D-036

# ソーシャルメディアの情動に着目した実世界事象予測手法の提案

## Predicting Real World Events based on Excitement of Social Media

池田 和史† 服部 元† 滝嶋 康弘† 麻生 英樹‡
Kazushi Ikeda Gen Hattori Yasuhiro Takishima Hideki Asoh

## 1．まえがき

近年、ソーシャルメディアを利用して株価の変動や商品の売り上げなど実世界の事象を予測する研究が注目されている。既存の研究では、ソーシャルメディア上のコメントの量および投稿されたテキストに含まれる肯否極性と予測対象の事象との相関関係に基づいて予測を行う。一方、ソーシャルメディア上では、言葉による肯否極性だけでなく、顔文字や特有の表現を用いて情動を表す投稿が多数出現する。本稿では、Web 上に公開されたニュース記事とそれに対するソーシャルメディア上のコメントから、当該ニュースがテレビ番組で放送されるか否かを予測するタスクを設定し、ソーシャルメディアの情動を考慮することが予測精度に与える影響を検証する。情動表現を用いた予測手法と、肯否極性のみを用いた予測手法の性能比較評価実験を行い、情動表現を用いることによる予測の有効性を確認した。

## 2．関連研究

ソーシャルメディアを利用して実世界の事象を予測する研究として、Bollen らは Tweet の肯否極性から世の中の雰囲気を POMS(Profile of Mood States）を元に 6 種類に分類した結果と、株式市場や主要な出来事との関係を調査している[3]。Asur らは Tweet から映画の興行収入を予測する手法を提案しており、映画に対する Tweet の量や肯否極性が予測の精度向上に貢献することを示している[4]。

これらの研究では、ソーシャルメディア上のコメントの量および投稿されたテキストに含まれる肯否極性と予測対象の事象との相関関係に基づいて予測を行う。一方、ソーシャルメディア上では、言葉による肯否極性だけでなく、顔文字や特有の表現を用いて情動を表す投稿が多数出現する。しかし、これらの表現が実世界の事象に与える影響は既存の研究では考慮されていなかった。

著者らはこれまでに Web 上に公開されたニュース記事の内容がテレビで放送されるか否かを、記事に対する Twitter の反応の特徴を素性として識別器に与えることで予測する手法を提案した[3]。本稿では、Twitter の反応の特徴として情動表現を追加することにより、予測精度向上が見込まれるかを検証する。

## 3．提案手法

### 3.1 提案手法の概要

はじめに、文献[3]で提案したニュース記事からテレビ放送の有無を予測する手法について説明し、次に本稿で提案する情動表現の抽出に基づく予測手法について説明する。

†（株）KDDI 研究所, KDDI R&D Laboratories, Inc.
‡独立行政法人産業技術総合研究所, AIST

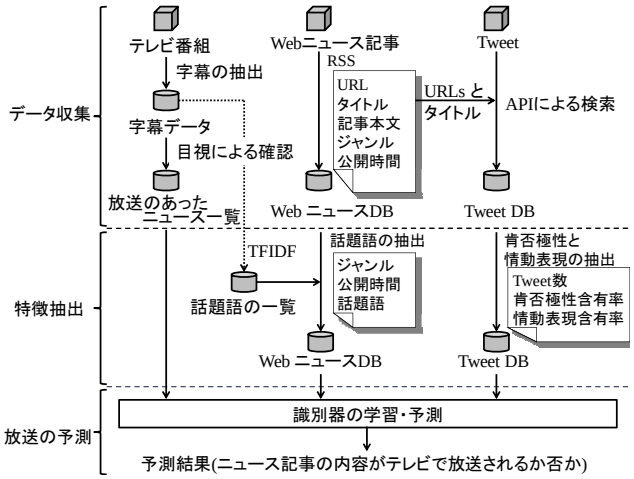


図 1 提案手法における処理の概要

表 1 ニュース記事と Twitter から抽出する特徴

| | |
|---|---|
| ニュース記事の特徴 | |
| | ジャンル(エンタメ、スポーツなど計 12 種) |
| | 公開時刻 |
| | 記事に含まれる話題語の数 |
| Tweet の特徴 | |
| | 総投稿数 |
| | 肯否極性の含有率(肯定、否定、中立) |
| | 情動表現の含有率(提案手法のみ) |

### 3.2 従来のテレビ放送予測手法

速報を除いた多くのニュースはテレビで放送される前に Web 上で同内容のニュース記事が公開され、記事に対する閲覧者の反応が Twitter 上に投稿される。この性質を利用することで、Web 上にニュース記事が公開された後、テレビ放送されるよりも前に、当該ニュース内容がテレビ放送されるか否かを予測することが可能となる。

本稿で対象とする問題を、ニュースをテレビで放送されるか否かに分類するという分類問題と考える。ニュース記事と Twitter 上の反応の特徴を素性として識別器の学習および予測を行う。識別器による予測を高精度に行うには、テレビ放送の有無に関連性の高い特徴を抽出することが重要であり、ニュース記事と Twitter 上の反応の特徴として、表 1 に示す特徴を取得する。話題語とは、ニュース記事が公開される直前の 1 週間におけるテレビニュースの放送字幕から TFIDF を用いて抽出した単語である。話題語を多く含むニュース記事はテレビ上で放送されやすいと考えられる。肯否極性については、ニュースに対する Tweet に人手で肯否極性を付与したものを教師データとし、任意の単語について出現の有無を特徴として識別器を学習させることにより取得する。

表 2　ソーシャルメディアの情動表現パターンと具体例

| |
|---|
| パターン 1: 顔文字を含むもの |
| 嬉しいニュースですね。優勝おめでとうございます(^_^)<br>え！…怖い((((；゜Д゜)))))))<br>いや、これは…(；´Д｀) |
| パターン 2:文字が連続しているもの |
| いやあああああああああああああああああ<br>うおおおおおおおおおおお<br>おいおいｗｗｗｗｗｗｗｗｗｗ |
| パターン 3:記号による感情表現 |
| マジ！？でも観たい！！<br>公式発表きたのか！！！欲しい！！<br>えっ？どういうこと？意味が分からない！！ |

### 3.3　情動表現抽出によるテレビ放送予測手法

ソーシャルメディア上では、ユーザの情動を表す投稿が多数出現する。代表的な情動表現を次の 3 パターンに分類し、ヒューリスティックなルールを導入して検出する。具体的な Tweet の例を表 2 に示す。

・パターン 1: 顔文字を含むもの。顔文字表現は多様なため、全ての顔文字表現を正確に抽出することは困難である。一方、多くの顔文字は携帯電話やパソコンなどに登録された入力辞書から選択されている場合が多く、頻繁に用いられる表現は有限であることから、Tweet 中に頻出の顔文字約 20,00 件を人手で辞書登録することで検出する。

・パターン 2: 文字が連続しているもの。対象 Tweet の形態素解析を行い、同一の形態素が複数連続して出現することを検出する。

・パターン 3: 記号による感情表現を含むもの。本稿では、「！」および「？」の出現の有無を検出した。

## 4. 性能評価実験

### 4.1　実験環境と手順

提案手法による予測の実現性を検証するため、(1)テレビ放送の有無における各特徴の出現傾向の比較、(2)テレビ放送されるニュース記事の予測精度評価、を実施した。Yahoo! ニュースを対象に平日 2 日分のニュース記事計 734 件と各記事に対する Tweet を実験データとして用意した。このうち、49 件のニュースが実際にテレビ上で放送されたことを、放送字幕を抽出し、人手で確認した。肯否極性については、事前に収集したニュース記事 50 件を対象に当該ニュースに対する Tweet 各 100 件に対して人手で肯否判定を行ったデータを用いて Naiive Bayse を学習したものを利用した。テレビ放送の有無の予測には、識別器として Support Vector Machine を利用し、1 日分のニュースを教師データ、もう一方の日のニュースを予測対象データとする 2-hold の評価を実施した。

### 4.2 実験結果

提案手法における各特徴の出現傾向を放送の有無ごとに比較した結果を表 3 に示す。放送されたニュース記事の方が Twitter 上の反応が大きい傾向や、話題語を多く含む傾向が確認された。肯否極性については、含有率が高いニュースの方が、放送される傾向はみられたがその差は小さい。これは、肯否極性の判別精度が十分に得られないことに起因していると考えられる。本実験条件における教師データを対象に 5-hold による肯否極性の推定精度を評価したところ 73.1%であった。肯否極性の判定誤りのノイズによって、放送の有無ごとの肯否の差が明確に現れない結果となったものと思われる。一方、情動表現パターン 1 から 3 については、それぞれ放送の有無と含有率に明確な差が見られた。どの指標も放送があったニュースの方が放送の無い場合と比較し、含有率が多いことが確認された。これらは、社会的なインパクトの大きいニュースに対して、より多くのユーザが情動的な反応を示すことを示唆している。

表 3　放送の有無における各特徴の出現傾向の比較

| | 放送有 | 放送無 |
|---|---|---|
| 平均 Tweet 数 | 49 | 685 |
| 話題語を 5 件以上含む記事 | **131 件** | 22 件 |
| 肯定表現の平均含有率 | **89.8 %** | 31.7 % |
| 否定表現の平均含有率 | 37.1 % | 34.8 % |
| 情動表現パターン 1 の含有率 | 42.1 % | 36.2 % |
| 情動表現パターン 2 の含有率 | **16.1 %** | 7.62 % |
| 情動表現パターン 3 の含有率 | **13.4 %** | 3.2 % |

表 4　テレビ放送されるニュース記事の予測精度

| | 再現率(Recall) | 適合率(Precision) |
|---|---|---|
| 従来手法 | 34.7 % | 85.0 % |
| 提案手法 | **59.2 %** | **87.8 %** |

放送される記事を提案手法で予測した際の再現率(正解記事数/放送記事数)、適合率(正解記事数/放送されると予測した記事数)を表 4 に示す。従来手法では、適合率は 85.0%と高いが、再現率は 34.7%にとどまっており、改善の余地があると言える。提案手法を用いることで新規に 12 件のニュースを正しく放送有りとして検出することができ、適合率は 87.8%に、再現率は 59.2%にそれぞれ向上した。これらの 12 件のニュースの特徴として、12 件のニュースに対する平均的な Tweet 数は 67 件であり、これは放送のあったニュースの平均的な Tweet 数 131 件と比較して明らかに少ないことが分かる。従来手法では、ニュースのジャンル、話題語、Tweet 数など表層的な情報の取得にとどまっており、肯否極性のような内容に基づく分析も精度が十分に得られないため、識別器による重み付けが表層的な情報に偏る傾向があった。提案手法を用いることで、Tweet 数は少ないながらも内容的にはユーザの興味が大きいようなニュース記事に対する予測が可能となり、再現率の向上につながったものと考えられる。

## 5　まとめ

本稿では、ソーシャルメディアを利用して実世界の事象を予測するため、顔文字や特有の表現など、ソーシャルメディア上の情動表現に着目した手法を提案した。性能評価実験によって、情動表現を考慮しない従来の手法と比較して再現率および適合率の向上を確認した。今後は商品の販売実績データなどと紐づけた分析を行うことで、提案方式の応用先の検討や適用可能性の検証を進める。

### 参考文献

[1] J. Bollen et.al, "Twitter mood predicts the stock market." Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

[2] S. Asur et.al, "Predicting the future with social media." In Proc. of WI-IAT, vol. 1, pp. 492-499, 2010.

[3] 池田和史 他, "Twitter の反応を用いたニュース記事のインパクト予測手法", 信学総大, D-4-3, 2012.