1K3-1

Twitterにおける候補者の情報拡散に着目した国政選挙当選者予測

Predicting Japanese General Election in 2013 with Twitter: Considering Diffusion of Candidates' Tweets

那須野薫 *1　松尾豊 *1
Nasuno Kaoru　Matsuo Yutaka

*1東京大学
The University of Tokyo

Election result prediction using mciro blogging service Twitter is now very active these days. The research approach so far is mostly classified into two way: one is focusing on voters and the other is focusing on candidates. Most of the research focusing on candidates use only the number of followers to predict election results. Considering on rapid progress in Twitter analysis these days , we should be able to predict election results with more sophisticated analysis. In this paper, we predict who pass the Japanese general election in 2013 using not voters' features but candidates' features: 6 features on candidate's account (followers_count, friends_count, number of tweets, etc.) and 3 features on candidates' information diffusion (size, variety and loyalty of information diffusion). We conduct a prediction experiment with Random Forest through 10-fold cross validation. The result is that f-measure with the features on information diffusion is higher by about 12% than that with only the features on candidate's account. The result also indicates desirable state of candidates' Twitter account for success in election.

1. はじめに

近年，ソーシャルメディアを用いた予測研究が活発である．マイクロブログサービスの Twitter は分析のデータソースとして広く用いられており，140 文字以下の投稿であるツイートを解析することで，インフルエンザの流行予測 [1] やストックマーケットの動向予測 [2] が可能であることが示されている．

Twitter を用いた選挙結果の予測研究も同様に行われており，先行研究は有権者に焦点を当てる研究と候補者に焦点を当てる研究に大きく分けられる．有権者に焦点を当てた先行研究として，[3] や [4] などでは政党名や政治家名に言及した有権者のツイート数やツイートの感情分析による選挙結果の予測が行われた．しかし，有権者に焦点を当てた分析は相反する結果を出す研究も多く，その予測可能性に疑問が呈されている．候補者に焦点を当てた先行研究として，[5] ではソーシャルネットワーキングサービスの Facebook と Twitter における候補者の投稿の購読者数から選挙結果を予測したが，候補者に焦点を当てた先行研究は他にほとんどない．これまでの Twitter 分析の進展を踏まえると，より高度な分析による予測研究が可能であると考えられる．

さて，選挙の投票に際して，有権者は自分がよく知っている候補者の中から投票する候補者を選択する可能性が高く，従って，候補者にとって有権者に対する認知度向上は重要な課題である考えられる．Twitter 活用による認知度向上という課題に着目すると，候補者の投稿が拡散される規模や投稿を受け取る有権者の多様性，また，有権者が他の候補者の投稿を受け取る度合い等は非常に重要な要素であると考えられるが，これらを考慮した研究はまだ行われていない．

そこで，本稿では，候補者の Twitter における情報拡散に着目して国政選挙の当選者予測を試みる．まず，候補者の投稿の拡散を再投稿 (リツイート) により支援するユーザ (以下，情報拡散支援者) を定義する．次に，候補者の情報拡散を評価するため，情報拡散支援者を考慮した情報拡散の規模，多様度，候補者への忠誠度の 3 つの指標を提案する．Twitter から直接取得できる 6 つのアカウントの状態に関する指標（フォロワー数，フレンド数，選挙期間中のツイート数，被登録リスト数，アカウント承認の有無，存在日数．以下，指標 A.）に加え，本稿で提案する 3 つの情報拡散に関する指標 (以下，指標 B) を素性として教師あり学習により当選者予測する．候補者の選挙期間中のツイート 42,645 とそのリツイート 368,694 から指標 B を作成し，教師あり学習には学習後に素性の重みを確認でき，また広く用いられ良好な結果が得られている Random Forest[6] を用いる．

指標 A と指標 B の合わせて 9 指標を素性として，選挙の当選 (当選を 1, 落選を 0) を予測する．10 分割交差検定による予測モデルの評価の結果，指標 A と指標 B を同時に用いる提案手法は，候補者のフォロワー数のみを素性とする従来手法と比較して予測性能 (F 値) が約 70%高かった．また，指標 A と指標 B を同時に用いた予測では指標 A のみを用いた予測よりも F 値が約 12%高く，本稿で提案の情報拡散に関する指標が予測精度向上に寄与していることが示された．また，提案手法による予測における各素性の重みや選挙当落との相関から，候補者が登録されているリストの数の多さが選挙当選に大きく関わっていることやフレンド数は少ない方が当選しやすいこと，アカウント認証の有無は選挙当落に無関係であること，情報拡散の規模や忠誠度は重要であるが多様度は選挙当選にあまり寄与しないことが示唆された．

本稿の構成は以下の通りである．まず次章で選挙予測分析に関する関連研究について概説する．3 章で本研究のアイデアについて，4 章でその実装である指標 B について説明する．5 章で予測実験に用いるデータの取得方法やデータの概観について述べ，6 章で予測実験を通して従来手法より提案手法が優れていることを示し，7 章で予測性能向上のための課題を整理し，8 章でまとめる．

2. 関連研究

Twitter を用いて選挙結果の予測を行う先行研究は有権者に焦点を当てる研究と，候補者に焦点を当てる研究に大きく分けられる．有権者に焦点を当てる先行研究では，相反する分析結

連絡先: 那須野薫 東京大学 nasuno@weblab.t.u-tokyo.ac.jp

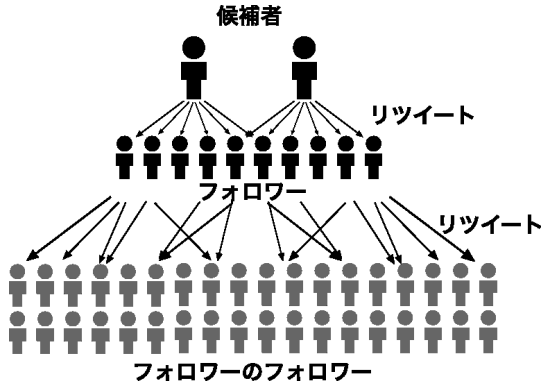

図 1: Twitter における情報拡散のイメージ図．ツイートはフォロワーとリツイートしたユーザのフォロワーに拡散する．

果を出すものも多く議論が紛糾している．2010 年に [3] では 2009 年のドイツ議会選挙において政党名に言及するツイート数が比較され，その数で各政党の得票数を予測できると結論づけた．しかし，2012 年に [8] では再実験の末，ツイート収集期間や選択政党が恣意的であり予測性能はないとして [3] を否定した．感情分析を利用した予測研究について，2010 年に [9] では米大統領選における候補者名や選挙関連語などを含むツイートの感情分析の結果と選挙に関する世論調査の結果を相関づけることに成功した．しかし，2011 年に [10] では米議会選挙を対象に候補者名を含むツイートの感情分析の結果を選挙予測に利用したが良い結果は得られず，ソーシャルメディアのデータを予測のためにブラックボックスとして利用すべきでないとした．これらを受けて，2011 年に [11] は同様に感情分析を米国の上院議員選挙の予測に利用したが良い結果は得られず，語の極性だけでは選挙結果を予測できないと結論づけた．このように，有権者に焦点を当てる先行研究は相反する結果を出すものも多く，予測可能性に疑問が呈されている．

候補者に焦点を当てる先行研究として，2013 年に [5] では候補者の投稿の購読者数から選挙結果を予測した．Facebook と Twitter における候補者の投稿の購読者数推移を同時に用いずにそれぞれ別の予測モデルの素性として利用し，線形回帰やロジスティック回帰により候補者の選挙当落を予測した．しかし，選挙結果とソーシャルメディアにおける購読者数は統計的に有意な関係があるものの，利用による効果は小さく選挙当落に影響を与えるのは僅差で争っているときだけであろうと結論づけた．候補者に焦点を当てた先行研究は他にほとんどなく，また，近年の Twitter 分析の進展を踏まえるとより高度な分析が可能であると考えられる．

3. 本研究のアイデア

本章では，本研究のアイデアについて説明する．まず，Twitter における情報拡散について，そのイメージを図 1 に示す．ユーザによるツイートの投稿はまずユーザのフォロワーに拡散され，次にユーザのフォロワーのうちツイートをリツイートしたユーザのフォロワーに拡散していく．従って，ユーザの情報拡散の規模はフォロワー集合の大きさだけでなく，リツイートしたユーザの数やそのユーザのフォロワー数にも依存している．人気ユーザによっては 1 ツイートあたり 100 以上リツイートされるもの多く，リツイートによる情報拡散への影響は小さくないと言える．

第二に，情報拡散の規模が等しい 2 ユーザについて考える．ツイートを受け取るユーザが知り合い同士である割合が高い場合は，ユーザが同じコミュニティに所属している可能性が高く，逆に，知り合い同士でない場合はユーザが異なるコミュニティ所属している可能性が高いと考えられる．情報拡散による認知度向上の点では，より多様なユーザに対する露出が多い方が望ましく，また，共通の興味関心によって成長するネットワークは粗な状態で拡大しやすいとする研究 [7] を考慮すると，ソーシャルネットワークの拡大という点からも構成ノードであるユーザの多様である方が良いため，ツイートを受け取るユーザの多様性は重要である考えられる．

第三に，情報拡散の規模と多様性が等しい 2 ユーザのついて考える．有権者は投票に際して，自分が良く認知している候補者の中から投票する候補者を選択する可能性が高いため，ある候補者 A にとって，そのツイートを受け取るユーザは競争相手の候補者 B のツイートを受け取らない状態の方が望ましい．得票率の向上という点で，候補者にとっては情報拡散は排他的である方が良く，すなわち候補者への忠誠度が高い方が良く，情報拡散を支援する情報拡散支援者の候補者への忠誠度（以下，情報拡散の忠誠度）は重要であると考えられる．

以上の考察から，ユーザの情報拡散について，情報拡散の規模，情報拡散の多様度，情報拡散の忠誠度を考慮することで，選挙当落に関する候補者の状態をより精度高く捉えることができると考えられる．

4. アイデアの実装

本章では，前章の議論に基づき 3 つの情報拡散に関する指標 (指標 B) の実装について説明する．情報拡散は候補者と候補者のツイートをリツイートすることで支援する情報拡散支援者によって行われると考えられるため，まず情報拡散支援者を定義し，その上で指標 B を定義する．

情報拡散支援者はユーザが候補者のツイートを拡散する度合いに基づいて定義する．候補者 C の期間中のツイート数を N，C のツイートをリツイートしたユーザを u_i，u_i がリツイートした C のツイート数を n_i とすれば，C のツイートにおける u_i のリツイート率は $RTrate_i = n_i/N$ となる．u_i のフォロワー数を fc_i とすれば，u_i が候補者の 1 ツイートを拡散するユーザ数の期待値 $reach_i$ は $reach_i = RTrate_i \times fc_i$ となる．ここで，$reach_i \geq \alpha$ を満たす u_i を C の情報拡散支援者と定義する．評価実験での計算結果から $\alpha = 100$ とした．

次に指標 B を定義する．前章の議論に基づき情報拡散の規模，情報拡散の多様度，情報拡散支援者の忠誠度を定義する．

- 情報拡散の規模：情報拡散支援者のリツイートを考慮した候補者アカウントのツイートを受け取るユーザ数の期待値と定義する．
- 情報拡散の多様度：候補者 C とその情報拡散支援者の集合を A，$\{a_i \in A\}$ の a_i が候補者の 1 ツイートを拡散するユーザ数を $reach_i$，a_i が $\{a_j \in A; i \neq j\}$ の a_j と相互にフォローしている関係でない割合を $variety_i$ とし，情報拡散の多様度を $\sum_i (reach_i \times variety_i) / \sum_i reach_i$ と定義する．
- 情報拡散の忠誠度：a_i が拡散する全候補者のツイートに対する C のツイートの割合を $loyalty_i$ とし，情報拡散の忠誠度を $\sum_i (reach_i \times loyalty_i) / \sum_i reach_i$ と定義する．ただし，a_i が候補者の場合 $loyalty_i$ は 1 とする．

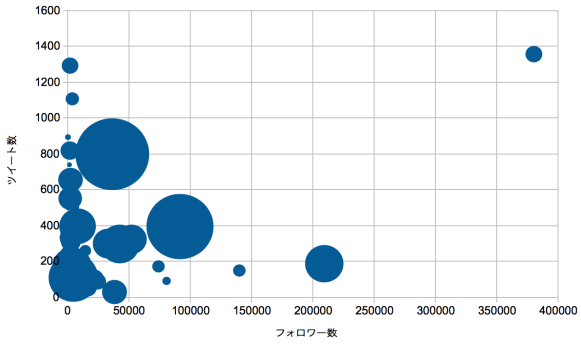

図 2: 各候補者のフォロワー数と選挙期間中のツイート数とツイートをリツイートされた回数（被リツイート数）の関係．バブルの大きさは候補者の被リツイート数を表す．

表 1: 10 分割交差検定による予測モデルの評価．

手法	accuracy	precision	recall	F-measure
ランダム予測	0.607	0.268	0.268	0.268
従来手法	0.702	0.508	0.280	0.335
RF 指標 A	0.766	0.573	0.455	0.507
RF 指標 AB	**0.780**	**0.658**	**0.499**	**0.568**

5. データセット

本章では，分析に用いるデータの取得方法と取得したデータの概観について述べる．まず，分析に用いるデータの取得方法について述べる．候補者のアカウントの状態に関する指標 (指標 A) や，情報拡散に関する指標 (指標 B) の算出に用いるデータを Twitter の REST API を用いて取得する．候補者のツイートとそのリツイートについては，候補者がインターネットを活用した選挙運動を行える期間 (以下，選挙期間) に投稿されたものを取得する．指標 A に用いるデータは REST API を用いて直接取得することができるものを選挙期間開始時と選挙期間終了時の 2 時点で取得する．具体的には，フォロワー数，フレンド数，被登録リスト数，選挙期間中のツイート数，アカウント認証の有無，存在日数の 6 つの指標を指標 A として採用する．被登録リスト数は候補者アカウントを含むリストの数で，存在日数はアカウントを作成してからの経過した日数である．本稿では，2013 年の参議院議員選挙を対象としてデータを収集し，選挙期間中に Twitter を利用していた 287 人の候補者のツイートを 42,645，そのリツイートを 368,694 取得した．また，対象データとなる 287 人の候補者のうち当選した候補者は 77 人であった．

次に，取得したデータの概観を示す．各候補者のフォロワー数と期間中のツイート数，ツイートがリツイートされた回数（被リツイート数）の関係を図 2 に示す．フォロワー数と被リツイート数の相関係数およびツイート数と被リツイート数の相関係数はそれぞれ 0.283，0.312 と小さく，必ずしもフォロワー数や期間中のツイート数が大きければ，より多くのリツイートによる情報拡散を期待できるわけではないことが分かる．

6. 予測実験

本章では，予測実験を通して従来手法より提案手法が優れていること示す．

表 2: RF 指標 AB の予測結果における各素性の重みと選挙当落との相関係数．

カテゴリ	素性	素性の重み	相関係数
指標 A	フォロワー数	**0.102**	0.124
	フレンド数	**0.235**	-0.0376
	選挙期間中のツイート数	0.0838	-0.0632
	被登録リスト数	**0.242**	0.236
	アカウント認証の有無	0.00154	0.0563
	存在日数	0.0790	0.0383
指標 B	規模	**0.100**	0.114
	多様度	0.0592	0.0815
	忠誠度	**0.0970**	0.113

各候補者を指標を組み合わせて素性ベクトルとして表現し，予測モデルに利用する．予測は教師あり学習で行い，学習後に各素性の重みを確認でき，また広く用いられ良好な結果が得られている Random Forest を利用する．Random Forest には機械学習ライブラリの Scikit-learn[12] を用い，10 分割交差検定により予測モデルを評価した．

予測結果を表 1 に示す．ランダム予測は 77/287 の確率で当選と予測するものでベースラインとして設けた．従来手法は [5] の予測モデルのデータセットに選挙期間開始時と投票前日のフォロワー数を推移データとして利用し，かつ，教師あり学習にランダムフォレストを用いるという条件下での予測実験の結果である．また，RF 指標 A は指標 A のみを利用した予測の結果で，RF 指標 AB は指標 A と指標 B を同時に利用した予測（提案手法）の結果である．RF 指標 AB の予測結果では予測性能を表す F 値が従来手法よりも約 70%高く，提案手法が従来手法より優れていることがわかる．また，RF 指標 AB の F 値は RF 指標 A の F 値よりも約 12%高く，本稿提案の情報拡散に関する指標もまた予測精度向上に寄与していることが分かる．

次に，RF 指標 AB の予測について，Random Forest の学習から得られた各素性の重みを表 2 に示す．素性の重みだけでは，素性が大きい方が当選に寄与するのか小さい方が当選に寄与するのかが分からないため，各素性と選挙当落（当選を 1，落選を 0）への相関分析を行い，相関係数 [*1] も併せて記載した．素性の重みと相関係数の絶対値の大小は概ね一致している．候補者をリストに登録するということは他のユーザとは分けてツイートを受け取るということであり，そのような熱心なユーザに関心を持たれる方が当選しやすいということが推察される．また，Twitter では歌手やタレントなど人気のあるユーザはしばしばフォロワー数が大きい一方で，フレンド数が非常に小さいということがあるが，そのようなユーザの方が当選しやすいということが示唆されている．一方で，アカウント認証の有無は重みが最も小さく候補者アカウントが Twitter により本人の認証がされているかどうかは選挙の当落とはほとんど関係ないと言える．情報拡散の多様度の重みもネットワークの規模や忠誠度と比べると小さく，情報拡散が多様であるか否かよりは，どれだけ多くの人に声が届くかや，より情報拡散支援者が候補者に対して忠誠であることの方が当選への貢献は大きいと考えられる．従って，従来手法と比べ予測性能が高かったのは，指標 A のフレンド数や被登録リスト数，指標 B の情報拡散の規模や忠誠度などが候補者の状態と当選の関係をよりよく捉えていたためではないかと考えられる．

*1 相関係数の方が精度が低い

7. 考察

本稿で作成した指標 B は予測性能を向上させたが，指標 A と指標 B を同時に利用した予測における F 値は 0.568 と高くはなかった．そこで，本稿の提案手法を日本の国政選挙に利用する際の課題について考察する．本稿の分析には Twitter の利用率，予測モデルの設計について課題があると考えられる．第一に Twitter の利用率について，当該選挙において候補者 433 人のうち 287 人（66%）が，当選者 121 人のうち 77 人（64%）が Twitter を利用した．インターネット選挙運動が解禁された初めての国政選挙であったことや候補者の多くの方が年配の方（候補者の平均年齢は 51 歳）であったことで全体の利用率が高くなかったのではないかと考えられる．年配の人よりは若い人の方がより Twitter を利用していたと考えられ，今後の国政選挙ではより多くの候補者の Twitter 利用が期待できると考えられる．

第二に予測モデルの設計について，本稿では全ての候補者に対して同様に当落の予測を行い，当該選挙の選挙方式を考慮しなかった．参議院選挙では選挙方式が選挙区制と比例区制で異なり，選挙区制に出馬する候補者は出馬した選挙区内で他の候補者と票を競い，比例区制に出馬する候補者は比例区制に出馬する全国の候補者と票を競う．また，比例区制では候補者の票だけでなく所属する政党の票も当落に考慮される．このように，候補者によって選挙方式や票を競う対象が異なるが，本稿では候補者全体における Twitter 利用率が十分高くなかったため，選挙方式を考慮せずに予測を行った．今後行われる選挙については選挙方式を考慮したモデル設計をすることで，予測性能を向上できる可能性があると考えられる．

8. まとめ

本稿では，候補者の Twitter における情報拡散に着目して国政選挙の当選者予測を行った．予測モデルの評価の結果，提案手法は候補者のフォロワー数のみを素性とする従来手法と比較して予測性能 (F 値) が約 70%高く，提案手法が従来手法よりも優れていることが示された．また，Twitter から直接取得できる指標（指標 A）と本稿提案の情報拡散に関する指標（指標 B）を同時に用いた予測では指標 A のみを用いた予測よりも F 値が約 12%高く，情報拡散に関する指標が予測精度向上に寄与していることが示された．また，当選するために Twitter における望ましい状態について，候補者が登録されているリストの数の多さが選挙当選に大きく関わっていることやフレンド数は少ない方が当選しやすいこと，アカウント認証の有無は選挙当落に無関係であること，情報拡散の規模や忠誠度は重要であるが多様度は選挙当選にあまり寄与しないことが示唆された．また，予測実験の結果を踏まえ，本稿の提案手法を日本の国政選挙に利用する際の課題について考察した．

本稿が今後のインターネット選挙運動の活性化に貢献すれば幸いである．

参考文献

[1] 荒牧 英治，増川 佐知子，森田 瑞樹：Twitter Catches the Flu:事実性判定を用いたインフルエンザ流行予測，情報処理学会研究報告. SLP，2011.

[2] Johan Bollen, Huina Mao, Xiaojun Zeng: Twitter mood predicts the stock market, Journal of Computational Science, Vol.2, Issue 1, March 2011, Pages 18, 2011.

[3] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe: Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 2010.

[4] Erik Tjong Kim Sang, Johan Bos: Predicting the 2011 dutch senate election results with Twitter, Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 5360, 2012.

[5] Michael P. Cameron,Patrick Barrett,Bob Stewardson: Can Social Media Predict Election Results? Evidence from New Zealand, Working paper in economics; 13,08, 2013.

[6] 波部 斉，ランダムフォレスト，情報処理学会研究報告 Vol.2012-CVIM-182 No.31，2012.

[7] Sanjay Ram Kairam, Dan J. Wang, Jure Leskovec: The life and death of online groups: predicting group growth and longevity, Proceeding WSDM '12 Proceedings of the fifth ACM international conference on Web search and data mining Pages 673-682. 2012.

[8] Andreas Jungherr, Pascal Jurgens, and Harald Schoen: Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. ‘ ‘ Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment ’’, Social Science Computer Review 30(2) 229-234 2012.

[9] Brendan O 'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, Noah A. Smith: From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 2010.

[10] Daniel Gayo-Avello, Panagiotis T. Metaxas and Eni Mustafaraj: Limits of Electoral Predictions Using Twitter, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011.

[11] Jessica Chung and Eni Mustafaraj: Can collective sentiment expressed on twitter predict political elections?, Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, 2011.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research Vol.12, 2011.