SNS 上に表れる個人感情を用いた社会トレンドについての研究

山口 和宏　　杉山 歩
Ho,Tu Bao　　Dam,Hieu Chi
北陸先端科学技術大学院大学 知識科学研究科
{ Kazuhiro_YAMAGUCHI,a-sugiya
bao,dam}@jaist.ac.jp

1　初めに

近年 SNS やミニブログに代表されるソーシャルメディアが急速に普及してきている。ソーシャルメディアの特徴は、利用者は誰もが容易に情報を発信することができ、基本的にオープンな双方向のコミュニケーションが可能な点にあり、発信される情報は日々の出来事や個人的な意見や感情など多様性に富んでいる。このような特徴を持つソーシャルメディアというリソースを利用し、企業内での知識共有 [1] や商品に対する消費者の意見抽出 [2]、流行語抽出といった社会トレンド分析 [3] などの分野で研究が活発となっている。

ソーシャルメディアのように書き手が不特定多数となる文書集合において分析をおこなう場合は、文書や単語の重要性だけではなく書き手の感情面にも注意を払う必要がある。そのため単語の感情極性や共起確率を利用し、書き手の感情に注意した分析が多く試みられている [4][5]。これらはコンピュータを用いて文章から感情を推定する場合に用いられるが、一方人間が文章から感情を抽出する場合においては個人の感性に依る所が大きく、文章のみで行われるコミュニケーションにおいては書き手と読み手の間に齟齬が生じる場合がある。この齟齬を解消する一手段として顔文字がタイプライターの時代より考案され、現代では爆発的に普及してきている。顔文字教室[1] によると、現在までに 1 万種類以上の顔文字が考案されており、ウェブページなどで欠かすことのできないものとなっている。

顔文字の主な種類については昨年報告した [6]。顔文字を用いることで、言葉で説明することが困難な感情を表現したり、言葉で表現すると冗長な表現を簡易的に表現することができるというメリットがある。この特徴のため、メールや twitter などインフォーマルなコミュニケーションや使用できる文字数に制限がある場合において顔文字は頻繁に利用される。本研究では上述した顔文字の利点を生かし意味論的テキストマイニングの視点から、感情表現として顔文字に注目し、twitter において日本語のツイートと共に頻繁に使用されている東洋式の顔文字を利用して分析をおこなう。

顔文字を目・口・頬などの要素に分解し、このデータと Web 上の顔文字辞書を用いて決定木分析をおこなった。顔文字は文脈により表す感情が決まる、ひとつの顔文字で複数の意味を持つ、という文脈依存性を表現するため、決定木による感情ラベルの適合確率を顔文字の感情ベクトルとした。

日本時間を基準とし 24 時間単位での感情量の推移と比率を考察した結果、顔文字を含むツイート数の割合がほぼ一定であることが明らかになった。これより、顔文字のみの感情解析であっても安定した結果を得られることが示唆された。

また、ポジティブな外的要因と感情の間に関係があることを昨年報告した [6]。しかし顔文字を含まないツイート数を考慮に入れておらず、論理的根拠が希薄であった。そこで本研究では、顔文字を含まないツイート数も考慮に入れ、ポジティブ・ネガティブな外的要因と SNS 上の感情にどのような関係があるのか調査をおこなった。分析の結果、ポジティブな外的要因ではツイート件数の増加が、ネガティブな外的要因では限定的なツイート件数の増加が確認された。

[1] http://kaomoji.kyo-situ.com/

表 1: twitter から収集するデータの概要

ツイート ID	266482307072937984
時間	2012-11-08 19:08:27+09
ツイート	帰宅 (′ω‘)

2 手法

2.1 利用データ

本研究では twitter のツイートデータを取得し、顔文字を抽出し考察を行う。Michal Ptaszynski らの先行研究では、顔文字を次のように定義している [7]。顔、姿勢などを表し、ユーザの感情を伝えるために頻繁に使われる文字列・記号列である。本研究ではこの定義に次の条件を付け加えた。顔文字とは、”(”と”)”で囲まれた文字列・記号列である。

本研究では Streaming API[2] によりツイートデータを収集した。ツイートを一意に示すツイート ID、発信された時間 (世界標準時)、ツイート内容を収集し、世界標準時から日本標準時への変換、全角・半角記号の統一を行い、ツイート ID の比較から重複するデータのないようにした。収集したデータの一例を表 1 に示す。

取得したデータから顔文字を抽出し、URL や”(笑)”といった文字列を取り除いた。考察を行うにあたり、決定木により顔文字の感情をベクトルで表現した。以降で行った比較に際しては、2012/10/01~12/26 までのデータを使用している。

2.2 決定木による顔文字の感情推定

顔文字ステーションの感情系顔文字辞書[3]を利用し、決定木による分析をおこなう。まず辞書データの内驚く、照れる、怒る、喜ぶ、泣く、落ち込む、汗・恐怖、笑うのラベルを持つ顔文字を使用した。そして”ハズカシィ・・(*p ω q*)”の”ハズカシィ・・”ようなテキストを削除した。また、2.1 で定義した顔文字に当てはまらないものを削除し、全角・半角記号を統一した。

次に中島氏が公開している顔文字辞書[4]を利用し、フリーの形態素解析器である mecab [5]により各顔文字を左目、口、右手などの各要素に分割し、これを教師データとした。なお、口や右目などが重複する場合は該当する顔文字を教師データから取り除いた。これにより

表 2: 教師データ一例

顔文字	ラベル	右目	右輪郭	右頬	口	左目	左輪郭	左頬	動線
Σ(o’д’o)	驚く	’	)	o	д	’	(	o	Σ
Σ(。・д・。)	驚く	・	)	。	д	・	(	。	Σ
Σ(゜д゜*)	驚く	゜	)	*	д	゜	(		Σ

表 3: 教師データ内のラベル内訳

汗・恐怖	喜ぶ	泣く	驚く	照れる	笑う	怒る	落ち込む	合計
13	21	8	16	26	18	35	10	147

147 件の教師データを得た。教師データの一例とラベル内訳をそれぞれ表 2 と表 3 に示す。なお、該当する記号がない場合は該当列を省略している。そしてフリーの統計処理ソフトウェアである R[6]および mvpart パッケージ[7] を利用し、分割した要素を説明変数、教師データの感情ラベルを非説明変数として、決定木による分析をおこなった。

2.3 解析

2.1 で収集したデータを利用し、時系列や社会イベントと SNS 上の感情との関係について考察をおこなうため、総量と比率の観点から解析をおこなった。

2.3.1 手順

解析は以下の手順でおこなった。

- 10 分毎にデータを区切り、出現した顔文字を集計する
- 各顔文字に 2.2 で作成した決定木のルールに基づき、感情ベクトルを付与する
- 感情ベクトルの時間推移から考察する

2.3.2 時間帯、曜日単位での感情比較

日本時間 (JST) を基準とし、24 時間の感情推移を曜日ごとに考察をおこなった。

2.3.3 社会イベント等、外的要因に対する感情表現の依存性

先行研究 [6] から、ポジティブな外的要因と感情の間に関係がある可能性が示唆された。そこで本研究で

[2]https://dev.twitter.com/docs/streaming-apis
[3]http://kaosute.net/jisyo/index.shtml
[4]http://www.haroperi.info/emoticon/annotated.html
[5]http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[6]http://www.r-project.org/
[7]http://cran.r-project.org/web/packages/mvpart/index.html

表 4: 決定木の分類精度

分類ラベル	汗・恐怖	喜ぶ	泣く	驚く	照れる	笑う	怒る	落ち込む
汗・恐怖	12	0	0	1	0	0	0	0
喜ぶ	0	18	0	0	0	3	0	0
泣く	0	0	8	0	0	0	0	0
驚く	3	2	1	7	2	0	0	1
照れる	2	1	3	0	16	2	0	2
笑う	1	1	0	0	3	13	0	0
怒る	1	2	5	3	1	0	18	5
落ち込む	0	0	0	1	1	0	0	8
正解個数	12	18	8	7	16	13	18	8
正解率	92.3%	85.7%	100.0%	43.8%	61.5%	72.2%	51.4%	80.0%

行:予測結果、列:教師データのラベル

はポジティブ/ネガティブな外的要因によって感情がどのように反応するのかについて考察した。なおポジティブな要因としてクリスマスを、ネガティブな要因として、東北地方でマグニチュード 5.0、最大震度 4 以上の地震が発生した 2012 年 10 月 3 日、25 日、11 月 3 日、9 日、24 日、12 月 7 日を例として考察をおこなった。

3 結果

3.1 顔文字の感情推定

決定木により得られたルールの分類精度を表 4 に示す。総合での分類精度は 68.0%であった。表より、概ね高精度で分類できているが、驚く・怒るについては分類精度が低くなっている。これは、これらの感情を表すために使用される記号がほぼ共通しているためと考えられる。

次に利用データからランダムに選択した 1000 件のデータより、2.1 での定義に合致する顔文字を抽出して得られた 147 種類の顔文字について、決定木により感情ラベルの予測をおこなった。なお決定木による予測をおこなうために、教師データに存在しない記号を用いている顔文字は除外した。これを人手により正誤の判断をおこない、各感情毎に正答率を算出した。この結果を表 5 に示す。

顔文字が表す意味は文脈により異なるため、顔文字に対応する感情のベクトル表現を試みた。決定木による予測確率を感情ベクトルとし、一例を表 6 に示す。これにより顔文字の特徴である感情の曖昧さが表現可能となり、部分的には感情を高精度で表現可能となったと考えられるが、統計学的手法により精度について検証する必要がある。

表 5: ランダム抽出したデータでの正解率

	汗・恐怖	喜ぶ	泣く	驚く	照れる	笑う	怒る	落ち込む	合計
正解個数	5	14	6	1	5	31	4	10	76
分類個数	6	22	7	7	17	42	29	17	147
正解率	83%	63%	86%	14%	29%	74%	14%	59%	52%

表 6: 決定木による予測確率

顔文字	汗. 恐怖	喜ぶ	泣く	驚く	照れる	笑う	怒る	落ち込む
(' ∇ ‘*)	0.06	0.06	0.00	0.00	0.17	0.72	0.00	0.00
(>ω<)	0.08	0.00	0.15	0.08	0.00	0.00	0.62	0.08
(' д ‘)	0.00	0.00	0.17	0.00	0.00	0.00	0.17	0.67

3.2 解析

3.2.1 時間帯、曜日単位での感情比較

日本時間基準で 24 時間単位で感情の推移を比較した。図 1 に 10 分当たりツイート数、感情の総量、顔文字を含むツイートの割合を示す。

図 1 より、顔文字を含むツイートの割合は 20%から 30%程度の範囲で安定しており、顔文字のみを使用した感情解析においても安定した結果を期待できる。平日と土日祝日の感情推移を図 2,3 に示す。昨年報告したとおり、平日は通勤通学時間・昼休憩・帰宅後の時間帯に特徴がある。

3.2.2 外的要因に対する感情表現の依存性

図 4 にクリスマスの感情推移を、図 5 に地震発生日の感情推移を示す。クリスマスでは地震発生日だけでなく平日・休日と比較しても、感情の総量が増加している。また、顔文字を含むツイートの割合は平日・休日と比べて変化が見られない。これより、ポジティ

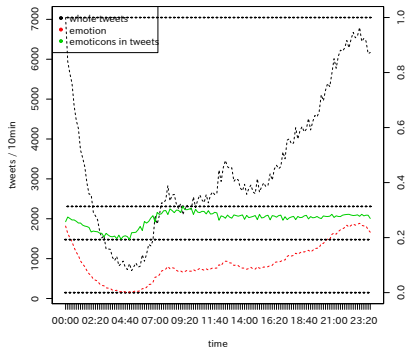

図 1: ツイート数と顔文字を含むツイートの割合

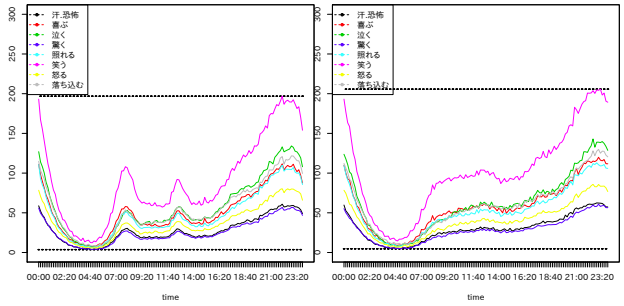

図 2: 平日の感情の推移

図 3: 休日の感情の推移

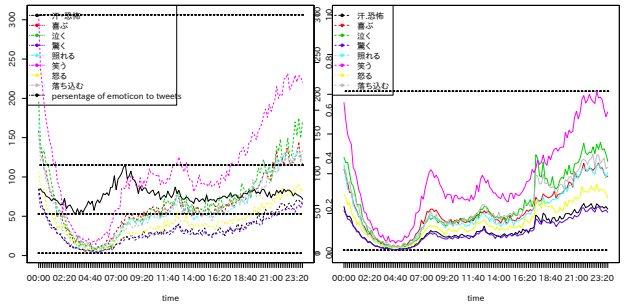

図 4: クリスマスの感情推移 (ポジティブな外的要因の例)

図 5: 地震発生日の感情推移 (ネガティブな外的要因の例)

ブな外的要因は顔文字を含むツイートの発信数を増加させることが示唆される。地震発生日については、17 時から 18 時にかけて泣く、落ち込むの感情量が急増している。Yahoo!Japan が提供する地震速報[8] によると、12 月 7 日 17 時 29 分に三陸沖にて最大震度 5 弱、マグニチュード 7.3 の地震が発生した。他の地震発生日で同時刻に地震が発生した日はないため、この結果は平均化されたにも関わらず当該日の地震が SNS 上の感情表現に影響を与えたことを示している。これらのことより、ポジティブな外的要因ではツイート件数の増加が、ネガティブな外的要因では限定的なツイート件数の増加が確認された。

4 終わりに

本研究ではテキストを利用しない感情解析として顔文字に注目し、決定木により感情ベクトルとして表現することで顔文字の持つ微妙な感情を表現を試みた。24 時間単位での感情の推移について比較をおこない、ポジティブ/ネガティブな外的要因は共に感情表現に影響を与えることが明らかになった。

本研究では顔文字以外のテキストを利用していない点に課題がある。また顔文字は種類が非常に豊富で、種類はなお増加しており、すべての顔文字に対して感情ラベルが付いた教師データを得るのは不可能である。故に顔文字の解析では未知語が数多く表れるが、未知後への対応は言語処理のみならずデータマイニングにおいても重要な課題であるため、この問題について取り組みたい。

[8] http://typhoon.yahoo.co.jp/weather/jp/earthquake/list/

参考文献

[1] 林幹人. イノベーション・プロセスにおける組織内ソーシャル・メディアの意義. 桜美林論考. ビジネスマネジメントレビュー, Vol. 1, pp. 33–45, 2010-03.

[2] 那須川哲哉. テキストマイニングを使う技術/作る技術:基礎技術と適用事例から導く本質と適用法. 東京電機大学出版局, 2006.

[3] 白木原渉, 大石哲也, 長谷川隆三, 藤田博, 越村三幸. Twitter における流行語先取り発言者の検出システムの開発. 情報処理学会 研究報告 [データベースシステム (DBS)], Vol. 2010-DBS-150, , 2010.

[4] 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. 第 4 回データ工学と情報マネジメントに関するフォーラム, 2012.

[5] 高野憲悟, 萩原将文. 感情関連語を用いた感情推定法の提案とニュースサイトのアクセス解析への応用. 日本感性工学会論文誌, Vol. Vol.11 No.3 pp.495-502, , 2012.

[6] 山口和宏, 杉山歩, 鈴木健之, 藤田哲也, Ho Bao Tu, Dam Hieu Chi. データマイニングを用いた顔文字表現の定量的評価による感情解析. 言語処理学会第 18 回年次大会, 2012.

[7] Michal Ptaszynski, Pawel Dybala, Rafal Rzepka, and Kenji Araki. Towards fully automatic emoticon analysis system (^o^). *In Proceedings of The Fifteenth Annual Meeting of The Association for Natural Language Processing (NLP-2010)*, pp. pp. 583--586, 2010.