ニュースサイトのコメントを用いた災害に対する社会的関心の抽出～新型コロナウイルス感染症を例として～

公共システム研究室　田中哲哉

1. はじめに

感染症や大規模な自然災害といった事態に対し政府は早急に様々な政策を講じる必要がある．その際，地域の実情や人々の関心を把握したうえで政策に反映させることが望ましいが，大規模な社会調査を行うことは困難である．ここで，Web 上には日々多様なニュースが投稿され，人々がそれらに対しコメントをしている実態に着目すると，ニュースサイトのコメントから人々の関心を把握できる可能性がある．しかし，コメントの数は膨大であり，さらにその内容や質は様々であることから，これらの背後にある社会的関心を見つけることは容易ではない．

そこで本研究では，コメントに対する人々の賛同に着目し，賛同を考慮したうえで大量のコメントから社会的関心を抽出することを試みる．自然言語処理の手法である BERT を用いて，2020 年 6 月から 2021 年 5 月末までの新型コロナウイルス感染症に関する記事に寄せられたコメントで実証する．

2. 本研究の考え方

Yahoo!ニュースでは，コメントに対し第 3 者が「そう思う（賛同）」または「そう思わない（反対）」を投票できる．ここで，コメントに対する賛同を賛同率と純賛同数で定義する．賛同数と反対数の合計を反応数とし，賛同数を反応数で除した値を賛同率とする．純賛同数は賛同数から反対数を引いた値である．まず，日本語学習済み Sentence-BERT を用いて，コメントの分散表現を算出する．次に，コメントを賛同に応じて「多くの賛同が得られたコメント」「賛同が得られなかったコメント」「その他」の 3 つに分ける．多くの賛同が得られたコメントに対して，x-means 法を用いてクラスター分析を行う．そして，クラスター内のコメントから代表的／特徴的なコメントを抽出しそれを社会的関心とする．最後に，クラスターに含まれるコメント数の時系列の変化を明らかにする．

3. 分析手法

3.1　分散表現の算出

BERT は，ディープラーニングを用いた自然言語処理モデルである．Sentence-BERT は，似ている文章は似た文章ベクトルになるように，既に学習済みのモデルに新たな層を追加し再学習させたモデルである．分析では，日本語学習済みのパッケージを用いて，各コメントの 768 次元の分散表現を算出する．

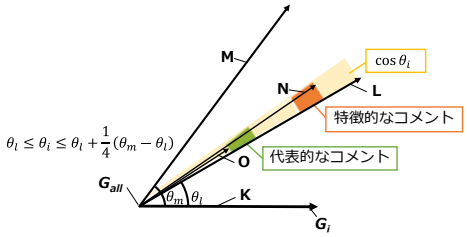

図 1　代表的なコメントと特徴的なコメント

3.2　クラスター分析

多くの賛同が得られたコメントの分散表現 768 次元に対して，x-means 法を用いてクラスター分析を行う．x-means 法は BIC 分割停止基準をもとに最適クラスターを自動で推定するアルゴリズムである．

3.3　代表的なコメントと特徴的なコメント

各クラスターの内容を最もよく表しているコメントを抽出し社会的関心を明らかにする．ここで，全てのコメントから見たときに，クラスター内の平均的なコメントと平均的ではないコメントを取り出す．それぞれ代表的／特徴的なコメントとする．図 1 にコメントの抽出方法を示す．コメントのコサイン類似度とユークリッド距離から代表的／特徴的なコメントを抽出する．

すべてのコメントの重心ベクトルを $\boldsymbol{G_{all}}$，クラスターi の重心ベクトルを $\boldsymbol{G_i}$，$\boldsymbol{G_{all}}$ から $\boldsymbol{G_i}$ へのベクトルを $\boldsymbol{K}$ とする．クラスターi に含まれるコメントの中で，$\boldsymbol{K}$ とのコサイン類似度が最大となるベクトルを $\boldsymbol{L}$，コサイン類似度が最小となるベクトルを $\boldsymbol{M}$ とする．これら 2 つのクラスターi のコサイン類似度が取りうる範囲の上位 25%以上に入るコサイン類似度を $\cos\theta_i$ とすると，$\cos\theta_i$ の取りうる範囲を次の式で表す．

$$\frac{\boldsymbol{K}\cdot\boldsymbol{L}}{\|\boldsymbol{K}\|\cdot\|\boldsymbol{L}\|} - \frac{1}{4}\left[\frac{\boldsymbol{K}\cdot\boldsymbol{L}}{\|\boldsymbol{K}\|\cdot\|\boldsymbol{L}\|} - \frac{\boldsymbol{K}\cdot\boldsymbol{M}}{\|\boldsymbol{K}\|\cdot\|\boldsymbol{M}\|}\right] \le \cos\theta_i \le \frac{\boldsymbol{K}\cdot\boldsymbol{L}}{\|\boldsymbol{K}\|\cdot\|\boldsymbol{L}\|} \tag{1}$$

$\cos\theta_i$ を満たしたクラスターi に含まれるコメントの中で，クラスターi の重心からのユークリッド距離が最大となるベクトルを $\boldsymbol{N}$，ユークリッド距離が最小となるベクトルを $\boldsymbol{O}$ とする．$\min_i d(\boldsymbol{K}, \boldsymbol{O})$ から $\max_i d(\boldsymbol{K}, \boldsymbol{N})$ までの下位 25%以下の距離を $d(X_i)$ とすると，$d(X_i)$ の取りうる範囲は次の式で表される．$d(X_i)$ を満たすコメントを代表的なコメントとする．

$$\min_i d(\boldsymbol{K},\boldsymbol{O}) \le d(X_i) \le \min_i d(\boldsymbol{K},\boldsymbol{O}) + \frac{\mathbf{1}}{\mathbf{4}}[\max_i d(\boldsymbol{K},\boldsymbol{N}) - \min_i d(\boldsymbol{K},\boldsymbol{O})] \quad (2)$$

また，$\min_i d(\boldsymbol{K}, \boldsymbol{O})$から $\max_i d(\boldsymbol{K}, \boldsymbol{N})$までの上位25%以上の距離を $d(Y_i)$とすると，$d(Y_i)$の取りうる範囲は次の式で表される．$d(Y_i)$を満たすコメントを特徴的なコメントとする．

$$\max_i d(\boldsymbol{K},\boldsymbol{N}) - \frac{1}{4}[\max_i d(\boldsymbol{K},\boldsymbol{N}) - \min_i d(\boldsymbol{K},\boldsymbol{O})] \le d(Y_i) \le \max_i d(\boldsymbol{K},\boldsymbol{N}) \quad (3)$$

4. 分析結果

4.1 クラスター分析の結果と解釈

本研究では，反応数が 10 以上のコメント 471,890 件を分析対象とし，768 次元の分散表現を算出した．図 2 に賛同数と賛同率のヒストグラムを示す．賛同率の最頻値は 0.8 となっていることから，賛同率 0.8 以上かつ純賛同数 5000 以上のコメントを多くの賛同が得られたコメントと定義した．該当するコメントは 2,605 件であり，最適クラスター数は 11 となった．表 1 にクラスターの解釈の結果を示す．

表 2 にクラスター5 から一部抜粋したコメントを示す．表 2 よりクラスター5 の代表的なコメントは「ワクチン（運用・副反応・効果）に関する意見」と解釈でき，これらの意見に多くの人々が賛同していたことがわかった．特徴的なコメントは「ワクチン（副反応・運用）に関する意見」と解釈でき，これらの意見に多くの人々が賛同していたことがわかった．本手法により，クラスターの話題のみならずどのような意見があったのかを明らかにできたと考えられる．

4.2 クラスター分析の可視化と時系列変化

図 3 にクラスターの UMAP の結果を示す．クラスター毎にコメントが分布していることがわかる．また，政治に関するクラスター1 と 2 は第 1 象限に，海外に関連するクラスター9 と 11 は第 4 象限に，特にコロナ対策と関連するクラスター3，4，5，7，10 は第 3,4 象限に位置することが明らかとなった．

図 4 にクラスターの時系列変化を示す．クラスター3（感染拡大に対する政策等への批判）は，2020 年 7 月，11 月，2021 年 4 月において前月より割合が大きく増加した．クラスター5（ワクチン）は，2020 年 7 月，9 月，2021 年 2 月～5 月にかけて出現しやすかった．クラスター9（五輪開催に対する批判）は，2020 年 10 月，2021 年 2 月～5 月に出現しやすかった．これらのクラスターの変遷は，Go To キャンペーンや東京オリンピック開催といった世の中の動きと関連していることが示唆された．

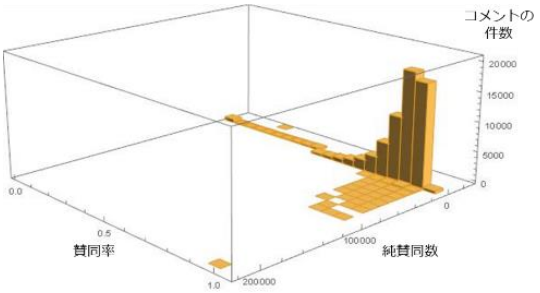

図 2　賛同数と賛同率のヒストグラム

表 1　クラスターの解釈

	件数	代表的なコメント（件数）	特徴的なコメント（件数）
1	362	菅総理に対する批判(5)	自民党に対する批判(5)
2	337	菅総理の説明責任(3)	説明に対する意見(5)
3	350	感染拡大に対する政策への批判(7)	発言に対する批判(1)
4	168	医療崩壊に関する意見(2)	医療崩壊に関する意見(4)
5	151	ワクチン（運用・副反応・効果）に関する意見(7)	ワクチン（副反応・効果）に関する意見(3)
6	169	人材に関する意見(4)	労働環境に関する意見(3)
7	162	飲食業の自粛・支援に関する意見(10)	制度のズルを批判(1)
8	253	税金の使途に対する批判(8)	税金の徴収・使途に対する批判(9)
9	194	五輪開催に対する批判(4)	五輪中止の意見(3)
10	259	GoTo や緊急事態宣言に対する批判(9)	緊急事態宣言の時期に対する批判(3)
11	200	日本と外国との関係に対する批判(15)	日本に対する敗北感(1)

表 2　クラスター5 のコメント（一部抜粋）

代表的なコメント	特徴的なコメント
・ワクチンの副作用が怖いです ・皆さんには，積極的なワクチン接種について考えていただきたいです ・集団ワクチンの目的は「迅速な集団免疫の獲得」なので，公平性も重視しつつ，効率よく接種することが重要だと思います	・ワクチン接種にどのような副反応(副作用)が存在し，どれだけのリスクがあるか現時点では分からない ・高齢者のワクチン接種を担当する医師と看護師が，先にワクチン接種してから，高齢者に打つようにすべき

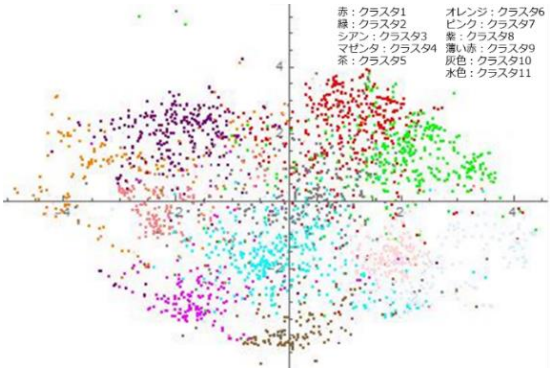

図 3　クラスターの UMAP の結果

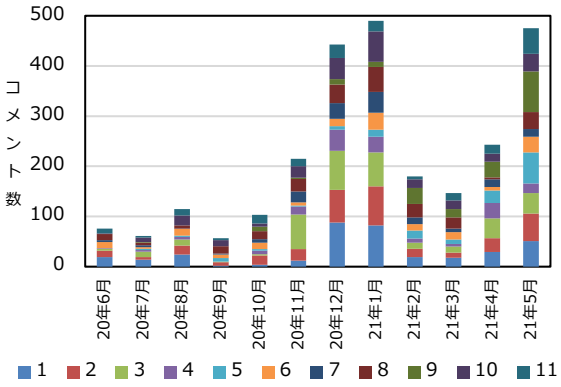

図 4　クラスターの時系列変化