ツイートの発信形式の使用割合に基づくユーザ探索支援

User Search Support Based on The Usage Rate of Tweets' Transmission Format

坂田 駿允 [1*]　砂山 渡 [2]
Toshinobu Sakata[1]　Wataru Sunayama[2]

[1] 滋賀県立大学大学院工学研究科
[1] Graduate School of Engineering, The University of Shiga Prefecture
[2] 滋賀県立大学工学部
[2] School of Engineering, The University of Shiga Prefecture

Abstract: In recent years, the widespread use of social networking services (SNS) has made it possible to communicate with a variety of people through networks. Especially in Twitter, which is used by many users as a broad and shallow communication tool, it is meaningful to support the search for users who match the interests of users.
In this study, we aim to extract features from the sentences posted by users, the usage rate of the tweet's transmission format, classify users based on the features using a classification tree, and estimate the user's personality and characteristics from the classification destination. Through experiments, we verified whether the labels given to the user sets classified in the classification tree well represent the personality and characteristics of the user sets.

1　はじめに

近年，Twitter をはじめとした，ソーシャル・ネットワーキング・サービス (SNS) が広く普及している．SNS の普及によって，近しい人間にとどまらず，遠く離れた人間や，顔も知らない人間，さらには海外の人間とも気軽にネットワークを介して，コミュニケーションが取れるようになった．特に，Twitter の利用特性について，“広く浅いコミュニケーションツールとして利用”，“特定ユーザのコミュニケーションツールとして利用”，“コミュニケーションツールとして利用せず” の 3 パターンに分類し，”広く浅いコミュニケーションツールとして利用”に 40%以上を占めていることを報告されている [1]．このことから，ユーザがより広く浅いコミュニケーションを行えるユーザの探索を支援することことは有意義である．

そこで，SNS を利用するユーザが過去に発信したツイートデータから，相手の性格や特徴を読み取ることができれば，相手とコミュニケーションを取る前に，対象ユーザがどんな人物なのかを把握し，利用者の興味に合うユーザの探索が可能になると考えられる．

そこで本研究では，ユーザの投稿した文章である，ツイートの発信形式の使用割合から特徴を抽出し，それらの組み合わせから，ユーザの性格や特徴といった，属性を特定することを目的とする．

*連絡先：滋賀県立大学大学院工学研究科電子システム工学専攻
〒 522-8533 滋賀県彦根市八坂町 2500
E-mail:oi23tsakata@ec.usp.ac.jp

2　関連研究

2.1　SNS ユーザの特徴抽出に関する研究

ツイートから，心理学において，人間が持つ様々な性格は，5 つの要素の組み合わせで構成されると考えられている，5 因子モデルに基づいた性格分析を行う研究がある [2]．この研究では，ツイートの発言を取得し，ニューラル言語モデルによる学習を用いて 5 因子モデルに基づいた性格分析を行っている．

ツイート集合を入力として，深層学習を活用し，特定の趣味に興味を持つユーザの投稿文の特徴を学習し，利用者が設定した趣味を文章から自動抽出する研究がある [3]．この研究では，人間の行動を 5 つの自我の状態に分類することで，個人の人格を表現するエコグラムを用いて，文章に現れる特徴から性格要素を推定している．

上記のようにツイートの文章や単語を，機械学習を用いて推定する研究が行われている．しかし，機械学習を用いた学習は，文章を入力データとして処理する

際に，形態素解析や正規表現を用いたストップワードの除去などにより，ツイートに現れる，記号やツイートの長さなどの情報が取得できない．本研究では，このような情報にもユーザの特徴や性格が現れると考え，文字数や感嘆符などの発話形式から得られる特徴によってユーザを推定する．

2.2 SNS ユーザの探索支援に関する研究

SNS の投稿内容を用いて，面識のない 2 者を引き合わせるための仲介者を探索する手法を提案した研究がある [5]．この研究では，SNS の投稿内容を解析し，SNS 登録者との関連の強さを表す関連度を付与した関係データを用いて，仲介者を探索している．上記研究では，友人関係，投稿内容，コメント応答の 3 つの関連データを用いて探索する手法を用いているが，この計算を全ユーザに対して行う必要があり，大量のユーザが存在する SNS においては探索が困難である．そこで，本研究では，多数のユーザに対応できるように，分類木を用いて多数のユーザが有する性格や特徴の集合に分類し，利用者の好みに合った集合から，好みのユーザを探索する手法を取る．

3 ツイートの発信形式の使用割合によるユーザの特徴抽出

ツイートデータを用いた性格推定には，有効とされているナイーブベース法を用いたもの [4] や，機械学習を用いたもの [2][3]，など多くの手法が行われている．これらの手法では，文章を形態素解析し，その際に不要とされる記号は取り除かれ，性格推定に考慮されない．しかし，Twitter において句読点を用いた年代推定の有効性 [8] や，画像や URL，他人のツイートを引用したツイートなども性格推定に有用である可能性が示されている [6]．

そこで本研究では，形態素解析で除去されることが多い，句点や感嘆符などの発話形式や，ツイートに現れる文字数や漢字の使用量などの特徴を用いて，ユーザの特徴の抽出を試みる．

3.1 ツイートの発信形式から得られる特徴

SNS の一つである Twitter で用いられるツイートは，ユーザによって様々である．ユーザによって変化する様々な発信形式から得られる特徴を組み合わせによって，ユーザの性格や特徴といった，属性が判定できるのではないかと考えた．

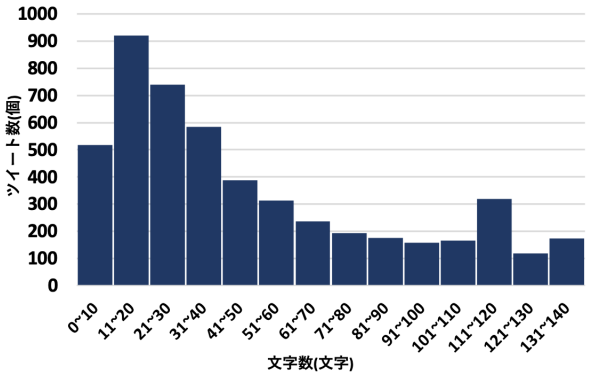

図 1: 無作為に集めた 5000 ツイートの文字数の分布

本研究では，特に多くのユーザのツイートで発現される特徴を抽出し，それらの組み合わせでユーザの様々な属性を判定する．また，本研究では，ユーザの有する性格や特徴を，属性と定義する．

3.2 本研究で用いるツイートの発信形式

本研究で，ツイートによる発話形式とは，ツイートの内容以外の情報，記号，文字数や使用される文字の種類など，ツイートの内容以外から得られる特徴のことと定義する．従来の性格推定では，文章を正規表現を用いた特徴の含まれない単語の処理，形態素解析を用いた文章の分割などにより文章の前処理が行われる．この前処理において，記号は単語の処理で排除され，文字数や文の長さなどの情報は形態素解析で失われる．しかし Twitter という口語体で投稿が行われることが多い SNS では，ユーザの属性推定に有意義なのではないかと考えた．本節では，本研究で用いるツイートの発信形式と，表れる特徴について述べる．

3.2.1 文字数

図 1 に無作為に集めた 5000 ツイートの文字数の分布を示す．0 文字から 40 文字でツイートするユーザが多く，以降の文字数では徐々に減っていくことがわかる．次に，文字数の違いによってどのような特徴が表れるかを，文字数分布の上位 1/3 である 30 文字，上位 2/3 である 60 文字を閾値として確認した．30 文字以下のツイートでは，自身のその時の感情をツイートしていることが多い．対して，60 文字以上の時は，自身の出来事をまとめてツイートしていて，30 文字以下のツイートと比べると，数ツイートに分割して共有することも可能な程，情報量が多いことがわかった．本研究では，60 文字以上のツイートを，文字数の特徴を有する，すなわち情報量の多いツイートとして定めた．

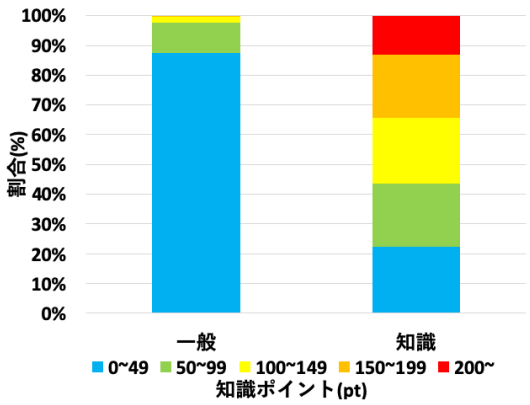

図 2: 一般ユーザと知識ユーザの知識ポイントの分布

3.2.2 漢字の難度と使用率

豊富な知識を有するユーザは，一般のユーザがツイートする際には使わないような難度の高い熟語や漢字を用いることが多い．そこで，政治家や評論家ユーザと，無作為に選択したユーザのツイートの漢字の使用量と難度について調べた．政治家や評論家ユーザなどの，知識を有するユーザを知識ユーザ，無作為に抽出したユーザを一般ユーザと定義する．本研究では，漢字の難度と使用量を日本漢字能力検定（漢検）の出題範囲である漢字を用いてポイント形式で判別を行った．

1 ツイートに含まれる漢字について，漢検 10 級から準 2 級まで，各 1pt から 10pt まで付与した．このポイントの合計をそのツイートの知識ポイントと定義する．図 2 に知識ユーザと，一般ユーザについて，各 20 名，各 100 ツイートの 2000 ツイートを用いて知識ポイントの分布を示す．一般ユーザのツイートの 8 割以上が知識ポイント 50pt 以下に対して，知識ユーザのツイートの 5 割以上が 100pt 以上の知識ポイントを有していることがわかる．これにより，知識ポイントによって，そのユーザが知識を有したツイートをしているかどうかを判別することが可能であると示された．知識ポイントが 100pt 以上のツイートを，知識の特徴を有する，すなわち知識が豊富と捉えられるツイートとして定めた．

3.2.3 感嘆符

感嘆符は，感動，驚き，強調を表す時によく用いられる．また，直接読者に話しかける、あるいは訴えかけることを目的とする，読者との共感を求める表現を用いる時にも用いられる [7]．つまり，自分の感情をより相手に伝えたい時に使われる．そこで，感嘆符が 1 つ以上含まれるツイートを，感嘆符の特徴を有する，すなわち自身の感情をよく表すツイートとして定めた．

3.2.4 句点

ツイートは口語体で投稿されることが多く，句読点を使われることが少ない．その中で，句読点を使うユーザは，年代が高いという研究結果がある [8]．よって，句読点を使うユーザは年代が高いといった特徴を有することがわかった．しかし，読点は句点よりも使われる頻度が高く，年代における読点の出現頻度の差異も低い [8] ことから，本研究で抽出したい特徴を表しにくい要素だと考え，句点のみを用いることにした．そこで，句点が 1 つ以上含まれるツイートを，句点の特徴を有する，すなわち年代の高いツイートとして定めた．

3.2.5 URL

本研究で用いたツイートの収集方法において，URL は Web ページだけでなく，画像や動画，相手のツイートを引用して表示する，引用リツイート（以下，引用 RT）としても取得される．画像や動画は文だけでは伝えられない内容を他ユーザに共有することができ，引用 RT は他ユーザのツイートを自身の文章に添えて他ユーザに共有することができる．文章だけのツイートと比べると，相手により多くの情報を共有したい時に使われる．そこで，URL が 1 つ以上含まれるツイートを，URL の特徴を有する，すなわちより相手に情報共有をしたがるツイートとして定めた．

3.2.6 ハッシュタグ

ハッシュタグを付けたツイートは，そのハッシュタグを検索したユーザに表示される．つまり，自身のフォロワー以外にハッシュタグを検索したユーザにもツイートを共有することができる．自身のツイートにハッシュタグを用いることは，より多くのユーザに自身のツイートを見てもらいたい，自己顕示の特徴を有している．そこで，ハッシュタグを 1 つ以上含むツイートを，ハッシュタグの特徴を有する，すなわち他ユーザに見てもらいたいツイートとして定めた．

4 ツイートの発信形式の使用割合に基づくユーザ探索支援システム

本研究で提案するユーザ支援システムは，利用者がコミュニケーションを行いたいユーザを，発話形式による特徴から推定した属性により，ユーザの探索支援

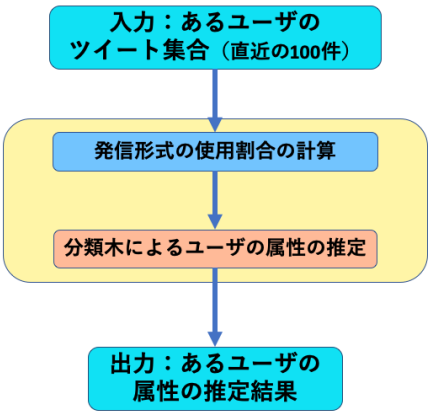

図 3: 提案するユーザ探索支援システムの概要図

を行うシステムである．本システムで様々な属性に分類されたユーザ集合から，利用者の好みの属性を選び，ユーザを選択することができる．また，気になったユーザを本システムを用いてどのような属性を有するのか推定することも可能である．図 3 に提案するユーザ探索支援システムの概要図を示す．本研究で想定する利用方法として，入力として，あるユーザのツイート集合の直近 100 件のツイートを用いる．この 100 件のツイートのうち，前章で挙げた発話形式の使用割合を計算し，あるユーザがどのような特徴をどの程度有するかを数値で表す．この数値を用いて，分類木によりユーザを属性に分類し，その結果をあるユーザの属性推定結果とする．

4.1 ユーザ探索支援システムの構成

4.1.1 システムに用いるデータ集合

本研究のシステムでは，1 ユーザから 100 ツイート取得し，100 ツイート中，前章で定義した特徴を有するツイート数を調べ，その数をユーザの持つ特徴量とした．また，ツイートに URL が含まれていた場合，URL を除いた文章から文字数をカウントした．この処理を施したユーザを 785 人集め，これを入力データとした．各発話形式の特徴量について，最大の特徴となる発話形式を正解ラベルとした．正解データを作成する際，偏差値を用いることで，他ユーザと比較して特徴量が最大となる特徴に正解ラベルを付与することができる．

4.1.2 Twitter ユーザの分類木

本研究では，分類木の分類結果を用いて，分類されるまでルートノードから，どのような属性を有するユーザか推定する．前項で示した入力データと正解ラベルを用いて，決定木を学習した．学習には，Python のオープンソース機械学習ライブラリである scikit-learn を用いた．決定木を学習させる際，分類される葉ノードの数が多くなると，中間ノードによる分岐が多くなり，ユーザの個性的な属性が大量に現れてしまう．これを避けるため，決定木の最大ノード数を 6 に設定した．決定木を学習した結果，39 の葉ノードが存在した．このうち，32 ノードは，20 ユーザ以下が分類されるノードであった．このノードのうち，実際に分類されたユーザのツイートを確認し，分類されたユーザ集合が共通の特徴を持っていると考えた，20 ユーザ以上所属するノードを抽出し，そのノード以降に 20 ユーザ以下が分類された子ノードは親ノードにまとめることで，学習した決定木を再構築した分類木を作成した．再構築した分類木を図 4 を示した．図 4 の再構築した分類木のルートノードに示される条件式によって，ユーザの特徴量でユーザが分類される，通ったルートノードの特徴の組み合わせにより，各葉ノードに分類されたユーザの属性を考察し，表 1 に示した，9 つのユーザ属性に分類した．

表 1: 分類したユーザ属性

大人型	年代の高さや真面目さを表す
自己顕示型	自身の情報や感情をより多くの人に伝える
内輪共有型	自身の情報や感情は伝えたいが ハッシュタグを用いて共有はしない
趣味特化型	画像などを用いて 自身の趣味を高頻度で共有する
丁寧・几帳面型	自身の出来事をまとめてツイートする
感情共有型	自身の感情を即座にツイートする
生活感共有型	自身の生活の中の出来事をまとめてツイート
趣味共有型	自身の趣味に関するツイートを低頻度で行う
内気型	自身のツイートを淡々とツイート

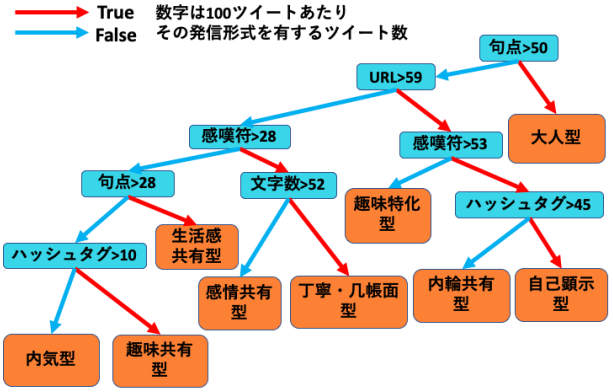

図 4: 再構築した分類木

表 2: 属性に属するユーザの第 1 候補の評価結果

		システムによる属性の推定結果								
		大人型	自己顕示型	内輪共有型	趣味特化型	丁寧・几帳面型	感情共有型	生活感共有型	趣味共有型	内気型
被験者が回答したユーザ属性	大人型	1	0	0	0	0	0	1	0	0
	自己顕示型	1	4	5	4	2	2	0	0	4
	内輪共有型	0	0	0	0	1	0	1	1	0
	趣味特化型	1	1	0	0	3	3	0	3	2
	丁寧・几帳面型	1	0	2	0	0	0	0	0	0
	感情共有型	0	2	2	3	1	5	3	4	2
	生活感共有型	3	1	0	3	1	0	3	0	0
	趣味共有型	3	3	2	1	3	1	2	3	3
	内気型	1	0	0	0	0	0	1	0	0

表 3: 属性に属するユーザのすべての候補の評価結果合計

		システムによる属性の推定結果								
		大人型	自己顕示型	内輪共有型	趣味特化型	丁寧・几帳面型	感情共有型	生活感共有型	趣味共有型	内気型
被験者が回答したユーザ属性	大人型	5	2	2	1	2	1	3	0	1
	自己顕示型	3	4	7	9	5	8	3	3	7
	内輪共有型	2	3	2	1	1	0	4	2	2
	趣味特化型	2	2	0	3	5	4	0	7	3
	丁寧・几帳面型	4	4	3	3	2	0	0	1	0
	感情共有型	1	6	5	5	6	9	7	8	6
	生活感共有型	8	3	4	4	2	2	8	2	5
	趣味共有型	5	4	5	5	6	8	5	9	4
	内気型	1	1	0	0	1	0	2	0	1
	正答率	0.45	0.36	0.18	0.27	0.18	0.82	0.73	0.82	0.09

5 ユーザ探索支援システムの評価実験

5.1 実験方法

本研究で得た，分類木によって 9 種類の各属性に分類されたユーザ 9 名のツイート，各 100 ツイートを被験者に読んでもらい，各ユーザがどの属性に分類されるかを第 3 候補まで (第 2 候補以降は自由回答) 評価する実験を行った．分類の対象となるユーザは，被験者ごとに各属性に分類されたユーザの中で異なるユーザを用いて，属性に分類された複数のユーザからその属性を評価できるようにした．評価を終えた後に，設定した属性は直感的にわかりやすかったか，予め用意したラベルづけを正しい属性ごとに分けられたと思うか，アンケート形式で回答してもらった．本実験は，理系学生 11 名に対して行った．

5.2 実験結果と考察

表 2 に，各属性に属するユーザの，第 1 候補の評価結果，表 3 に，各属性に属するユーザの，全ての候補の評価を足した結果を示す．表 2 を見ると，感情共有型，生活感共有型，趣味共有型，自己顕示型において，高い評価結果を得た．それぞれの結果に注目する．

感情共有型は，表 2 から，第 1 候補における正答率が高い事がわかる．これは，文字数が少ない時の特徴である，今の感情を短文でツイートするという推測が正しかったと言える．生活感共有型は，表 2 を見ると，第 1 候補において，感情共有型と同じ回答数になっていることがわかる．この結果から，生活感共有型の属性に分類されたユーザは，感情共有の特徴も有すると考えられる．この結果から考えられる事として，ツイートの中に起きた事や考えに対する感情をツイートに含めていたために，感情共有の特徴が得られたと言える．趣味共有型は，表 2 を見ると，第 1 候補において，感情共有型に 4 人，趣味特化型に 3 人，趣味共有型に 3 人回答している．この結果から，趣味共有型の，自身の趣味をハッシュタグを用いて低頻度で共有するユーザと，趣味特化型の，自身の趣味を画像を用いて高頻度で共有するユーザとの違いは，この属性からは判断できない事が言える．自己顕示型は，表 2 を見ると，第 1 候補において，一番高い回答数を得た．次に多かった回答数は趣味共有型であった．この結果から考えられ

る事として，趣味は，自身の楽しみとして愛好する事なので，自身の楽しみを共有するということは，自己顕示に似た特徴を得られると言える．よって，自己顕示型は，趣味共有型の一部として存在すると言える．

次に，表 2 から，低い評価結果の，内気型，内輪共有型，丁寧・几帳面型，大人型，趣味特化型に注目する．内気型は，表 3 を見ると，1 つしか回答を得られなかった．さらに，内気とは逆の特徴と考えられる，自己顕示型の回答が 7 つと多くの回答を得た．これは，ハッシュタグの特徴を有さないユーザに，相手に自身のツイートを共有したがらない特徴はないからだと考えられる．内輪共有型は，表 2 を見ると，第 1 候補において，自己顕示型との回答を一番多く得た．これは，内輪共有型とは逆の分類先の属性で，ハッシュタグの特徴によって分類されている．内気型と同じく，ハッシュタグの特徴が小さい場合に，相手に共有したがらないユーザではないためと考えられる．丁寧・几帳面型は，表 2 を見ると，感情共有型，趣味共有型，自己顕示型の回答を多く得た，これは，内輪共有型と同じく，逆の分類先の属性が感情共有型で，文字数の特徴による，自分の出来事や考えをまとめてツイートする，という特徴が適していなかったと考えられる．大人型は，表 2 を見ると，1 人しか回答を得られず，生活感共有型で 3 人，趣味共有型で 3 人という回答数を得た．これは，大人型が句点のみの分類なので，句点による，真面目や年代が高い特徴よりも，生活感共有型や，趣味共有型の特徴の方が発現しているためと考える．趣味特化型は，表 2，表 3 のどちらにおいても，自己顕示型の回答が 1 番高い結果となった．これは，自己顕示型での考察と同じく，趣味に特化したツイートは，自己顕示に似た特徴を得る事ができ，趣味共有型よりも強い趣味の共有ツイートなので，強い自己顕示の特徴を得たと考える．

低い評価結果の考察として，発話形式の特徴が適切でなかった，他の特徴より弱い特徴が強い特徴に消されてしまった，他のユーザ属性と似た特徴を持っていたなどが挙げられる．これらの改善案として，その発話形式の再調査や取捨選択を行い，各発話形式に独立した特徴を付与する必要がある．また，利用者がよく理解し，差別化しやすいユーザ属性を定義し，判別されたユーザの違いを明確にする必要がある．

6 おわりに

Twitter のユーザ探索支援を行うために，ツイートの発信形式の使用割合から特徴を抽出し，分類木からユーザを探索するシステムを作成した．

作成した分類木に分類されたユーザ集合につけたラベルが，そのユーザ集合をよく表すラベルか検証する評価実験を行った．設定した 9 つのラベルのうち，4 つのラベルは，抽出した特徴とその組み合わせから得られるラベルとして適切であったが，5 つのラベルについては，特徴を抽出できなかったり，他の特徴に埋もれてしまった特徴があった．今後の目標として，結果の得られなかった特徴の再検討，新たな特徴の探索をし，ラベルの検討に幅を持たせるとともに，結果の得られた特徴やラベルの類似点を見直し，より正確なユーザ探索を目標としていきたい．

参考文献

[1] 鳥海不二夫，神谷達幸，石井健一郎：Twitter におけるつぶやきを用いたユーザ特性分析，第 7 回ネットワーク生態学シンポジウム，pp.1234–1237，(2011)．

[2] 塚野駿，柴田千尋，政倉祐子，田胡和哉：ニューラルネット言語モデルによる Twitter 上の発言からの 5 因子モデルに基づく性格分析，情報処理学会第 78 回全国大会，pp3–4，(2016)．

[3] 若宮悠希，砂山渡，畑中裕司，小郷原一智 : 深層学習を用いた Twitter からの趣味情報の抽出，人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会 (第 24 回)SIG-AM-24-03，pp12–19，(2020)．

[4] 杉山 美智子，小田浩一：ナイーブベイズ法を用いた Twitter による性格推定，言語処理学会第 20 回年次大会発表，pp1123-1125，(2014)．

[5] 小松恭子，中澤昌美，池田和史，服部元，滝嶋康弘 ：円滑な人脈形成のための SNS 投稿内容に基づく仲介者探索手法，情報処理学会第 76 回全国大会講演論文集，pp455–456，(2014)．

[6] 山田康輔，笹野遼平，武田浩一：「いいね」「シェア」 をした投稿のテキスト情報を利用した SNS ユーザの性格推定，人工知能学会論文誌，Vol.35，No.4，pp1–12(2020)．

[7] 村田年, LossaRoma：異なる文章ジャンルの判別可能性に関する調査 / ブログ本文、新聞社説、文学作品、論文を対象として，日本語と日本語教育，No.42，pp125–135，(2014)．

[8] 江口大賀，菊池浩明：ツイートの文章に使われている句読点に基づく属性推定，情報処理学会第 82 回全国大会，pp431–432，(2020)．