえる, 許すまじ, 建設阻止, 今すぐ廃案, 反対だ, はいりません, されてしまいます, する国絶対反対, に反対しましょう, などしている場合ではない, も反対, に巻き込まれる, は廃案しかない, の危険性, 反対します, は許されない, 反対デモ行進に参加しよう, に反対しよう, を阻止しよう, したがる総理はいらない, は許せない, はさせない, の都合のよい口実となります, は断固反対, いやだ, をなくそう, する国にさせない.