

市民意見収集のための「保育園」に関するツイート投稿者の 属性表現の抽出と属性の推定

柏野 和佳子¹ 平本 智弥 関 洋平²

¹ 国立国語研究所 音声言語研究領域 ² 筑波大学 図書館情報メディア系

1 はじめに

ソーシャルメディアのマイクロブログ (Twitter を利用) の投稿から市民意見を収集し、自治体の行政の課題を改善するための意見抽出の研究を進めている (関 2017, 柏野ほか 2017, 柏野ほか 2018, 雨澤・関 2018, 安藤・関 2018, 柏野ほか 2020, 石田ほか 2021).

柏野ほか(2019), 柏野ほか(2020)では, 2015 年～2018 年の間に投稿された, つくば市, 水戸市, 横浜市の三つの都市に住むユーザのツイートのうち, 「保育園」に関するものを対象に, 人手によりその内容がポジティブであるのか, ネガティブであるのかを判定し, ポジティブな表現, ネガティブな表現を抽出した結果を報告した. それら一連の評価表現の抽出に際し, 評価者であるツイート投稿者が, どのような立場から評価しているのかが問題になった. たとえば, 保育園の職員であるのか, 保育園に子供を通園させている親であるのか, これから子供を通園させたいと思っている親であるのか, あるいは, それらには該当しないが, 保育園問題に関心を寄せる市民であるのか.

投稿者の属性は, アカウント名やユーザ名, プロフィール情報からある程度自動で抽出可能ではある. しかし, 筆者らは, それらが不明瞭な場合や不十分な場合もあるため, ツイート本文のみで投稿者の属性判断がどの程度可能であるかを人手により試行することとした. そこで, 収集したツイートを対象に, ツイート本文以外の情報を見ずに, 投稿者の属性の特定を可能にする表現を抽出し, それに基づき属性を推定し, 本文以外にある情報と推定結果を照合した. 本稿ではその試行作業について報告する.

また, 親であることの特定を可能にする表現に我が子の呼びかけ表現があるが, ツイート中には様々なバリエーションが見られる. それについてもくわしく報告する.

2 分析対象のツイート

柏野ほか(2020)で用いた「つくば市保育園データ」(つくば市民の Twitter アカウント 69 個による「保育園」に関する 2017 年 7 月 1 日～2018 年 3 月 31 日分のツイート 150 件)に加え, 新たに収集したツイートの一部 (横浜市民の Twitter アカウントによる「保育園」に関する 2010 年 2 月 1 日から 2020 年 4 月 30 日分のツイート, 6,202 件) のうち, さらに次の順で絞ったものを本稿の分析対象用ツイート「横浜市新保育園データ」として用意した.

1. 2019 年 4 月 1 日～2020 年 4 月 30 日までに投稿されたツイート, 1,743 件.
2. 分析対象外と判断したアカウントのツイートを除外. 計 1,215 件.

1 は, 人手で作業可能な範囲として最近の 1 年間に対象を絞ったものである. 1 年のうち, 4 月は入園, 3 月は卒園, 10 月～12 月は次年度 4 月入園申し込み次期, 1 月～2 月はその入園の合否が出る時期である. また, 2019 年 10 月より, 「幼保無償化」(幼児教育無償化) が開始された. そういった期間に投稿されたツイートが対象ということである.

2 は, 今回の分析対象を個人の属性表現抽出にするため, ユーザ名よりあきらかに個人でない判断できるアカウントはあらかじめ除外した. たとえば, 政治団体や商業施設等である. ただし, 非個人アカウントのうち, 幼稚園・保育園のアカウントは, そこで園長などの個人的な意見が書き込まれるケースがあるため残した. また, ユーザ名が個人商店名の場合は, 店主が個人的なツイートをすることもするため, 属性判断作業の段階で対象外か否かの判断を行うこととし, それらアカウントも残した. その結果, ここで削除したアカウントは 34 個, 残したアカウントは 94 個であった.

つまり, 属性判断作業の対象ツイートは次の表 1 のとおりである.

表 1 分析対象ツイート (件)

データ名	アカウント数	ツイート数
つくば市保育園データ	69	150
横浜市新保育園データ	94	1,215

3 属性表現の抽出と属性の推定

3.1 作業手順

「つくば保育園データ」と「横浜市新保育園データ」とを対象に、それぞれについて、人手により、次の手順にて、属性表現の抽出と、属性の推定、推定結果の照合を行った。

1. プロフィールがたどれる情報（ユーザ名など）を非表示にし、ツイートをシャッフルする。
2. 属性表現（＝属性を推定可能にしている語句）をツイート本文から抽出。1 ツイートから複数得られる場合は、複数抽出。（今回は最大で3つだった）
3. 2 から推定した属性を付与。※以下、「推定属性」と呼ぶ。推定する属性候補が複数ある場合は、第1候補、第2候補として付与した。（今回は最大で2候補に収めた）
4. 非表示にした情報（ユーザ名など）を再表示し、ツイートをユーザ名ごとの日付順にし、付与した属性と照合。

3.2 作業例

たとえば、図1では、「保育園の見学に行ってきた」や「保育園落ちた」を属性表現として抽出し、そこから、投稿者の属性を「保活中の親」（保活＝保育園入園のための活動）や「保育園に落ちた親」と推定した例である。

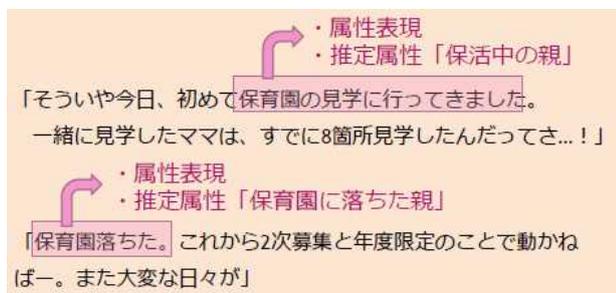


図 1 属性表現抽出と属性推定の例 2

3.3 属性表現の抽出例

たとえば、「保育園児の親」と推定させる語・表現の抽出例は次のとおりである。

①保育園（に）預ける

ツイート：保育園預けてちょっと電車に乗って帰ってきて保育園お迎えに行ってきただけでぐったり疲れちゃった。身体が鈍ってる。もりもり仕事したいけど。

②保育園（に／へ／まで）お迎えに行く

ツイート：早めに保育園へお迎えに行くと、まだ担任の保育士さんがいらっしやるので、園での様子を結構細かく聞けるというメリットはあるんですね。まあだいたいトラブルを聞く羽目になるんですが。

③〔時期・時間表現〕が／は保育園の〔イベント名〕（だ）

ツイート：今日は保育園のお楽しみ会 美桜 ver 今年で最後のお楽しみ会。相変わらず全開で楽しんでたなあ。ほんとよく笑う

「保活中の親」と推定させる語・表現の抽出例は次のとおりである。

①保育園（の）見学（に）行く

ツイート：雨すごいけど、職場近くの保育園見学行くんですけど。。。やまないかな？

②保育園（に）受かりました、保育園（に）入れました

ツイート：横浜市保育園第一希望受かったー！！！！みんなが第一に選びそうなほうをあえて第二にして、保育方針に差異はない姉妹園を第一希望に書いてよかった！どうせ場所もたいして変わらん、先生たちも毎年シャッフルされるから保育方針も変わらん！

ツイート：【報告】保育園入れました。今日、2次募集の結果が来ました。ご心配くださったみなさん、ありがとうございます。受かったのは小規模保育だしいろいろ思うことはありますが、ひとまずほっとしています

③保育園（を）保留（する／中）

ツイート：保育園保留中。年度限定対象の保育園を見学に行った。ほかに2組の親子が来てたーママと一緒に入れたらいいですねって言われた。ほんと。

④保育園（に）落ちた

ツイート：すいません、いいねくれたので気になりました。保育園どうでしたか？うちは落ちました…

3.4 属性候補が複数ある場合

属性候補が複数ある場合は、第1候補、第2候補として推定属性を付与した。たとえば、「今日からまた保育園がんばろう！」は、「保育園児の親」と「保育園関係者」の両方の可能性があり、両方を付与した。細かなことを言うと、「親」と考えられるものが「親」ではなく「祖父母」の可能性もあるのだが、今回の対象ツイートで祖父母とはっきり確定できたものではなく、そこは「親」としてのみ推定している。

先の図1で、「保育園落ちた」という属性表現に対し、「保育園に落ちた親」の推定属性を付与している例を示した。ツイート上には、多く、「#保育園落ちた」とハッシュタグ付きでも多用されるが、こちらは属性候補が2つ出てくる。

「保育園落ちた」は、2016年2月に匿名で投稿された「保育園落ちた日本死ね」というタイトルのブログが話題になったことをきっかけに、実際に入園できなかったことを直接指すだけでなく、保育園不足問題を表すフレーズとしても使われるのである。そのため、「#保育園落ちた」という表現は、「保活中の親」の可能性に加え、「待機児童問題に関心がある人(親以外)」の可能性もある属性表現ということになる。よって、図2のハッシュタグの2例は、2つの属性を候補として付与している。

「今日も保育園のことで区役所へ。身分証明を持っていけば、自分のランクを教えてくれるらしい #保育園落ちた #横浜市」

- ・属性表現
- ・推定属性「保育園に落ちた親」

「待機児童問題に関心がある人(※保育園に落ちた親以外)」

「#保育園落ちた #待機児ゼロ <https://t.co/Wq4fLtdpPC>」

- ・属性表現
- ・推定属性「保育園に落ちた親」
- 「待機児童問題に関心がある人(※保育園に落ちた親以外)」

図2 属性表現抽出と属性推定の例2



図3 図2の2例目ツイートの参照先

図2の1例目は、図には示していないが、「保育園のことで区役所へ」(推定属性1「保育園児の親」、推定属性2「保活中の親」とあわせると、「保育園に落ちた親」と判断可能な例である。「#待機児ゼロ」というハッシュタグの並びや、さらに、リンク先(図3)まで見ると、保育園不足問題を意図していることがわかり、「待機児童問題に関心がある人(親以外)」と判断できる例である。

3.5 作業結果

属性表現の抽出数、付与した推定属性数および、属性判断作業時に非表示にしていた情報(ユーザ名など)と照合した結果を、表2に示す。

表2 属性表現抽出と属性推定の照合結果

データ名	属性表現	推定属性	正解	不正解	保留	正解率
つくば市 保育園データ	872	1,068	901	129	38	84.4%
横浜市 新保育園データ	127	163	124	24	15	76.1%

8割前後と、高い正解率を得ることができた。

4 我が子の呼び方表現

親であることの特定を可能にする、我が子の呼びかけ表現を抽出した。これには、柏野ほか(2020)で用いた「つくば市保育園データ2」(2018年4月1日～2018年10月31日分のツイート109件)、「水戸市保育園データ」(2017年7月1日～2018年3月31日分のツイート139件)も対象ツイートに加えて行った。抽出結果は次のとおりである。

(1) 「子」「子供」を含むタイプ

うちの子、我が子、わが子
上の子、下の子、末っ子
子供たち

(2) 生まれた順(序列)がわかるタイプ

長男、長女、二番目

※(1)の「上の子」「下の子」「末っ子」も、ここにも含まれる。

(3) 「娘」「息子」を含むタイプ

娘、息子、うちの娘、うちの息子、娘ちゃん、息子くん、娘氏

(4) 「娘」「息子」のカタカナ表記や変形

ムスメ、ムスコ、むちゅめ、ムスッコ

※ほかの呼び方と比べて、投稿者＝その子供の親である可能性が高い印象。

(5) ニックネーム「+ちゃん」

例：ふうちゃん

(6) 本名「+ちゃん」「+君」

例：楓音，太良君

(7) 「姫」「王子」を含むタイプ

姫，王子

なお、「うちの姫」「王子様」といった例も，分析対象外のツイートを広く検索すれば見つかりそうではあるが，ざっと検索したところ，ペットの呼びかけ例が多い印象を得た。

また，今回の調査対象中にあった，あるアカウントの他ツイートも追跡調査したところ，次のような変遷例も得られた．あわせて報告する。

[2014年] 息子

[2015年] 息子氏、息子達

[2017年] 長男

[2018年] 末っ子、上の子ら、赤子

[2019年] 息子たち、息子 s、息子 1st、息子 2nd

「息子が増えていった結果，それぞれの「息子」に「s」「1st」「2nd」などを付けて区別している．ツイートならではのユニークな表現方法と言える。

5 おわりに

ツイートの投稿者の属性は，アカウント名やユーザ名，プロフィール情報に，「1児の母」「2児の父親」「保育士」「市会議員」などと示されていることがあり，「保育園」のツイートがどのような立場でされたものか判断できる場合はある．しかしながら，それらの情報だけでは不明瞭であったり不十分であったりする場合もあるため，ツイート本文のみで投稿者の属性判断がどの程度可能であるかを人手により試行した．その結果，8割近く判断可能であることを示した．

しかしながら，人手作業には時間と量の限界があるため，今後は，これらの人手による試行結果を，どのように自動抽出に結び付けていけるかを検討する必要がある．今後の課題である．

謝辞

本研究は，JSPS 科研費 19H04420 の助成を得ています。

参考文献

- 安藤有生，関洋平(2018)「市民のツイートをを用いた分散表現に基づく地名に対する市民の関心の傾向の可視化」『知能と情報』vol.30, no/6, pp.804-814.DOI: 10.3156/jsoft.30.6_804
- 石田哲也，関洋平，柏野和佳子，神門典子(2021 予定)「複数の属性の関連性に着目したソーシャルメディアからの市民意見抽出」『第 13 回データ工学と情報マネジメントに関するフォーラム (DEIM 2021)』.
- 柏野和佳子，立花幸子，平本智弥，関洋平(2017)「市民意見の収集システムで得られたツイートからの「保育園」「教育」に関する意見抽出」『言語処理学会第 23 回年次大会発表論文集』pp.533-536.
- 柏野和佳子，平本智弥，関洋平(2018)「市民意見の収集システムで得られたツイートからの道路・交通に関する意見抽出」『ことば工学研究会資料』vol. 57, pp.13-21.
- 柏野和佳子，平本智弥，関洋平(2019)「市民意見収集のためのツイート表現の分析」『電子情報通信学会技術研究報告』vol.118, no.516, pp.95-98.
- 柏野和佳子，平本智弥，関洋平(2020)「市民意見収集のための「保育園」に関するツイートからの評価表現の抽出」『言語処理学会第 26 回年次大会発表論文集』pp.1313-1315.
- 関洋平(2017)「自治体における市民意見を利用した新たな気づき」『スマートシティ「都留市モデル」シンポジウム』
- 南澤亜樹，関洋平(2018)「市民のツイートを行政課題ごとに分類するための関連語の抽出」『第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM 2018)』J6-2