E-035

マイクロブログを用いた感情表現収集
Automatic Collection of Emotional Expressions using Microblogs

水岡 良彰† 鈴木 優†
Yoshiaki Mizuoka Masaru Suzuki

1. はじめに

近年，自分の考えや感想などを Twitter[1]に代表されるマイクロブログに投稿する人が増加している．投稿されたメッセージには投稿者の感情を述べたものが含まれることから，メッセージが持つ感情情報を抽出して口コミ分析などに利用できれば価値が高いと考えられる．感情情報の抽出には手掛かりとなる表現を用いることが考えられるが，マイクロブログなどの CGM(Consumer Generated Media)では多様な表現が出現するため，人手で網羅することは困難である．

マイクロブログのうち，同時に同じ事象に関して投稿がなされたログでは，複数のユーザが同じ事象に対して同じ感情を持つことが多いため，同じ感情を持つメッセージが時間的近傍に出現しやすいと考えられる．そこで本研究では，そのようなログの特性を利用して，感情表現を収集する手法を提案する．本稿では特に，テレビ番組に関する Twitter ログを対象に，感情表現を収集する手法を提案する．加えて実データを用いた実験から提案手法の有効性を示す．

以下，2 章では人手による感情表現収集を行った結果について考察する．3 章で感情表現かどうかを判定する手法の原理について述べ，4 章で 3 章の原理を利用した提案手法のアルゴリズムを説明する．5 章で実験と考察，6 章でまとめと今後の課題を述べる．

2. 人手による感情表現収集

どのような感情表現がマイクロブログに出現するか調べるため，テレビ番組に関する Twitter ログから感情を表す表現を人手で収集した．なお Twitter に投稿されたそれぞれのメッセージはツイートと呼ばれるため、以下ではメッセージとツイートを区別せずに用いる。

感情表現の収集対象は，表 1 に示す番組の放送時間中の対応するハッシュタグを含むツイートとした．表現の収集対象とする感情は「かっこいい」「かわいい」「泣ける」「笑える」の 4 種類とした．収集手順として，まず対象のツイートから各感情を表す表現を人手で抽出し，そのうちの明確な感情表現を人手で選択した．明確な感情表現とは，その表現単体で感情が判断できる表現とした．収集された表現中には前後の流れを考慮することで感情を判断できる表現も存在したが，今回は感情表現であるかのあいまい性を排除するために明確な感情表現のみを収集した．

収集結果の一部を表 2 に示す．表 2 を見ると，まず代表的な感情表現から派生した言い換え表現があることが分かる．例えば「かっこいい」を表す感情表現には，「かっこいい」から派生した「かっけ～」「ｶｯｺﾖｽ」といった表現や，「かっこいいいいい」といった同じ文字が続く数の違い，「カッコイイ」と「ｶｯｺｲｲ」といった全角と半角の違いがあることが分かる．一方，代表的な感情表現とは表記が異なる感情表現も存在することが分かる．例えば「かっこいい」を表す「イケメン」や「美男子」といった表現や，「泣ける」を表す顔文字などである．

以上のように，感情表現は言い換え表現の幅が広く，また代表的な感情表現と表記が全く異なる表現も存在することが分かった．よって，このように多様な感情表現を収集できる手法が必要といえる．

表 1 人手による感情表現抽出の対象ツイート

放送局	放送日時	番組名	ハッシュタグ	ツイート数
NHK総合	2010/09/26 20:00～20:45	大河ドラマ「龍馬伝」第39回「馬関の奇跡」	#ryomaden	5474
フジテレビ	2010/09/20 21:00～22:00	夏の恋は虹色に輝く 第10回「もう会えない」	#FujiTV	843
日本テレビ	2010/09/20 20:00～21:24	ボクシング ダブル世界タイトルマッチ	#NTV	143

表 2 人手で収集した感情表現（一部）

かっこいい	かわいい	泣ける	笑える
かっけ～	かわいかったー	泣きそう	wwwww
ｶｯｺﾖｽ	可愛すぎ	泣けてきた	笑った
かっこいいいいい	かわえー	悲しいな	ワロタ
カッコイイ	ツンデレ	ウルウル	(笑)
ｶｯｺｲｲ	かぁーいい	(泣	爆笑ww
恰好良すぎ！	かわゆす	(T_T)	おもろいな
カッコよ過ぎ	カワイイ	(´;ω;`)	面白かった
カコイイ	カワエェ	･ﾟ･(ノД`)･ﾟ･	可笑しい
美男子	可愛い	切ない	バリうけ
イケメン	きゃわ！	感動した	おんもろ！

3. 感情表現を判定する手法の原理

ある表現が感情を表すかを判定する方法として，その表現が代表的な感情表現と同一文書内で共起するかどうかを手掛かりとする方法が考えられる．しかし Twitter は投稿文が短いため，一つの投稿内に複数の感情表現が含まれることが少ない．

そこで本研究では，同時に同じ事象に関して投稿されていると考えられるログを利用し，時間的近傍のメッセージを見ることで，感情表現の判定を行うアプローチをとる．時間的近傍の各投稿者がそれぞれ異なった表現を用いている場合には，複数のメッセージを見ることで多様な表現の判定が可能になると期待できる．本稿では，リアルタイムに番組の感想が投稿された，テレビに関する Twitter ログを利用する．

テレビに関する Twitter ログでは，テレビ番組を視聴中の人によってシーンに同期したツイートが多くなされている．複数のテレビ視聴者は基本的にテレビ番組の同じシーンに関して同時に感想を抱くと考えられるが，その感想をツイートするまでにはラグがあるため，同じシーンに関するツイートは時間的近傍にばらつきを持って出現すると考えられる．感情についても同様と考えられる．例えばテレビ番組で笑えるシーンが流れた場合には，笑えるという感

†株式会社東芝 研究開発センター 知識メディアラボラトリー, Knowledge Media Laboratory, Corporate Research and Development Center, Toshiba Corporation.

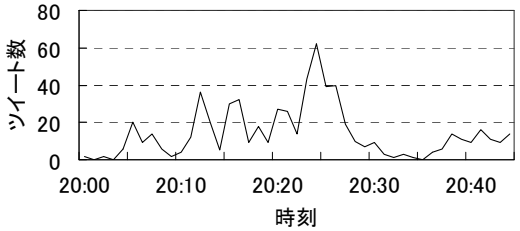

図 1　龍馬伝放送中の「かっこいい」という感情を表すツイート数の推移

表 3　感情表現の出現頻度検定のための 4 分表

	表現Bの時間的近傍	表現Bの時間的近傍外	計
表現Aを含む	*a*	*b*	*a+b*
表現Aを含まない	*c*	*d*	*c+d*
計	*a+c*	*b+d*	*N=a+b+c+d*

情を持つツイートがなされる．図 1 は，表 1 の龍馬伝に関するツイートについて，人手で「かっこいい」という感情を持つと判定されたツイート数の時間的推移をグラフに示したものである．例えばツイート数が特に多い 20:24 分付近では，かっこいいシーンとして盛り上がった戦のシーンが放送されていた．このように同じ感情を持つツイートは，テレビ番組のシーンに同期して時間的近傍に偏って出現する．

感情表現は基本的に投稿者が持つ感情が表現されたツイートに含まれているので，例えば「かっこいい」という感情表現を含むツイートは，「かっけ～」「イケメン」といった同じ感情を表す別の表現を含むツイートと，近い時間に出現しやすいといえる．

感情表現の収集では上記の現象を利用する．すなわち，ある表現が与えられたときにその表現が感情表現であるかを，代表的な感情表現を含むツイートが時間的近傍に偏って出現するかどうかで判定する．このため，本稿で提案する感情表現の収集手法は，近い時間で同じ感情の投稿がなされるものであれば良く，テレビに関するログでなくとも構わない．例えばニコニコ動画[2]のように，動画の特定の再生時間上に投稿するコメントを活用することなどが考えられる．

本手法では出現の仕方に偏りがあるかどうかの判定に，カイ二乗検定の手法を利用する．代表的な感情表現を表現 A，感情表現か判定したい表現を表現 B とすると，表現 B を含むツイートの書き込み時間の時間的近傍とそれ以外で，表現 A を含むツイートの出現頻度が等しいという仮説を帰無仮説としてカイ二乗検定を適用する．そしてカイ二乗値が大きい値を示す，時間的近傍に表現 A が偏って出現する表現 B を感情表現と判定する．

本稿ではカイ二乗値の計算にイエーツの修正[3]を用いる．a を表現 B を含むツイートの時間的近傍に存在する表現 A を含むツイート数，b を表現 B を含むツイートの時間的近傍外に存在する表現 A を含むツイート数，c を表現 B を含むツイートの時間的近傍に存在する表現 A を含まないツイート数，d を表現 B を含むツイートの時間的近傍外に存在する表現 A を含まないツイート数とする．4 分表は表 3 で示され，カイ二乗値 χ^2_{Yates} は次の式(1)で計算される．

$$\chi^2_{Yates} = \frac{N\left(\max\left(0,|ad-bc|-N/2\right)\right)^2}{(a+b)(c+d)(a+c)(b+d)} \qquad (1)$$

なお $N = a+b+c+d$ である．

本来のカイ二乗検定では，カイ二乗値 χ^2_{Yates} によって帰無仮説を棄却するか否かを判断するが，ここではこの値を感情表現か否かを判定するために用いる．さらに，表現 B の時間的近傍に表現 A が偏って出現するか考慮する必要があるため，感情表現の判定では次の条件式(2)を満たす場合に感情表現と判定する．

$$\frac{a}{a+c} > \frac{b}{b+d} \qquad (2)$$

4. 感情表現の収集手法

本章で提案する感情表現の収集手法は，大きく以下の 2 ステップを踏む．

1. 収集元の Twitter ログから，出現頻度が上位の表現を感情表現の候補として収集する．
2. 時間的近傍のツイートを利用して，収集した候補の表現が感情を表すか判定する．

以下 4.1 節ではステップ 1 で利用する Twitter ログの表現分割について説明する．4.2 節で感情表現候補の収集について説明し，4.3 節で感情表現の判定について説明する．

4.1 Twitter ログの表現分割

Twitter ログのような CGM は，必ずしも文法的に正しく記述されているとは限らないため，既存の形態素解析器では適切に単語分割することが難しい．そこで本稿では文字間接続確率による単語分割[4]を利用する．この手法は，ある文字列が単語のようなまとまりを構成している場合，まとまりの前後には様々な文字が出現するという仮定を用いて，前途の文字との接続確率が低い場合に単語分割を行う手法である．すなわち，まとまりの構成文字間の接続確率は相対的に高くなる．なお本手法は単語分割の手法であるが，本稿では表現分割手法として扱う．

本稿では[4]にある方式のうち，3-gram の積を用いる．この方式は計算式(3)が閾値 T 以下のときに $i-1$ 番目と i 番目の部分を分割する．

$$P(C_i \mid C_{i-2}, C_{i-1}) \times P(C_{i-1} \mid C_i, C_{i+1}) \qquad (3)$$

ただし，$P(C_i \mid C_j, C_k)$ は j 番目と k 番目の文字がそれぞれ C_j と C_k の場合に i 番目の文字が C_i となる条件付き確率を表している．

4.2 感情表現候補の収集

感情表現候補の収集では，表現で分割した Twitter ログについて，各表現数をカウントして出現頻度上位を収集する．ただし短い表現はノイズが多いため，本稿では 3 文字以下の表現は対象外とする．3 文字以下の短い感情表現には数に限りがあるため，人手を介した別の方法で感情表現を収集しても良いと考えられる．また，次のステップにおける統計的な手法を用いた感情表現の判定では，ある程度以上の出現頻度が無いと正しい判定が期待できないため，出現頻度の低い表現を感情表現の候補から外すことで計算量を削減している．

感情表現候補の収集の手順は次の通りである．

1. 表現収集元の Twitter ログのテキストから，Twitter 固有の文字列や URL，引用部分を除去する．
2. 4.1 節の手法を用いて，Twitter ログのテキストを表現単位に分割する．
3. 分割した各表現に対して，Twitter 固有の文字列の除去，スペースの除去，NFKC 形式による Unicode 正規化，同じ文字が 3 文字以上続いている場合に 2 文字に短縮を行う．
4. 分割した表現のうち，3 文字未満の表現を除去する．
5. 各表現の出現数をカウントし，出現数が閾値 n 回以上の表現を収集する．

手順 3 において収集した表現に施している処理は，感情に関係無い表現の除去と，簡単な表記揺れに対する対処である．表記揺れをある程度抑えることで，表記揺れによる出現数の低下が抑えられ，次のステップで処理がうまくいく可能性を高めることができる．なお，Twitter 固有の文字列とは，”RT”や”QT”，ハッシュタグ(#～)，ユーザ名(@～)である．

4.3 感情表現の判定

感情表現の候補が，感情表現かどうか判定するアルゴリズムは次の通りである．以下で，代表的な感情表現を表現 A，感情表現か判定したい候補表現を表現 B とする．

1. Twitter ログのテキストを，メッセージ部分と引用部分に分割する．
2. メッセージ部分と引用部分の両方について，Twitter 固有の文字列や URL，スペースを除去し，NFKC 形式による Unicode 正規化を行い，同じ文字が 3 文字以上続いている場合に 2 文字に短縮する．
3. メッセージ部分について表現 B を含むか調べ，このようなツイートの前後 m 分の時間区間を求める
4. 求めた時間区間内と時間区間外のそれぞれについて，メッセージ部分あるいは引用部分に表現 A を含むツイート数と含まないツイート数をそれぞれカウントする．ただし，メッセージ部分の表現 B と被る場合はカウントしない．
5. カウントした数を利用して式(2)の条件を満たす表現 B のうち，式(1)で計算される χ^2_{Yates} が上位のものを感情表現と判定する．

上記のアルゴリズムでは，代表的な感情表現 A を含むかどうかについてはメッセージ部分と引用部分を調べている(手順 4)が，感情表現候補の表現 B を含むかどうかについてはメッセージ部分のみ調べている(手順 3)．これは次の理由による．一般的でない使い方をされた表現 B を含むツイートが多く引用された場合，この一般的でない使い方が高頻度に出現してしまうことになり，誤った判定が起こる可能性がある．感情表現の候補である表現 B は，必ずしも確度の高い感情表現とは限らないため，このような事が起こる可能性が高い．一方，代表的な感情表現である表現 A は確度の高い表現であると想定して良いため，このような事が起こる可能性は低いからである．

手順 4 で，メッセージ部分の表現 B と被る表現 A はカウントしない．これは，時間的近傍を求めるために使用した部分を再度カウントに使用することを防ぐためである．

表 4　感情表現の収集に用いた Twitter ログ

テレビ局	ハッシュタグ	ツイート数
NHK総合	#NHK	711,313
NHK教育	#ETV	89,035
日本テレビ	#NTV	363,158
TBS	#tbs	323,228
フジテレビ	#FujiTV	526,447
テレビ朝日	#tvasahi	400,367
テレビ東京	#tvtokyo	187,976

表 5　収集対象とする感情と対応する代表的な感情表現

感情	代表的な感情表現
かっこいい	かっこいい
かわいい	かわいい
泣ける	泣ける，悲しい
笑える	笑える，面白い

5. 感情表現の収集実験

本章では前章で述べた感情表現の収集手法を用いて，テレビ番組に関する Twitter ログから感情表現を収集する実験を行う．

5.1 実験条件

感情表現の収集に用いる Twitter ログは，表 4 に示すテレビ局の 2010 年 4 月～10 月の 7 ヶ月分のテレビ番組に関する Twitter ログとした．ツイート数は合計で 2,601,524 ツイートである．収集の対象とする感情と，対応する代表的な感情表現は表 5 とした．表現分割に使用する閾値は $T = 0.001$，表現の出現頻度の閾値は $n = 200$ 回，時間的近傍の範囲を決める時間幅は $m = 1$ 分とした．

5.2 実験結果

感情表現として収集された上位 30 件は表 6 の通りとなった．表 6 より，「かっこいい」に対して「かっけぇ」「カッコイイ」といった表記揺れに加えて「イケメン」のように，代表的な感性表現とは表記が異なった表現も収集できていることが分かる．また，「泣ける」に対して「・゜・(ノД`)・゜・。」のように，記号などで構成される顔文字も収集できていることが分かる．

収集した 30 表現の適合率と再現率を計算した結果を表 7 に示す．適合率は，収集した 30 表現のうち，明確に感情を表すと思われる表現の割合とした．再現率は、真の正解集合が分からないので擬似的に 2 章にて人手で収集した感情表現に 4.2 節の手順 3 と 4 を行ったものを正解集合とし、そのうち明確に感情を表すと思われる収集した表現を含むものの割合とした．表より，適合率については極端に低い感情は無いが，再現率については「泣ける」「笑える」が特に低くなっている．

5.3 考察

表 6 の収集された感性表現の結果を観察すると、「かっこいい」という代表的な感性表現に対して「かっこよす」「カッコイイ」といった表記揺れといえるバリエーションに加え，「イケメン」といった表記が異なる表現も収集できている．このような表現はルールなどで収集することは難しく，時間的近傍の利用がうまく効果を発揮した例といえる．また代表的な顔文字をいくつか収集ができているこ

表 6　収集された感情表現

かっこいい	かわいい	泣ける	笑える
かっこよ	かわええ	(;ω;)	おもしろ
かっこよす	かわいい!	泣ける。	おもしろい
カッコイイ	かわいいな	(T_T)	面白かった
かっこよすぎ	かわいすぎ	(;ω;)ブワッ	(ﾟ∀ ﾟ)
カッコいい	可愛すぎ	・ﾟ・(ノД`)・ﾟ・。	キター――
かっけえ	カワイイ	(;_;)	――!!
かっけぇ	かわいすぎる	..(;ω;)	ﾟ∀ ﾟ)
かっこいいな	かわゆす	泣きそう	面白い!
かっこよすぎる	可愛いな	武市さん	おもろい
かっこいい!	かわいいよ	T^T)	ｷﾀｰ―(ﾟ∀ ﾟ)―
イケメン	可愛すぎる	Mother	―(ﾟ∀ ﾟ)―!!
かっこいいなぁ	かわいい。	(;∀;)	ｷﾀｰ―(ﾟ∀ ﾟ)―!!
かっこいい。	可愛い!	何度見ても	きたああ
かっこい	かわいいなぁ	(ToT)	面白い。
かっこいい	可愛いよ	泣いてる	面白かった。
ジェネラル	カワユス	イベント	キングオブコント
かっこよかった	ちゃんかわいい	イイハナシダナー	ーー!!
のシーン	かわいい	のシーン	キンコメ
向井くん	可愛いなぁ	(;ω;)	おもしろかった
いい!!	ファッション	亡くなった	キングオブコメディ
キムタク	かわいいなあ	エピソード	なかなか
速水先生	ちゃんか	シーンは	始まるよ
いいなぁ	赤ちゃん	(T-T)	もうすぐ
このシーン	サッカー	mother	面白すぎ
高杉晋作	可愛い。	ダイヤモンド	もうすぐ始まるよ!
アレンジ	(*´∀`*)	はやぶさ	面白いな
(*´д`*)	(*´Д`)	追跡!AtoZ	コメディ
ハァハァ	ダーウィン	フルーツ	おもしろい!
」なう!	シュート	アニソン	キターー
ルージュ	ハァハァ	カプセル	ハーバード

表 7　感情表現収集の適合率と再現率

	かっこいい	かわいい	泣ける	笑える
適合率	53%	70%	50%	67%
再現率	54%	44%	14%	20%

とが分かる．表現の分割に[4]の手法を用いることで代表的な顔文字について適切に分割ができ，さらに代表的な顔文字であれば出現数もある程度期待できたため，収集に成功したと考えられる．

収集された表現には，人名や番組名の一部なども含まれている．今回の目的は感性表現の収集のため誤りとなるが，今回のアプローチによって，テレビ番組にて「かっこいい」「かわいい」「笑える」とよく評される人物や番組の抽出も可能であると期待できる．

表 7 の上位 30 件を感性表現として収集した場合の適合率を見ると、50～70%を示している．収集結果を辞書としてそのまま利用するには厳しいかもしれないが，手作業による選別を併用することで十分活用できると考えられる．提案手法では人名や番組名などが収集されてしまうことに対する対策を一切していないため，改善の余地があるといえる．

代表的な感情表現は，出現頻度が高く，感情が明確なものを選ぶと良いと考えられる．出現頻度が低かったり曖昧な表現であったりする場合，統計的な処理を行うため，うまく判定ができなくなるためである．予備実験として，出現頻度が低い表現のみを代表的な感性表現として収集したところ，適合率が下がる傾向を確認している．

一方，再現率を見ると，どの感情も高くないが特に「泣ける」「笑える」が低くなっている．ここで分析のために表 7 の再現率を，感情表現候補の収集の再現率と、感情表現の判定の再現率に分解してみたところ表 8 となった．

表 8　感情表現の収集の再現率および
感情表現の判定の判定率

	かっこいい	かわいい	泣ける	笑える
候補収集	59%	48%	26%	55%
感情判定	92%	92%	55%	37%

「かっこいい」や「かわいい」は，感情表現の判定の再現率は高く，感性表現候補の収集が再現率の低い原因となっていることが分かる．すなわち出現頻度が高い表現は収集ができているといえる．提案手法はある程度の出現頻度を持たないと収集できないため，この適合率の改善には出現頻度に依らない手法を別途考える必要があるだろう．一方「泣ける」「笑える」はどちらの再現率も高くない．「泣ける」については特に感情表現候補の収集の再現率が低くなっているが，これは顔文字の種類が多く低出現頻度な表現が収集できなかったためと考えられる．「笑える」については例えば「やりすぎｗ」「眉毛薄っｗｗｗ」などのように「ｗ」との組み合わせで初めて「笑える」という感情を表すものが多く，組み合わせた表現の出現頻度が低かったためと考えられる．従って再現率を改善するにはさらに多くのデータを用いる必要があると考えられる．

6. おわりに

本稿では，感情表現を判定する手法の原理として，同じ感情を表すメッセージが時間的近傍に偏って出現することを利用したアプローチを提案した．また，このアプローチを利用した感情表現の収集手法を提案した．実データを用いて感情表現の収集実験を行い，ある程度の出現頻度があれば多様な表現を収集できることを示した．

実験において，適合率および再現率が低かった根本的な原因は，出現頻度の低い表現が存在することにあり，より多くのデータを用いることで改善されると思われる．ただし今回の実験でも 7 ヶ月分のデータを利用しており，決して少ないデータ量ではない．そのため，出現頻度が低い表現に対応する手法を考える必要がある．

本研究は感情表現の収集が目的であったため，人名や番組名が収集された場合は誤りとした．しかし本研究のアプローチは感情表現の収集に限定されたものではないため，同じアプローチを用いることで，例えば面白いと言われる人や，泣ける番組などの判定への適用が可能であると思われる．本アプローチの応用は今後の課題である．

また，今回はあいまい性を排除するために人手で収集したもののうち明確な感情表現を正解として評価対象としたが，曖昧な感情表現も多く見受けられた．今後、このような表現の収集や評価についても検討していきたい．

参考文献

[1]Twitter, http://twitter.com/
[2]ニコニコ動画, http://www.nicovideo.jp/
[3]Yates F., “Contingency table involving small numbers and the χ2 test”. Supplement to the Journal of the Royal Statistical Society 1(2): 217–235 (1934).
[4]飯塚 泰樹, “接続確率最小法による教師なし単語分割”, 自然言語処理研究会報告 2000(86), 33-40 (2000).