明治大学総合数理学部

**2020** 年度

卒　業　研　究

# ツイートの長さと句読点に基づく年齢・性別の推定

学位請求者　先端メディアサイエンス学科

江口大賀

# 目次

# 第 1 章

# はじめに

Twitter 上で投稿される文章は，身の回りの出来事や趣味に関する事を口語体で投稿される事が多い．そこで使われる言葉使いや単語は，ユーザの年代や性別によって変わる事が予想される．代表的なツイート例を，図 1.1 に示す．

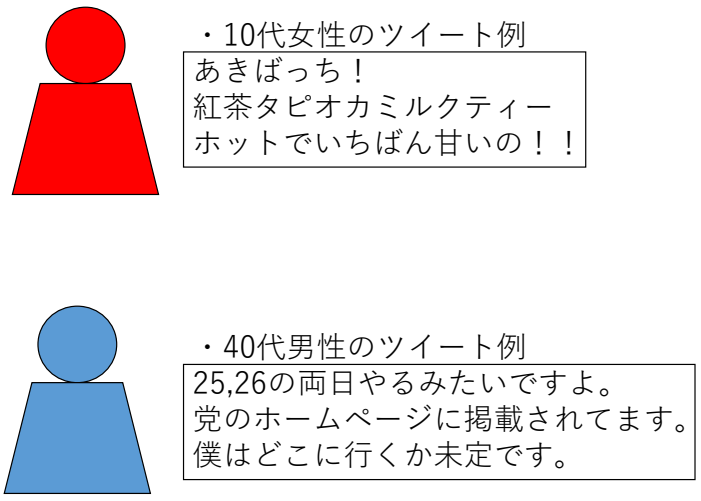


図 1.1　ツイート例

長浜らは，ツイートから得られた単語の $\chi$ 二乗値を用いるアルゴリズムを用いて，ユーザの性別の推定を行った [1]．男子は「僕，俺」などの名詞を多用し，女子は「*)，ω」などの記号を多用する傾向があった．また，品詞の出現割合では，男女間で大きな偏りがなかったことを報告している．

そこで本研究では，自然言語処理の段階でストップワードに指定されて削除されがちである句読点等に注目し，Twitter に投稿された文章から，ユーザの年齢と性別の属性推定を行う．10 代のユーザは句読点の使用回数が 20 代以上のユーザより少なく，年代が上がるにつれて句読点の使用回数が多くなるという傾向が，予想

されるからである．しかし，若い年代の句読点の使用回数が少ないのは，投稿する文章量が少ないからではないかという懸念点が挙げられる．そこで，文章の長さにも考慮し，句読点を使用する割合に注目した解析を行う．以上の本実験のシステム構成図を図 1.2 に示す．

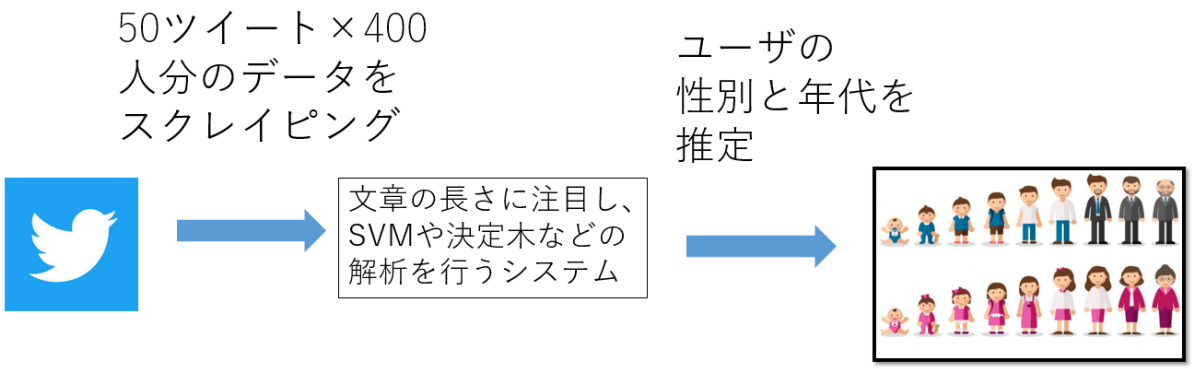


図 1.2　システム構成図

# 第 2 章

# 提案手法

## 2.1　データ収集

本研究のデータ収集では，プロフィールから Twitter のユーザを検索する「ツイプロ」[4] を用いる．このサービスは各アカウントのプロフィールの文章から，年齢・性別・地域・職業・趣味などを分類している．「ツイプロ」を用いて，プロフィールから年代と性別が分かるユーザのアカウント名を集め，TwitterAPI 等を用いて，収集したユーザのツイートを取得する．

収集した合計 400 人分のユーザの属性と人数を表 2.1 に示す．1 人当たり 50 個のツイート，1 年で計 20,000 個のツイートを収集する．2019 年度と 2020 年度で合わせて 40,000 個のツイートを収集する．本研究のデータ取得時期を表 2.2 に示す．

表 2.1　収集したユーザの属性の統計値

| 性別\年代 | 10 代 | 20 代 | 30 代 | 40 代 | 合計 |
|---|---|---|---|---|---|
| 男性 | 50 | 50 | 50 | 50 | 200 |
| 女性 | 50 | 50 | 50 | 50 | 200 |
| 合計 | 100 | 100 | 100 | 100 | 400 |

表 2.2　データ取得時期

| | 2019 年度 | 2020 年度 |
|---|---|---|
| 期間 | 2019 年 11 月 4 日<br>～2019 年 11 月 6 日 | 2020 年 11 月 30 日<br>～2020 年 12 月 2 日 |

## 2.2　用いた手法

本実験では，3 つの手法による検証を行う．各手法の仮説，用いたデータ数，その手法による検証を行った年度を表 2.3 に示す．

表 2.3　手法の仮説とデータ数一覧

| 手法 | 立てた仮説 | 学習データ数 | 評価データ数 | 検証した年度 |
|---|---|---|---|---|
| 1 | 各属性で最も差が出る単語は，句読点である | 200 | 200 | 2019<br>2020 |
| 2 | 若い年代は，句読点を使わない | 100 | 100 | 2019<br>2020 |
| 3 | 若い年代は，文章が短い | 100 | 100 | 2020 |

また，各手法の説明変数と目的変数を表 2.4 に示す．なお，手法 3 の説明変数の”生起確率”は，ユーザの全ツイートの文章における，句読点が含まれる割合を表した数値である．すなわち，以下の式で示される．

$$\text{生起確率} = \frac{\text{ユーザの句読点の使用件数}}{\text{ユーザの全ツイートの文字数の合計}}$$

表 2.4　手法の説明変数と目的変数

| 手法 | 説明変数 | 目的変数 |
|---|---|---|
| 1 | 400 人のユーザ内の誰かが 3 回以上使用した単語 (2019 年度は 1,2052 個, 2020 年度は 1,1147 個) の使用件数 | 10 代，20 代，30 代，40 代の 4 年代と性別の計 8 種類 |
| 2 | 「。」と「、」の 2 単語の使用件数 | ある年代とそれ以外の年代の 2 種類 |
| 3 | 「。」と「、」の 2 単語の生起確率 | |

## 2.3　実験方法

様々な機械学習を用いて年代や性別の学習を行い，各属性を推定する．手法 1 ではランダムフォレスト，手法 2・3 ではサポートベクタマシン (以下，SVM と称する) を用いる．

本実験は，python を用いる．ツイートの収集では，urllib，pyquery，TwitterAPI を用いる．得られたツイートの自然言語処理には janome を用いる．属性推定における決定木とサポートベクタマシンには，sklearn を用いる．

# 第 3 章

# 実験結果

## 3.1　手法 1：特徴語の抽出

手法 1 から抽出された特徴語の上位 10 語と、その重要度を表 3.1 に示す．なお，重要度の算出には，Random Forst Classifier 内の関数である feature_importance[5] を用いた．重要度は，各々の説明変数の値が，目的変数を算出するのにどれ位重要かを示す．句読点の 2 単語の重要度は，1 位と 3 位であった．また，特徴語の上位 5 語の各属性における平均使用件数を表 3.2 に示す．

表 3.1　特徴語と重要度

| 順位 | 単語 | 重要度 [%] |
|---|---|---|
| 1 | 。 | 0.84 |
| 2 | 私 | 0.74 |
| 3 | 、 | 0.62 |
| 4 | を | 0.61 |
| 5 | まし | 0.46 |
| 6 | 僕 | 0.42 |
| 7 | ない | 0.39 |
| 8 | です | 0.37 |
| 9 | ある | 0.37 |
| 10 | ！ | 0.36 |

表 3.2　各属性の特徴語の平均使用件数

| | 10 代 | | 20 代 | | 30 代 | | 40 代 | |
|---|---|---|---|---|---|---|---|---|
| | 男性 | 女性 | 男性 | 女性 | 男性 | 女性 | 男性 | 女性 |
| 。 | 7.1 | 7.0 | 33.7 | 21.4 | 45.4 | 41.2 | 65.7 | 37.3 |
| 私 | 0.4 | 2.0 | 0.3 | 2.3 | 1.5 | 4.3 | 2.0 | 6.4 |
| 、 | 10.6 | 20.4 | 34.4 | 20.8 | 39.3 | 39.3 | 60.9 | 41.5 |
| を | 7.8 | 7.9 | 21.3 | 13.7 | 27.9 | 23.2 | 38.2 | 26.7 |
| まし | 3.2 | 7.0 | 7.4 | 6.4 | 8.3 | 9.2 | 11.1 | 10.0 |

## 3.2　手法 2：句読点による年代ごとの分類

各年代の句読点の使用件数の平均と標準偏差を表 3.3 に示す．この結果から，年代が上がるにつれて，句読点の使用件数が多くなる事が分かる．

表 3.3　各年代の句読点の使用件数の統計値

| 年代 \ 統計値 | 平均 | | 標準偏差 | |
|---|---|---|---|---|
| | 。 | 、 | 。 | 、 |
| 10 代 | 7.1 | 15.5 | 11.8 | 15.9 |
| 20 代 | 27.5 | 27.6 | 43.1 | 29.4 |
| 30 代 | 43.3 | 39.3 | 39.5 | 32.5 |
| 40 代 | 52.0 | 51.2 | 40.6 | 37.9 |

句読点の使用件数から，SVM を用いて年代の推定を行う．その結果を表 3.4 に示す．この表での再現率は，100 個の評価データの属性を推定し，正解したデータ数の割合とする．10 代の推定の再現率が，特に高い事が分かった．4 年代の中で最も再現率が高かった 10 代の推定の散布図を図 3.1 に示す．

表 3.4　ある年代 100 人とそれ以外の年代 100 人の句読点による推定の再現率

| | 10 代 | 20 代 | 30 代 | 40 代 |
|---|---|---|---|---|
| 再現率 [%] | 70.6 | 56.0 | 62.3 | 67.4 |

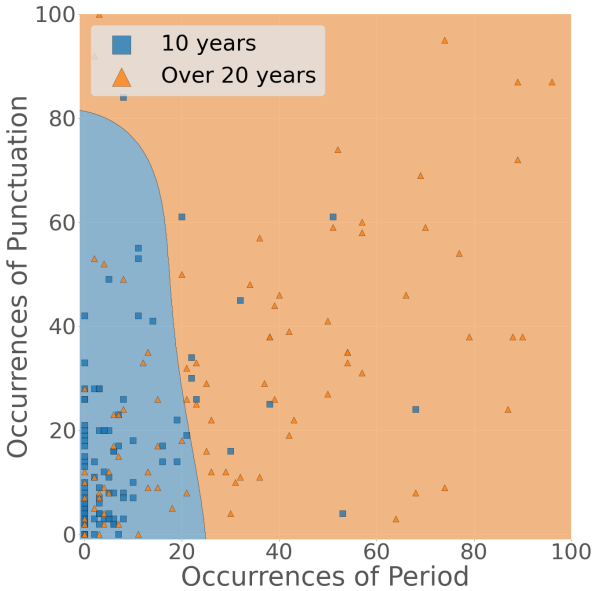


図 3.1　10 代 100 人と 20,30,40 代の 100 人の句読点による SVM の分類の散布図

## 3.3　手法 1・2：2019 年度と 2020 年度の実験結果の比較

手法 1・2 における 2 年間の実験結果を比較した．表 3.5 より，手法 1 の結果では句点 (「。」のこと)，読点 (「、」のこと) ともに重要度の値が，2019 年度より下がった．また，表 3.6 より，手法 2 の 10 代の推定の再現率が，2019 年度の結果よりも 5.6 %低下した．

表 3.5　手法 1:“特徴語と重要度”の比較

| | 2019 年度 | | 2020 年度 | |
|---|---|---|---|---|
| 順位 | 単語 | 重要度 [%] | 単語 | 重要度 [%] |
| 1 | を | 0.93 | 。 | 0.84 |
| 2 | 。 | 0.87 | 私 | 0.74 |
| 3 | 、 | 0.72 | 、 | 0.62 |
| 4 | 私 | 0.70 | を | 0.61 |
| 5 | まし | 0.57 | まし | 0.46 |

表 3.6　手法 2:“ある年代 100 人とそれ以外の年代 100 人の句読点による推定の再現率”の比較

| | 10 代 | 20 代 | 30 代 | 40 代 |
|---|---|---|---|---|
| 2019 年度の再現率 [%] | 76.2 | 54.5 | 61.8 | 62.9 |
| 2020 年度の再現率 [%] | 70.6 | 56.0 | 62.3 | 67.4 |

## 3.4　手法 3：文章の長さを考慮した句読点による分類

各ユーザ群の文字数 (1 ツイート当たり) の平均を表 3.7 に示す．文字数をカウントする上で，URL やメンション (Twitter のユーザ名を載せる事) 等は，正規表現を用いて削除した．

この結果から，年代が上がるにつれて，1 ツイート当たりの文章量が多くなる事が分かった．中でも，10 代男子の 1 ツイート当たりの平均文字数が 32 文字であり，特に少ない．

表 3.7　各属性の文字数 (1 ツイート当たり) の平均

| | 男子 | 女子 |
|---|---|---|
| 10 代 | 32.0 | 40.0 |
| 20 代 | 50.7 | 42.8 |
| 30 代 | 58.7 | 57.1 |
| 40 代 | 68.6 | 58.2 |

句読点の生起確率の平均と標準偏差を，各年代に分けて表 3.8 に示す．10 代の句点 (「。」のこと) と読点 (「、」) の生起確率は，0.36 %と 0.89 %であり，4 年代の中で最も低い値である．また，句読点の生起確率を用いた SVM による年代の推定結果を表 3.9 に示す．

表 3.8　各年代の句読点の生起確率 [%] の統計値

| 年代 \ 統計値 | 平均 | | 標準偏差 | |
|---|---|---|---|---|
| | 。 | 、 | 。 | 、 |
| 10 代 | 0.36 | 0.89 | 0.49 | 0.68 |
| 20 代 | 1.06 | 1.08 | 1.20 | 0.84 |
| 30 代 | 1.41 | 1.29 | 1.06 | 0.80 |
| 40 代 | 1.54 | 1.51 | 1.06 | 0.86 |

表 3.9　ある年代 100 人とそれ以外の年代 100 人の句読点の生起確率による推定の再現率

| | 10 代 | 20 代 | 30 代 | 40 代 |
|---|---|---|---|---|
| 再現率 [%] | 73.0 | 55.9 | 59.2 | 66.0 |

## 3.5　手法 2 と手法 3 の実験結果の比較

各属性の句読点の平均使用件数 (手法 2) と平均生起確率 (手法 3) から，年代毎に偏差値を求め，表 3.10 に示す．

表 3.10　句読点の平均使用件数と平均生起確率の偏差値

| 年代 \ 偏差値 | 使用件数 (平均) | | 生起確率 (平均) | |
|---|---|---|---|---|
| | 。 | 、 | 。 | 、 |
| 10 代 | 35.1 | 36.5 | 34.0 | 36.9 |
| 20 代 | 47.0 | 45.6 | 49.2 | 45.1 |
| 30 代 | 56.3 | 54.4 | 56.9 | 54.2 |
| 40 代 | 61.4 | 63.4 | 59.7 | 63.7 |

表 3.4 と表 3.9 より，10 代の推定の再現率においては，手法 3 が手法 2 よりも 2.4 %高い．一方で，20〜40 代の 3 年代の推定の再現率においては，手法 3 が手法 2 よりも低い．

# 第 4 章

# 考察

## 4.1　手法 1 の仮説「各属性で最も差が出る単語は，句読点である」

句読点の 2 単語の重要度は，表 3.5 より，2019 年度と 2020 年度の両年で 3 位内の順位という結果であった．この結果は，手法 1 の仮説を支持するものである．

手法 1 の実験結果の問題点として，重要度の値が上位の単語でも低い事が挙げられる．例えば，2020 年度の最も高い重要度は，「。」の 0.84 %であった．これは，説明変数の数が多い事が原因だと考えられる．

## 4.2　手法 2 の仮説「若い年代は，句読点を使わない」

表 3.3 から，年代が上がるにつれて，句読点の使用件数が多くなる傾向がある事が分かる．この結果は，手法 2 の仮説を支持するものである．

10 代の推定の再現率は，表 3.6 より，2019 年度が 76.2 %，2020 年度が 70.6 %であり，他の年代の推定よりも高い結果であった．従って，句読点の使用回数に注目することで，10 代か 20 代以上かの 2 択ならば，高い精度で推定する事ができると言える．

## 4.3　手法 3 の仮説「若い年代は，文章が短い」

表 3.7 より，10 代の 1 ツイート当たりの文字数が，他の年代よりも短い結果となった．この結果より，手法 3 の仮説は支持された．

10 代の句読点の生起確率は，表 3.8 と表 3.10 より，句読点の使用件数と同様に，他の年代よりも顕著に低い結果であった．

表 3.9 より，10 代の推定の再現率は 73.0 %であり，他の年代の推定よりも高い結果であった．更に，これは表 3.4 の手法 2 の再現率よりも 2.4 %高い．よって，10 代の推定には，句読点の使用回数よりも，句読点を使用する割合に注目する事が有効であると言える．

# 第 5 章

# おわりに

本研究では，従来の実験ではストップワードに指定される事が多い句読点に注目し，性別や年代等の属性推定を行った．その結果，句読点の使用回数によって，10 代か 20 代以上かの推定を行ったところ，再現率が 70.6 %であり，他年代の推定の再現率より 10 %ほど高いという結果が出た．

さらに，文章の長さを考慮し，句読点を使用する割合に注目した年代の推定も行った．その結果，「若い年代は，句読点を使用する頻度が低い」という特徴が示された．また，10 代か 20 代以上かの推定においては，句読点を使用する割合に注目する事が有効である事が分かった．

## 謝辞

本研究を行うにあたり，多くの方より御指導いただきました．特に，多大なる御指導を受け賜りました，指導教官である明治大学総合数理学部先端メディアサイエンス学科の菊池浩明教授に深く感謝申し上げます．また，研究の実験に協力して下さった菊池研究室の松本寛輝さん, 伊藤充司君, 研究室の皆様に深く感謝の意を表するとともに，謝辞とさせていただきます．

# 参考文献


[1] 長浜祐貴，遠藤聡志，當間愛晃，赤嶺有平，山田考治，“Twitter の投稿文章による人物像の推定”，2012 年度 JSiSE 学生研究発表会，2013.

[2] 岩朝史展，松本和幸，吉田稔，北研二，“Twitter ユーザの属性別感情推定の検討”，言語処理学会 第 22 回年次大会 発表論文集，pp.389-392，2016.

[3] 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，pp.3 _431-3 _432， 2020.

[4] s21g Inc.，“ツイプロ”，(https://twpro.jp/，2021 年 1 月参照)

[5] “sklearn.ensemble.RandomForestClassifier — scikit-learn 0.24.1 documentation”，(https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html，2021 年 1 月参照)