ツイートデータに基づく政党支持率の推定

花岡 見帆[†] 馬 青[†] 村田 真樹[††]

[†]龍谷大学理工学研究科数理情報学専攻 [††]鳥取大学大学院工学研究科情報エレクトロニクス専攻

1 はじめに

昨今、マスメディアでは政党支持率を中心に議論が行われている。政党支持率は世論調査[1]で予測される。世論調査方法としてマスメデイアが最も用いている方法は電話調査である。しかしながら質問者と回答者間でバイアスがかかることや、マスメディアの中には特定の政党を支持する傾向があることなどから、多くの問題が考えられており世論調査に対して疑念の声が多く上がっているのが現状である。

本研究では新たな政党支持率の予測を行うことを目標にシステム開発を行う。近年、SNS を用いた政党の支持率や選挙結果の予測を行う研究が多くなされている。従来の研究では候補者のフォロワー数や、候補者のツイートへの有権者の「いいね」やリツイートの数に着目した研究が多いが、最近、Twitter のツイート内容から政党の支持率を予測するという研究がなされている[2]。この研究は Twitter の各政党を含むツイートを対象に感情情報の抽出を行い、学習データとして目的変数にマスメディアが出した政党支持率を定めて重回帰分析を行い、説明変数の係数を求めることで政党支持率を予測することを可能にした。しかし本研究はマスメディアが出した政党支持率への疑問提起から始めたものなので、マスメディアが出した政党支持率を使用せずに予測を行う。

2 研究の概要

本研究では個々のユーザーに着目し、個々のユーザーが発したツイートから政党支持率を予測することを目的とする。個々のユーザーに着目するため、ユーザーごとにツイート取得を行う。Twitter には政党に対しての発言がないユーザーも多く存在するため、一度でも政党に対してツイートをしたことがあるユーザーに限定して取得を行う。政党をキーワードとして用いてツイートの収集を行い、政党を含んだツイートをしたユーザーを登録するユーザー辞書を作成する。登録ユーザーのツイートを取得し、ユーザーごとにツイートを管理するユーザーデータベースを作成する。登録ユーザー数が膨大なため、政党支持率の予測では調査対象とするユーザー数に制限を行う。ユーザーの中には、政党に対してのツイート件数が極端に多いまたは少ないユーザーが含まれているため、ツイート件数を考慮して調査対象ユーザーの決定方法を複数提案する。調査対象ユーザーのツイートからユーザーの政党支持・不支持判定を行い、任意月の政党支持率を予測する。ユーザーの中には任意月から最新ツイートでは特定の事象について否定的なツイートをしていても、過去のツイートから今でも肯定的な場合がある。よって政党支持・不支持判定に反映させるツイートの制限を行う。ツイート日時や政党を含む過去のツイート件数を考慮して判定対象ツイートの決定方法を複数提案する。そして判定対象ツイートを機械学習にて推定を行い、推定結果からユーザーの政党支持・不支持を判定し、任意月の政党支持率の予測を行う。提案した方法の有用性を確認するため、方法の比較および考察を行う。なお、本研究では政党を「自民党」に限定して研究を行う。

3 ユーザー辞書とユーザーデータベースの作成

Twitter API[3]を用いて2018年5月から10月末まで政党を含むツイート(ツイートID・アカウント名・ツイート日時が付与)を取得した。本稿では「自民党」に限定しているため、ツイート取得用のキーワードは「自民党」である。そして取得したツイートに付与されているツイート ID を登録するユーザー辞書を作成した。ユーザー辞書には約15万ユーザーが登録されている。また調査対象ユーザーの決定方法では、ユーザーの「自民党」を含むツイート回数を考慮する。よってユーザー辞書にはユーザーIDと共にそのユーザーの「自民党」を含むツイート件数も記した。そしてユーザー辞書に登録されたユーザーに対してツイート取得を行い、ユーザーごとにツイートを管理するユーザーデータベースを作成した。

4 調査対象ユーザーの決定方法

調査対象となるユーザーの決定方法として以下の3つの方法を提案する。いずれの方法もユーザー辞書を用いる。

- ランダムに調査対象ユーザーを決定する(以降、この方法を「**ランダム**」と呼ぶ)
- 「自民党」を含むツイート件数が多いユーザーから調査対象ユーザーとする(以降、この方法を「**ツイート頻度順**」と呼ぶ)

- 「自民党」を含むツイート件数を参考に層化を行い、各層からランダムに調査対象ユーザーとする。ツイート数を基準に以下の2つ方法で層化を行う。N を層化基準の個数とする。
  I. N＋1層に層化(以降、この方法を「**層化抽出法Ⅰ**」と呼ぶ)
  II. N 層に層化　最小の層化基準より少ない層を除く(以降、この方法を「**層化抽出法Ⅱ**」と呼ぶ)

層化抽出法の比率配分法を参考に「層化抽出法Ⅰ」と「層化抽出法Ⅱ」を作成した。層化の基準からユーザーの層化を行い、各層のユーザー比率を求める。比率を変えずに和が指定ユーザー数となるように各層の調査対象ユーザー数を決め、各層からランダムに調査対象ユーザーを抽出する。下記に層化抽出法Ⅰ・Ⅱの例を挙げる。1,000 ユーザーから 20 ユーザーを抽出する場合、表 1 のように設定し、ユーザーの層化が出来たとする。その時層化抽出法Ⅰ・Ⅱは表2ようにユーザー比率と各層の調査対象ユーザー数を求める。

表 1 ユーザーの層化

層化基準	{100 ツイート,20 ツイート}	指定ユーザー数	20
層化基準から ユーザーの層化	{100 以上, 100 より少なく 20 以上, 20 より少ない}		
	{200 ユーザー, 300 ユーザー, 500 ユーザー}		

表 2 層化抽出法Ⅰ・Ⅱのユーザー比率

方法	各層のユーザー数	各層の ユーザー比率	各層の調査対象 ユーザー数
層化抽出法Ⅰ	{200, 300, 500}	{2, 3, 5}	{4, 6, 10}
層化抽出法Ⅱ	{200, 300}	{2, 3}	{8, 12}

5　判定対象ツイートの決定方法

判定対象ツイート決定方法として以下の3つを提案する。いずれの方法もツイートのツイート日時情報を用いる。

- 指定月内のツイートのみを対象とする(以降、この方法を「**指定月分(specified month)**」と呼ぶ)
- 指定月内の最新ツイートから、指定ツイート件数分を対象とする(以降、この方法を「**指定ツイート件数分(several tweets)**」と呼ぶ)
- 指定月を含む、指定数ヶ月分のツイートを対象とする(以降、この方法を「**指定月間分(several months)**」と呼ぶ)

例えば2018年9月の政党支持率を予測する時、「指定月分」の場合、2018年9月の「自民党」を含むツイートが判定対象ツイートとなる。「指定ツイート件数分」の場合で指定ツイート件数を10件と指定していた時、2018年9月末から過去にさかのぼってツイート 10 件分が判定対象ツイートとなる。「指定月間分」の場合で指定月間数を3ヶ月と指定していた時、2018年9月, 8月, 7月のツイートが判定対象ツイートとなる。

6　政党支持率の予測

6.1 政党支持率算出処理の流れ

政党支持率の算出手順を下記に述べる。

1. 調査対象ユーザーの決定を行う(4節)。
   各調査対象ユーザーに対して以下の処理を行う。
   1.1 判定対象ツイートの決定を行う(5 節)。
   1.2 判定対象ツイートのベクトル化および機械学習を用いて政党支持の推定を行う。
   1.3 手順 1.2 の推定結果から平均を求める(以降、求めた値を「**ユーザー支持度**」と呼ぶ。)
   1.4 ユーザー支持度から、政党支持・不支持を判定する。ただし、支持判定方法は以下の二つを用いる。X をユーザー支持度とする。
   - X≦-1/3: 不支持, -1/3＜X＜1/3:言及していない, X≧1/3:支持(以降、この方法を「**3分割法**」と呼ぶ)
   - X＜0: 不支持, X＝0:言及していない, X＞0: 支持(以降、この方法を「**比率法**」と呼ぶ)
2. 手順 1.4 で支持と判定されたユーザー数/(対象ユーザー数 － 対象ツイートが無いユーザー数)を求め、政党支持率を算出する。

手順 1.2 については先行研究[4]で得られた手法を用いる。これについては 6.2 節に述べる。手順 1.2 では「肯定」,「否定」,「言及していない」の3クラス推定を行い、「肯定的」なツイートの評価値は+1、「批判的」なツイートの評価値は-1、「言及していない」ツイートの評価値は 0 とする。よって手順 1.3 のユーザー支持度は、-1 から+1 の範囲の実数値で表したものとなる。+1 に近い値であるほど肯定的なツイートが多く、-1 に近い値であるほど否定的なツイートが多いユーザーとなる。

手順1.2の判定対象ツイートの決定では、指定月付近にツイートを行っておらず判定対象ツイートがないため、手順 1.2 以降の処理を行えないユーザーが出てくる。そのユーザーは「無回答者」とみなす。よって手順2の政党支持率算出式の分母では対象ユーザー数と「無回答者」の差を使用する。

6.2 ツイートの政党支持推定

6.1節の「政党支持率算出処理の流れ」の手順1.2の機械学習によるツイートの政党支持推定は先行研究[4]で行ったものである。

ツイートのベクトル化は、辞書の作成と、辞書を用いた特徴ベクトルの作成からなる。辞書の作成では、本研究に特化した政党支持・不支持の観点から、反対勢力を批判した造語(例:ネトウヨ, パヨク)なども考慮した名詞辞書を作成した。さらに感情が関係しているため感情を表す形容詞辞書も作成した。ベクトルの作成では、辞書内の「単語の出現回数」を特徴ベクトルの要素とした。その単語は名詞辞書からトップ N 以内の単語、形容詞辞書からトップ M 以内の単語を抽出する。N と M の和は固定で 200 と設定し、N と M の比率を 1:3 として特徴ベクトルを作成した。そして機械学習 SVM にてグリッドサーチを行い、ハイパーパラメータの最適化を行った。N と M の数値と比率は開発データを用いて実験を行い、ツイート推定の平均精度が最も高いものを使用した(詳細は[4]を参照)。

7 実験

4, 5節に述べた方法と6.1節の支持判定方法を組み合わせて2018年の5月から11月の政党支持率を算出し比較および考察を行う。方法を表 3 にまとめる。方法の組み合わせは計24通りである。登録ユーザー数は156,121ユーザーであり、その中から10,000ユーザーを調査対象ユーザー決定方法で抽出する。

表 3 方法のまとめ

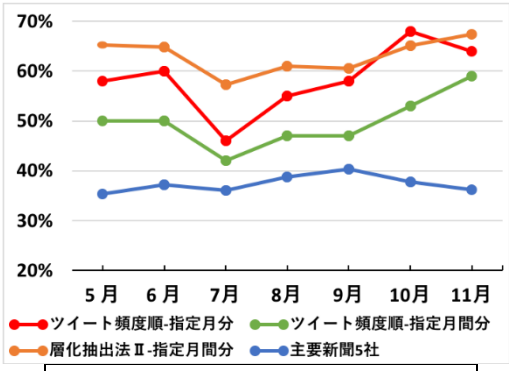
調査対象ユーザー決定方法	ランダム, ツイート頻度順, 層化抽出法, 層化抽出法Ⅱ
判定対象ツイート決定方法	指定月分, 指定ツイート件数, 指定月間分
支持判定方法	3分割法, 比率法

7.1 実験結果

実験結果から有効な方法の組み合わせを表 4 にまとめ、表4の組み合わせで予測した政党支持率の結果を図1に示す。

表 4 有効な組み合わせ

ツイート頻度順-指定月分(比率法)
ツイート頻度順-指定月間分(比率法)



層化抽出法Ⅱ-指定月間分(比率法)

図1:政党支持率の予測

青線の「主要新聞5社」は産経新聞, 読売新聞, 日経新聞, 毎日新聞, 朝日新聞の出した政党支持率の平均からなる。マスメディアの中には特定の政党を支持する傾向があることから平均を用いた。調査方法およびデータが異なっているため値ではなく政党支持率の変化に着目して比較および考察を行う。

5月から9月にかけて、いずれのグラフも同様の変化を見ることができた。しかし9月から変化に差が見られる。9月から10月にかけて「主要新聞5社」は下降しているのに対し、提案手法のいずれも上昇している。

このことは10月に取り上げられた記事の内容が原因であると考える。2018年の10月『東京都ヘイトスピーチ条例案　自民党以外賛成多数』という記事が多く見られた。ユーザーのツイート内容としては自民党に対して否定的または言及していないツイートであったが、肯定的なツイートと推定されていた。「賛成」というポジティブな単語が含まれていることにより、推定が正しく行えていなかったと考える。

また10月から11月にかけて「主要新聞5社」と「ツイート頻度順-指定月分」は下降しているのに対し、「指定月間分」を用いた曲線では上昇している。これは判定対象ユーザー決定方法の違いによるものだと考える。「指定月間分」では指定月間を3ヶ月と指定し、過去のツイートも考慮している。よって10月の推定が正しく行われていなかった「肯定」のツイートも判定対象ツイートに含まれてしまう。そのためユーザーの支持・不支持判定が上手く行えておらず、支持率が上昇してしまったと考える。よって「指定月間分」では指定月以外の推定ミスが反映されてしまうため、「指定月分」の方が有効だと考える。この問題の解決にはツイートの推定精度を向上させることが必要であり、先行研究[4]での特徴ベクトルの作成方法を改善する必要がある。

これ以降の節では 24 通りの実験結果を受けて、有効と有効ではない方法について比較および考察を述べる。

7.2 調査対象ユーザー

4種類の調査対象ユーザー決定方法の比較および考察を行う。「ツイート頻度順」では「自民党」に対するツイートが多い順に調査対象ユーザーとするため、政党支持・不支持判定をするのに十分なツイート件数が得られた。それに反して「ランダム」ではツイート件数が極端に少ないユーザーも調査対象ユーザーとしてしまう。極端にツイート件数が少ないため指定月付近にツイートされていない場合もあり、判定対象ツイートが得られず、判定するのに情報量が不十分となる。「層化抽出法 I」でも同様に、極端にツイート件数が少ないユーザーを含む層のユーザー数の比率が高くなってしまい、情報量の過疎化が見られたため有効ではなかった。そこで極端にツイート件数が少ないユーザーを調査対象ユーザーから除く「層化抽出法Ⅱ」を作成した。情報量を保ちつつ、統計に基づいた標本抽出が行うことができ有効であった。しかし情報量の制限から情報量を補うことが出来る判定対象ツイート決定方法「指定月間分」を用いる必要があること、そして判定対象ツイート件数を「ツイート頻度順」の方が多く得られることから、もっとも有効な方法は「ツイート頻度順」であった。

7.3 判定対象ツイート

3種類の判定対象ツイート決定方法の比較および考察を行う。「指定月分」では指定月の政党へのツイートを判定対象とするため、指定月の感情を顕著に得ることができ有効であった。「指定ツイート件数分」では頻繁に政党に対してのツイートを行うユーザーに対しては情報量を大幅に制限してしまう。また頻繁にツイートを行わないユーザーに対しては古すぎる過去のツイートも判定対象ツイートとしてしまうため、有効ではなかった。一方、ユーザーの中には指定月から最新ツイートでは偶然に政党に関係する事象について否定的なツイートをしているにすぎず、過去のツイートから政党に対して今でも肯定的な立場にあることを推測できることも考えられる。よって一定期間の過去のツイートを考慮する「指定月間分」を作成した。判定対象ツイート件数を増加させることができ、情報量の制限を行う「層化抽出法Ⅱ」にも有効であった。しかし実験結果より指定月以外の推定ミスも反映されることから、「指定月分」の方がより有効であった。

7.4 支持判定

2つの支持判定方法の比較および考察を行う。「3分割法」ではユーザー支持度-1 から+1 の実数値の範囲を3分割とし判定を行う。政党不支持ユーザーは、否定的なツイート件数が多いことからユーザー支持度が-1 となりやすい。それに対し政党支持ユーザーは肯定的なツイート以外も行っており、ユーザー支持度は 0 から+1 に偏りなく存在した。よって「3分割法」では極端なユーザーの判定しか出来ておらず、予測した政党支持率が極端に低くなった。よって「比率法」の肯定的なツイートと否定的なツイートの件数の比率で判定を行う方法が有効であった。

8 まとめ

本稿では個々のユーザーが発したツイートに着目し政党支持率を予測する手法を提案した。調査対象とするユーザーの決定方法では「自民党」を含むツイート件数を考慮して複数の方法を考案した。判定対象とするツイートの決定方法ではツイート日時や政党を含む過去のツイート件数を考慮して複数の方法を考案した。これらの方法を組み合わせ、ツイートの政党支持を機械学習で推定し、ユーザーの政党支持・不支持判定を行い、政党支持率の予測を行った。

予測結果および考察から、調査対象ユーザー決定方法には「ツイート頻度順」、判定対象ツイート決定方法には「指定月分」、支持判定方法には「比率法」が有効であった。

有効な方法の組み合わせによる政党支持率の予測は結果として、マスメディアの出した政党支持率と同様の変化が見られ、ある程度社会の流れをつかめた予測を行うことができたと考える。政党支持率自体が不確かなため、数値的に正当性を示すことはできない。しかし SNS でも政党支持率の予測や変化を見ることができると考える。各マスメディアの世論調査の電話調査では膨大な人手が必要なところ、本研究では機械学習を用いることで大幅な人員・コスト削減を可能にした。

参考文献

[1] 世論調査, https://ja.wikipedia.org/wiki/世論調査

[2] 増井, 藤野, 山本. 2017. Twitter の多軸的感情と政党支持率との関係について. 言語処理学会第 23 回年次大会, pp.222−225.

[3] Twitter Developer, https//developer.twitter.com/

[4] 花岡, 馬, 村田. 2019. SVM と LSTM を用いた政党支持に関するツイート推定. 言語処理学会第 25 回年次大会, pp.671−674.