

# センチメント分析を用いたテレビ番組における バースト期間のツイート分類

本多 波輝† 横井 健†

東京都立産業技術高等専門学校 創造工学専攻†

## 1. はじめに

近年、マイクロブログの1つである Twitter<sup>[1]</sup>では、テレビ番組を見ながら番組に対する自分の意見をリアルタイムで投稿するユーザが増加している。多数のユーザが同時に投稿を行った時間帯(バースト期間)は番組の重要場面であると考えられ、バースト期間に投稿された文章(ツイート)を解析することは有益である。しかし、重要場面の状況や登場人物に対して全てのユーザが同一の意見を投稿しているわけではないため、ツイートを似た意見ごとに分類することで、より詳細な解析を行うことができると考える。

そこで、本研究ではテレビ番組に対するツイートからバースト期間を特定し、番組の内容に依存しないセンチメント情報を用いることでバースト期間のツイート分類を行う。

## 2. 提案手法

本研究では、蝦名ら<sup>[2]</sup>の研究を基にバースト期間の検出を行い、「感情辞書の作成」・「ツイートの分類」の2つの手順よりツイート分類を行う。以後、バースト期間とバースト期間終了から1分間を合わせた期間を解析期間と呼び、ツイートを投稿したユーザを単にユーザと呼ぶ。

### 2.1. 感情辞書の作成方法

本研究では、解析期間のツイートを分類する際に、センチメント情報を用いる。センチメント情報は、ユーザがどのような感情に基づく意見を投稿したのかを判断する際に用いる値と定義する。センチメント情報は感情辞書を用いてユーザに付与する。本研究では、感情表現辞典<sup>[3]</sup>に収録されている言葉をデータ化し、ツイートとマッチングを行えるように構築した感情辞書を作成した。感情表現辞典には「嬉・怒・哀・怖・恥・好・厭・昂・安・驚」の10種類の感情表現を使用する際に用いる言葉(名詞、形容詞、副詞)が収録されている。各感情に対応した10種類の辞書を作成し、ツイートに各感情の辞書とマッチングする言葉の出現頻度を、そのツイー

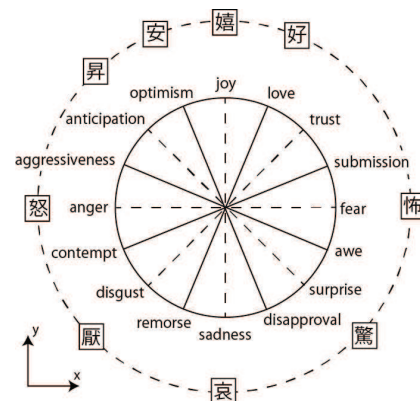


図1 感情表現辞典の感情を Plutchik の感情の輪に当てはめた図

トを行ったユーザのセンチメント情報として割り当てる。ただし、感情表現辞典には、ツイートで多くみられる、「可愛い」の漢字から片仮名への表記ゆれである「カワイイ」や、新語である「キター」などの言葉は収録されていない。そこで、感情表現辞典に収録されていない言葉は水岡ら<sup>[4]</sup>の手法を用いて、作成した各感情の辞書に追加することで補った。

### 2.2. ツイートの分類方法

本研究では短い解析期間内でユーザの意見は変わらないと仮定し、解析期間においてユーザが複数回行った投稿は結合し、1つの投稿とみなして処理する。

作成した各感情の辞書よりマッチングを行い、ユーザにアノテーションとしてセンチメント情報を付与する。ここで、アノテーションを付与するに当たり、「嬉」と「好」のように相関があると考えられる感情を考慮して、Plutchik<sup>[5]</sup>が提案した Plutchik の感情の輪を用いてツイートを分類する。Plutchik の感情の輪に本研究で用いるセンチメント情報を図1に示すとおりに当てはめる。ここで、「恥」については Plutchik の感情の輪に当てはまる感情が無いため、「恥」については検討しないことにする。図1の横軸を x 軸、縦軸を y 軸とし、円の中心座標を  $[0, 0]$  とし、Plutchik の感情の輪を範囲  $D: x^2 + y^2 \leq 1$  の円として考える。

ツイートごとに、式(1)より x 座標を、式(2)より y 座標を決定し、求めた座標をツイートの

Tweet Classification Using Sentiment Analysis in Burst for Television Program.

† Namiki Honda

† Takeru Yokoi

† Tokyo Metropolitan College of Industrial Technology

表1 テレビに対する意見を5つに分けた基準

クラス	テレビに対する意見	各クラスの範囲
クラス1	感情の高揚を表す意見	$y > x$ and $0 \leq x$ and $x^2+y^2 > 3.34$
クラス2	好意的な意見	$y \geq -x$ and $0 > x$ and $x^2+y^2 > 3.34$
クラス3	怒りの意見	$y < -x$ and $y < x$ and $x^2+y^2 > 3.34$
クラス4	哀しみの意見	$y > -x$ and $y \leq x$ and $x^2+y^2 > 3.34$
クラス5	恐れ	$y \geq -x$ and $y \leq x$ and $x^2+y^2 > 3.34$

意見を表す点とする。但し、式(1)、(2)の $\delta_i$ は感情*i*がツイートに出現する回数を表している。また、本研究ではテレビ番組に対する意見を表1に示す5つに区分し、ツイートの座標がどの範囲に所属するかを求めることで分類を行う。

$$\sum_{i=1}^9 \delta_i \times \frac{\text{感情}i \text{の}x \text{座標}}{\sum_{j=1}^9 \delta_j} \quad (1)$$

$$\sum_{i=1}^9 \delta_i \times \frac{\text{感情}i \text{の}y \text{座標}}{\sum_{j=1}^9 \delta_j} \quad (2)$$

### 3. 実験

#### 3.1. 実験方法

ツイートの分類を行うにあたり、解析期間の投稿を Twitter API<sub>[5]</sub>を用いて収集し、“RT”、“QT”、“@”、“#”の機能を用いた投稿を除外して時系列順に並べたものをデータセットとする。表2に示す解析期間のツイートを、提案手法より分類した際の正解率を式(3)より求め、分類の精度について考察を行う。

$$\text{正解率} = \frac{\text{クラスの正解ツイート数}}{\text{クラスに所属するツイート数}} \quad (3)$$

ここで、正解ツイートは表1に示した各クラスの意見を表していると考えられるツイートを人手で収集し、そのツイートの個数を数えた値である。

#### 3.2. 実験結果と考察

表3に評価実験の結果を示す。なお、表3の(1)は「FNS歌謡祭」を、(2)は「第7戦楽天×巨人」を、(3)は「ガリレオ 第一話」を表している。表3より、正解率が0.50以下になるクラスもあるが、所属ツイート数が50以上のクラスに着目すると全て0.70以上と高い値になっていることから、高い精度でツイートを分類できたと考える。

表3より、所属ツイート数が50以上のクラスの平均正解率が(1)と(2)は0.80以上の高い値になり、(3)は若干低い0.70となった。(3)のドラマ番組では番組の感想や場面の状況について投稿するユーザが多く、(1)のバラエティ番組や(2)のスポーツ番組と比較するとツイートが長文になることが分類の精度を下げる要因になった

表2 解析期間に関する情報

	(1)FNS歌謡祭	(2)第7戦楽天×巨人	(3)ガリレオ 第一話
バースト期間	21:19:11~21:20:27	21:50:24~21:51:54	21:53:00~21:53:10
解析期間	21:19:11~21:21:27	21:50:24~21:52:54	21:53:00~21:54:10
解析期間の投稿数	1,060件	1,660件	902件
投稿したユーザ数	759人	1,298人	702人

表3 評価実験の結果

	(1)の所属ツイート数	(1)の正解ツイート数	(1)の正解率	(2)の所属ツイート数	(2)の正解ツイート数	(2)の正解率	(3)の所属ツイート数	(3)の正解ツイート数	(3)の正解率
クラス1	152	129	0.85	408	372	0.91	12	3	0.25
クラス2	248	202	0.81	284	252	0.89	82	58	0.71
クラス3	130	100	0.77	21	9	0.43	98	69	0.70
クラス4	12	4	0.33	130	114	0.88	3	2	0.67
クラス5	2	0	0.00	2	0	0.00	2	1	0.50

表4 分類に失敗したツイート例

「ガリレオ 第一話」の代表ツイート
吉高あんまり好きじゃない・・・滑舌もあんまし良くないなあ

と考える。また、表4に示す「好きじゃない」という不快を表しているツイートを、本手法では、作成した「好」の感情辞書より「好き」という言葉で抽出を行ってしまい、好意的な意見を表すクラス2に分類してしまったことも分類の精度を下げる要因となった。これを解決するためには、ツイートを分類する際に否定語を考慮する必要があると考える。また、表4に示す「吉高」という人名のようなバースト期間の特徴語を抽出し、係り受け解析を用いて特徴語に対する感情語を抽出することで、より適切な分類ができると考える。

#### 4. おわりに

本研究ではテレビ番組に対するツイートからバースト期間を特定し、番組の内容に依存しないセンチメント情報を用いることでバースト期間のツイート分類を行った。評価実験を行った結果、所属ツイート数が50以上のクラスに着目すると全て0.70以上と高い値になっていることから、高い精度でツイートを分類できたと考える。

今後の課題としては、否定語を考慮して、分類の精度を上げることを考えている。

#### 参考文献

- [1] Twitter, <http://twitter.com>, (参照 2013-1-12)
- [2] 蝦名亮平, 中村健二, 小柳滋. リアルタイムバースト検出手法の提案. 日本データベース学会論文誌, Vol25, No.1, pp.174-182, 2010.
- [3] 中村明. 感情表現辞典. 東京堂出版, 1993.
- [4] 水岡良彰, 鈴木裕. マイクロブログを用いた感情表現収集. 第10回情報科学技術フォーラム, 2011(1), pp.291-295, 2011.
- [5] Twitter API v1.1, <http://dev.twitter.com>, (参照 2013-1-12)
- [6] R. Plutchik. Emotions and Life. Amer Psychological Assn, 2002.