

顔文字の役割を利用したツイートの感情極性推定

池川 知里[†] 新妻 弘崇^{††} 太田 学^{†††}

[†] 岡山大学工学部情報工学科 〒700-8530 岡山県岡山市北区津島中 3-1-1

^{††}, ^{†††} 岡山大学大学院自然科学研究科 〒700-8530 岡山県岡山市北区津島中 3-1-1

E-mail: [†], ^{†††}{ikegawa, ohta}@de.cs.okayama-u.ac.jp, ^{††}niitsuma@suri.cs.okayama-u.ac.jp

あらまし Twitter のようなテキストによる発話では、非言語情報を伝えることが困難であり、話し手・聞き手間で感情的なトラブルが生じやすい。よって、テキストの感情を推定し、そのようなトラブルの原因となる発話の検出に利用することができれば有用である。本研究では、ツイートのようなテキストによる発話において、非言語情報を表現する手段の一つである顔文字に注目し、それらがテキスト中に与える効果を分析する。そして、分析により得られた顔文字の役割を利用して、ツイート文の感情極性を推定する一手法を提案する。結果として、言語情報や顔文字の感情極性のみを利用するよりも、顔文字の役割を利用した方が、多くのツイート文の感情極性を正しく推定することができた。

キーワード 顔文字, 感情推定, Twitter, ツイート

Tweet sentiment polarity estimation using roles of emoticons

Chisato IKEGAWA[†], Hirotaka NIITSUMA^{††}, and Manabu OHTA^{†††}

[†] Department of Information Technology, Faculty of Engineering, Okayama University

3-1-1 Tsushima-naka, Kita-ku, Okayama, 700-8530 Japan

^{††} Graduate School of Natural Science and Technology, Okayama University

3-1-1 Tsushima-naka, Kita-ku, Okayama, 700-8530 Japan

E-mail: [†], ^{†††}{ikegawa, ohta}@de.cs.okayama-u.ac.jp, ^{††}niitsuma@suri.cs.okayama-u.ac.jp

Key words Emoticon, Sentiment Estimation, Twitter, Tweet

1 はじめに

近年、コミュニケーションツールとしてのソーシャルメディアの利用が広く普及している。とりわけ、簡単かつ気軽に情報を発信して交流することができる Twitter を利用するユーザは多い。しかし、Twitter のようなテキストによる発話では、対面コミュニケーションと比較して、話し手・聞き手間の感情的なトラブルが生じやすいという問題がある。これは、対面コミュニケーションにおいて、聞き手が話し手の感情を推測する手がかりとなる表情・声のトーン・ジェスチャーといった非言語情報を、テキストによる発話では伝えることが困難なためである。そのようなトラブルを回避するために、テキストの感情を推定し、トラブルの原因となる発話を検出することができれば有用である。

そこで、本研究では、Twitter におけるユーザの書き込みであるツイートの感情極性を推定するための一手法を提案する。まず、実際にユーザが投稿しているツイートを収集し、言語情報の感情極性と、顔文字の感情極性の組み合わせが、ツイート

文の感情極性に与える効果を分析することで、顔文字の働きについて検討する。ここで顔文字とは、文字や記号を組み合わせで顔になぞらえた文字列のことであり、文章の感情的側面を補足し、非言語情報を伝達する手段の一つとして用いられる。また、言語情報とは、ツイート文の顔文字以外のテキスト部分のことであり、ツイート文とは、ツイートを顔文字が付与されている文単位で切り出すことによって得られた文章のことであり、そして、分析の結果から得られた顔文字の働きを顔文字の役割と呼び、それを利用してツイート文の感情極性の推定を試みる。

本稿の構成は以下の通りである。2 節で関連研究について述べ、3 節で本研究に用いる感情分類体系について説明し、4 節で顔文字の役割を提案する。そして、5 節で顔文字と言語情報の感情極性の推定方法と、顔文字の役割を利用したツイート文の感情極性の推定方法を説明し、6 節でツイート文の感情極性を推定する。最後に 7 節でまとめと今後の課題について述べる。

2 関連研究

文章の感情推定に関する研究には、次のようなものもある。

菅原ら [1] は、Izard が定義した 9 種類（喜び・怒り・悲しみ・嫌悪・恐怖・罪・恥・興味・驚き）の感情からなる感情分類体系を用いて、テキストの感情抽出に取り組んだ。感情の強度推定では、語レベル解析、句・文レベルの解析を順に行った。まず、語レベル解析では、構文解析されたテキストに対して、あらかじめ用意した感情関係語データベースに基づき、感情スコアを割り当てる。そして、時制・人称・副詞・接続助詞等の文構造を考慮した句レベル・文レベルの感情スコアを計算し、それらを踏まえて文章全体の感情とそのスコアを示した。

堀宮ら [2] は、ツイートの感情によって、それに対する別ユーザーからの返信であるリプライの言語表現にパターンが存在するという仮説を立て、発言に対する他者反応を利用したツイートの感情推定手法を提案した。リプライの付いたツイートの感情推定において、リプライを考慮することで、ツイートのみからでは推定することが困難な感情推定が可能になると述べている。

近年では、顔文字といった非言語情報を、文章の感情推定へ利用することに、目が向けられ始めている。そのため、文章から顔文字の抽出を行う研究も盛んに行われており、その手法には、顔文字の目・口・輪郭の情報を用いるもの [3] や、顔文字のパターンを正規表現で定義するもの [4] などがある。汎用の Unicode の文字プロパティと Unicode ブロックの情報を用いて、多彩な顔文字を抽出できる手法も提案されている [5]。さらに、篠山ら [6] や若井ら [7] は、顔文字の感情推定を利用し、言語情報も踏まえた文章の感情推定手法を提案している。篠山らは、文章に顔文字が付与されることで感情がどのように変化するかを調査するために、アンケートを実施し、その結果を考察した。そして、言語情報と顔文字をそれぞれ感情推定することで、顔文字を考慮した感情推定を実現した。感情分類体系には、14 種類の感情（喜び・楽しみ・期待・悲しみ・驚き・感謝・平静・不安・好き・安心・憧れ・怒り・焦り・嫌悪）を用いている。若井らは、顔文字の役割には“強調”，“自嘲”，“弛緩”という 3 種類の役割があるとした。そして、役割の特徴を抽出し、役割ごとに重みを付け、文章の感情値と重みの積からツイート全体の感情値を算出した。感情分類体系には、熊本らが定めた 6 種類の感情（楽しい・悲しい・怒り・うれしい・緊迫・のどか）を用いている。

3 感情分類体系

感情推定に関する研究には、Ekman の基本感情 6 種類（幸福感・驚き・恐れ・悲しみ・怒り・嫌悪）を利用するものを始め、Plutchik の基本感情 8 種類（喜び・悲しみ・受容・嫌悪・恐れ・怒り・驚き・期待）を利用するもの、感情表現辞典 [8] を基にした 10 種類の感情（喜・怒・哀・怖・恥・好・厭・昂・安・驚）を利用するものなどがある。

しかし、感情を細かく定義すると、類似する感情を分類することが困難となる。そこで、本研究では、感情表現辞典 [8] で定義されている 10 種類の感情に“無感情”カテゴリを加えた 11 種類を、“positive”，“negative”，“neutral”の 3 種類に分類したものを、感情分類体系とする。この感情分類体系では、“喜”，“好”，“安”を“positive”，“怒”，“哀”，“怖”，“恥”，

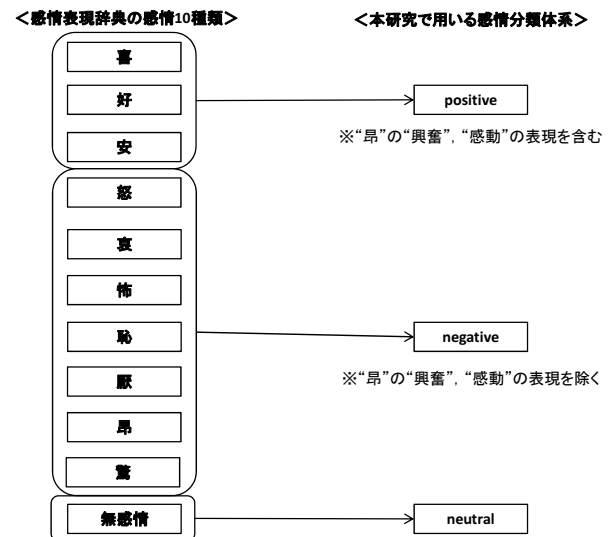


図 1 感情表現辞典で定義された感情分類体系と本研究で用いる感情分類体系の対応関係

表 1 若井らが用いた顔文字の役割の働きと特徴

役割	働きと特徴
強調	・ 顔文字が文の意味をより強める ・ 肯定的な文面で、文の感情と顔文字の感情が同じ
自嘲	・ 顔文字が自分に呆れて笑うさま、さげすんであざ笑うさまを示す ・ 文中に否定形があり、顔文字の感情が肯定的
弛緩	・ 否定的な文面に対して、顔文字が文の意味を和らげる ・ 文の感情と顔文字の感情が反対

“厭”，“昂”，“驚”を“negative”，“無感情”を“neutral”に統合している。ただし、例外として，“昂”カテゴリに属している“興奮”，“感動”の表現は，“positive”に分類する。図 1 に、感情分類体系の詳細を示す。このように、複数の感情を一つの感情極性にまとめることで、感情解釈の幅の広さにある程度対応することが可能となる。

4 顔文字の役割

4.1 既存研究で用いられた顔文字の役割

若井ら [7] は、顔文字の役割には“強調”，“自嘲”，“弛緩”の 3 種類があるとした。それぞれの役割の働きと特徴については、表 1 のようなものが挙げられている。

本研究では、実際にユーザーが投稿しているツイートを収集し、言語情報の感情極性と、顔文字の感情極性の組み合わせが、ツイート文の感情極性に与える効果を分析することで、顔文字の役割についてさらに詳しく検討する。

4.2 提案する顔文字の役割

本研究では、実験対象として、Twitter から“卒業”というクエリで、2014 年 1 月 26 日のツイートを 1,500 件抽出した。抽出には、Twitter の非公式ライブラリである Twitter4j [9] を利用した。その 1,500 件のツイートのうち、顔文字が付与されていたものは 396 件あり、出現した顔文字の総数は 479 個であった。それから得られた 479 文のツイート文について、言語

表 2 言語情報の感情極性・顔文字の感情極性・ツイート文の感情極性の分析結果

言語 + 顔文字\ツイート文	positive	negative	neutral
positive + positive	44	0	0
positive + negative	23	1	0
positive + neutral	0	0	0
negative + positive	1	11	0
negative + negative	0	40	0
negative + neutral	0	2	0
neutral + positive	155	18	4
neutral + negative	16	156	4
neutral + neutral	0	0	4

表 3 提案する顔文字の役割

言語 + 顔文字\ツイート文	positive	negative	neutral
positive + positive	強意	-	-
positive + negative	感動表現	-	-
positive + neutral	-	-	-
negative + positive	-	弛緩	-
negative + negative	-	強意	-
negative + neutral	-	-	-
neutral + positive	感情付与	自嘲	装飾
neutral + negative	願望	感情付与	装飾
neutral + neutral	-	-	装飾

情報の感情極性・顔文字の感情極性・ツイート文の感情極性を人手で分析した。表 2 に分析結果を示す。なお、この表の一番左側の列の左側は言語情報の感情極性、右側は顔文字の感情極性、一番上の行はツイート文の感情極性を表す。

表 2 の結果をもとに、言語情報の感情極性と、顔文字の感情極性の組み合わせが、ツイート文の感情極性に与える効果について検討した。その結果を、顔文字の役割の観点からまとめて、表 3 に示す通り提案する。それぞれの顔文字の役割の具体例を表 4 に示す。なお、この表の見方は、表 2 のそれと同様である。

5 感情極性推定

本節では、顔文字と言語情報の感情極性の推定方法について説明する。

5.1 言語情報の感情極性の推定

感情語辞書を利用した、言語情報の感情極性の推定方法について述べる。

まず、感情表現辞典 [8] に収録されている感情語を参考に、“positive”と“negative”を表す 2 次元ベクトルの感情極性値を割り当てた感情語辞書を作成した。感情極性の強度は、0.0～1.0 のスコアで表している。

そして、lucene-gosen [11] を用いて、言語情報を形態素解析し、その結果を作成した感情語辞書と比較することで、与えられた言語情報の中に、感情表現が含まれているかどうかを確認する。言語情報中に、感情表現が含まれている場合は、その語に割り当てられている 2 次元ベクトルの感情極性値を、その言語情報の感情極性値として加算していく。

最終的に、2 次元ベクトルの感情極性値の中で、より高い値

表 4 顔文字の役割の具体例

役割	効果	具体例
強意	顔文字が文の意味をより強める	きこちゃん好き (^ _ ^)
		本当に卒業しちゃうの寂しいです.... (' ・ ω ・ ')
感動表現	泣くほど嬉しいという感動の気持ちを表す	ほんと、優しい後輩で嬉しいよ (' ; ω ; ')
弛緩	否定的な文面に対して、顔文字が文の意味を和らげる	学校やだ n ^ ω ^ n
感情付与	直接的な感情表現のない文章に感情情報を付与する	サケパなう (≧▽≦)
		段々うちらも卒業がちかづいてるね (' ・ _ ・ ')
願望	自分の希望や、相手への応援の気持ちを表す	早く卒業して東京行きたい (* _ *)
		ふぁいと !! (` ω ´)
自嘲	顔文字が自分に呆れて笑うさま、さげすんであざ笑うさまを示す	基本法律科目ニガテすぎて埋まんけん多分いつまで経っても卒業できんなー (* ^ o ^ *)
装飾	疑問文の末尾に付与される付随情報	皆さん進学? ^ ^

となった成分が示す感情極性を、その言語情報が示す感情極性とする。なお、“positive”成分と“negative”成分の値が等しくなった場合、その言語情報の感情極性は“neutral”とする。

5.2 顔文字の感情極性の推定

本研究では、三好ら [4] が提案した手法を用いて、“positive”と“negative”の 2 種類の感情極性に顔文字を分類する。三好らは、正規表現を用いて顔文字を抽出し、抽出した顔文字を構成する各記号の感情別出現率の合計値から、顔文字の感情を推定した。ここで、記号の感情別出現率とは、顔文字に用いられている単一記号や接続する複数記号が、感情ごとに、顔文字の集合内でどの程度使われているか数え上げ、出現した割合を算出したものである。

5.2.1 顔文字の抽出

正規表現で定義した顔文字のパターンにマッチする文字列を、顔文字として抽出する。顔文字のパターンは、以下の 4 つの条件を満たすものと定義する [4]。

- (1) 半角または全角の丸括弧で両端がくくられている
- (2) 括弧内では、英数字・ひらがな・カタカナ・漢字が連続して 3 文字以上続かない
- (3) 2 文字以上の文字列である
- (4) 括弧のすぐ外側の英字等は顔文字に含める

ここで、三好らの顔文字を抽出する手法の性能を確認するための予備実験として、顔文字図書館 [10] から三好らの方法で顔文字を抽出した。なお、顔文字図書館は、シンプルな顔文字から複雑な顔文字までがジャンル別に紹介されているサイトである。抽出対象の顔文字 1,286 個のうち、87.3% の 1,123 個を正抽出することができた。ただし、顔文字の片端または両端が欠落した状態で抽出されたものも顔文字の感情極性の推定に利

表 5 予備実験の感情極性の推定結果

	正解率 (%)
単一記号	80.2
接続する 2 記号	81.3
接続する 3 記号	65.4

用できると考え、本実験では、それらも抽出できたとみなしている。

5.2.2 記号の感情極性別出現率を利用した顔文字の感情極性の推定

顔文字を構成する各記号の感情極性別出現率から、顔文字の感情極性を推定するために、顔文字を構成する各記号の感情極性別出現率を計算する。ただし、空白文字や丸括弧のみからなる記号の組み合わせの出現率は考えない。そして、顔文字を構成する単一記号や複数記号の感情極性別出現率を感情極性ごとに合計し、合計した感情極性別出現率が最も高い感情極性をその顔文字の感情極性とする。

ここで、5.2.1 節の予備実験で抽出した顔文字 1,123 個を対象として、5 分割交差検定により、顔文字の感情極性を推定し、正解率を算出した。1,123 個の顔文字データは、感情極性が“positive”であるもの 500 個、“negative”であるもの 623 個から構成されている。単一記号の感情極性別出現率、接続する 2 記号の感情極性別出現率、接続する 3 記号の感情極性別出現率を用いた感情極性の推定結果を表 5 に示す。なお、ここでの“正解率”とは、感情極性の推定対象である顔文字のうち、システムがその顔文字の感情極性を正しく推定することができたものの割合である。この表から、単一記号の感情極性別出現率を用いて感情極性を推定する場合と、接続する 2 記号の感情極性別出現率を用いて感情極性を推定する場合の正解率が高いことが分かる。ゆえに、本研究では、単一記号の感情極性別出現率と接続する 2 記号の感情極性別出現率を用いて、顔文字の感情極性を推定する。

5.3 顔文字の役割を利用したツイート文の感情極性の推定

4.2 節で提案した顔文字の役割を利用した、ツイート文の感情極性を推定方法について述べる。

まず、ツイートを収集し、言語情報と顔文字に分割する。そして、それぞれについて感情極性を推定し、言語情報の感情極性と顔文字の感情極性の組み合わせから、顔文字の役割のいずれが用いられているかを判断し、利用することで、ツイート文の感情極性を推定する。なお、本研究では、顔文字の役割のうち、“強意”、“感情表現”、“弛緩”の 3 種類を用いた。図 2 にこの手順を示す。

6 ツイート文の感情極性の推定実験

6.1 言語情報の感情極性とツイート文の感情極性

5.1 節で説明した方法により、言語情報の感情極性の推定を試みた。言語情報の感情極性を推定する実験を行った。実験データは、4.2 節で用いたツイート文 479 文である。479 文のツイート文のうち、言語情報の感情極性が“positive”のものは 68 文、“negative”のものは 54 文、“neutral”のものは 357

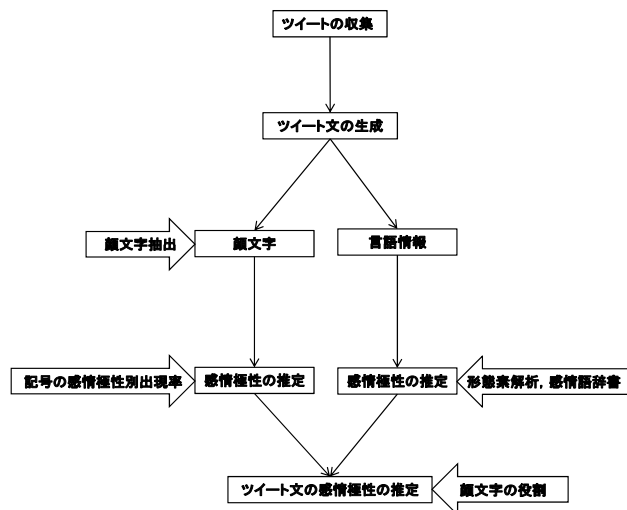


図 2 ツイート文の感情極性の推定手順

表 6 言語情報の感情極性の推定結果

	positive	negative	neutral
positive	16/68	0/68	52/68
negative	1/54	21/54	32/54
neutral	4/357	3/357	350/357

表 7 システムが言語情報の感情極性を正しく推定することができた割合

(正解数/分類数)	正解率 (%)
387/479	80.8

表 8 言語情報の感情極性がツイート文の感情極性と一致したものの割合

(正解数/分類数)	正解率 (%)
49/479	10.2

文あった。表 6 に言語情報の感情極性の推定結果を示す。また、表 7 にシステムが言語情報の感情極性を正しく推定することができた割合を示し、表 8 にシステムが言語情報の感情極性を正しく推定することができたもののうち、言語情報の感情極性がツイート文の感情極性と一致したものの割合を示す。

6.2 顔文字の感情極性とツイート文の感情極性

5.1 節で説明した方法により、顔文字の感情極性を推定する実験を行った。実験データは、6.1 節で用いたものと同様である。

まず、実験データのツイート文 479 文から、5.2.1 節の手法を用いて顔文字を抽出した。抽出対象の顔文字 479 個のうち、91.0%の 436 個を抽出することができた。ただし、5.2.1 節と同様に、顔文字の片端または両端が欠落した状態で抽出されたものも顔文字の感情極性の推定に利用できると考え、それらも抽出できたとみなしている。抽出することができなかった 43 個の顔文字について、5.2.1 節で示した顔文字が満たす 4 つの条件のうち、(1) を満たしていないものは 35 個、(2) を満たしていないものは 10 個あった。なお、抽出することができなかった顔文字の中には、複数の条件を満たしていないものも含まれる。抽出することができた顔文字 436 個のうち、顔文字の感情

表 9 抽出することができた顔文字の具体例

感情極性	具体例
positive	(^ _ ^), (≥ ∇ ≤), (* ' ∇ ' *)
negative	(' ; ω ; '), (' · _ · '), Σ (' □ ' ;)
neutral	(6 B 6), (· ∞ ·)

表 10 顔文字の感情極性の推定結果 (単一記号)

正解\推定	positive	negative	neutral
positive	165/214	45/214	4/214
negative	22/216	191/216	3/216
neutral	3/6	2/6	1/6

表 11 顔文字の感情極性の推定結果 (接続 2 記号)

正解\推定	positive	negative	neutral
positive	175/214	27/214	12/214
negative	98/216	105/216	13/216
neutral	1/6	2/6	3/6

表 12 システムが顔文字の感情極性を正しく推定することができた割合

	(正解数/分類数)	正解率 (%)
単一記号	357/436	81.9
接続 2 記号	283/436	64.9

表 13 顔文字の感情極性がツイート文の感情極性と一致したものの割合

	(正解数/分類数)	正解率 (%)
単一記号	299/436	68.6
接続 2 記号	245/436	56.2

極性が“positive”のものは 214 個，“negative”のものは 216 個，“neutral”のものは 6 個あった。抽出することができた顔文字の具体例を表 9 に示す。

そして、5.2.2 節の予備実験で用いた、顔文字図書館から抽出した 1,123 個の顔文字より各記号の感情極性別出現率を算出し、単一記号の感情極性別出現率と接続する 2 記号の感情極性別出現率を用いて、感情極性の推定対象である顔文字 436 個の感情極性の推定を試みた。表 10 に単一記号の感情極性別出現率を用いた顔文字の感情極性の推定結果を、表 11 に接続する 2 記号の感情極性別出現率を用いた顔文字の感情極性の推定結果を示す。また、表 12 に、システムが顔文字の感情極性を正しく推定することができた割合を示し、表 13 に、システムが顔文字の感情極性を正しく推定することができたもののうち、顔文字の感情極性がツイート文の感情極性と一致したものの割合を示す。なお、表 12、表 13 の“単一記号”、“接続 2 記号”とは、それぞれ、単一記号、接続する 2 記号の感情極性別出現率を用いたことを表す。

6.3 顔文字の役割とツイート文の感情極性

6.1 節の表 8、6.2 節の表 13 の結果から、言語情報の感情極性や顔文字の感情極性だけでは、ツイート文の感情極性を推定するには不十分である。

そこで、5.3 節で説明した方法により、ツイート文の感情極性を推定する実験を行った。実験データは、6.1 節で用いたツ

表 14 顔文字の役割を利用したツイート文の感情極性の推定結果

正解\推定	(正解数/分類数)	正解率 (%)
単一記号	302/436	69.3
接続 2 記号	247/436	56.7

weet 479 文のうち、6.2 節で顔文字を抽出することができた 436 文のツイート文である。この感情極性の推定結果を表 10 に示す。なお、表 14 の“単一記号”、“接続 2 記号”は、表 12、表 13 のそれと同様である。言語情報の感情極性や顔文字の感情極性のみを利用してツイート文の感情極性を推定した場合と比較すると、正解率の向上が見られた。

7 まとめと今後の課題

本研究では、ツイート文の感情極性を推定するために、テキストによる発話において、非言語情報を表現する手段の一つである顔文字に注目した。まず、実際にユーザが投稿しているツイートを収集し、言語情報の感情極性と、顔文字の感情極性の組み合わせが、ツイート文の感情極性に与える効果を分析することで、既存研究で提案された顔文字の役割についてさらに詳しく検討し、“強意”、“感動表現”、“弛緩”、“感情付与”、“願望”、“自嘲”、“装飾”の 7 種類の役割を提案した。言語情報の感情極性や顔文字の感情極性だけでは、ツイートの感情極性を推定するには不十分であったため、提案した顔文字の役割のうち“強意”、“感動表現”、“弛緩”の 3 種類をあわせて利用してツイートの感情極性を推定した。その結果、言語情報の感情極性や顔文字の感情極性のみを利用した場合と比較して、顔文字の役割も利用した場合の方が、多くのツイート文の感情極性を正しく推定することができた。

今後の課題としては、ツイートの感情極性の推定を正確に行うために、顔文字の役割のうち“感情付与”、“願望”、“自嘲”、“装飾”の 4 種類の利用方法の検討や、顔文字抽出の正解率の向上が挙げられる。また、感情極性ではなく、個々の感情を推定する方法についても検討の必要がある。

文 献

- [1] 菅原久嗣, アレナ ネピアロスカヤ, 石塚満, “日本語テキストからの感情抽出”, The 23rd Annual Conference of the Japanese Society for Artificial Intelligence, pp. 1-2, 2009.
- [2] 堀宮ありさ, 坂野遼平, 佐藤晴彦, 小山聡, 栗原正仁, 沼澤政信, “Twitter における発話者へのリプライを用いたユーザ感情推定手法”, DEIM Forum 2012, F2-1, 2013.
- [3] 山下諒, 谷謙治, 高見一正, “携帯メールの絵文字・顔文字解析による気分推定法と楽曲推薦法”, 情報処理学会研究報告 AVM, Vol.2008, No.69, pp. 51-56, 2008.
- [4] 三好辰明, 太田学, “ツイートに出現する顔文字等の文字と記号に着目した感情分類”, DEIM Forum 2013, D9-2, 2013.
- [5] 風間一洋, 榊剛史, 鳥海不二夫, 篠田孝祐, 栗原聡, 野田五十樹, “顔文字に着目したツイートの感情変化の分析”, WebDB Forum 2013, A2-3, 2013.

- [6] 篠山学, 松尾明子, “顔文字を考慮した対話テキストの感情推定に関する研究”, 香川高等専門学校紀要, 1, pp. 151-153, 2010.
- [7] 若井祐樹, 熊本忠彦, 瀬本明代, “ツイートの感情抽出の為の顔文字の役割分類”, 第2回 Web インテリジェンスとインタラクシオン研究会 (ARG SIG-WI2), 2013.
- [8] 中村明: 感情表現辞典, 東京堂出版, 1993.
- [9] Twitter4j: <http://twitter4j.org/ja/index.html>
- [10] 顔文字図書館: <http://www.kaomoji.com/kao/text/>
- [11] lucene-gosen: <https://code.google.com/p/lucene-gosen/>