インターネット上の当選運動・落選運動の分析

Analyzing the campaigns for the 2013 Japanese House of Councillors election on Twitter

船木 洋晃*1 佐々木 彬*1 岡崎 直観*1*2 乾 健太郎*1 深田 陽介*3 Youkou FUNAKI Akira SASAKI Naoaki OKAZAKI Kentaro INUI Yosuke FUKADA

> 竹下隆一郎*³ 田森 秀明*³ 野澤 博*³ Ryuichiro TAKESHITA Hideaki TAMORI Hiroshi NOZAWA

 *1 東北大学 *2 JST 戦略的創造研究推進事業「さきがけ」 Tohoku University PREST, JST

*3株式会社朝日新聞社 メディアラボ

Media Lab, The Asahi Shimbun

In Japan, the ban for election campaigns on the Internet has recently been lifted. Now, citizens have the right to campaign on the Internet. In response to this situation, we analyze the campaign of the 2013 election by the members of the House of Councilors by using data on Twitter. More specifically, we generate a classifier which classifies the campaigns automatically, and we evaluate its accuracy. Finally, we create a network of retweet information and analyze it, visual the discussion and connection of users participating in the campaigns, and draw the behavior of the election campaigns on Twitter.

1. はじめに

2013 年夏に行われた第 23 回参議院議員通常選挙は、いわゆるネット選挙解禁後初の国政選挙として注目を浴びた。ネット選挙解禁により、候補者や政党関係者以外の一般の市民がウェブサイト等を利用して選挙運動を行うことが認められる(改正公職選挙法第 142 条の 3 第 1 項)。ウェブサイトを利用する方法とは、インターネット等を利用する方法のうち、電子メールを利用する方法を除いたものことで、ホームページ、ブログ、ソーシャル・ネットワーキング・サービス(SNS)、動画共有サービスなども含まれる。

本研究では、2013 年夏の参議院選挙時のツイートを分析し、選挙運動にツイッターがどのように活用されたのかを調査する。具体的には、ツイッター上での当選運動・落選運動を計算機が自動的に認識する手法を構築し、その精度を評価する。ツイートから当選運動・落選運動を自動的に認識することで、各候補者に対する当落運動の時系列推移をモニタリングしたり、ツイッター上の当落運動と実際の選挙結果の相関を調べることができる。さらに、各候補者に関するツイートからリツイートネットワークを作成し、その特徴を分析することで、当落運動に参加しているユーザのつながりや議論を可視化し、ツイッター上における選挙運動の勢力図を描く。

すでにネット選挙が解禁されていた海外を中心に、関連した研究が盛んに行われている。Tumasjan らは、ドイツの連邦選挙に関するツイートをテキスト分析ソフト LIWC を用いて分析し、Twitter が政治的な意見交換の場としての有効な手段であるとともに、Twitter 上での政党に対する単なる投稿数が選挙結果に影響をあたえることを示した [Tumasjan 10]。Larssonらは、Twitter 上の投稿データが分析のための理論的な根拠となることを示し、また Twitter の機能の一つであるリツイートがユーザとそのユーザをフォローしていないユーザとのつなが

連絡先: 船木洋晃, 東北大学工学部情報知能システム総合学科, 宮城県仙台市青葉区荒巻字青葉 6-3-09, 022-795-7140, 022-795-4285, b2tb2205@s.tohoku.ac.jp

りを分析するために効果的であることを示した [Larsson 12]. 湯淺は、公職選挙法の改正により、ウェブサイト等を利用した 選挙運動が解禁されたことを受け、参議院議員選挙や中間市議 会委員選挙を通じて浮上した改正公職選挙法の問題点を考察 し、今後の課題についての検討を行った [湯淺 13].

2. 機械学習による当選運動・落選運動の分類

ツイッター上に投稿される候補者に関するツイートは膨大であり、それら全てを人手で当選運動・落選運動に分類するコストは非常に大きい。そのため本節では、当選運動・落選運動を自動的に判別するための分類器を構築する。

2.1 データ作成

本研究では、2013 年 7 月 4 日から 20 日までに投稿された ツイートのうち、第 23 回参議院議員通常選挙の立候補者名を 本文中に含む 1,327,907 ツイートを取得した。実験では、当選 運動もしくは落選運動が起こっていた候補者に焦点を当てるため、朝日新聞デジタルの「注目の当落」ページ*1に含まれる候補者の中で、候補者名を含むリツイート回数の合計が 100 以上ある、34 名の候補者を対象とした。

教師有り学習で分類器を構築するため、被リツイート数が20回以上のツイートに対して「当選運動」、「落選運動」、あるいは「その他」のラベルを付与する。また、その時点でラベルが付与されたツイートの種類数が100未満であった候補者については、その数が100となるように被リツイート数の多いツイートから追加でラベルを付与する。ここで、各ラベル付与の基準は以下のように定義した(論文掲載の都合上、候補者名に該当する部分は[候補者名]として表記する)。

当選運動は,(1)のような候補者自身の呼びかけ,(2)や(3)のような有権者からの好意的なツイート等,候補者が当選することを期待する内容を含むツイートに対して付与する.

(1) 東京選挙区は、 [候補者名] にチャンスを下さい!

^{*1} http://www.asahi.com/senkyo/senkyo2013/chumoku/

- (2) [候補者名] の選挙事務所を訪問。本当に老若男女がボランティア活動をして熱気を感じた。このエネルギーは本物。
- (3) 立派だ、この [候補者名] と言う人物は。頭が下がる。こ ういう人が日本にいるというだけで、まだ日本は救われ る、と言う気になる。

落選運動は、(4)、(5)のように候補者が落選することを期待する内容のツイートや、(6)のように候補者に対して否定的な意見を述べているツイートに対して付与する。

- (4) [候補者名] が当選したら、日本人の負けだから。
- (5) [候補者名] が当選なんかしたものなら、6 年間も国会に 居座ることになります。
- (6) [候補者名] みたいに嘘とデマで大衆扇動しようとした人 は許しちゃいけないと思うのよね。

その他は,(7),(8)のように単に事実を述べているだけのツイートや,(9)のように当選を期待する内容とも落選を期待する内容とも判断できないツイートに対して付与する.

- (7) [候補者名] 本日の街宣予定 新宿西口 12:00~13:00
- (8) 僕のとこに来た情報では、[候補者名] 6 位。
- (9) [候補者名] の街頭演説を見に来ました!

今回の実験で用いた 34名の候補者名を含むツイートの総数は 915,392 ツイートであり、その中の 662,332 ツイート (72.4%) はリツイートであった。ここで、候補者による分割交差検定 (後述)を公平に行うため、複数の候補者に言及しているツイートを削除する。この制約により、34名の候補者名を含むツイートの総数は 779,823 ツイートとなり、その中の 553,486 ツイート (71.0%) はリツイートであった。また、ラベルを付与したツイートの種類数は 4,902 ツイートであり、リツイートによる重複を考慮すると 308,334 ツイート (39.5%) に対して正解のラベルを付与したことになる。このうち、「当選運動」は 1,193 ツイート (72,249 リツイート)、「落選運動」は 621 ツイート (52,225 リツイート)、「その他」は 3,088 ツイート (183,860 リツイート) であった。

2.2 分類器の構築

分類器の構築には、多クラスロジスティック回帰(最大エントロピー法)を用いた。分類器の素性として、ツイート本文中の単語ユニグラム、単語バイグラム、日本語評価極性辞書(名詞編)[東山 08] と日本語評価極性辞書(用言編)[小林 05] のエントリとの一致、ツイートを発信したユーザ名、プロフィールに分類学習の対象候補者名が含まれるか否かを用いた。本手法では、ラベルを付与したツイートを訓練データとして用いて分類学習を行い、生成された分類器を用いて訓練データ以外のツイートのラベルを推定する。生成したモデルについて、訓練データと評価データに同じ候補者の事例が入らないように34分割交差検定により評価したところ、精度は71.02%であった。

2.3 失敗事例・考察

分類器がラベル付けを失敗したツイートの例を以下に挙げる.

- (10) あえていいます [候補者名] を当選させれば東京都民の恥 である
- (11) [候補者名] さんご本人がただいま小岩駅前で演説されています。参議院議員選挙東京選挙区に立候補されています。

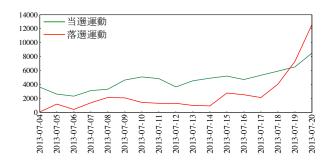


図 1: 候補者 A の当選運動・落選運動の推移

- (12) [候補者名] さん「今、本気で取り組まなければ。(途中省略) 市民の力で」
- (10) は「落選運動」ツイートであるが、分類器は「当選運動」と推定した。これは、文中に含まれる「当選」という形態素が分類器内で「当選運動」を強く示唆する素性となっていたためである。(11) は当選を期待する内容とも落選を期待する内容とも判断できないため「その他」ラベルが付与されるべきツイートであるが、分類器はどちらも「当選運動」と推定した。(12) はかぎ括弧内に該当候補者自身の話した内容が入っていて、このツイートをしたユーザはその事実のみを述べているだけである。このツイートには「その他」と推定すべきであるが、分類器は「落選運動」と推定しており、かぎ括弧内(引用部分)は素性生成の範囲外とするなどの改良が必要である。

2.4 当選運動・落選運動の推移

今回の実験で対象にした34名の候補者のうち,3名の候補者(候補者A,候補者B,候補者Cとする)について,構築した分類器を用いて選挙期間中の当選運動・落選運動の推移をプロットしたものを図1,図2,図3に示す。このように,当選運動・落選運動の推移を可視化することにより,選挙期間中の当選運動・落選運動の盛り上がりを知ることができる.

候補者 A に対する当選運動・落選運動の推移(図 1)を見ると、2013 年 7 月 18 日までは当選運動に比べて落選運動のほうが少ないものの、19 日から 20 日の選挙期間終了間際に落選運動が増加し、当選運動を上回る勢いとなっている。ただ、選挙日間際の落選運動に負けることなく、この候補者は当選した。候補者 B に対する当選運動・落選運動の推移(図 2)を見ると、13 日に落選運動が急増したことが分かる(この候補者は落選した)。候補者 C に対する当選運動・落選運動の推移(図 3)を見ると、選挙期間全体を通して当選運動に比べて落選運動が盛んであったことが分かる。このことから、この候補者は落選するのではないかと予測されるが、実際の選挙結果は当選であった。これは、4.3 節でも述べるように、選挙では様々な要因の複合で候補者の当落が決まるため、ツイッター上の当選運動・落選運動のみから当落を判断できないためである。

3. リツイート・ネットワークの分析

今回の選挙において、各候補を支持もしくは不支持のユーザはツイッター上でどのように繋がっていたのであろうか?このことを調べるため、リツイートネットワークを可視化し、似たような意見を持つユーザのクラスタを発見する。なお、本節の分析は岡崎らの手法 [岡崎 13] を踏襲している。

リツイート・ネットワークの構築方法は以下の通りである. あるユーザ A のツイートを別のユーザ B がリツイートした際

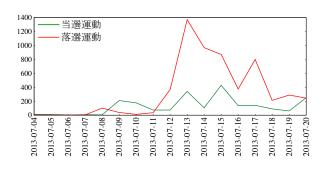


図 2: 候補者 B の当選運動・落選運動の推移



図 3: 候補者 C の当選運動・落選運動の推移

に、ユーザ A からユーザ B にリンクを作成する。リンクの重みはリツイートの回数とする。このような処理を、今回収集した全ツイートデータに適用すると、ユーザをノード、リツイート関係をリンクとしたネットワークができる。ネットワークはGephi*2というソフトウェアを用いて、可視化を行う。可視化では Gephi に実装されている ForceAtlas2 というアルゴリズムを用い、重みの高いリンクの距離が短くなるようにノード配置の最適化を行う。さらに、2. 節で述べた分類学習の結果を利用し、各ユーザ(ノード)の立場(候補者に対する支持・不支持)を色付きで表現する。ユーザ u が候補者に対してに関して当選運動の側に立つユーザか、落選運動の側に立つユーザかを、(1) 式で定量化する。

$$z(u) = \frac{1}{|T(u)|} \sum_{t \in T(u)} y(t)$$
 (1)

ここで,T(u) はユーザu が発信したツイートの集合,y(t) はツイートt の内容が候補者に対する当選運動であれば+1,落選運動であれば-1,どちらでもなければ0 を返す関数で,2. 節で説明した分類器の出力そのものである.したがって,z(u) はユーザu が候補者に対して当選運動を行うユーザであれば+1 に近くなり,落選運動を行うユーザであれば-1 に近くなる.リツイート・ネットワークの可視化では,z(u) が-1 に近いノードを赤色,+1 に近いノードを緑色で彩色する.

2. 節において対象となっていた 34 名の候補者のうち、2.4 節で説明した 3 名の候補者のリツイート・ネットワークを可視化した.候補者 A,候補者 B,候補者 C のリツイート・ネットワークを,図 A,図 B にそれぞれ示す.候補者 A のネットワーク図(図 B)では,当選運動を行うユーザ(緑色のノード)からなるクラスタと,落選運動を行うユーザ(赤色のノード)からなるクラスタに,二分されている.これは,当選

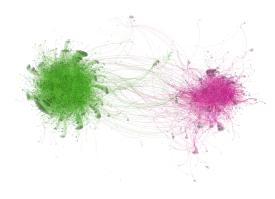


図 4: 候補者 A のリツイート・ネットワーク図

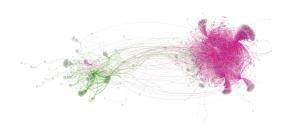


図 5: 候補者 B のリツイート・ネットワーク図

運動・落選運動を行うユーザはそれぞれ自分と同様の意見を持つユーザのツイートをリツイートする一方、当選運動を行うユーザから落選運動を行うユーザへのリツイート、またその逆のリツイートは非常に少ないということを表している。この特徴は、候補者 B のネットワーク(図 5)と候補者 C のネットワーク(図 6)にも見受けられる。ここで特筆すべきは、ツイートの内容から当落を予測する処理と、リツイート関係からクラスタを可視化する処理が独立であることで、ツイートの内容だけでなくユーザ間のつながりから当落運動を認識できることは興味深い現象と言えよう。

4. 当選運動・落選運動と実際の当落の相関

これまでに説明した分析に基づくと、各候補者の当落をどのくらい予測できるのであろうか? 本節では、2. 節で構築した分類器による当選運動・落選運動の盛り上がりと、候補者の実際の当落の相関を調べる。

4.1 分析手法

はじめに、2. 節で説明した手法を用いて、34名の候補者の名前を含む全てのツイートに「当選運動」「落選運動」「その他」のラベルを付与する。そのうえで、各候補者cに対して(2)の計算を行い、 $\operatorname{score}(c) \geq 1$ の候補者に対して「当選」と予測し、 $\operatorname{score}(c) < 1$ の候補者を「落選」と予想する。

$$score(c) = \frac{pos(c)}{neg(c)}$$
 (2)

ここで、pos(c) は候補者 c の名前を含むツイートのうち、「当選運動」のラベルが付与された数、neg(c) 「落選運動」のラベルが付与された数である。なお、選挙区選挙に関して、同一選挙区で立候補している候補者に式 2 で当落を予測する際に、その選挙区の定数を超える候補者に対して「当選」を予測しないようにするため、score(c) の大きい順に定数分だけ「当選」と

^{*2} https://gephi.org/

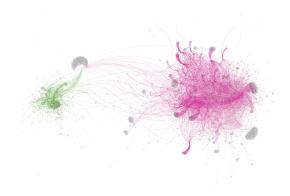


図 6: 候補者 C のリツイート・ネットワーク図

表 1: 分類器が推定した当落運動と実際の当落との相関

ラベル	適合率	再現率
当選	75.00% (18/24)	81.82% (18/22)
落選	60.00% (6/10)	50.00% (6/12)

予測し、残りの候補者に対しては「落選」と予測した.

4.2 実験結果

以上の方法で予測される「当選」「落選」と、候補者の実際の当落結果を比較し、評価を行った。表1に、「当選」「落選」についての適合率と再現率を示す。対象の34名の候補者中24名について提案手法の予測と実際の当落が一致し、当落の予測精度は70.59%であった。この実験結果により、当落注目の候補に関して、ツイッター上の当落運動と実際の当落との間には、一定の相関があることが分かった。特に、「当選」についての適合率・再現率は約8割という優れた性能を示した。

4.3 考察

これまでの実験結果から、当落が注目される候補に対して、 実際の選挙結果をある程度予測できることが分かった。しかし ながら、このようなアプローチが全ての選挙区で通用する訳で はなく、様々な要因により当落の予測に失敗する恐れがある。

政党政治が浸透した現代では、各候補者の当落が所属する政党の影響を受けることが多い。例えば、同一の選挙区に立候補している2人の候補者では、より支持層の厚い政党に属する候補者のほうが当選しやすい傾向にある。また、圧倒的な知名度を持つ候補者は、その他の候補者と比較すると有権者から注目されるので、当落運動が盛り上がりやすい。以上のことから、各候補者の所属政党や知名度を考慮しつつ、当落運動の盛り上がり度合いから当落を予測することが望ましい。

さらに、インターネット上で当選運動・落選運動を行うユーザと、実世界の有権者の層が乖離することがあるため、インターネット上の選挙運動の盛り上がり度合いから当落を予測するには限界がある。例えば、インターネット上で当選運動が盛んに行われている立候補者と、同一選挙区内でそれほどインターネット上での選挙運動が行われていない立候補者がいたとしても、後者がインターネットを利用しない有権者層から厚い支持を受けている場合、インターネット上の選挙運動から当落を予測するのは困難となる。

以上のような様々な理由により、本手法のみで全ての選挙区の当落を予測するには限界があると考えられる。しかし、当選か落選かで注目を浴びるような候補者に関しては、メディアの取材情報や世論調査の分析に加えて、インターネット上での選

挙運動の分析結果も、当落予測の有用な判断材料になり得る. 2. 節の当落運動の分類器の交差検定は候補者による分割という厳しい条件を用いたが、それでも約7割の分類精度を達成している. したがって、本研究で構築した学習データや分類器は、今後の選挙の際にもそのまま転用可能であると期待される. 今後の政治の情勢次第では、インターネット上の選挙運動がより一層活発となり、選挙期間中に候補者や有権者から発信される情報は増加する. それに伴い、人手のみによる選挙運動の分析はさらに困難になると思われ、本研究がネット選挙の情勢分析の一助となると期待している.

5. **おわり**に

本研究では、ツイッター上での当選運動・落選運動の分析を行い、またそれらと実際の選挙結果との相関を考察した。ツイッター上での当選運動・落選運動は、ツイートの内容の分類やネットワーク分析に基づき、ある程度の精度で認識が可能であることが分かった。本研究の延長として、候補者や有権者はツイッター上での選挙運動の状況をモニタリングしたり、候補者の当落が語られる際のトピックを抽出するなどの応用が考えられる。また、ツイッターを用いた分析に加え、メディアの取材情報や世論調査と組み合わせることは、選挙期間中の世の中の動きを知る上で、重要課題であると認識している。今後、選挙における SNS の活用がさらに進むことが予想される。候補者の公式ツイッターの利用状況やコミュニケーションなどから、候補者の考え方や政策などを抽出し、有権者がネット選挙のメリットを享受できるようなシステムを構築していきたいと考えている。

謝辞

本研究は、東北大学工学部 情報知能システム総合学科「Step-QI スクール」の支援を受けた。

参考文献

- [東山 08] 東山昌彦, 乾健太郎, 松本裕治: 述語の選択選好性に 着目した名詞評価極性の獲得, 言語処理学会第 14 回年次 大会論文集, pp.584-587 (2008)
- [小林 05] 小林のぞみ,乾健太郎,松本裕治,立石健二,福島 俊一:意見抽出のための評価表現の収集,自然言語処理, Vol.12, No.3, pp.203-222 (2005)
- [Larsson 12] Larsson, A, O. and Moe, H.: Studying political microblogging: Twitter users in the 2010 Swedish election campaign, New Media & Society 14.5, pp.729-747 (2012)
- [岡崎 13] 岡崎直観, 佐々木彬, 乾健太郎, 阿部博史, 石田望: ツイッター分析に基づく福島県産桃に対する風評の実態解明とその対策, 第 26 回日本リスク研究学会年次大会, B-5-3 (2013)
- [Tumasjan 10] Tumasjan, A., Sprenger, T, O., Sandner, P, G., and Welpe, I, M.: Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment, ICWSM 10, pp.178-185 (2010)
- [湯淺 13] 湯淺墾道:インターネット選挙運動の解禁に関する 諸問題, 情報セキュリティ総合科学 5 (2013)