DEIM Forum 2019 C3-4

マイクロブログにおいて論争化する議論の予測

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　〒 153–0041 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　マイクロブログにおいてセンシティブなトピックについて議論を行うとき，同じ意見の持ち主同士でのみ話し合い，違う意見の持ち主とは交流をしなくなる極化と呼ばれる現象がしばしば発生する．極化が進行すると議論は論争状態となり，建設的な議論を行うことが難しくなる．そこで，本研究では論争を予防することを目的とし，その第一段階としてある議論が論争になるかどうかを予測する手法を提案する．具体的には，所与の議論が論争へと発展するかどうか予測する問題を，その議論の参加者たちの極化の度合いを示す論争度 [6] が将来的に一定以上増えるか，減るか，変わらないかを分類する 3 値分類問題として定式化する．この問題を議論の初期段階における，ツイートの内容や議論参加ユーザの性質，およびユーザ間のつながり等の手がかりを用いて訓練した分類器で解くことを試みる．実データを用いた実験を行うため，まず大規模な Twitter データから既存研究で得られた賛否表明パタンを用いて論争候補トピックのデータセットを構築した．更に，論争が発生したかどうかを人手で注釈付けした小規模データを用いて実際の論争と論争度の関係を分析し，論争度の変化予測に有効と考えられる特徴量を考案した．最後に，分類器の訓練と評価を行う実験により，定量的な評価として提案した特徴量の寄与率を分析し，また予測の成功事例と失敗事例を分析することで定性的な評価も行った．

キーワード　マイクロブログ，エコーチェンバー，極化

1 はじめに

情報技術の発展により遠く離れた人とのやり取りが容易になり，マイクロブログ上では多様なトピックについての意見交換（議論）が行われるようになったが，議論のトピックによってはしばしば同じ意見の持ち主同士でのみ話し合い，違う意見の持ち主とは交流をしなくなる極化と呼ばれる現象（いわゆるエコーチェンバー）が発生し，建設的な議論が阻害されてしまう．例えば，議論のトピックが「アメリカ大統領選」や「原発再稼働」のようなセンシティブで論争になりやすいトピックであった場合，人々は極化により自分と同じ意見の人とのみ交流し，自分と違う意見に触れようとしなくなる傾向にある．同じ意見の人が集まったグループ内では特定の意見のみがどんどん先鋭化していき，建設的な議論が行えなくなることが広く知られている [8]．

一度論争化した議論は解消すること自体が困難であるとの報告 [2] もあり，極化に対しては予防的なアプローチが必要となる．極化が発生した議論を正常な状態に戻す研究として，異なる意見の持ち主同士をユーザ推薦などにより引き合わせることで極化の緩和を狙った研究 [5] や，議論に参加しているユーザに極化の発生している様子を見せ，自省を促す研究 [7] 等が存在する．しかし，前者の研究は理論評価にとどまっており，後者の研究は長期的な有効性が薄いことが示されている．

そこで本研究では極化を予防するために，論争化（極化が進むこと）する議論を機械学習により訓練した分類器で予測することを目指す．具体的には，マイクロブログ上の所与の議論に対して計算される，極化の度合いを示す論争度（RWC）という指標 [6] を用いて，所与のトピックに関する議論の論争度が将来的に一定値以上増加するかどうかを予測する．また，極化が収まる場合に対する理解を深めるためにも，論争度が将来的に一定値以上減少する場合かどうかも予測対象とする．すなわち，所与のトピックに関する議論の論争度が一定値以上増加するか，減少するか，または変化しないかを予測する三値分類問題を解くことを試みる．

実データを用いた実験を行うため，まず大規模な Twitter データから論争候補トピックを収集・分析し，データセットを作成するとともに分類に有効な特徴量を提案した．具体的には，既存研究で用いられた賛否表明パタンリスト [12] を用いて論争化した可能性のある議論で取り扱われているトピック（論争候補トピック）のデータセットを構築した．加えて，その一部をサンプリングして論争が実際に発生しているかどうか注釈付けして分析した．そして，この分析から得られた知見を基に論争化の予測に有効と考えられる特徴量を設計した．この特徴量を作成したデータセットから抽出し，分類器の訓練・評価を行った．

実験では作成したデータセットを用いて分類器を訓練して識別性能を評価し，定量的な評価として提案した特徴量の有効性を，定性的な評価として識別の成功例・失敗例を確認した．結果として，議論の初期段階における議論参加者間のインタラクションに基づいて構築されるグラフの統計量と，投稿されたツイートのテキスト特徴量の組み合わせが手がかりとして有用であるとわかった．また，実際の識別の成功例・失敗例を確認したところ，提案手法は偶然トピックが複数の意味を持つ場合などを論争化と捉えてしまうこともあったもの，概ね論争化や非論争化を捉えられていることが確認できた．

2 関連研究

マイクロブログ上の論争に関連する研究としては，主に論争の検知を行う研究，論争を分析・定量化する研究，および論争を解消することを試みた研究がある．以下ではこれらの研究を順に示す．

2.1 論争の検知

論争を検知対象とした研究は少なく，更に検知対象に制限がある．Popescu ら [11] は Wikipedia から取得した著名人一覧を用いて各著名人について言及したツイートを集め，人手で論争が起きているかどうか注釈付けしたデータセットを作成した．更に，著名人への言及ツイートのテキスト特徴量や，言及ツイートをしたユーザたちのつながりから計算されるグラフ特徴量，また外部ニュースサイトで同時期に対象の著名人が言及されたニュースが存在するかという外部ニュースから得られる特徴量をデータから抽出して，所与のツイートセットで論争が起きているかどうかを判定する識別モデルを訓練し，実際のツイートに適用した．この研究では高い精度で論争の識別を実現しているものの，現実には著名人が事象の中心に登場していないものの論争が発生している事象も多く [1]，そういった事象をカバーできていない．

本研究では，議論に共通して現れやすい賛否表明のパタンを用い，論争が起きそうなトピックについてのツイートを自動抽出することで，著名人が議論の中心にいる場合などに限定されない，多様なトピックの抽出を試みている．

2.2 論争度：議論の極化の定量化

論争について分析した研究は数多く存在し [1,3,4,10,13]，これらの知見を用いてマイクロブログ上における所与の議論がどの程度論争化しているのかを示す指標を提案した研究がある．Kiran ら [6] は Adamic らのブロガネットワークの分析で報告されている「グループ間交流がグループ内交流に比べて非常に疎となる」という論争事例の特徴 [1] に着目することで，マイクロブログ上のあるトピックがどの程度論争を引き起こしているかを Random Walk Controversy (RWC) という指標で定量化した．具体的には，あるトピックに関心をもつ Twitter ユーザのインタラクション（ここでは RT）をエッジとしたグラフ（以下，インタラクショングラフ）を作って 2 つのグループへ既存のクラスタリングアルゴリズム（Metis [9]）によって分割し，各グループのユーザの中からランダムに選んだユーザを始点・終点とするランダムウォークを複数回行い，グループを横断するランダムウォークと，横断しないランダムウォークの割合の差を計算することで RWC を算出した．更に，所与のトピックにおいて論争が進行しているほど論争度が大きな値を示すため，R どの程度論争が進行しているかの指標として RWC を用いることを提案した．式 (1) に RWC の計算式を，また Fig. 1 に RWC が高くなるトピックと低くなるトピックそれぞれに関するインタラクショングラフの例を示す．ただし，X と Y は

1：例えば国際捕鯨委員会 (IWC) を脱退したことの是非を問う「IWC 脱退」や，不登校児童への対処法について議論を引き起こした NHK の「不登校特集」など．

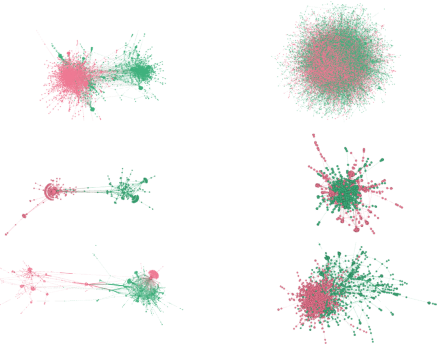

Fig. 1: RWC が高くなるトピック（左）と低くなるトピック（右）のインタラクショングラフの例．左は上から順に「保育園落ちた日本死ね」，「築地移転」，「スパコン詐欺」で，右は「コミックマーケット開催」，「関ジャニ昴脱退」，「豊田真由子議員謝罪」についてのインタラクショングラフ．左列のグラフは赤グループと緑グループが大きく分離しているのに大して，右列のグラフは比較的密接に絡み合っている．

2 つに分割されたグループであり，P_{AB} はグループ A から始まったランダムウォークがグループ B で終了する確率を示す．

$$RWC = P_{XX}P_{YY} - P_{XY}P_{YX} \quad (1)$$

本研究は Kiran らと同様にマイクロブログを対象としているため，彼らの RWC を基に対象のトピックが将来論争となるかを予測するタスクを設定した．そのためデータセット構築の際も，論争候補トピックの収集後に各事例について RWC を算出し付与している．付与した RWC を参考に各トピックを分析し，論争化予測の手がかりとなる特徴量の考案を行っている．

2.3 論争化した議論の正常化とその限界

一度論争化した議論を正常な状態に戻すのは困難であるという報告も有り，論争化した議論を正常な状態に戻すことを目指した研究はいくつか存在するものの，やはり予防的なアプローチが必要と考えられる．

Kiran ら [5] は極化した 2 つのグループの影響力の大きいユーザ同士を結びつけることで極化の緩和を狙った．具体的には，2 つのグループ内の影響力の大きいユーザの中から，反対側の意見にも耳を傾ける可能性の高いユーザを今までのインタラクション（RT など）を基に選び出し，その人物らのつながりを促進する (Twitter におけるユーザ推薦のような機能によって) ことで極化を緩和することを目指した．しかし，この提案手法の評価は理論的な評価に留まっており，現実の論争化した議論をどれだけ解決可能であるかは不明である．

また，Gillani ら [7] は，ユーザに自分と異なる意見にも耳を傾けさせる意識を高めるための手法を開発した．具体的には，論争に参加しているユーザたちのつながりをグラフにして可視化して異なる意見をもつユーザグループに分割したうえで，ユーザの一部にそのグラフを提示し自身の情報獲得がグラフの一部しかカバーしていないことを自覚させることでユーザの意識を高めることを狙った．評価においては，対象ユーザのその後のフォロー傾向などを観察することで手法の妥当性を検証し

た．しかし，その結果，長期的には提案手法の有効性が薄いことが確認された．

また，Vo ら [14] はウェブ上でのフェイクニュースの共有により生じる誤報の拡散や議論の不健全化を防ぐための新たなアプローチを提案している．具体的には，ガーディアンと呼ばれる，ウェブ上に投稿された情報の事実確認を積極的に行っているユーザを発見し，検証が必要な情報の URL を彼らに推薦を行う．これにより，ガーディアンらが情報の真偽をより積極化に検証することを狙っており，評価実験によって推薦精度の向上が確認されている．しかし，このアプローチには議論を健全化させることが可能なガーディアンの数には限りがあるという制約が存在する．

このように，既存研究は論争の解消にあたって一定の成果を出しているものの，各手法に改善点もある．また，Kiran らや，Gillani らの手法を用いるとしても，解決の対象となるトピックを人手で選定しなければいけないという問題が残されている．一方で本研究は人手の選定無しに将来論争が起きるトピックを探し出す手法を提案しているため，本研究によって今後の深刻な論争化が生じると予測されたトピックに対して，これらの研究のアプローチを重点的に用いることで論争の極化を抑止するという相互補完的な関係を実現できる．

その他にも，論争の解消自体がそもそも困難であることも示唆されている．Bail ら [2] は対立する 2 つのグループのそれぞれの構成員に対して，対立している側の情報を継続的に摂取してもらうことの影響を調査した結果，対立が解消されるわけではなくむしろより対立が激化する傾向にあるということを報告している．これは深刻化してしまった論争への対処が困難であることを示唆している．

以上の研究を踏まえると一度論争化した議論を正常な状態に戻すための有効な方策は限られていると考えられる．そこで本研究では，論争化に対する予防的なアプローチに取り組むこととした．具体的には，論争化予防の第一歩として，将来論争化しうる議論を早期段階で検知するという問題に取り組む．

3 論争候補トピックデータセットの構築

本章では論争化する議論の分類器を訓練するために用いる，議論が論争化したトピック・しなかったトピック（以下，論争候補トピック）の収集方法と，そのトピックに言及ツイートからなるデータセットの構築方法について述べる．本研究では，議論はトピックを指すキーワードと，そのトピックに言及したツイートの集合で表されるとする．そのため，論争候補トピックが収集できれば，そのトピックに言及したツイートを大規模ツイートデータから抽出することで，論争化した可能性のある議論を収集可能である．まず Twitter API によって収集したツイートから，既存研究で使用された賛否表明パタンリストを用いて賛成または反対の対象となる名詞連続をキーワードとして抽出する．更に，論争を起こすトピックに特有の品詞パタンにマッチし，新しく出現した，一定以上のツイートがなされているキーワードを抽出し，これを論争化候補トピックとした．最後に，論争度変化を見るために，この論争候補トピックに言及したツイートをトピック発生初日とトピック発生初日から 10 日間の 2 種類の期間に渡って収集した．

Table 1: 賛否表明パタンとヒットしたツイートと抽出される名詞連続（太字はフィルタを通過したもの）の例．

パタン（賛否）	マッチしたツイート文	名詞連続
してくれ（賛）	都営地下鉄って言ってんだから，全部地下にしてくれよ	都営地下鉄，全部地下
好評（賛）	13 巻アニメイト限定セット」好評予約受付中!	**13 巻アニメイト限定セット**
してください（賛）	1 日 1 回プレイできますので，是非プレイしてくださいね	**1 日 1 回プレイ**，是非プレイ
やめて（反）	それをもう金輪際やめて欲しいですね．	それ，金輪際
しない（反）	大井町線止まってりゃ諦め付いたんですけど下手に動いてるもんだから出社しないとならないという	大井町線，諦め，出社
の危険性（反）	コラム:世界で高まる「ドル不足」の危険性	コラム，世界，**ドル不足**

データセットのもととなるツイートデータは筆者らの研究室で継続的に収集しているツイートデータから抽出した日本語ツイート約 18 億件を用いた．このツイートデータは最初に 30 名程度の著名人をターゲットユーザとして選び，2011 年の 3 月以降のターゲットユーザのツイートを収集するとともに，ターゲットユーザとインタラクション（RT・メンション・クオート）を行ったユーザを新たにターゲットユーザリストに加え，同様の処理を繰り返し行うことで拡大したものである．2017 年 11 月から 2018 年 10 月までのツイートデータ約 100 億件を抜き出し，その中から日本語ツイート約 18 億件を抽出して使用した．

まず，論争の対象となる事物は少なくとも賛成または反対の対象となっていると考えたため，このデータに対して人が何かに賛成または反対を表明するときに用いる言語パタンを用いてパタンマッチを行い，Twitter 上で賛否の対象となっているキーワードを収集した．具体的には，日本語ツイートの URL を除去して正規化を行い，句読点などに基づく文分割を行った各文に対して，Sasaki らが作成した「賛否表明パタン」 [12] を用いて Twitter ユーザによって賛成または反対されている名詞連続を収集し，論争候補トピックの収集を試みた．ただし，Sasaki らは本来「X 反対」のようにパタンの直前に来る名詞のみを賛否の対象としているが，本研究では文中のパタンより前にくる名詞連続すべてを賛否の対象として扱った．パタン・マッチするツイート文・抽出される名詞連続の例を Table 1 に示す．

次に，収集したキーワードを観察し論争の対象となりそうな事物に多く見られる品詞パタンを見つけたので，この品詞パタンにマッチするキーワードのみを抽出した．具体的には，賛否パタンにマッチした文の名詞連続を確認した結果，「築地市場移転」や「玄海原発再稼動」などのように，論争化を引き起こしそうな名詞連続には「実世界の事物」が「何らかの行動をする」ことを示すパタンが多く見られた．「実世界の事物」は殆どの場合「固有名詞」であり，「何らかの行動をする」ことを示す名詞も殆どの場合「サ変活用名詞」であるため，「固有名詞」と「サ変活用名詞」の両方を含んでいる名詞連続のみを抽出対象とした．

更に，既に出現したトピックが論争化するかどうかは比較的

容易であることから，過去 30 日以内に出現したキーワードは除外した．例えば，一定期間経過しても論争が発生していないトピックはその後も論争が発生する確率が非常に低く予測の必要性が薄い．そのため今回は予測の必要性の高い，新たに出現したトピックのみを対象とする．

最後に，ここまでの処理を施したキーワードを観察したところ論争化する可能性が非常に低いと考えられるキーワードが多く見られたため，ストップワードを導入してヒューリスティクスによるフィルタリングを行った．具体的には，「XX 誕生日おめでとう」や「XX 発売」などのパタンが多く見られたため，これらのキーワードに共通する単語として，「発売，販売，誕生，生誕，周年，記念」をストップワードとして用いた．今使用したデータ中でストップワードにヒットしたキーワードは 240 件で，「南京大虐殺記念館訪問」や「安倍晋三記念小学院」など議論が論争化すると考えられるトピックもわずかに見られたが，その殆どは論争になる可能性が非常に低いと考えられるトピックであった．

ここまでの手順に従ってツイートデータから論争候補トピックを収集し，更に各トピックに言及したツイートを，トピック出現当日と，出現当日から予測対象の未来である 10 日後までの間に渡って収集した．ただし，このときトピック出現当日のツイート数が 50 未満であったトピックは重要性が低いとして除外した．ここまでの考察を踏まえた論争候補トピックの抽出手法を以下に示す．

（1） ツイートを正規化し，句読点に基づき文分割．

（2） パタンマッチした文だけを抽出．

（3） 抽出した文のパタンより前に出現する名詞連続を抽出．

（4） 「固有名詞」と「サ変活用名詞」を両方含む連続名詞を抽出．

（5） 直近の 30 日間に出現していないものを抽出．

こうして獲得したトピックとツイートのセットを収集対象の期間によって訓練・開発・評価データへと分割した．このとき，未来のデータを使用して過去の事例を予測してしまわないように，訓練データよりも開発データの時系列が後に，開発データよりも評価データの時系列が後に来るようにデータ分割を行った．最終的な論争候補トピックの統計量を Table 2 に示す．ただし，最右列の $|E|/|V| > 2$) の列については 4 章で述べる．

4 論争度指標の分析

本章では，2 章で述べた論争度が実際の論争・非論争トピックに対してどのように振る舞うのかが不明であるため，3 章で作成したデータセットを一部サンプリングして，実際に論争が発生しているかどうかを人手によって注釈付けした小規模データセットを作成し，実際の論争に対して論争度指標がどのように振る舞うか分析を行った結果について述べる．更に，いくつかのトピックの論争度が特定の値に集中していたため，これらのトピックについて調査したところインタラクショングラフが木構造などの特殊な構造になっていることがわかったため，論争度指標とインタラクショングラフの構造との関係の分析を行ったので，これについても述べる．最後に，これらの分析内容から，データセットの要件や論争度変化予測に有効と考えられる特徴量について述べる．

Table 2: 収集した論争候補トピック数．

用途	期間	トピック数	トピック数 ($\|E\|/\|V\| > 2$)
訓練	2017/12/1〜2018/07/31	11008 件	3399 件
開発	2018/08/1〜2018/09/20	2436 件	742 件
評価	2018/10/1〜2018/10/31	1302 件	406 件

Table 3: 論争・非論争を注釈付けしたトピックのサンプル．

論争トピック		非論争トピック	
総裁選延期	北海道節電率	軍事的緊張緩和	青葉モカ誕生祭
不登校特集	商業捕鯨再開案	北方領土放棄	産業廃棄物扱い
P 優勝	日露平和条約締結	サムスピ新作	15 日発売

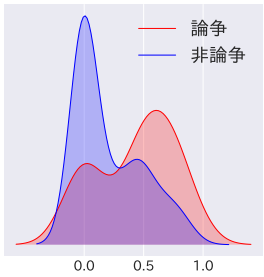

Fig. 2: 注釈付けした 100 トピックの RWC の密度分布．

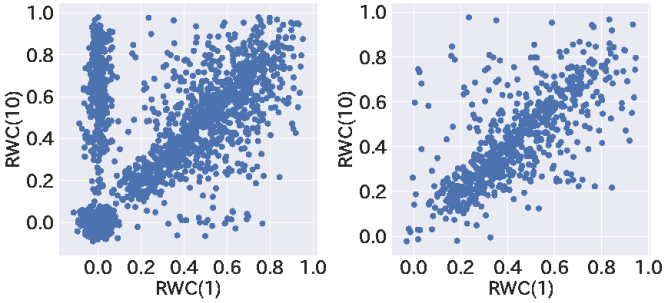

Fig. 3: （左）トピックキーワード 2436 個の RWC プロット．（右）その内，インタラクショングラフの平均次数 $2|E|/|V|$ が 2 以上である（木構造でない）トピックの RWC プロット．

まず，3 章で作成した開発データから 100 件のトピックをサンプリングして実際のツイートを確認し，各トピックにおいて論争が発生しているかどうかの注釈付けを行った．ただし，ここでは論争が発生したかどうかの判断基準として，あるトピックの発生初日の全ツイートを RT された回数でソートして RT 回数の大きい方から数十件観察し，その中に好意的なツイートと否定的ツイートが両方見られる場合に，そのトピックを論争トピックとした．結果，100 件中 20 件で論争が発生しており，80 件のトピックでは論争の発生は確認されなかった．注釈付けしたトピックのサンプルを Table 3 に示す．以下，トピックの登場初日のツイートだけから算出した RWC を RWC(1)，トピック登場初日から 10 日間の全ツイートから算出した RWC(10) と表記する．

次に RWC が実際の論争トピックに対してどのように振る舞

うか不明であるため，先に作成した注釈付けした 100 件のトピックに対する論争度の振る舞いを確認した．4 節で注釈付けした 100 件のトピックについて，ラベルごとの RWC の密度分布を Fig. 2 に示す．図中の赤色が論争トピック 20 件の RWC 密度分布で，青色が非論争トピックの RWC 密度分布である．これを見ると，論争トピックと非論争トピックでピークに差が存在することが確認できる．これは，RWC が論争トピックに対しては高い値に，非論争トピックに対しては低い値になる傾向にあることを意味しており，RWC が論争の度合いの指標として人間の感覚に合うように振る舞うことが確認できた．

最後に，グラフの構造が木構造であるなどの特殊な構造であるときの論争度の振る舞いについて述べる．3 章で収集した開発データ中の論争候補トピックについて，RWC(1) と RWC(10) のプロットを Fig. 3 の左に示す．Fig. 3 を見ると，殆どのトピックが直線 $x = 0$，直線 $y = x$，または直線 $y = 0$ の 3 つの直線の近傍に位置していることがわかる．

ここで，直線 $x = 0$ 近傍に位置するトピックについて RWC 値とインタラクショングラフを観察したところ，RWC 値については実際には論争が観察されてないのに RWC が高くなってしまうケースが存在した．そのようなトピックのインタラクショングラフはほとんどまたは完全に木構造グラフとなっていた．このため，インタラクショングラフが木構造のような特殊な形状であるときには RWC が正しく機能しないことが予測される．これを確認するために，インタラクショングラフがほぼまたは完全に木構造であるトピックとそうでないトピックの RWC 密度をプロットした．具体的には，インタラクショングラフの平均次数 [2]$|E|/|V| > 2.0$ となるグラフとそうでないグラフに分けた．Fig. 4 にこの密度プロットを示す．開発データ中，インタラクショングラフが木構造になったトピックは 1558 件，そうでないトピックは 878 件であった．

Fig. 4 から，グラフが木構造である場合，該当するトピックで論争が起きているかどうかによらず RWC が 0 に集中していることがわかる．よって，予測の対象トピックは，そのトピックから作られるグラフが木構造であるものは除去することが望ましい．そのため，予測の対象とするトピックはトピック登場初日のインタラクショングラフが $|E| > |V|$ となるものだけを取り扱う．ここで，Fig. 3 の右に，開発データのトピックから，インタラクショングラフが木構造となるものを除去したトピックについて，RWC(1) と RWC(10) のプロットを示す．左右で比べると，直線 $x = 0$ 近傍のトピックが大きく減少したことがわかる．

ここまでの分析結果から，RWC が適切に機能しないためにインタラクショングラフが木構造となるトピックは除去することが望ましいことと，トピックの RWC にはインタラクショングラフの特徴が大きく関係していることが裏付けられた．すなわち，3 章で作成したデータセットに更にフィルタリングを行う必要があると考えられる．また，将来の RWC の予測の際にはグラフ特徴量が有効であると考えられる

2 : 完全に木構造のグラフでは $|E| = |V| - 1$ であるため 2.0 未満となる．

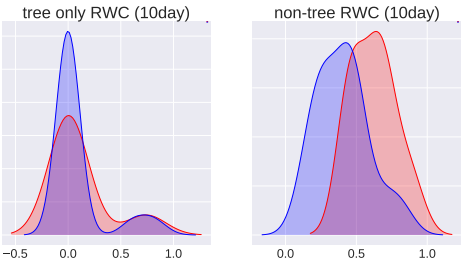

Fig. 4: （左）インタラクショングラフが木構造であるトピックと（右）それ以外のトピックそれぞれの RWC(10) 密度分布．

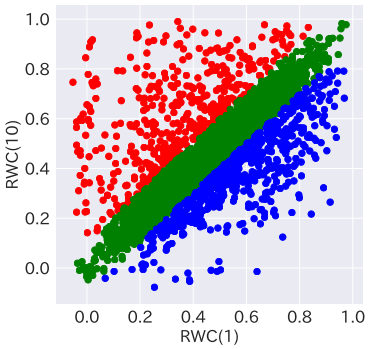

Fig. 5: 学習・開発データ 4141 トピック RWC(1)，RWC(10) のプロット．$\mathrm{RWC}_{\mathrm{diff}}$ の大きさで 3 クラスに分けている．

5 論争化予測手法

本研究ではある議論の論争が激化するか収束するか変わらないかを識別する分類器を学習し，これを用いて論争度が上がる議論の検知を行う．分類器には，線形カーネルのサポートベクターマシン（SVM）を用いた．

$$\mathrm{RWC}_{\mathrm{diff}} = \mathrm{RWC}(10) - \mathrm{RWC}(1) \qquad (2)$$

具体的なタスクとしては，論争度の変化について，あるトピックの RWC(10) と RWC(1) を用いて式 (2) のように $\mathrm{RWC}_{\mathrm{diff}}$ を定義し，あるトピックが所属するクラスを以下の 3 クラスから判断する識別タスクを解く．

- UP クラス： $\mathrm{RWC}_{\mathrm{diff}} > 0.1$
- DOWN クラス： $\mathrm{RWC}_{\mathrm{diff}} < -0.1$
- REMAIN クラス： それ以外

このタスクの説明図を Fig. 5 に示す．この図は，後述する 4387 個のトピックの RWC(1)，RWC(10) をプロットし，UP クラスを赤，REMAIN クラスを緑，DOWN クラスを青で塗り分けたものである．本研究で解くタスクは，所与のトピックがどの領域に来るのか予測するタスクといえる．

ただし，0.1 というクラス境界の値は，論争度計算の際に生じるブレよりも大きな値になるように設定した．論争度はランダムウォークによって算出される値であることから試行によってブレが生じるため，ブレを測定するためにいくつかのトピックにおいてランダムウォークを 1 万回実行した．結果として論争度の標準偏差が 0.05 未満であったことから，論争度が 0.1 以上変化していればブレではなく，確かに議論が論争化した（議論の論争化が収まった）とした．

この分類問題を解くにあたっては，議論に投稿されたテキストの特徴量と，議論参加者間のつながりのグラフ特徴量を用いた．まずどのような単語が使用されるかが手がかりとなる．例えば，「ドナルド・トランプ」や「集団的自衛権」などの単語が使われている場合はおそらく議論が論争になりやすいと考えられる．よって，議論に投稿されたツイートのテキスト特徴量を使用する．また，4 節で見たようにグラフの構造と RWC には相関があることがわかったので，議論参加者間のネットワークのグラフ特徴量も使用する．更に，議論の参加者の特性が論争化に関係していると考えられるため，それも用いる．Table 4 に今回使用した特徴量の一覧を示す．

5.1 テキスト特徴量

議論で使用されている単語は論争化予測に有効であると考えられるため，議論のテキスト特徴量として，議論に投稿されたツイートで使用されている語句の埋め込みベクトルの平均を用いた．あるトピックに言及したツイートの集合があるとき，それらのツイート中に特定の語句が出現すると，そのトピックはその後論争になる可能性が高くなる．こういった語句の特徴を捉えるために，対象のトピックに関するツイート集合の単語埋め込みベクトルを平均し，分類器の特徴量として用いる．単語ベクトルとしては，訓練済みベクトルが鈴木ら [15] によって公開されている [3]ため，それを使用する．

ツイートからの語の抽出方法としては，対象のツイートをすべて形態素解析し，名詞・動詞・形容詞と判断された単語のみを抽出する．ただし，RT によって同一の語句が複数回出現することが考えられるため，RT や引用されたツイートは複数回出現していても 1 回のみの出現として扱う．また，論争予測の分類器の訓練の際は，データセット中で出現頻度が 5 回以上の単語のみを対象とする．

5.2 グラフ特徴量

論争度は議論参加者間のインタラクショングラフに基づいて計算されるため，インタラクショングラフの基本的な統計量や，議論参加者間のこれまでの関係性，および各参加者の性質などを特徴量として使用した．

まず，所与のトピックのインタラクショングラフの構造を用いた特徴量としては，まずノード数やエッジ数，クラスタリング係数や平均次数などの単純なグラフ統計量が存在するため，これらを使用する．

他にも，単純なグラフ統計量の他に議論に参加しているユーザたちの普段の関係性をグラフ構造から獲得することが考えられる．例えば仲の悪い二人のユーザが同じ議論に参加した場合，その議論のインタラクション中においても仲の悪さが反映され論争化すると考えられる．そこで，議論に参加しているユーザのもともとの関係性を手がかりとして用いるために，議論に参加する前の一定期間内に行ったインタラクション（RT・メンション・クオート）からグラフを構築し，そのグラフの統計量も用いた．このように構築したグラフをリレーショングラフと呼ぶ．具体的には，現在のインタラクショングラフに，議論参加ユーザが過去 30 日間にした・されたインタラクションのうち，5 回以上同じ人との間になされたインタラクションをエッジとして加えることでリレーショングラフを構築した．このときもとのインタラクショングラフに存在しないユーザも新たに付け加える．

また，議論参加ユーザの性質を捉えることは議論の論争度変化予測に有効と考えられるため，議論参加ユーザが過去に参加した議論の論争度の統計量も特徴量として使用する．例えば，論争そのものが好きなユーザが参加している議論は論争化する可能性が高いと考えられる．具体的には，所与の議論に参加している各ユーザに対して過去の参加議論の論争度を算出し，その和と平均（いずれもスカラ値）を特徴量として使用する．

3 : http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/

Table 4: 使用した特徴量．記載のない場合はすべてスカラ値．

種類	使用特徴量
テキスト特徴量	単語埋め込みベクトルの平均 (200 次元)
グラフ特徴量（リレーショングラフ（後述）についても求める）	ノード数，エッジ数，クラスタリング係数，平均次数，RWC，議論参加者が過去に参加した議論の RWC の総和・平均

Table 5: 評価データでの識別性能．

model	uniform	Text only	Graph only	Text+Graph
UP	0.165	0.199	0.308	**0.316**
DOWN	0.258	0.268	**0.512**	0.508
REMAIN	0.480	**0.521**	0.385	0.444
macro	0.301	0.329	0.402	**0.423**
micro	0.337	0.384	0.404	**0.433**

6 実　　験

本章では 2017 年 11 月から 2018 年 9 月までの議論を用いて分類器を学習し，2018 年 10 月の議論の分類性能を測ることによって，提案した特徴量が議論の論争化・収束・無変化の予測にどの程度寄与しているのかを評価する実験について説明する．

なお，4 章で見たように実験データからは，トピック登場初日のインタラクショングラフの平均次数が 2.0 を下回るトピックは除外することが望ましい．実際に各データから除去した後のデータ数の統計量を Table 2 の最右列に示す．

6.1 実 験 手 順

分類器の学習の際，グラフ特徴量は 0 から 1 の範囲内で正規化を行った．また，本研究で用いる分類器である線形カーネルの SVM はペナルティ項の係数 C をハイパーパラメータとしてもつので，上述のすべての特徴量を使用し，マクロ平均 F1 を最大とする 5.0 を C として使用した．なお UP クラスと DOWN クラスは REMAIN クラスに対してサンプル数が少ないため，訓練の際は全クラスの数がサンプル数最少の DOWN クラスと同じになるようにアンダーサンプリングを行った．

本研究で提案するタスクは新しいものであるため比較の対象となるベースラインをどのように設定するかが問題となる．本研究では UP クラス，DOWN クラス，REMAIN クラスをランダムに予測する識別器をベースラインとして用いる．更に，テキスト特徴量とグラフ特徴量をそれぞれ単独で訓練に用いた場合の精度も比較対象とする．

6.2 実験結果

Table 5 にテストデータでの予測精度を示す．結果として，DOWN クラスや REMAIN クラスでの予測精度は他の特徴量のほうが大きくなったが，クラスサンプル数の比率を無視するマクロ平均においては提案した特徴量を用いたモデルが最良の結果を示した．これにより，提案した特徴量の有用性が示された．また，テキスト特徴量やグラフ特徴量のみを単独で使用した場合に比べると両者を組み合わせたほうがマクロ平均 F1 とマイクロ平均 F1 の両方においてベースラインを上回っている．

更に，各特徴量の寄与率を測るため，Ablation テストを行った．具体的には，寄与率を測りたい特徴量を除いてモデルを学習し，すべての特徴量を使用して訓練を行った場合と比べてその予測精度の増減を調べた．特徴量除去によって精度が減少した特徴量は予測に有効な特徴量であると言える．テキスト特徴量のみの予測精度，ユーザグラフ特徴量のみの予測精度は計算済みであるため，ここではユーザグラフ特徴量の各項目について，その寄与率を計算する．結果を Table 6 に示す．この結果，トピックが発生した初日のグラフ構造がもっとも寄与していることが分かった．

7 事例分析

本章では，提案した手法によって，クラスを正確に予測できたトピックの分析を行うことで，提案手法の能力と限界点について考察する．

7.1 成功事例

クラス予測に成功したトピックについて，実際にどのような議論が Twitter で行われていたかを分析した．そのようなトピックのうち，実際に論争が発生している，あるいは論争が収束していることが確認された事例を Table 7 に示す．

論争の解消傾向が実際に確認されたトピックの一例として，「憲発議」の出現当日および 10 日後における RWC の値とインタラクショングラフの可視化結果を Fig. 6 に示す．本トピックに関する Twitter での議論を分析したところ，トピックが出現した当日には，安倍議員の支持層の議論と石破議員の支持層との議論に極化が生じていることが観測された．実際に，インタラクショングラフにおける両支持層の間にはエッジが少なく，RWC(1) の値も 0.877 と高いことが同図より分かる．本トピックの出現から 10 日が経過した後では，RWC(10) の値は 0.407 へと減少した．同図右のインタラクショングラフからも分かるように，この時点ではどちらの派閥のツイートも RT する人が増えており，極化が多少解消されていることが実際に確認できた．提案手法は，このようなトピックに対して，正しいクラス DOWN を予測することができた．

また，Fig. 7 には論争に発展すると予測し実際に論争にまで繋がった事例「新潮 45 編集部」の出現当日および 10 日後におけるインタラクショングラフを示す．このトピックに関する Twitter での議論を分析したところ，トピックが出現した当日には，新潮 45 という週刊誌への批判意見がほとんどであった一方，徐々に賛成意見が増えていた．これは，新潮社の公式アカウントが新潮 45 に対する批判ツイートを連日 RT したことに対する賞賛の意見が増えたことに起因している．

Table 6: Ablation test の結果の F1 値（開発データ）．

ablation	UP	DOWN	REM	macro	micro
ablation なし	0.309	0.459	0.459	0.409	0.423
- グラフ特徴量（初日）	0.282	0.444	0.321	**0.349**	0.348
- グラフ特徴量（過去）	0.316	0.444	0.395	0.385	0.389
- RWC（初日）	0.261	0.346	0.462	0.356	0.384
- RWC（過去）	0.297	0.455	0.413	0.388	0.395
- 過去議論 RWC	0.307	0.450	0.422	0.393	0.402

Table 7: 予測の成功したトピックの実例．

論争発生	論争収束
自民総裁選	ネトウヨ体験
人種差別撤廃委員会	衆議院通過
新潮 45 休刊	福島市議聴取
獣医学部構想	築地市場先行破壊

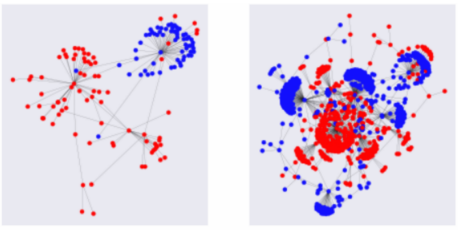

Fig. 6: 成功トピック「憲発議」の出現当日（左）と 10 日後（右）のインタラクショングラフ．RWC 値は，$(\mathrm{RWC}(1), \mathrm{RWC}(10)) = (0.877, 0.407)$.

Table 8: 予測の失敗したトピックの実例．

論争発生と予測	論争収束と予測
歌番組出演	強化拡張パック
ドクターヘリ緊急救命	サンダース大学付属
SHOCK 公演 1600 回	散華行ブルース
城崎広告	渡辺麻友卒業

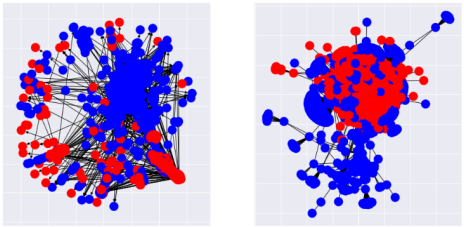

Fig. 7: 成功トピック「新潮 45 編集部」の出現当日（左）と 10 日後（右）のインタラクショングラフ．RWC 値は，$(\mathrm{RWC}(1), \mathrm{RWC}(10)) = (0.080, 0.382)$.

7.2 失敗事例

提案手法によって正しいクラスを予測できたものの，Twitter での実際の議論を確認してみると，UP クラスであるにも関わらず論争が発生していない事例や，DOWN クラスであるにも関わらず論争が収束していない事例が存在した．その実例を Table 8 に示す．

提案手法によって正しいクラス DOWN が予測されたものの，実際には論争の収束が確認されなかったトピック「コラボ装備」について，RWC(1)・RWC(10) と出現当日および 10 日

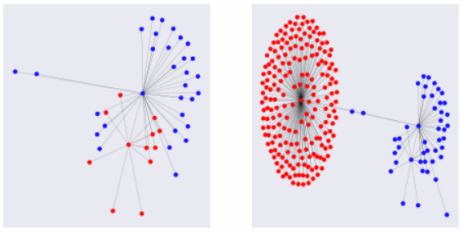

Fig. 8: 失敗トピック「コラボ装備」の出現当日（左）と 10 日後（右）のインタラクショングラフ．RWC 値は，$(\mathrm{RWC}(1), \mathrm{RWC}(10)) = (0.099, 0.814)$．

後におけるインタラクショングラフの可視化結果を Fig. 8 に示す．同図のインタラクショングラフを見ると，確かにトピック出現から 10 日までの間で論争が激化しているように見える．しかし，実際の議論の内容を分析してみたところ，複数のソーシャルゲームがコラボ装備と呼ばれる機能 [4] を同時期にリリースしており，それぞれのゲームによってプレイヤー層が異なるため，インタラクションに分離が生じていただけであり，論争は生じていなかった．この結果は，事前に選定したトピックの中に複数の異なるトピックが混在していると，予測の結果に悪影響を及ぼすことを示唆している．将来の論争度のより正確な予測のための課題として，均一な粒度のトピックを同定することが考えられる．また，本研究では平均次数に基づき木構造グラフを事前に省いていたが，本分析の結果から，信頼性の高い予測の実現のためには，その閾値を適切に選択する必要があると考えられる．

8 おわりに

本論文では，人々が同じ意見を持つ人同士の間でのみ交流する事態を避けることを目的として，所与の議論が論争化するかどうかを予測するタスクを提案し，このタスクを論争度が変化する議論の識別タスクとして定式化した．更に，大規模 Twitter データを用いて，このタスクを解くにあたって必要な論争候補トピックのデータセットを作成した．次に収集したデータセットの一部に人手で注釈付けを行い分析することで，論争度変化に有効と思われる特徴量について考察した．最後に論争度変化予測問題を議論の初期の段階におけるツイートの内容やユーザ間のインタラクショングラフの特徴量などの手がかりを用いた分類器で解いた．更に，分類器の識別精度や各特徴量の寄与率について検証し，また予測成功例・失敗例については検証も行った．結果として，議論の初期段階における議論参加者間のインタラクショングラフの統計量と，投稿されたツイートのテキスト特徴量の組み合わせが手がかりとして有用であることがわかった．

文　献

[1] Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM, 2005.

[2] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.

[3] Mauro Coletto, Kiran Garimella, Aristides Gionis, and Claudio Lucchese. Automatic controversy detection in social media: A content-independent motif-based approach. *Online Social Networks and Media*, 3-4:22–31, 10 2017.

[4] Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066. Association for Computational Linguistics, 2017.

[5] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Reducing controversy by connecting opposing views. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 81–90, New York, NY, USA, 2017. ACM.

[6] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy on social media. *Trans. Soc. Comput.*, 1(1):3:1–3:27, January 2018.

[7] Nabeel Gillani, Ann Yuan, Martin Saveski, Soroush Vosoughi, and Deb Roy. Me, my echo chamber, and i: Introspection on social media polarization. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 823–831, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee.

[8] Kathleen Hall Jamieson and Joseph N Cappella. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press, 2008.

[9] George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing*, 20(1):359–392, 1998.

[10] Zhe Liu and Ingmar Weber. Is twitter a public sphere for online conflicts? a cross-ideological and cross-hierarchical look. In *SocInfo*, 2014.

[11] Ana-Maria Popescu and Marco Pennacchiotti. Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1873–1876. ACM, 2010.

[12] Akira Sasaki, Kazuaki Hanawa, Naoaki Okazaki, and Kentaro Inui. Other topics you may also agree or disagree: Modeling inter-topic preferences using tweets and matrix factorization. In *ACL*, 2017.

[13] Benjamin Timmermans, Tobias Kuhn, Kaspar Beelen, and Lora Aroyo. Computational controversy. In Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri, editors, *Social Informatics*, pages 288–300, Cham, 2017. Springer International Publishing.

[14] Nguyen Vo and Kyumin Lee. The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '18, pages 275–284, New York, NY, USA, 2018. ACM.

[15] 鈴木正敏, 松田耕史, 関根聡, 岡崎直観, and 乾健太郎. Wikipedia 記事に対する拡張固有表現ラベルの多重付与. **言語処理学会第 22 回年次大会**, pages 797–800, 2016.

4：あるゲーム内において別ゲームのキャラクターの装備が利用可能になること．