深層学習を用いた客観指標との比較による個人特有の嗜好の抽出

Extraction of Individual-Specific Preferences by Comparing with Objective Indicators using Deep Learning

渡邉 有輝 [1*]　砂山 渡 [2]　畑中 裕司 [2]　小郷原 一智 [2]
Yuki Watanabe[1]　Wataru Sunayama[2]　Yuji Hatanaka[2]　Kazunori Ogohara[2]

[1] 滋賀県立大学工学部電子システム工学科
[1] Department of Electronic Systems 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, studies to capture the trends of the world and the personal characteristics of SNS users have been promoted from SNS data. However, simply trying to capture the characteristics of individuals is mixed with the general characteristics of people around the world, and it is difficult to extract individual characteristic features.Therefore, in this research, we propose a method to extract personal preferences from the difference with this index, using the taste of many people in the world as an objective index.In other words, we construct a network that classifies comment sets in SNS into items within the field by deep learning, targeting one field.Then, personal preference is evaluated based on which item the individual comment is classified into many items.

1 はじめに

近年，SNS（Social Networking Service）のデータから世の中の傾向を捉える研究 [1][2][3] やユーザの個人的特徴を捉える研究 [4][5][6] が進められるようになってきた．しかし，単純に個人の特徴を捉えようとすると，世の中の人々がもつ一般的な特徴と混ざってしまい，個人特有の特徴を抽出することができない．そこで本研究では，世の中の多くの人の嗜好を客観指標として，この指標との差分として個人の嗜好を抽出する方法を提案する．すなわち，ある嗜好を表す分野を対象として，SNS におけるコメント集合をその分野内の項目（構成要素）に分類するネットワークを深層学習により構築した上で，個人のコメントがどの項目に多く分類されるかをもとに，個人の嗜好を評価する．

2 深層学習による個人特有の嗜好抽出システム

この章では，構築したシステムについて，システムの構成と個人特有の嗜好抽出までの流れについて述べる．

*連絡先： 滋賀県立大学工学部電子システム工学科
〒 522-8533 滋賀県彦根市八坂町 2500
E-mail: oi23ywatanabe@ec.usp.ac.jp

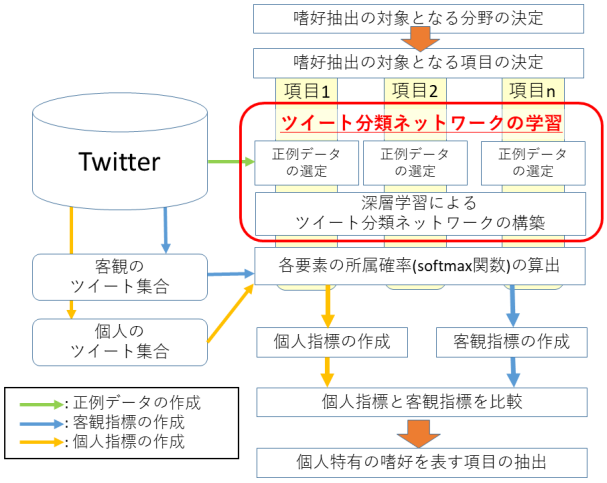

図 1: 提案する個人特有の嗜好抽出システムの構成

2.1 システムの構成

本研究で提案するシステムの構成を図 1 に示す．

分野内の項目（構成要素）の正例データを深層学習の 1 つである DNN（Deep Neural Network）を用いて学習し，分類ネットワークの重み付けを行う．そして，構築した分類ネットワークに，全体のテキスト集合を入力データとした場合の分類結果から，一般の嗜好度

を表す客観指標の作成を行う．また，個人のテキスト集合を入力データとした場合の分類結果から，個人の嗜好度を表す個人指標の作成を行う．最後に，客観指標と個人指標の比較を行うことで，個人特有の嗜好抽出を行う．

2.2 抽出項目となる分野，項目の決定

本研究では，特定の分野における嗜好の抽出を行っている．そのため，嗜好抽出の対象となる分野と項目をあらかじめ決める必要がある．今回は，料理に関する嗜好抽出システム，アイドルグループ「乃木坂 46」のメンバーに関する嗜好抽出システムの構築を行った．

2.3 各項目の正例データの選定

本研究では，Twitter で投稿されたメッセージであるツイートを用いる．スマートフォンの利用と旅行消費に関する調査 (2017) では，各 SNS での使用目的に関する調査をおこなっており，Twitter は「本音で語る」，「ちょっとした思い付き」，「趣味の活動」などが高く，等身大の自分自身を表現する目的の利用者が多いという結果が出ている．また，日本での利用者が 4500 万人を超えているという点からツイートを入力データとして用いている．

深層学習に用いる正例データを，以下の条件で収集した．料理の正例データとして各料理 1000 件，合計 3.5 万件を収集した．また，乃木坂 46 の正例データとして各メンバー 500 件，合計 2 万件を収集した．今回は，他の項目の正例データを負例データとして取り扱うため，負例データの収集は行わない．

- 項目名を含むツイート
- 他の項目名を含まないツイート
- 重複したツイートは 1 ツイートのみ使用
- 料理の場合は,「おいしい」「うまい」を含むツイートを条件に追加

2.4 文章分類ネットワークの構築

本研究では，文章分類を深層学習を用いて行う．深層学習とは，中間層を 2 層以上に多様化したニューラルネットワークである．深層学習による文章分類の構成を図 2 に示す．学習ネットワークの構造は，基本的に入力層と 2 つ以上の中間層，出力層で構成されており，各層は情報を格納するノードで構成されている．また，ノード同士は情報を伝える線であるエッジで連結しており，情報伝達時にはエッジに伝わる情報を強化したり，減衰させたりする値である重みが与えられる．

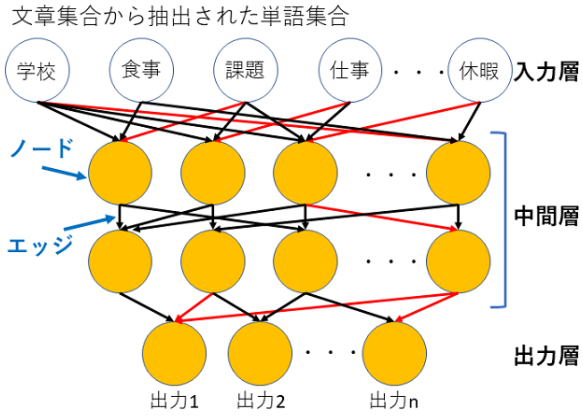

図 2: 深層学習による文章分類の構成

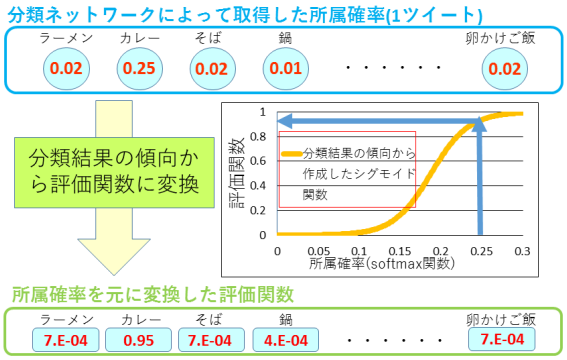

図 3: 分類結果を用いた指標の作成

本研究では，深層学習のライブラリのひとつである Deep Learning 4j(DL4j)[7] を利用して深層学習の実装を行う．

深層学習の入力データや抽出対象の文章は，データを数値化する必要がある．そのため，文章を形態素解析し，BoW(Bag of Words) への変換を行う．BoW とは，各文章での単語の出現頻度をベクトルで表現したものである．また，形態素解析には Igo[8] を用い，形態素解析用の辞書として MeCab[9] で使用される ipadic を使用した．

2.5 分類結果による指標の作成

分類結果を用いた指標の作成の流れを図 3 に示す．

本研究では，一般的な嗜好度を表す客観指標と，個人の嗜好度を表す個人指標の作成を行った．客観指標には，料理の客観指標には,「おいしい」「美味しい」「うまい」「美味い」のいずれかを含むツイートを 150 万件,

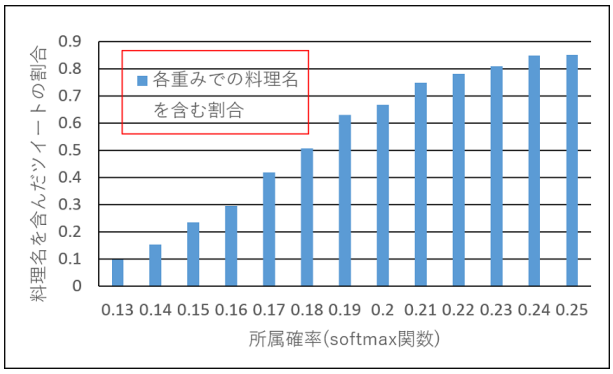

図 4: 料理名を含んだツイートの割合

表 1: ツイートの文章が料理名のみだった場合の softmax 関数値

料理名	所属確率	料理名	所属確率	料理名	所属確率
ラーメン	0.2067	ピザ	0.2551	ハンバーガー	0.2556
カレー	0.2397	おでん	0.2351	肉まん	0.2542
そば	0.1724	ハンバーグ	0.2021	牛丼	0.2121
鍋	0.2212	みそ汁	0.2110	コロッケ	0.2214
寿司	0.2195	刺身	0.2173	とんかつ	0.1949
うどん	0.2164	たこ焼き	0.2501	サンドイッチ	0.2581
からあげ	0.1926	ステーキ	0.2160	豚汁	0.2115
焼肉	0.2070	チャーハン	0.2400	卵焼き	0.1759
パスタ	0.1945	オムライス	0.2238	肉じゃが	0.2380
おにぎり	0.1864	天ぷら	0.2094	親子丼	0.1855
ポテト	0.2427	お好み焼き	0.2175	卵かけご飯	0.2754
餃子	0.1981	シチュー	0.2455	所属確率の平均値	0.2201

乃木坂 46 の客観指標には，各メンバー名を含むツイート 90 万件の分類結果を用いて指標の作成を行う．また，個人指標には，個人のツイートの分類結果を用いて指標の作成を行う．

文章分類時の各出力層での所属確率を表す softmax 関数の値と，ツイート内容が各項目名のみだった場合の softmax 関数の値から指標の作成を行う．客観指標に用いたツイートの分類結果を softmax 関数の値ごとに，各項目名につき 100 件確認し，ツイート内に各項目名が含まれている割合を調べた．料理名を含んだツイートの割合を図 4 に示す．また，ツイートが各料理名のみだった場合，乃木坂 46 の各メンバーのみだった場合の softmax 関数の値を表 1，表 2 に示す．

以上の結果から，料理に関する分類ネットワークでは，softmax 関数値が 0.2201 の時に，乃木坂 46 に関する分類ネットワークでは，softmax 関数が 0.1433 の時に評価関数が 0.8 となるようなシグモイド関数を作成する．

softmax 関数の値から評価関数への変換をツイートごとに行い，評価関数の合計値をそれぞれ客観指標，個人指標として用いる．作成した客観指標の上位 10 項目を表 3，表 4 に示す．

表 2: ツイートの文章が乃木坂 46 のメンバー名のみだった場合の softmax 関数値

メンバー名	所属確率	メンバー名	所属確率	メンバー名	所属確率
秋元真夏	0.1227	斉藤優里	0.1273	樋口日奈	0.1556
生田絵梨花	0.1350	阪口珠美	0.1528	星野みなみ	0.1186
伊藤かりん	0.1104	桜井玲香	0.1380	堀未央奈	0.1297
伊藤純奈	0.1218	佐々木琴子	0.1589	松村沙友理	0.1479
伊藤理々杏	0.1396	佐藤楓	0.1360	向井葉月	0.1385
井上小百合	0.1406	白石麻衣	0.1533	山崎怜奈	0.1777
岩本蓮加	0.1500	新内眞衣	0.1498	山下美月	0.1420
梅澤美波	0.1415	鈴木絢音	0.1494	吉田綾乃クリスティー	0.1634
衛藤美彩	0.1530	高山一実	0.1625	与田祐希	0.1288
大園桃子	0.1486	寺田蘭世	0.1395	若月佑美	0.1467
川後陽菜	0.1407	中田花奈	0.1653	渡辺みり愛	0.1447
北野日奈子	0.1418	中村麗乃	0.1400	和田まあや	0.1299
久保史緒里	0.1584	西野七瀬	0.1391	所属確率の平均値	0.1433
齋藤飛鳥	0.1448	能條愛未	0.1474		

表 3: 料理に関する客観指標

順位	料理名	客観指標	順位	料理名	客観指標
1	ラーメン	29568.07	6	寿司	10332.29
2	カレー	19388.25	7	からあげ	8646.55
3	鍋	14959.47	8	パスタ	8585.22
4	そば	12834.28	9	焼肉	7340.67
5	うどん	11158.08	10	ポテト	7293.95

2.6 個人特有の嗜好抽出

前節の方法を用いて客観指標と同様に個人指標の作成を行い，客観指標と比較することで個人特有の嗜好抽出を行う．

今回は，個人のツイート 100 件を用いて個人指標の作成を行った．また，客観指標と個人指標をそれぞれ割合に置き換え，客観指標と個人指標での各項目の割合差から嗜好抽出を行う．本研究では，割合差の偏差値が 55 以上の項目を個人特有の嗜好として抽出している．料理と乃木坂 46 に関する嗜好抽出の一例を表 5，表 6 に示す．表 5 から，この利用者の個人特有の嗜好として「ハンバーグ」「たこ焼き」「ピザ」が抽出された．また，表 6 から「阪口珠美」「渡辺みり愛」「能條愛未」が抽出された．

個人特有の嗜好を抽出することにより，自分のツイートを入れた際は，主観視した場合と客観視した場合の結果を比較することが可能となり，主観だけでは分からなかった新たな嗜好の発見につながることが期待できる．また，自分以外の利用者のツイートを入れた場合は，共通の嗜好を持つ人を探すツールとして利用できるのではないかと考えている．

表 4: 乃木坂 46 のメンバーに関する客観指標

順位	メンバー名	客観指標	順位	メンバー名	客観指標
1	西野七瀬	115306.07	6	松村沙友理	30514.93
2	齋藤飛鳥	107105.05	7	与田祐希	27300.86
3	白石麻衣	77860.25	8	梅澤美波	24729.78
4	山下美月	34051.79	9	星野みなみ	22543.10
5	生田絵梨花	33150.39	10	山崎怜奈	21395.45

表 5: 料理に関する嗜好抽出

料理名	客観指標の割合	個人指標の割合	偏差値	抽出結果
ハンバーグ	2.19%	17.07%	81.05	○
たこ焼き	1.97%	15.87%	79.01	○
ピザ	2.31%	15.42%	77.34	○
親子丼	1.45%	2.30%	51.77	
豚汁	1.18%	1.75%	51.19	
肉じゃが	1.06%	1.41%	50.73	
卵焼き	1.49%	1.83%	50.69	
コロッケ	1.44%	1.76%	50.66	
ハンバーガー	1.20%	1.45%	50.53	

3 個人特有の嗜好抽出結果の妥当性の検証実験

本章では，個人の嗜好抽出結果の妥当性の検証実験について述べる．本研究では，深層学習を用いて Twitter 利用者個人のツイート集合を分類し，分類結果から客観指標と個人指標を作成し，比較することで個人特有の嗜好抽出を行うことを目的としている．そのため，指標を用いた抽出結果が客観的に評価されている必要がある．今回は，本研究で構築した料理の嗜好抽出システムと乃木坂 46 の嗜好抽出システムを用いて検証を行った．

3.1 抽出結果が収束するために必要なツイート件数の検証

本実験で構築したシステムによる，個人特有の嗜好度を示す偏差値が収束するために必要なツイート件数の検証を行った．5 人の Twitter 利用者のデータを用意し，偏差値が何件で収束するのか検証を行った．今回は，ツイート数を 20，50，100，200，500，1000，2000 件にした場合に分けて検証を行った．

料理と乃木坂 46 に関するシステムでのツイート件数と嗜好度を示す偏差値の関係をそれぞれ図 5，図 6 に示す．

図 5，図 6 の結果から，1000 件が妥当なツイート件数であると判断した．また Twitter 利用者一人当たりの平均ツイート数が約 1400 件であるという結果も理由の一つである．個人の特徴を捉えることが必要なため，ツイート件数が多いに越したことはないが，必要なツイート件数を増やしてしまうと一部の利用者にしか利用できないシステムになってしまう．よって，嗜好抽出に必要なツイート件数は 1000 件と結論付けた．

現在の手法における問題点として，大きく二つ挙げられる．一つ目は，ツイート件数が 1000 件に満たない場合の対処法である．解決策としては，利用者がリツイートといいねを行った他の利用者のツイートを入力データに追加することが挙げられる．リツイートは他の人のツイートに対して，良い情報や気に入った情報などを引用する場合に用いられており，また，いいね機能はそのツイートをお気に入り登録できる機能である．よって，個人特有の嗜好との関連性は下がる可能性はあるが，双方ともに利用者が興味を持っているツイートであるといえるので，入力データを増やすという点では利用できるのではないかと考えている．

二つ目は，ツイートの投稿時間を考慮していないという点である．今回のシステムでは，最新のツイートと 1000 件前のツイートを同等に扱っている．しかし，嗜好は常に同じではないため，最新のツイートの方が個人の嗜好抽出としては重視すべきである．そのため，分類結果の所属確率だけでなく，ツイートの投稿時間を評価関数へ変換する際の要素に追加することで，システムの精度向上と，抽出結果の収束に必要な件数の減少が期待できると考えている．

表 6: 乃木坂 46 に関する嗜好抽出

メンバー名	客観指標の割合	個人指標の割合	偏差値	抽出結果
阪口珠美	1.17%	21.99%	99.40	○
渡辺みり愛	1.35%	5.14%	58.99	○
能條愛未	1.20%	3.56%	55.62	○
斉藤優里	1.30%	2.87%	53.71	
向井葉月	0.71%	2.06%	53.20	
伊藤かりん	0.86%	1.97%	52.64	
星野みなみ	2.46%	3.32%	52.04	
中村麗乃	0.98%	1.77%	51.88	
吉田綾乃クリスティー	0.60%	1.34%	51.75	
和田まあや	0.88%	1.59%	51.70	

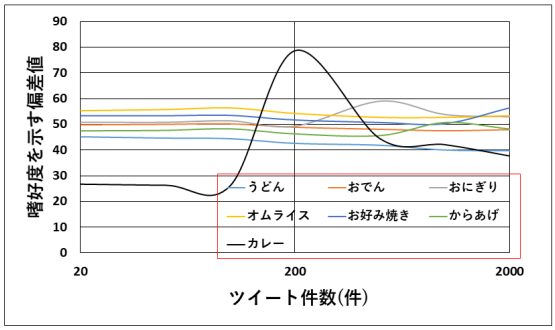

図 5: 料理に関するツイート件数と嗜好度を表す偏差値の関係

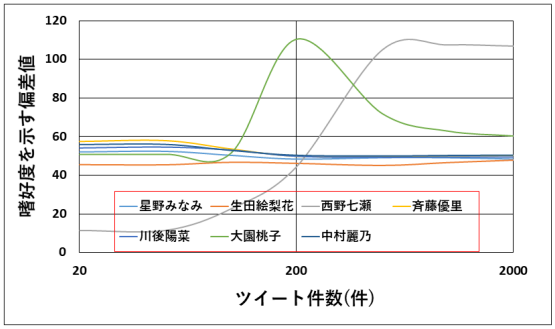

図 6: 乃木坂 46 に関するツイート件数と嗜好度を表す偏差値の関係

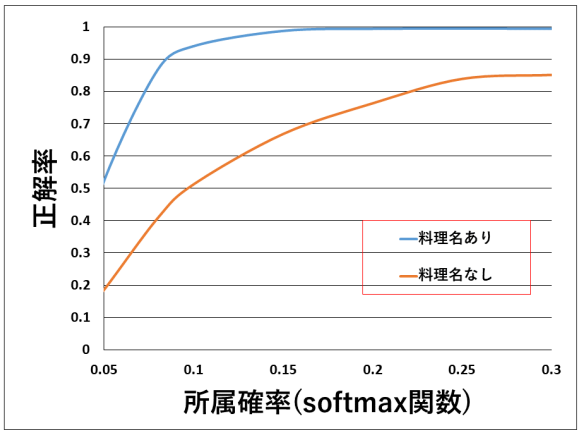

図 7: 料理に関する所属確率別での分類率

3.2 各項目の関連語に関する検証

分類ネットワークにおいて，項目名以外の単語から好きな項目を判定できるか検証を行った．各項目の関連語が学習されていると確認できれば，深層学習を用いた効果があると判断できるためである．

テストデータとして，各料理の正例データ 300 件 (合計 10500 件)，各乃木坂 46 メンバーの正例データ 250 件 (合計 1 万件) を用いて「醤油ラーメンおいしい」のような項目名を含む場合と,「醤油おいしい」のような項目名を削除した場合での分類率を比較した．分類結果をそれぞれ図 7，図 8 に示す．また,「ラーメンおいしい」などの項目名以外に関連語が含まれていないツイートを除くため，分類時の所属確率（softmax 関数値) 別での分類率から比較を行っている．

図 7 から，料理名を含まない場合でも 6〜8 割の分類は可能であるという結果が得られた．また図 8 から，乃木坂 46 のメンバー名を含まない場合でも 9 割の分類は可能であるという結果が得られた．

次に各料理の関連語を抽出し，関係のありそうな単語が学習されているのか検証を行った．ツイート内容が，入力層に登録している単語 1 単語のみであった場合の分類結果から関連語の抽出を行った．料理の分類ネットワークでは学習データ 3.5 万件の内，20 回以上出現している 2438 単語を入力層に登録している．また，乃木坂 46 の分類ネットワークでは学習データ 2 万件の内，10 回以上出現している 2436 単語を登録している．深層学習で学習された各料理と関連が高い単語，乃木坂 46 各メンバーと関連が高い単語をそれぞれ表 7，表 8 に示す．

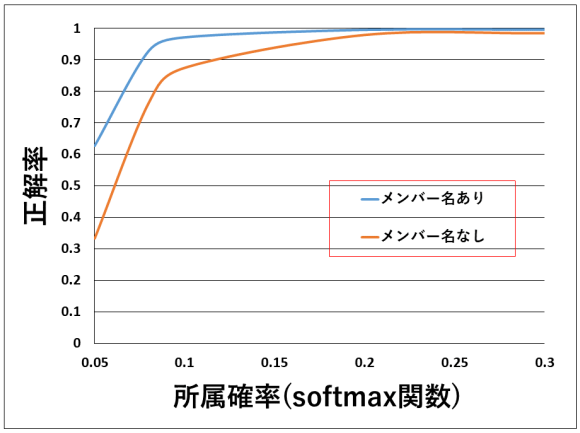

図 8: 乃木坂 46 に関する所属確率別での分類率

表 7: 深層学習で学習された各料理と関連が高い単語（上位 10 単語）

ラーメン	カレー	寿司	餃子	ハンバーガー
カップ	インド	回転	王将	サンドイッチ
天ぷら	ナン	ネタ	宇都宮	バーガー
麺	キーマ	はま	皮	マクドナルド
チャーシュー	無印	シロー	浜松	マック
煮干	スパイス	回る	包む	ピエロ
揚げたて	レトルト	魚	水	ラッキー
蕎麦	グリーン	サーモン	モチ	アメリカ
タンメン	ステーキ	すし	焼ける	モス
こってり	チキン	北海道	祭り	バンズ
コシ	ルー	マグロ	クリームチーズ	コーラ

表 7，表 8 を見ると，結果から，関連している単語が学習されていることが確認できた．しかし，上位に他の項目が入っている場合も見られる．関連語として他の項目があるということは，文章分類によって選別する項目が独立しておらず，他の項目と強い関係があるということである．そのため，深層学習を用いる理由の一つである各項目の特徴語の学習が難しくなっている．

この問題を解決する手法としては，深層学習を用いて関連の深い項目をまとめるという手法が挙げられる．各項目の関連語に共通の単語が含まれている場合，そ

表 8: 深層学習で学習された乃木坂 46 各メンバーと関連が高い単語（上位 10 単語）

秋元真夏	生田絵梨花	伊藤かりん	西野七瀬	佐々木琴子
真夏さん	いくちゃん	将棋	卒業	ダム
佐藤	ピアノ	甥	年間	アニメ
まなったん	アンコール	純奈	年内	松村
諦める	ロミオ	寺田	なぁちゃん	塩
さんま	グルメ	有能	発表	小百合
技術	葵	フォーカス	エース	ルックス
真	ミュージカル	手塚	みなみ	真夏
松村	ジュリエット	蘭世	久保	物語
楓	陽菜	理々杏	中村	ガン
若月	川後	話	居る	芋

れぞれの項目は関連度が高いと予測できる．そこで，これらの項目を 1 つにまとめて学習を行うことで，ネットワークの簡潔化と分類の精度向上が期待できる．この手法では，現在の手法では抽出できなかった特徴語を取得できるのではないかと考えている．

4 結論

本研究では，Twitter 利用者を対象として深層学習による文章分類を用いた個人特有の嗜好抽出を行うシステムを構築した．また，システムによる抽出結果の妥当性の実験として各項目の関連語からも嗜好が抽出できるのか検証を行い，6〜8 割のツイートでは分類が可能であったことから，抽出結果の妥当性が確認された．これにより，目視では困難であった個人特有の嗜好抽出が可能となり，個人の嗜好について客観的に見ることが可能となった．

今回のシステムは文章分類時の softmax 関数値を用いている．そのため，抽出する項目が独立しているほど有用なシステムであるため，個人特有の趣味を抽出するシステムの構築などに応用できると考えられる．

これを受けて今後の課題として，項目に関連した単語をより強く抽出する方法の検討や，時系列を考慮した抽出方法の検討を行い，より汎用性の高いシステムの構築を目標としていきたい．

参考文献

[1] 谷 季恵, 松村 嘉之: Twitter 上の情報拡散がもたらす商品販売効果推定モデルの提案，精密工学会学術講演会講演論文集，pp. 3-4(2016)

[2] 福井 一喜: 東京大都市圏に居住する若者の観光・レジャーにおける SNS 利用―「SNS 映え」を超克する若者たち―，E-journal GEO，14 巻 1 号 pp. 1-13(2019)

[3] 澤山 郁夫, 三宅 幹子: 大学生の独り言的ツイートは独り言なのか――発話傾向との関連から，パーソナリティ研究，27 巻 1 号 pp. 31-41

[4] 山根 宏彰, 萩原 将文: SNS における統計情報による文章の嗜好推定，日本知能情報ファジィ学会ファジィシステムシンポジウム講演論文集，pp. 780-783(2013)

[5] 原 侑平, 原 元司: トピックモデルを用いた Twitter ユーザの性別判定，第 80 回全国大会講演論文集，pp．537-538(2018)

[6] 浅妻 佑弥，山下 晃弘，松林 勝: SNS 上への発言の特徴分析に基づくユーザの属性推定，第 80 回全国大会講演論文集，pp．541-542(2018)

[7] DeepLearning4j，(URL)https://deeplearning4j.org/index.html

[8] Igo，(URL)https://github.com/sile/igo

[9] MeCab，(URL)http://taku910.github.io/mecab/