Web コメントを用いた社会の風潮の測定手法に関する研究

公共システム研究室　戸田雅生

1. はじめに

近年, 感染症や自然災害といった未曾有の災禍が発生しており, 2019 年 12 月中国で確認された新型コロナウイルス感染症（COVID-19）は世界的に大流行し, 人々の生活や経済活動に対し多大な影響を及ぼしている. このような緊急事態において, 政府は感染拡大を抑止するための対応と, 生活状況に応じた対策を迅速に行う必要がある.

感染が拡大している中で, 人々は行動制限を強いられ, 以前の日常生活を続けられない状況にある. このような状況下においては, 人々には不安や不満などの様々な感情が広がることが予測される. 本研究では, このような人々の感情の総体を「風潮」と呼ぶ. 風潮には一過性のものもあれば, 解消されずに人々の心理や行動に悪影響を与えかねないものもある. そこで, 風潮の状態を定量的に把握することができれば, 対策を講じる手立てになると考える.

人々の心理的状態を把握する 1 つの手段として質問紙調査がある. しかし, 緊急事態下においては, 時間と予算の制約により, 大規模な社会調査を実施することは難しい. 一方で, インターネット上には, 日々のニュース記事に対して人々が様々なコメントを投稿しており, さらにコメントに対する賛否を確認することができる. つまり, 大量のテキストに表れる感情と, それらに付随する情報を用いて, 風潮を明らかにすることが可能である.

本研究では, Web サイトに投稿されたコメントを対象として, テキスト解析により社会の風潮を測る手法を開発する. 1 年間の新型コロナウイルス感染症に関する記事で実証する.

2. 本研究の考え方

2.1　風潮の定義

人々の感情を定量化する方法に感情分析(Sentiment analysis)がある. テキストから特徴的な単語あるいは表現を抽出し, それらを総合的に評価することにより, 文書の書き手の感情や態度などを明らかにする方法である. 製品やサービスに対するレビューの評価に用いられている. ここで, 単語の特徴を表す指標として単語感情極性対応表がある. 単語に–1 から+1 の実数値が与えられ, +1 に近い単語ほど良い印象(positive)であることを示す.

本研究では, この対応表を用いてまずコメント 1 つ 1 つの感情値を算出する. 次に感情値を総合的に評価した値を風潮と考え, 日にち毎の感情の平均値および標準偏差を指標とする. ここで, 感情値を算出するにあたり, コメントの投稿者だけでなく閲覧者の反応も考慮する. 具体的には, コメントの内容に対する同意・非同意の数を用いる.

2.2　風潮の評価

風潮の評価の方法について述べる. 平均と標準偏差の 2 軸に図示し I から IV に分類する.

I　ポジティブで標準偏差が大きい. 記事に対してネガティブやポジティブ, さまざまな感情がある中で, 比較的ポジティブな感情が多い.
II　ネガティブで標準偏差が大きい. 記事に対してネガティブやポジティブ, さまざまな感情がある中で, 比較的ネガティブな感情が多い.
III　ネガティブで標準偏差が小さい. 記事に対してネガティブな感情ばかりである.
IV　ポジティブで標準偏差が小さい. 記事に対してポジティブな感情ばかりである.

ここで, 望ましい状態は, どんな内容の記事やコメントにもポジティブ（肯定的）な感情の状態（IV), またはネガティブ（否定的）とポジティブ（肯定的）な感情がどちらも存在している状態である(I および II). 一方で望ましくない状態は, どんな内容に対してもネガティブ（否定的）な感情の状態である（III). 本研究では, 1 日ごとにコメントの感情の状態をプロットし, 月別および話題別に比較することを試みる.

3. 分析手法

d 日目における i 番目のコメントの感情値 e_{di} を定義する. 以下では d 日目における i 番目のコメントをコメント di と表記する. コメント di に含まれる単語 j は感情極性 s_j を持つ. コメント di において単語 j が f_j 回出現したとき e_{di} を以下のように定義する. コメント di に出現する単語の集合を A_{di} で表す.

$$e_{di} = \sum_{j \in A_{di}} f_j s_j \qquad (1)$$

ここで, コメント di の「そう思う」の数を g_{di}, 「そう思わない」の数を b_{di} とする. これらを考慮した感情値を次のように定義する.

$$E_{di} = e_{di}(1 + g_{di} - b_{di}) \qquad (2)$$

本研究では 1 日ごとに世間の感情の状態を求める. 世間の感情の状態は, その日における世間の感情値の平均と標準偏差という 2 つの指標で代表する. まず, ある d 日のコメント数の合計を N_d とすると, d 日における世間の感情値の平均は次式で表される.

$$\overline{E_d} = \frac{\sum_{i=1}^{N_d} E_{di}}{N_d} \qquad (3)$$

同様に, 標準偏差は次式で表される.

$$\sigma_d = \sqrt{\frac{\sum_{i=1}^{N_d}(E_{di} - \overline{E_d})^2}{N_d}} \qquad (4)$$

人々の感情の状態が大きく変化した日を把握する．まず，平均感情値がポジティブ（またはネガティブ）な方向に変化したのかは，次式から判定する．

$$\bar{E}_d - \bar{E}_{d-1} = \begin{cases} > 0 & positive \\ < 0 & negative \end{cases} \qquad (5)$$

次に，感情値が偏った（またはばらついた）のかは，次式から判定する．

$$\sigma_d - \sigma_{d-1} = \begin{cases} > 0 & variability \\ < 0 & uniformity \end{cases} \qquad (5)$$

4. 分析結果

2020 年 6 月 1 日から 2021 年 5 月 31 日までの 1 年間を対象期間として，「コロナ」と「政府」の 2 つの単語をタイトルまたは記事内容に含む国内の記事を収集した．記事数は 9,089 件，コメント数は 1,887,994 件であった．図 1 に記事数と新規感染者数の推移を示す．相関係数は 0.61 であり，記事数と新規感染者数には相関があることがわかった．

2021 年 5 月の感情の状態を図 2 に示す．縦軸は感情値のばらつきを表す標準偏差，横軸は平均感情値である．II と IV に分類される日にちが多い．ポジィティブに集中する状態および，ネガティブにばらついた状態が多いことがわかった．

2021 年 5 月において，平均感情値が最も大きく変化したのは 5 月 7 日から 5 月 8 日であり，ネガティブな方向に変化した．平均感情値および標準偏差が最も大きく変化したのは，5 月 24 日から 5 月 25 日である．ポジティブな感情に集中したことがわかった．

5 月 8 日の記事を確認すると，15 道県で新規感染者数が過去最多で大都市圏から地方へ感染が広がっている内容の記事に対し，国民の行動が制限されている中で議員などが会食をしている点やオリンピックを開催する方向など政府側の行動や方針などをネガティブな表現で批判しているコメントが見られた．そのコメントに対して約 7 万件の「そう思う」の投票があった．

図 3 には，Go To キャンペーンに関する感情の状態を示す．1 年間を通して，平均感情値および標準偏差が最も大きく変化したのは，2020 年 7 月 18 日から 7 月 19 日であった．Go To キャンペーンに対する首相の説明責任を追及する記事に対し，ネガティブなコメントが見られた．そのコメントに対し約 9 万件の「そう思う」の投票があった．Go To キャンペーンのトラベル事業の開始は 7 月 22 日であり，事業開始前に社会にはネガティブな感情に集中していたことがわかった．

5. おわりに

本研究では，Web サイト上のニュース記事に投稿されたコメントを感情分析することで社会の

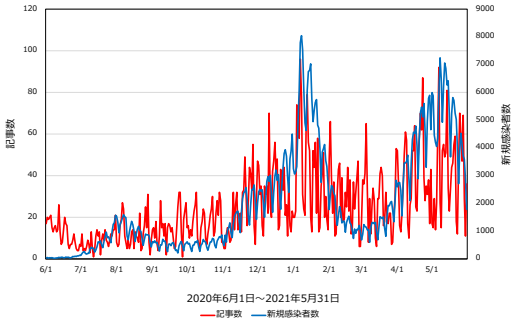

図 1　記事数と新規感染者数

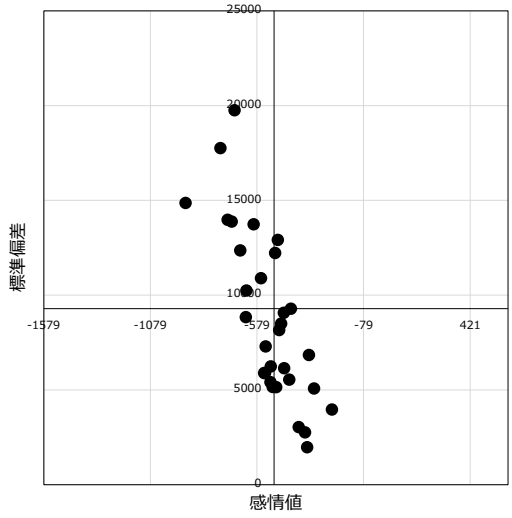

図 2　2021 年 5 月

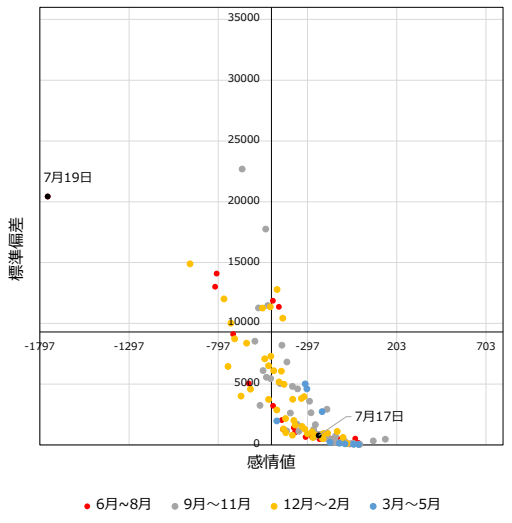

図 3　Go To キャンペーン

風潮を測る手法を開発した．平均感情値と標準偏差により 1 日の社会全体の感情の状態をプロットし，どのような変化が生じたのかを明らかにした．前日と比較することにより，感情の状態が変化する様子を可視化することができた．今後の課題としては，記事の内容への賛否を風潮に反映させることである．