SNS のテキスト情報を利用したユーザの年代推定システムの提案

藤田 晃太朗† 前田 亮‡

†立命館大学情報理工学研究科 〒525-8577 滋賀県草津市野路東 1-1-1
‡立命館大学情報理工学部 〒525-8577 滋賀県草津市野路東 1-1-1
E-mail: †is0325pf@ed.ritsumei.ac.jp, ‡amaeda@is.ritsumei.ac.jp

あらまし 近年，主に Twitter，Instagram を始めとするソーシャルネットワーキングサービス（SNS）の情報を使用して，企業が市場分析や広告宣伝に活用している．その中で，年齢と性別は重要な属性とされている．しかし，実際に市場分析や広告宣伝を行う際に，SNS から API やスクレイピング等を用いて取得したユーザの属性が正確なものであるかどうかを判断するのは困難である．本研究では，年齢に着目し，Twitter のデータセットを使用して，各年代のユーザが発言する文章から特定の年齢層が使う単語を抽出する．抽出した単語の特徴量を用いて，機械学習を使用して分類を行い，対象ユーザの年代推定を行うことを目的とする．実験の結果，顔文字が年代の推定に一定の影響を及ぼすことがわかった．

キーワード Twitter，SVM，機械学習，年齢推定

1. はじめに

近年，主に Twitter，Instagram を始めとするソーシャルネットワーキングサービス（SNS）の情報を使用して，企業が市場分析や広告宣伝に多く活用している[1]．このような情報を使用する背景として，スマートフォンの普及により SNS の利用者数が急増したことが挙げられる．企業にとっては，利益を増やすことができる機会でもある．大量にある SNS の情報において重要視されていると考えられるのは，「年齢」と「性別」である．マーケティングを行う際に年齢別の嗜好や性別による嗜好を判別することができれば，商品を売り込む際に効果的なマーケティングができるのではないかと考える．

しかし，実際に市場分析や広告宣伝を行う際に，SNS から API やスクレイピング等を用いて取得したユーザの属性が正確なものであるかどうかを判断するのは困難である．その理由として，必ずしも SNS を使っているユーザが本人であるか確証が取れないことが挙げられる．SNS は基本的に年齢，性別問わずに誰でも自由に使えるため，SNS のプロフィール上では「20 代男性」と書いてあるが，実際にそのアカウントを使用しているのは，違う年齢層の人間であることも考えられる．

また，SNS を使用するユーザが人間でないことも考えられる．これは，Bot システムと呼ばれる，投稿者が人間でなく，プログラム上で自動的に発言内容を投稿するシステムである．Bot システムを使用している場合では，発言が同じ内容を何度も投稿するため，実在する人間の発言内容ではない．これらの理由より，市場分析等を行う際に，実在する人物でない発言内容や年齢や性別を偽っている人間の発言内容が障害となることが考えられる．

本研究では，正確にユーザの属性を推定することを目標とする．性別推定については，既存研究において多くされているが，年齢推定は性別推定に比べて多くされていない，その理由として，性別は男女の 2 つしかないため，推定を行うのは比較的簡単である．しかし，年齢は性別に比べて，比較する対象が多いため，単純ではない．そこで，本論文では顔文字に着目して年代推定を行う手法を提案する．顔文字は人間の感情をテキストメッセージ上で表現することができる記号である．テキストメッセージ上では，会話や電話とは違い，相手の感情を推測することはできない．しかし，顔文字を用いることによって，相手に感情を伝えることができる．若者であるほど，顔文字を使用する割合が高いという報告[2]があるため，今回顔文字を特徴量として年齢推定を行う．

2. 関連研究

ユーザの属性を推定する研究は数多く存在している．特に，性別を推定する研究は多くの報告例がある．

佐古ら[3]は，Twitter ユーザの性別を自動判別するシステムを考案した．Twitter から短文テキストを収集し，前処理を終えたデータの中から特徴を抽出し，サポートベクターマシン（SVM: Support Vector Machine）を用いて性別判定を行っている．佐古らの研究では，代名詞と終助詞の 2 つの素性が性別判定に有効であることを示している．

Burger ら[4]は，Twitter ユーザの性別を判定するために，言語に依存しない特徴について研究を行った．Burger らは，性別を分ける特徴が 4 つ（アカウント名，本名，職業や家庭環境等の詳細な情報，ツイート文）にあると考え，上記の 4 つの特徴を入れた場合において精度が 92.0%と非常に高い値となった．また，本名を特徴として使った場合には 89.1%の精度，アカウント名を使った場合は 77.1%の精度となった．性別を推

定する際には，名前が大きな特徴となっていることがわかり，特に本名が重要であることを示している．

年齢を推定する研究も存在している．財津ら[5]は，犯罪者プロファイリングを実現するために，ブログの著者の年代を推定している．この研究では，ランダムフォレスト（RF: Random Forest）と SVM を用いて年代を推定している．使われている特徴量としては，「ずっと（副詞）」，「は（係助詞）」が使用されている．財津らの考察において，年代推定に有効な特徴は文体的特徴だと結論付けられていることから，本研究においてもこの特徴を用いる．

3. 提案手法

3.1. 使用する用語の定義 [6]

本研究において使用する用語について説明を行う．

Ⅰ. TL（タイムライン）: ユーザがフォローしているアカウントのツイート文だけを見ることができる．基本的に時系列順となっている．

Ⅱ. ツイート: ユーザが Twitter に投稿する内容のことである．ユーザは一回の投稿で，140 文字の文章を投稿でき，画像や動画も同時に投稿可能である．

Ⅲ. RT（リツイート）: Twitter ユーザが他人のツイートを自分のツイートのように改めて投稿する機能である．基本的な機能はツイートと同じである．

Ⅳ. @（メンション）: ある特定のユーザに向けて，送信する投稿内容である．ユーザ名の前に（@）がついているツイートである．

3.2. 提案手法の概要

本研究における提案手法の処理の流れを図 1 に示す．本研究では，主に 4 つの処理が行われている．第一段階では，Twitter-Get-Old-Scraper [7] を用いて，対象となるツイートを取得し，アカウント毎に取得したツイート内容をひとまとめにする．第二段階では，取得したツイートに対して形態素解析を行う．第三段階では，形態素解析の結果に対して，前処理を行い特徴量を抽出する．第四段階では，取得したアカウントを正解データとして機械学習の手法の一つである SVM を使用して評価を行う．ここで，Twitter API [8] を使用しない理由として，Twitter API は，起動した時点から 1 週間前までのツイートしか取得できないという制限があり，長期間にわたってツイートを取得する本研究においては Twitter API を使用するのは不適だと考えたからである．

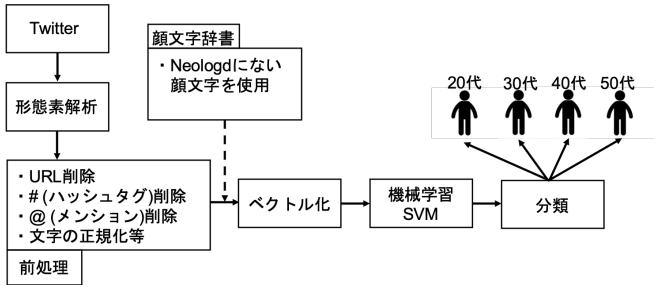

図 1 提案手法の処理の流れ

3.3. 前処理

本研究における形態素解析では MeCab [9] を使用する．また，本研究は年齢について着目した研究であり，年齢によって使用する単語が変化する可能性が高い．そのため，できるだけ多くの単語を収集する必要があると考えられる．そこで，MeCab に用いる辞書として新語・固有表現に対応した mecab-ipadic-NEologd [10] を使用する．SNS 上においては，話題となった流行語や新語など，従来の MeCab の辞書（ipadic）に載っていない単語が数多く存在すると考えられる．そのような従来の辞書では適切に分かち書きをすることができなかった文章に対して，mecab-ipadic-NEologd を用いることで適切に分かち書きをすることができる．

たとえば「トランプ大統領が再選に向けて動き出した」という文に対して，ipadic を用いて形態素解析を行うと，「トランプ/大統領/が/再選/に/向け/て/動き出し/た」と分割され，トランプという言葉が「人名」ではなく一般名詞という結果となる．一方，mecab-ipadic-NEologd を用いて形態素解析を行うと，「トランプ大統領/が/再選/に/向け/て/動き出し/た」と分割され，「トランプ大統領」という言葉となり「人名」として結果が出力される．

形態素解析を行った結果に対して，テキストの前処理を行う．今回は，neologdn [11] を用いてテキストの前処理を行う．neologdn を使用した前処理では「全角・半角の統一」，「桁区切りの”,”の対処」，「空白の対処」を行うことができる．

「全角・半角の統一」は，SNS 上の文章では人によって同じ文字であっても，人によって書き方が異なるいわゆる表記揺れが多く発生している．具体的には，文章の表現として，「ﾊﾝｶｸｶﾅ」や「ハンカクカナ」のように，同じ言葉であるが，これら 2 つの言葉は異なる単語として分けられてしまう．そのため，neologdn を用いることによって，半角カナから全角カナに表現を変更する．一方，「！？」等の記号については，全角カナから半角カナに表現を変更する．

「桁区切りの”,”の対処」は，金額を表現する際に使われる「3,000 円」等の数字と数字の間にある「,」を

削除することである．

「空白の対処」は「　PRML 副読本　」等の意図的に空白を空けている文章が存在するが，この文章について前処理を行わずに形態素解析を行うと，空白部分が「空白」という形で分かち書きされてしまう．そのため，neologdn を用いることによって，「PRML 副読本」のように表現を変更する．

加えて，@（メンション），http（URL 文），pic（画像部分），RT（リツイート）等の個人の発言部分とは関係ない内容については，削除を行う．

3.4. データセット

本研究で使用するデータセットとして，Twitter より取得したデータを用いる．データ取得においては，図 2 に示すようにツイプロ[12]より取得を行う．ツイプロは，日本人の Twitter 利用者のプロフィールを検索することができる検索エンジンサイトである．図 3 に示すように，ツイプロの年齢の項目を選択して，年齢の項目から上位 300 件を対象にしてデータ取得を行う．図 3 より，年齢の項目からこの時点で年代に分けられているのではないかという意見があるが，ツイプロはプロフィール上の 20 歳等の年齢を表す単語や西暦により年代を判断しているため，正確に年代に分けられているとは言い難い．そのため，データ取得を行う際に，プロフィール欄を確認し，明らかに対象とする年齢でない場合や広告目的のアカウントである場合には，そのアカウントを取得しない．

今回，取得するデータの対象としては，20 代，30 代，40 代，50 代のアカウントデータを対象とする．データの取得範囲としては，20 代は 23~27 歳，30 代は 33~37 歳，40 代は 43~47 歳，50 代は 53~57 歳を対象とする．その他の世代のデータに関しては，1 つのアカウントから取得できるデータが少ないため，今回の研究では除外した．また，データの取得期間については，2016 年 10 月 1 日から 2019 年 10 月 1 日の期間に投稿されたツイートを取得した．このような，データの取得に条件をつけた理由としては，年齢に関する研究であるため，むやみにデータの取得期間を長くしてデータを取得した場合，年代がずれてしまう可能性があったためである．また，データの取得範囲を広くした場合には，31 歳のアカウントデータと 39 歳のアカウントデータでは，31 歳は 20 代後半とほぼ違いがなく 39 歳は 40 代前半とほぼ違いがないと考えられるため，データの取得を制限する．最終的な各年代別のアカウント数は，20 代が 541 個，30 代が 679 個，40 代が 709 個，50 代が 815 個となった．また，最終的に取得した総ツイート数は，20 代が 1,069,948 ツイート，30 代が 1,390,817 ツイート，40 代が 1,178,429 ツイート，50 代が 1,204,985 ツイートとなった．

図 2 ツイプロのトップページ

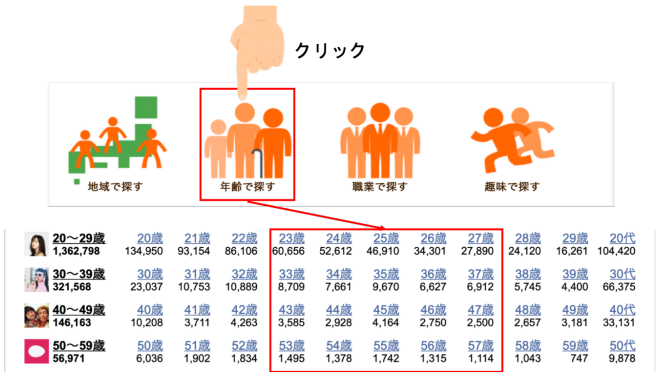

図 3 ツイプロの年齢の項目

3.5. 特徴量

用いる特徴量は，前述した財津らの研究において，年代推定に有効とされた文体的特徴である，「ずっと（副詞）」，「は（係助詞）」を用いる．また，著者の主観に基づき，年代推定に有効な文体的特徴として，「名詞」，「動詞」，「感動詞」を用いる．「感動詞」とは，感動，呼びかけ，応答，あいさつを表す語句の品詞である．それぞれの語句例を以下に示す．

⑴「感動」：ああ（嗚呼），あら，あれ
⑵「呼びかけ」：もしもし，さあ
⑶「応答」：はい，いや，うん
⑷「あいさつ」：こんにちは，こんばんは

3.6. 顔文字辞書

本研究では，年代推定を行うために顔文字辞書を作成する．顔文字辞書を作成するにあたり，使用する顔文字を定義する必要がある．既存の顔文字に関する研究は多くされているが，集めた顔文字を公開しているケースは少ない．また，顔文字は多種多様な記号や数字を組み合わせて作られたものであるため，文字列中のどこまでが顔文字とされるかが議論となる．渡邉ら[13]の研究では，系列ラベリングを使用して「平仮名」や「数字」等の文字種に素性値を与え Twitter の記事コーパスより顔文字を抽出している．

今回は,Twitter よりデータを収集しているため,SNS でよく使われる顔文字を集める必要がある．そこで，今回は「TL で集めた顔文字とか」[14]より顔文字を抽出して，顔文字辞書を作成する．本サイトは，Twitter を始めとする SNS における TL より顔文字を抽出し，纏めたサイトである．このサイトより，675 個の顔文字を抽出する．また，抽出対象となる顔文字は mecab-ipadic-NEologd の辞書に存在していない顔文字を対象としている．

3.7. 分類手法

本研究における分類手法として SVM を使用した．SVM は 2 クラス分類問題を解くための教師あり学習手法である．今回，使用した分類手法は主に 2 クラス分類問題を解くための手法であり，分類する組み合わせも 2 クラスとした．そのため，今回実験を行う組み合わせは 20 代～50 代をそれぞれ総当たりにした組み合わせ（例：20 代と 30 代，20 代と 40 代等）にして分類を行う．

4. 評価実験

4.1 評価方法

検証は，まずデータセットを学習用データ 8 割，テスト用データ 2 割に分割する．加えて，学習用データに対して，scikit-learn の GridSearch ライブラリを使用して最適なパラメータを求める．求めたパラメータの値を元々の学習用データの学習に用いて，学習したモデルをテスト用データに用いて評価を行った．

4.2 評価実験の尺度

本研究における提案手法の評価は，混同行列（Confusion Matrix）を用いて各 2 値分類問題において（1）正解率（Accuracy），（2）適合率（Precision），（3）再現率（Recall），（4）F 値（F-measure）による評価を行う．混同行列とは，機械学習における 2 値分類問題に対して分類結果を行列の形で表現したものであり，機械学習におけるモデルの評価として使用される．本研究では，各実験対象に対して，適合率，再現率，F 値をそれぞれ算出した．適合率は，学習モデルがある年代だと推定したもののうち，実際にその年代である割合を表し,再現率は実際にある年代であるもののうち，その年代だと推定された割合を表し，F 値は適合率と再現率の調和平均によって示される値を表す．今回の実験で用いる混同行列の例を表 1 に示す．

表 1 において TP（True Positive）は学習済みのモデルに対して，モデルの予測が 20 代のアカウントデータであった場合に，実際の結果も 20 代のアカウントデータである件数である．FP（False Positive）は，予測が 20 代のアカウントデータであった場合に，実際の結果が 50 代のアカウントデータである件数である．FN（False Negative）は，予測が 50 代のアカウントデータであった場合に，実際の結果が 20 代のアカウントデータである件数である．TN（True Negative）は，予測が 50 代のアカウントデータであった場合に，実際の結果も 50 代のアカウントデータである件数である．

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (4)$$

表 1 混同行列の例

	実際：20 代	実際：50 代
予測：20 代	TP	FP
予測：50 代	FN	TN

本研究における，実験の比較年代層は以下の通りとする．

実験 1. 20 代と 30 代
実験 2. 20 代と 40 代
実験 3. 20 代と 50 代
実験 4. 30 代と 40 代
実験 5. 30 代と 50 代
実験 6. 40 代と 50 代

4.3 実験結果

本研究における実験結果を表 2, 3, 4 ,5, 6, 7 に示す．全ての年齢層の実験結果について説明を行う．実験 1~6 の全ての実験において，顔文字なしの場合と顔文字ありの場合において比較を行う．

実験 1（20 代と 30 代）の条件においては，顔文字を特徴量として取り入れる前後において，適合率，再現率，F 値それぞれの変化は微小であった．それぞれの変化は 0.01 が最大であったため，20 代と 30 代の顔文字による変化はそれほど見られなかった．

実験 2（20 代と 40 代）の条件においては，結果に大きく違いが見られた．適合率は顔文字がある場合とない場合において，双方とも 0.10～0.11 ポイントの向上が見られた．一方，再現率，F 値は 20 代ではそれぞれ値の向上が見られたが 40 代では値の低下が見られた．

実験 3（20 代と 50 代）の条件においては，実験 2 と

同様の傾向が見られた．適合率では，それぞれ 0.05～0.06 ポイントの向上が見られた．再現率と F 値の結果についても同様に，20 代に値の向上が見られたが 50 代においては値の低下が見られた．

実験 4（30 代と 40 代）の条件においては，実験 1，実験 2，実験 3 の結果とは異なる結果が見られた．顔文字なし，顔文字ありの結果を確認すると，適合率，再現率，F 値全てにおいて，双方とも 0.02～0.04 ポイントの向上が見られた．

実験 5（30 代と 50 代）の条件においては，実験 1，実験 2，実験 3 の結果とは異なる結果が見られた．結果を確認すると，30 代の結果において適合率は，0.01 の低下が見られた．また，50 代の結果においては変化が見られなかった．一方，再現率と F 値においては，30 代では値の向上が見られ，50 代では値の低下が見られた．

実験 6（40 代と 50 代）の条件においては，実験 5 とほぼ同じ傾向が見られた．適合率は 40 代の結果では，変化はなく 50 代の結果では 0.02 ポイントの向上が見られた．一方，再現率と F 値は実験 5 と同じ傾向となり，40 代では，再現率と F 値両方ともに値の向上が見られ，50 代では，値が低下する傾向が見られた．

表2 実験1の結果

年齢層	Precision	Recall	F1	Accuracy
20代 顔文字 なし	0.47	0.56	0.51	0.53
30代 顔文字 なし	0.59	0.51	0.55	
20代 顔文字 あり	0.47	0.57	0.52	0.53
30代 顔文字 あり	0.60	0.49	0.54	

表3 実験2の結果

年齢層	Precision	Recall	F1	Accuracy
20代 顔文字 なし	0.32	0.27	0.30	0.50
40代 顔文字 なし	0.58	0.64	0.61	
20代 顔文字 あり	0.42	0.76	0.54	0.50
40代 顔文字 あり	0.69	0.33	0.45	

表4 実験3の結果

年齢層	Precision	Recall	F1	Accuracy
20代 顔文字 なし	0.37	0.32	0.34	0.55
50代 顔文字 なし	0.63	0.68	0.65	
20代 顔文字 あり	0.43	0.55	0.49	0.57
50代 顔文字 あり	0.68	0.57	0.62	

表5 実験4の結果

年齢層	Precision	Recall	F1	Accuracy
30代 顔文字 なし	0.44	0.47	0.46	0.47
40代 顔文字 なし	0.50	0.47	0.48	
30代 顔文字 あり	0.46	0.47	0.47	0.49
40代 顔文字 あり	0.52	0.51	0.52	

表6 実験5の結果

年齢層	Precision	Recall	F1	Accuracy
30代 顔文字 なし	0.43	0.48	0.46	0.54
50代 顔文字 なし	0.62	0.57	0.60	
30代 顔文字 あり	0.42	0.52	0.47	0.52
50代 顔文字 あり	0.62	0.53	0.57	

表7 実験6の結果

年齢層	Precision	Recall	F1	Accuracy
40代 顔文字 なし	0.47	0.74	0.57	0.48
50代 顔文字 なし	0.53	0.26	0.35	
40代 顔文字 あり	0.47	0.78	0.59	0.49
50代 顔文字 あり	0.55	0.24	0.33	

5. 考察

実験 1, 実験 2, 実験 3 の結果より, 20 代とその他の年代における比較では，20 代の適合率，再現率，F 値の値が全て向上または同じ値となり，下落している値はなかった．この結果より，20 代では顔文字を多く使っており，特徴量として有効であるため，20 代における適合率, 再現率, F 値が向上したものと考えられる．

傾向として，20 代よりも 40 代，50 代において適合率，再現率，F 値が高い傾向が見られた．特に，実験 5 の 40 代顔文字ありにおける結果では，再現率が 0.78 という値となり, 本研究における実験結果においては，最も高い値が得られた．また，再現率においては，顔文字なしの場合において値が高い傾向にある．具体的には，実験 2 の結果の 40 代顔文字なしにおける再現率は 0.64，実験 3 の結果の 50 代顔文字なしにおける再現率は 0.68 となっている．これらの 2 つの結果は，顔文字ありの場合では，0.33，0.57 となる．これは，顔文字を特徴量として取り入れたことにより，逆に予測に影響を及ぼしたと考えられる．F 値については，再現率の変化に連動して，数値が変化している傾向にある．本研究では，顔文字が年齢推定にどのような影響を及ぼすかについて実験を行ったが，再現率の変化が大きく F 値に影響を及ぼしているため，今後は，再現率を向上できるような組み合わせで推定を行っていくことが重要であると考える．

6. まとめと今後の展望

本研究では，Twitter のテキスト情報を用いて，ユーザの年代を推定するシステムの提案を行った．また，ツイートにおける顔文字に着目して，年代推定システムの精度向上について検討を行い，実験を実施した．結果は，年齢の組み合わせによっては，適合率，再現率，F 値の値が向上する傾向がみられ，特に再現率，F 値が大きく値が向上した組み合わせがあった. しかし，逆に適合率，再現率，F 値が低下する組み合わせもあった．再現率，F 値については，一方の年代で値が上昇すると，もう一方の年代で値が低下する傾向が見られた．これは，トレードオフの関係になっていると考えられる．

今後の展望としては，現時点では全体的な精度があまり高くないため，引き続き精度向上を検討したい．また，年代推定だけではなく性別等の他の要素を含めた推定システムの構築を進めていき，最終的には年代だけを推定するシステムではなく，性別や居住地等を推定するシステムを開発していきたいと考えている．

参 考 文 献

[1] 経済産業省，“平成 27 年度商取引適正化・製品安全に係る事業（ソーシャルメディア情報の利活用を通じた BtoC 市場における消費者志向経営の推進に関する調査）”, https://www.meti.go.jp/press/2016/04/20160411002/20160411002.html（参照 2019/12/13）

[2] Yukiko Nishimura, “A sociolinguistic analysis of emotion usage in Japanses blogs: variation by age and topic”, AoIR, 2015

[3] 佐古 龍，原 元司，“Twitter 利用者の性別判定システムの構築”，第 29 回ファジィシステムシンポジウム, 2013

[4] John D. Burger, John Henderson, George Kim and Guido Zarrella, “Discriminating Gender on Twitter”, Proc. of conference on EMNLP，pp. 1301-1309，2011

[5] 財津亘，金明哲，” 機械学習を用いた著者の年齢層推定―犯罪者プロファイリング実現に向けて―”, The harris science review of Doshisha University, Vol. 59，No. 2，pp. 58-65, 2018

[6] WANG Yu，前田 亮，“トピックモデルを用いた Twitter 関連情報に基づくユーザ嗜好の推測手法の提案”，第 11 回データ工学と情報マネジメントに関するフォーラム，C8-3，2019

[7] PJHRobles/Twitter-Get-Old-Tweets-Scraper - Github, https://github.com/PJHRobles/Twitter-Get-Old-Tweets-Scraper（参照 2019/11/25）

[8] Twitter API, https://developer.twitter.com/content/developer-twitter/ja.html（参照 2019/12/11）

[9] MeCab: Yet Another Part-of-Speech and morphological Analyzer https://taku910.github.io/mecab/（参照 2019/11/25）

[10] neologd/mecab-ipadic-neologd - Github, https://github.com/neologd/mecab-ipadic-neologd（参照 2019/11/25）

[11] ikegami-yukino/neologdn – Github, https://github.com/ikegami-yukino/neologdn（参照 2019/11/27）

[12] ツイプロ, https://twpro.jp（参照 2020/01/07）

[13] 渡邉 謙一, 高橋 寛幸, 但馬 康宏, 菊井 玄一郎, “系列ラベリングによる顔文字の自動抽出と顔文字辞書の構築”, 言語処理学会第 19 回年次大会, pp. 866-869，2013

[14] “TL で集めた顔文字とか”，http://kaomoji.n-at.me（参照 2020/01/08）