# 絵文字を用いた感性情報の推定

05001401　中央大学　　*津村　大悟　TSUMURA Daigo
　　日本 IBM　　北嶋　弓月　KITAJIMA Yuzuki
05000907　東海大学　　大竹　恒平　OTAKE Kohei
01405390　中央大学　　生田目　崇　NAMATAME Takashi

## 1. はじめに

近年，携帯電話端末の爆発的な普及により SNS などのインターネットを介した非対面のコミュニケーションが普及している.

このような非対面形式では, 表情や声の調子, 身振りなどの非言語情報がないため，対面コミュニケーションに比べて，微妙な感情伝達が難しい場合も多い. 日本で生まれた絵文字は, こうした感情を表現するのに広く使われるようになったが，種類が数千となり，感情がわかりづらいものも少なくない.

そこで本分析では, Q&A サイトに投稿された文章を対象に，文章中に使われる絵文字の感情推定を行い，文章をどの程度感情を推定できるのかについて検証する.

## 2. データ概要

今回の分析では，乳幼児を持つ親向けポータルサイトの質問データと回答データを分析に用いる.データは経営科学系研究部会連合協議会主催令和 4 年度データ解析コンペティションによって提供された, コネヒト株式会社が運営する情報サイトママリのデータを用いる.

なお本分析では，質問データと回答データ 3 年分のうち，2019 年に該当するデータを使用した．データの行数は約 1388 万行.質問データ,回答データの絵文字総数はそれぞれ約 2203 万件,約 610 万件であった.

## 3. 分析手法

本研究では, 文章の感情分析, 絵文字のクラスタリング，二項ロジスティック回帰分析による感情の推定を行った. 分析には, 使用頻度の高い上位 100 個の絵文字が含まれている文章を抽出し用いた.

感情分析では，asari [1]を用いてポジティブ，ネガティブのポジネガ値，sentimentja [2]を用いて幸福，悲しみ，怒り，嫌悪，驚き，恐れの 6 感情値を文章ごとに算出した. ここで, 特定の絵文字が使われている文章全体のポジネガ値, 6 感情値の平均値をそれぞれ算出し，それを特定の絵文字の持つ感情値とした.

絵文字のクラスタリングでは，絵文字ごとの 6 感情値を K-means [3]によるクラスタリングを行った.

感情の推定では，クラスタリングに用いた 100 文章を抽出し，ポジネガのラベル付けを行ったものを使用した. ラベル付けは 4 人で行い, 回答が異なるものに関しては多数決とした．絵文字を用いてどの程度文章の感情を推定できるのか検証するため，絵文字とテキストの感情値を用いて二項ロジスティック回帰分析を行い精度評価を行った. また, 説明変数に絵文字とテキストの感情値を用いたものと，絵文字の感情値のみを用いたもので比較を行う.

## 4. 分析結果と考察

asari による全文章のポジネガの平均値を以下に示す.表 1 の通り，全体にポジティブと判定される文章が多いことがわかる.

表 1　全文章のポジネガ値平均

| | 質問データ | 回答データ |
|---|---|---|
| ポジティブ | 0.713 | 0.866 |
| ネガティブ | 0.287 | 0.135 |

sentimentja による 6 感情値の平均値は以下に示す.

表 2　全文書の感情値平均

| 感情 | 質問データ | 回答データ |
|---|---|---|
| 幸福 | 0.398 | 0.571 |
| 悲しみ | 0.605 | 0.565 |
| 怒り | 0.303 | 0.265 |
| 嫌悪 | 0.278 | 0.235 |
| 驚き | 0.535 | 0.549 |
| 恐れ | 0.562 | 0.468 |

全文書の感情値平均より，幸福, 驚きは回答データの方が平均値が高く，悲しみ，怒り，嫌悪，恐れは質問データの方が、平均値が高いことがわかる.

また, K-means によるクラスタリング結果を以下に示す.

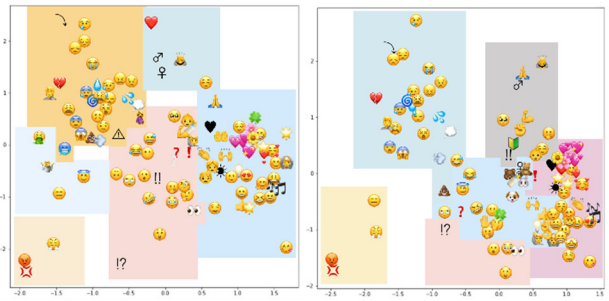

図 1　絵文字のクラスタリング
（左：質問データ，右：回答データ）

横軸は正の方向に笑顔の顔文字などが増加していき，負の方向に涙や怒りの絵文字が位置している縦軸は正の方向に頻出順位が高い絵文字，負の方向に頻出順位が低い絵文字が位置している．このことより、横軸はポジネガ，縦軸は使いやすさを表していると考えられる．

ロジスティック回帰分析による結果を以下に示す．

表 3　混合行列（絵文字＋テキスト）

| | | 予測結果 | |
|---|---|---|---|
| | | ネガティブ | ポジティブ |
| 正解のデータ | ネガティブ | 5 | 3 |
| | ポジティブ | 0 | 12 |

表 4 精度評価（絵文字＋テキスト）

| | |
|---|---|
| 正解率 | 0.85 |
| 適合率 | 0.80 |
| 再現率 | 1.00 |
| F1 スコア | 0.89 |

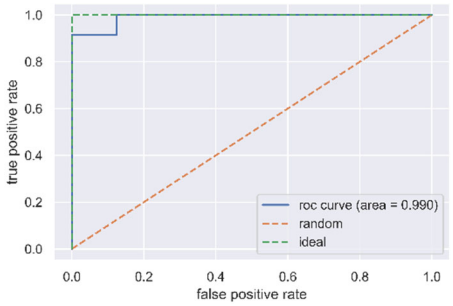


図 2　ROC 曲線（絵文字＋テキスト）

結果から，ネガティブなテキストを予測することが難しいということがわかる．

表 5　混合行列（絵文字）

| | | 予測結果 | |
|---|---|---|---|
| | | ネガティブ | ポジティブ |
| 正解のデータ | ネガティブ | 3 | 5 |
| | ポジティブ | 1 | 11 |

表 6 精度評価（絵文字）

| | |
|---|---|
| 正解率 | 0.70 |
| 適合率 | 0.69 |
| 再現率 | 0.92 |
| F1 スコア | 0.79 |

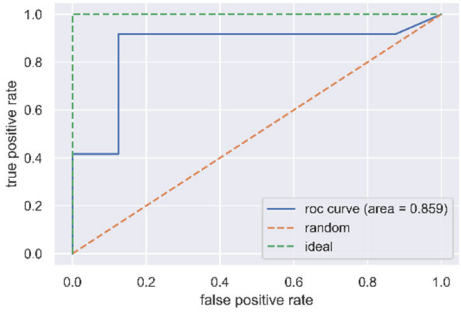


図 3　ROC 曲線（絵文字）

ネガティブなテキストにおいては半数以上の予測が間違っている．ただし，精度評価をみると全体としては比較的に高い精度で予測できていることがわかる．

## 5.　おわりに

本分析では，質問データと回答データの絵文字を抽出し，その絵文字が使われているテキストに関して、感情分析を行った．

本分析では，２つのライブラリを用いて，ポジネガと 6 感情の感情値を算出し，質問データと回答データで絵文字の使われ方の比較を行った．感情値を用いた絵文字のクラスタリングでは，特定の感情と共に使われる絵文字を相当数特定することができた．また，感情推定の検証を二項ロジスティック回帰分析で行い，絵文字とテキストの感情値を説明変数することで，高い精度で予測することができた．

## 参考文献

[1] asari, https://github.com/Hironsan/asari
[2] sentimentja, https://github.com/otodn/sentiment_ja_js
[3] J. MacQeeen, “Some methods for classification and analysis of multivariate observation,” *Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281-297, 1967