

# 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

山口 和宏      杉山 歩      鈴木 健之  
 藤田 哲也      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-sugiyama, s1050027  
 s1050045, 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では利用者は誰もが容易に情報を発信することができ、双方向のコミュニケーションが可能な点に特徴がある。発信される情報は日々の出来事や個人的な意見など多様性に富んでいる。この情報発信が容易で双方向のコミュニケーションが可能なプラットフォーム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] などの分野で研究が活発となっている。

意見抽出において注意すべき点は、特定のキーワードの意味が文脈により決定される場合があることである。このことは特にウェブページの集合から知識・意見の抽出を試みる際に問題となっており、文脈によりキーワードの意味を特定する潜在意味解析が注目を浴びている。

感情解析においてはテキストマイニングを利用した潜在意味解析によりテキストから感情を抽出する試みが成功しており、アンケートデータの解析や株価予測などに役立てられている。Johanらの株価予測における研究では、twitterのツイートデータと気分解析ツールである Google-Profile of Mood States(GPOMS)を利用して感情の総和とダウ平均株価が相関関係にあることを示している [3]。ツイートデータから GPOMS により Calm, Alert, Sure, Vital, Kind, Happy の 6 次元の感情情報を抽出し、Calm の増減からダウ平均株価を 86.7% の精度で予測している。しかし日本語の構文の曖昧さや意味の多様性から、日本語のテキストから感情情報を抽出することは今日まで課題となっている。

コンピュータで文章から感情を抽出する手法として

は大きく分けて次の二通りの方法がある。すなわち、感情コーパスを利用して文単位に感情を抽出する方法と感情語辞書を利用して品詞単位に感情を推定し、構文・形態素解析を行って感情のスコアを算出する方法である。

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。この祖語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<sup>1</sup>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。

顔文字は数個の記号の組み合わせで構成され、表 1 に示すような種類に分けられる。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行う。分析の結果、twitter への投稿が生活リズムと密接な関係があること、利用デバイスやイベント時期においては感情表現に明確な違いが見られることが明らかになった。

<sup>1</sup><http://kaomoji.kyo-situ.com/>

表 1: 顔文字の種類	
種類	顔文字
西洋式	:)
東洋式	(^o^)
	$\begin{matrix} \wedge & \wedge \\ \text{c} & \text{c} \\ \text{c} & \text{c} \end{matrix}$ アッ
日本式	$\begin{matrix} \text{c} & \text{c} \\ \text{c} & \text{c} \\ \text{c} & \text{c} \end{matrix}$

表 2: twitter から収集するデータの概要	
ツイート ID	49139402818462
時間	23:49:45
ツイート	行かせていただきやす (・ ω ・)
クライアント	Tween

また、本研究で使用するデータの概要を先行調査 [5] と比較を行った。

## 2 手法

本研究では以下の手順で分析を行った。

1. twitter からデータを収集し 顔文字を抽出する
2. 顔文字に感情ラベルを割り当てる
3. 時系列やイベントの有無、投稿元デバイスなどの視点から考察する

### 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 [4]。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。

本研究では顔文字の意味の多様性に関与する要素として口の表現に着目した。ツイートデータの取得には tweepy ライブラリの Search API を利用し、検索キーワードは「 $\sqcap$  or  $\sqcup$  or  $\omega$  or  $\vee$  or  $\Sigma$  or  $\sum$  or  $\nabla$  or  $>$  or  $<$  or  $\geq$  or  $\leq$  or  $\_$ 」とし、10 分毎にツイートデータを収集した。Search API によりツイートを一意に示すツイート ID、ツイートが発信された時間 (世界標準時)、ツイート内容、ツイートに使用されたアプリケーション情報 (クライアント情報) 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。この結果を表 2 に示す。

上記の手順によりデータを収集し、取得したツイートデータから顔文字を抽出し、出現回数を記録した。1 日という期間で出現頻度が上位 40 位以内の顔文字を高頻度出現顔文字とし、時系列の視点から考察を行った。考察を行うにあたり、40 個の顔文字に喜怒哀楽の感情ラベルを割り当てた。以降で行った比較に際しては、2011/12/3~28 日までのデータを使用している。

### 2.2 解析

2.1 で収集したデータを利用し以下の 3 つの観点から分析を行った。

1. 時間帯、曜日単位での感情比較
2. 投稿ツール依存性
3. 社会イベント等、外的要因に対する感情表現の依存性

#### 2.2.1 時間帯、曜日単位での感情比較

まず一日の内で表現される感情の総数の推移を明らかにするため、2011/12/3 28 までのデータの平均を取り、時間別の顔文字の出現数を算出し感情ラベルごとの集計を行った。

#### 2.2.2 投稿ツール依存性

twitter への投稿の手段として使用される媒体によって表現される感情の総数の推移を明らかにするため、デスクトップパソコン (以降 PC)、モバイル端末 (以降携帯端末) のいずれから投稿されたかによってツイートデータを分類し、クライアント情報を元に PC または携帯電話、スマートフォン、iPad などの携帯端末のいずれから投稿されたかを分類し、データの平均を取り、時間別の顔文字の出現数を算出し感情ラベルごとの集計を行った。

#### 2.2.3 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 [3] で示されたように、感情とダウ平均株価の間には相関関係がある。本研究では利用データを日本語に限定していることから、日経平均株価の変動との相関関係について明らかにするため、日経平均株

表 3: 利用デバイスの割合

	先行調査	収集データ
PC	12%	29%
携帯端末	29%	60%

価の上昇日、下降日について比較を行った。また福岡がうつ病を季節性感情障害の一種として捉えた考察を行っている [6]。この中で自殺件数と気象と精神状態の関係について考察をしており、気象条件が精神に影響を与えることが示唆された。このことから本研究では天気と感情の関係について考察するために東京都の天気について比較を行った。さらにイベントの有無に対して感情がどのように反応するのかを明らかにするため、クリスマス为例として 2011/12/24,25 についてデータの平均を取り、時間別の顔文字の出現数を算出し感情ラベルごとの集計を行った。

## 3 結果

### 3.1 利用データ

本研究で使用するデータについて PC、携帯端末それぞれのデバイス毎の利用実態を分類した結果を表 3 に示す。なおクライアントによっては PC、携帯端末の両方に対応しており、分類することが困難なものがあり、これらは集計から除外した。表に示したとおり、先行調査<sup>2</sup>とほぼ同等の割合で PC、携帯端末からツイートが発信されていることが確認され、良い一致度を示している。

### 3.2 解析

#### 3.2.1 時間帯、曜日単位での感情比較

時間帯ごとのツイート数の推移を図 1 に示す。図 2 について、平日では 8 時頃にツイート数の盛り上がりがあるが土日祝日では見られないことから、8 時頃のピークは通勤・通学によるものと考えられる。また平日・土日祝日に共通して正午頃以外の日中はツイート数が減少していること、平日の終業、帰宅時刻である 18 時以降にツイート数が増加傾向にあること、睡眠時間である 1~6 時の間はツイート数の減少が顕著であることが読み取れる。この結果から図 1 のツイートの増減は生活リズムに沿ったものであると言える。

<sup>2</sup>先行調査においては携帯端末の中に携帯電話は含まれていない

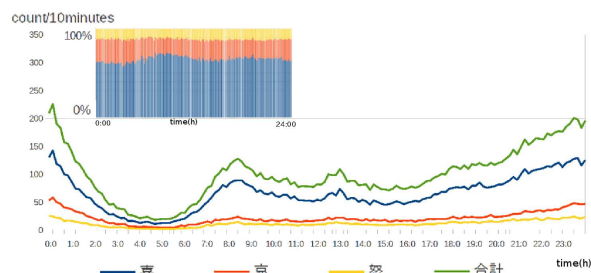


図 1: 1 日平均ツイート数の推移と感情の割合

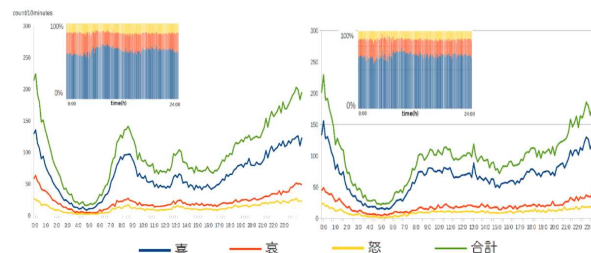


図 2: 曜日に着目した 1 日平均ツイート数の推移、割合の比較 (左:平日、右:土日祝日)

#### 3.2.2 投稿ツール依存性

ツイートの投稿元を PC・携帯端末で分類し比較した結果を図 3 に示す。この結果 PC を使用したツイートの方が喜の感情の割合が平均して約 20%程度高いことが確認された。PC と携帯端末ではツイートを行うと判断してからツイートを発信するまでの時間が異なる。このツイートを発信するまでの時間に発信する意欲が減少し、実行されないことがありうることが示唆され、特に喜以外の感情を含むツイートで顕著であることが示唆される。

#### 3.2.3 社会イベント等、外的要因に対する感情表現の依存性

YAHOO!ファイナンス<sup>3</sup>から 2011/12/3~28 の期間の中で日経平均株価が前日と比較して上昇した日と下落した日を調査し、比較を行った結果 (図 4) において日経平均株価が上昇した日と下落した日の感情に有意な差が見つからなかった。日本気象協会<sup>4</sup>から 2011/12/3~28 の東京都の天気を調査し比較を行った結果 (図 5) において、天気と感情に有意な差は見られなかった。全体とクリスマスのイベント時期とを比較した結果 (図 6) においては、クリスマスの時期には全体と比較して約 30%ツイート数が増加することが確認

<sup>3</sup><http://finance.yahoo.co.jp/>

<sup>4</sup><http://tenki.jp/>

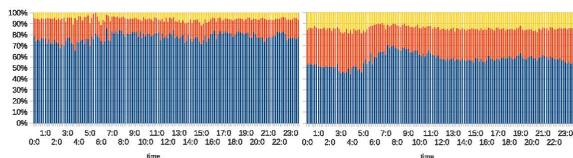


図 3: 利用デバイスに着目したツイート の喜怒哀の割合比較 (左:PC、右:携帯端末)

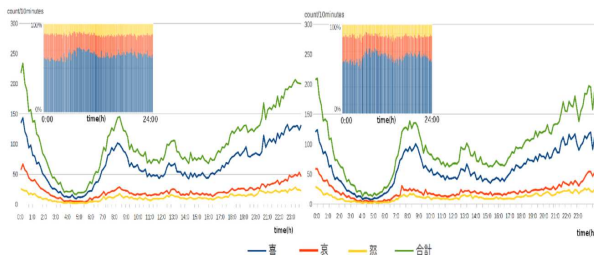


図 4: 株価に着目した 1 日平均ツイート 数の推移、割合の比較 (左:上昇、右:下落)

された。またすべての感情が一律に増加するのではなく、喜の感情のみが増加することも確認された。

## 4 終わりに

本研究では意味論に基づくテキストマイニングとして個人が発信する感情表現である顔文字に注目し、喜怒哀の感情ラベルに分類を行い時系列の視点から考察を行った。平日と土日祝日のツイート 数の推移の比較から、twitter の投稿が生活リズムに基づいていることを明らかにした。また利用デバイスにより差異が生じるかを明らかにするために PC、携帯端末のどちらから発信されたかでツイートを分類し分析を行った。この分析からツイートを発信するという判断した時から実行するまでの時間の長さが実際にツイートを実行するか否かに影響を与えうるという示唆を得た。外的要因の影響については日経平均株価と天気に着目した比較では有意な差は見られなかったが、クリスマスに着目した比較では全体よりも喜の感情が増加することを確認した。

今後はツイートを発信したユーザー情報を活用したユーザー属性の絞り込みによる分析や、返信やリツイートを対象とした分析、顔文字の感情ラベルへの振り分け手法の高度化など、より多角的な視野で分析を行いたい。

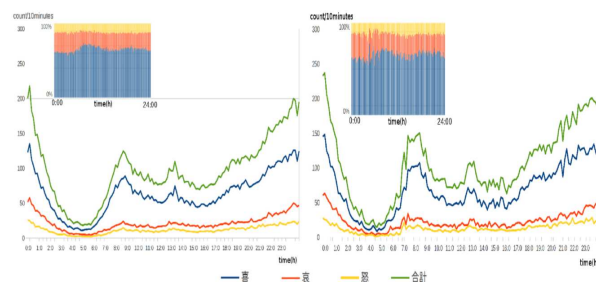


図 5: 天気に着目した 1 日平均ツイート 数の推移、割合の比較 (左:晴、右:曇)

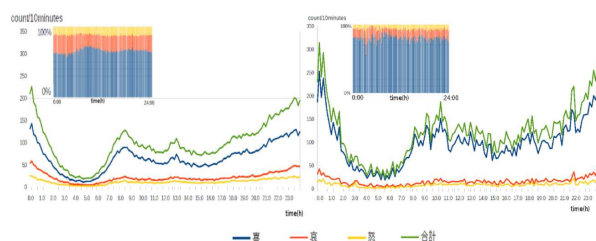


図 6: 全体とイベント時を比較した 1 日平均ツイート 数の推移、割合の比較 (左:全体、右:クリスマス)

## 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] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2010.
- [4] 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.
- [5] ネットレイティングス株式会社. スマートフォンメディア利用実態調査レポート vol.2, 2011.
- [6] 福岡義隆. 気象・季節の感情障害への影響. 国際環境研究協会, 2003.