Jpn. J. Health Phys., **55** (1), 15 〜 22 (2020) DOI: 10.5453/jhps.55.15

ノート

ソーシャルビッグデータを活用した放射線被ばくに対する不安意見の解析システムの開発

峰松　優*1，藤淵　俊王*2,#，有村　秀孝*2

（2019 年 10 月 25 日受付）
（2019 年 12 月 20 日採択）

Development of Method using Sentiment Analysis for Anxiety Opinion of Radiation Exposure by Social Big Data

Yu MINEMATSU,*1 Toshioh FUJIBUCHI*2,# and Hidetaka ARIMURA*2

A lot of people concerns about radiation exposure by Fukushima nuclear accident. But no method that we effectively understand people's anxiety because we grasped people's needs about radiation exposure with individualism (ex. Telephone, questionnaire). In this paper, we suppose that it is possible to efficiently collect more people's needs by utilizing twitter, which is one of Social Network Service, easier than before. Firstly, we search tweet including words of ("radiation" and "exposure"), normalize unicode, process pointless words (ex. URL, html tags, @NAME). Secondly, we extract tweets with anxious words dictionary that there are low dependency on texture and annotate 999 tweets to Anxious or Nonanxious. Thirdly, we extract tweets by sentiment analysis. Finally, we analysis tweets by utilizing KHCoder. In this dataset, about 40 % opinion data are anxiety about medical exposure. We infer that people are concerned about radiation exposure related to medical treatment, which is relatively close to everyday life. In this paper, we developed the system, which from collecting anxious opinion about health effects due to radiation exposure by using Twitter to doing dataset tendency analysis on about 60% Precision and Recall, by utilizing anxiety dictionary and sentiment analysis.

KEY WORDS: radiation, exposure, anxiety opinion, social big data, Twitter, sentiment analysis, text mining.

I　緒　　論

2011 年 3 月 11 日に発生した東日本大震災後の福島第一原子力発電所の事故以降，多くの市民が放射線の被ばくに不安や懸念を持つようになった。市民が放射線被ばくに対する懸念の対象はさまざまである。例えば，医療機関では被ばくに関して不安を抱いている人々（市民，患者，医療従事者等々）がすでに高い割合でいるという報告がある[1–3]。実際に，各医療機関は，国際放射線防護委員会（International Commission on Radiological Protection; ICRP）による勧告に基づき，各被検者の被ばく線量を管理している。ただ，各市民の被ばくに対するリスクの認識は多様であるため，医療従事者は患者の医療被ばくに対する不安を適切に和らげる必要性がある。しかしながら，放射線被ばくの健康影響に対する異なる認識があることや放射線被ばくの専門家の数が限定的なためマンパワーの限界があること[4]，リスクコミュニケーションの難しさ[5]が課題となり，被ばくに対する市民のニーズに適切に応えることが難しい現状があるとい

*1 九州大学大学院医学系学府保健学専攻医用量子線科学分野；福岡県福岡市東区馬出 3−1−1（〒 812−0054）
Department of Health Sciences, School of Medical Sciences, Kyushu University; 3−1−1, Maidashi, Higashi-ku, Fukuoka-shi, Fukuoka 812−0054, Japan.
*2 九州大学大学院医学研究院保健学専攻医用量子線科学；福岡県福岡市東区馬出 3−1−1（〒 812−0054）
Department of Health Sciences, Medical Research Institute, Kyushu University Graduate School; 3−1−1, Maidashi, Higashi-ku, Fukuoka-shi, Fukuoka 812−0054, Japan.
Corresponding author; E-mail: fujibuch@hs.med.kyushu-u.ac.jp

える。

われわれは，膨大なデータから市民の健康不安を把握することができれば，対策を打ち出しやすくなると考えた。従来は，市民の被ばくによる健康不安に対する意見を集める際に，放射線事故や放射線診療などに対する被ばく相談では各医療機関や行政などが電話対応やアンケート，掲示板投稿などで市民のニーズを知り得てきた。しかし，これには手間や時間もかかり限られた内容しか集めることができず，一般的な質問への対応になっていることが多かった。

本研究では，ソーシャルネットワークサービス（Social Network Service; SNS）の一つである Twitter を用いることで，市民の放射線による健康不安の意見を幅広く効率的に抽出できるのではないかと考えた。Twitter には，膨大なデータがあり，老若男女問わず幅広い年齢層の利用者が存在し，基本的に匿名で使用されるため，ユーザーの本音が多く投稿される傾向にあるのも特徴である[6, 7]。このような Twitter の特徴を踏まえると，Twitter から市民が被ばくに対してどのような不安を抱いているかを幅広く詳細に知ることで，個人の被ばくへの不安を解消する対策を立てやすくなると期待できる[8, 9]。具体的には，被ばくへの不安意見自体にどのような特徴や傾向があるのかを把握することができれば，重点的に対策を打ち出す領域の根拠を示すことに繋がる[10]。

Twitter の投稿から被ばくに対する不安意見を抽出するにあたっては，不安用語を投稿内の文章に含んでいるが，実際に不安を抱いている文章かどうかを判定したうえでデータをいかに抽出するかが問題となる。Twitter には 1 投稿あたり 140 文字以内という制約がある中で複数の文章を書くことや複雑な日本語で書かれていることが多い。そのため，投稿内容自体が不安を抱いているのかどうか判別することが難しい。

そこで本研究では，（1）「放射線」と「被ばく」を含んだツイートに対して不安用語を含むツイートを抽出する（2）不安用語を含むツイートがネガティブな感情を表しているかを感情分析で判定する，放射線による健康不安の意見収集および解析手法を提案する。

II 方　　法

本研究における提案手法の枠組みを **Fig. 1** に示す。まず，「放射線」と「被ばく」（表記ゆれを含む）というキーワードでツイートを収集およびスクリーニングを行った。期間は，2018 年 7 月 1 日～ 2018 年 12 月 31 日までとした。次に，上記のデータセットから三浦氏ら[9]によるソーシャルビッグデータにおける災害情報の伝播と感情の調査の際に用いられた不安用語辞書を採用して不安用語を含むツイートを抽出した。さらに，抽出したツイートに感情分析を使って分析することで不安意見かどうかを判定した。この時，999 件のツイートが不安を意味しているか否かを判別するラベル付けを行うことでデータセット内の不安意見かどうかの感情値の閾値を評価した。最後に，テキストデータ分析ツールである KHCoder を使って不安意見とみなしたデータセットの分析を行った。

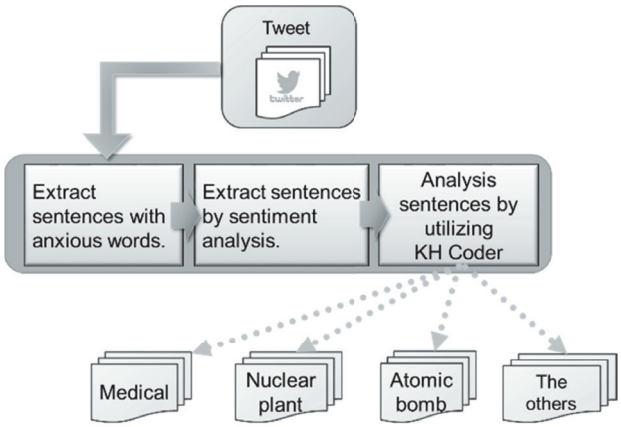

Fig. 1　An overview of the proposed method.

1. ツイートの収集

はじめに，2018 年 7 月 1 日～ 2018 年 12 月 31 日の期間中に「放射線」と「被ばく」という単語を文章中に含むツイートを収集した。この時，各単語のひらがな，カタカナ表記と被ばくに関しては「被爆」と「被曝」も同様に表記ゆれとしてツイートを検索し，計 395,423 件収集した。

2. 不安用語辞書による文章抽出

次に，ツイートに含まれる不安という感情反応を抽出するため，不安用語辞書を用いた。この不安用語辞書は，東日本大震災を元に災害発生後の感情語の出現傾向等を Twitter から分析した研究によって作成されている。本辞書を用いて不安用語を含むツイートを抽出し計 2,983 件収集した。

3. 感情分析による文章の極性評価

感情分析は，書かれた文章の意見を判断することで，意見自体が元気で活動的で生き生きとしたポジティブな

ものか，反対にネガティブなのか，それともニュートラルなものなのかを，一連の単語から判断し分類することができる。実際に，感情分析は単語のポジティブ，ネガティブを数値化した極性辞書が公開されており，SNSや顧客の声のレビュー等にある世間の意見を感情分析することで，人々が実際に何を考えているかを突き止めるのに役立てられている。

Twitterのデータでは，不安用語を含むだけで不安意見とみなして分析を行ったとしても，Twitterの140字以内であれば複数文同時に書くことができることや複雑な日本語の使用方法が原因となり不安意見でないツイートを含んでいる可能性が高い。そのため，実際にネガティブな感情を含んだツイートをいかに収集できるかが重要である。そこで，文章の感情の値を定量的に分析することができる感情分析を活用し，ツイート上の不安意見だとみなすことができる感情値を定めることで不安という感情を含む文章の抽出を試みた。具体的な評価実験は，後方に記す。この時，ツイートは1,528件になった。

4. KHCoderによるテキスト分析

今回作成したデータセットを分析することで，Twitterを活用した放射線による市民の不安意見を効率的に把握できる可能性を探った。詳細に傾向を分析する上で，人が手作業で分析を行うと恣意的な分析になる可能性が存在する。その場合客観性が損なわれた分析になるため，今回はデータを要約し掲示する際に，手作業が省かれるため分析者の持つ考えや理論からなる偏見を排除することができるためKHCoderを用いた[12]。今回は，データセットから得られた頻出後の出現パターンの類似性や語同士の繋がりを可視化することができる共起ネットワークを用いた。本研究では，データ量が最低必要とされる100以上を上回ること，定量的な分析をクラスタリング分析で補うことが可能と考えて本手法をとった。共起ネットワークにより単語の結びつきを視覚化し，単語同士の結びつきの強さを計算することでどのような被ばくに対して不安を抱いているのか文章群を推測した。さらに，分けられた文章群それぞれに対して単語の出現頻度パターンを把握することで，更に詳細に被ばくへの対象を探った[13]。本研究では，データ数が2,000以下と中規模であること，被ばくへの不安からまずは複数のカテゴリの分類傾向を把握することが目的であることから，比較的細かいネットワークを書くことが必要であると考えた。さまざまな条件で繰り返し分析を試みた結果，有意かつ明確なネットワークを書くと考えられた出現頻度34以上の語で，共起するJaccard係数を0.095に設定して描出した。

また，共起ネットワークでは分析結果が相関関係であるため，定量的な判断ができない。そのために，階層型クラスタリングを用いることで定量的に出現が似通っている語群を分析し，共起ネットワークと併せて不安文章の文章群を推察した。階層型クラスタリングは，分析の対象となる個体を，お互いの類似度に従っていくつかのグループに分割する手法である。クラスタリング分析には，クラスターの結合方法がWard法，群平均法および最遠隣法の3つ，対象間の類似度の定義が，Jaccard係数，Euclid距離およびCosine係数の3つが用意されている。本研究では，語と語の共起関係を重視することで文章群の推察を行っていることから，クラスターの結合方法は最も明確なクラスターを作り，分類感度が高い特徴を有し，実用性が高いWard法を，距離係数については，今回用いた共起ネットワーク分析における共起関係の強弱についてJaccard係数によって計算されていることから，Jaccard係数を用いて分析を行った。

III 評 価 実 験

1. 評価対象

今回収集したデータセットの不安用語辞書によって抽出された2,983件のうち999件を使用する。また，同研究室内の学生3人の人手で不安の対象が被ばくや放射線であるツイートを不安意見とし，それ以外を非不安意見と二値に分類した。以下に一例を示す。

（不安例）

「被ばくって大体5–10年後に発症するらしいじゃん，めっちゃ怖い…」

（非不安例）

「花火は，光るだけでもいやだから見ないという被爆者のかたも多いらしい。花火の音や光るだけでも怖いって，戦争って本当に恐ろしかったのですね。」

2. 評価方法

感情分析を正解データと不正解データに実行することで，放射線や被ばくに対する不安意見を効率的に抽出できる感情値を評価した。日本語の言語処理における事前処理として用いられる形態素解析と構文解析にはMeCabとCabochaを用いた[8]。今回の感情分析による感情値の設定には，日本語評価極性辞書[11]を採用し，ポジティブな感情を+1.0，ネガティブな感情を−1.0，中立な感情を0として+1.0 〜 −1.0の範囲でツイート全体

の感情値を求めた。各データの感情値を計算後，正解データである不安意見であるデータセットが比較的 −1.0 側に寄った分布を示すだろうと仮定して，度数分布にて表示し，適切な感情値を式（1）より定めた。感情値の閾値を設定した後に，正解データの抽出精度の評価結果の考察に適合率（Precision）式（2）と再現率（Recall）式（3）を使用した[8]。それぞれの計算方法について，以下に示す。また，上記の結果から ROC 曲線を求め，AUC（Area Under the Carve）を算出した。

（ツイートの感情値）

$$= \frac{\text{ツイート中の感情語} \times \text{単語の極性値の総和}}{\text{感情語数の総和}} \quad (1)$$

（例文）あの映画の主人公哀れだった。もっと愛情があると和んだはず。

（パースした例文）あの 映画 主人公 哀れ だ った 。もっと 愛情 が ある と 和んだ はず 。

* 下線有が感情語に該当

（結果例）例文の感情値 $= (-1 + 1 + 1)/3 = 0.333$

$$\text{Precision} = \frac{\text{抽出した正解ツイート数}}{\text{抽出したツイート数}} \quad (2)$$

$$\text{Recall} = \frac{\text{抽出した正解ツイート数}}{\text{全ての正解ツイート数}} \quad (3)$$

IV 結 果 と 考 察

1. 感情分析による文章抽出

III 評価実験による正解データと不正解データ群の感情値ごとの頻度分布を **Fig. 2** に，ROC 曲線を **Fig. 3** に示す。感情値（−0.3）の時の適合率は 0.584 および再現率は 0.577 になった。また，不安意見と非不安意見への分類結果は，3 人の分類の一致度は約 6 割になった。**Fig. 2** より正解データが感情値の負側に偏っていることがわかる。そこで，今回感情値の（−0.3）ポイント以下に属するツイートを不安意見として定義した。この時の AUC は，**Fig. 3** より 0.570 であった。要因として，2 つのことが考えられる。一つ目は，人手で正解データと不正解データにタグ付けし分別を行った際に，3 人の間で不安意見かどうかの判断に差があったことが挙げられる。明確に不安意見かどうか判断できるルールを設けることで正解データの作成が可能ではある。しかしながら，本データセットではデータが少なく抽出できる文章数が低下することから分析が困難になると懸念される。二つ目は，ツイートの文脈を理解することが難しいことが挙げられる。ツイート中で不安用語と放射線や被ばくという単語が主従関係にあるかを判断するモデルを構築することで抽出精度が向上すると考えられる。しかしながら，ツイートという複雑な日本語の特性上，不安用語と放射線や被ばくという単語が主従関係にないが放射線や被ばくに対して懸念を抱いていると考えられる文章も多々存在するため，困難が生じる。一例を下記に示す。

（主従関係にない例）「放射能被爆は遺伝子を傷つけるらしい。恐ろしいよ。」

（主従関係にある例）「がん検診でかなり被曝するって怖いよ。検査し続けたら相当なものなのか？」

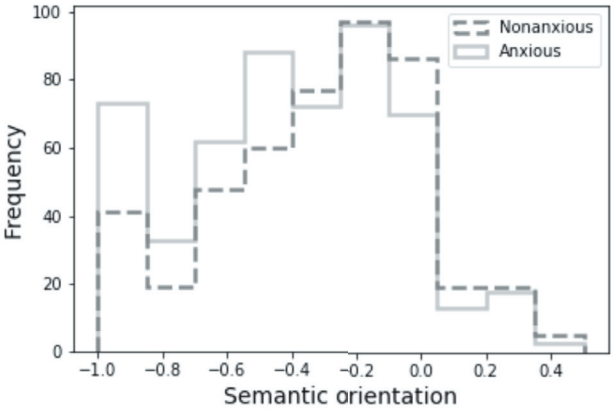

Fig. 2 Frequency distribution of sentiment analysis of the data set.

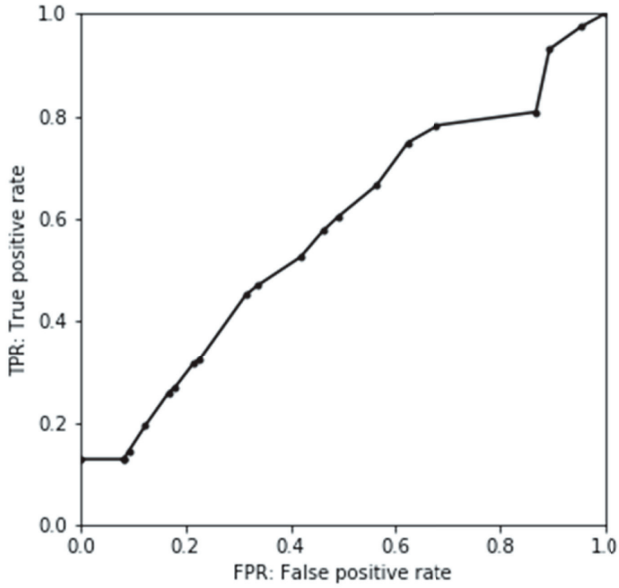

Fig. 3 ROC curve of sentiment analysis of the data set.

2. KHCoder によるテキスト分析

KHCoder を活用し III 評価実験により抽出したデー

タセットを 2 パターンの分析をした。

i　共起ネットワークによるカテゴリ推測

共起ネットワークによる結果を **Fig. 4** に示す。**Fig. 4** より大きく 4 つのカテゴリ（医療，原発，原爆，電磁波）に分類することができた。電磁波による健康不安を抱いている市民がいることがここでわかる。市民の新たなニーズを把握する可能性があることも考えられる。さらに，データを増やすことやツイート内の地理的条件等の情報を活用することで幅広いニーズかつリアルタイムでの分析も可能になる[14]。

（1）【医療】からは，放射線によるがん治療時，一般撮影などの X 線検査時に対するリスクを不安に思う人々がいることが明らかになった。

（2）【原発】からは原発事故による放射線影響に不安を想う人々がいることが明らかになった。放射線による内部被ばくや土壌等の汚染に関心があった。

（3）【原爆】からは，戦争経験者や被爆 2 世による影響により，不安を感じる人々がいることが明らかになった。

（4）【電磁波】からは，新たなニーズとして電磁波に対する身体影響を不安に想う人々がいることが明らかになった。

ii　階層型クラスタリングによる分析結果

階層型クラスタリングによる結果を **Fig. 5** に示す。その時の各クラスタの単語出現頻度を **Table 1** に示す。**Fig. 5** より医療（A，C），原発（D），原爆（B）に関心があると推測することができる。医療に関して A と C の 2 群にグルーピングされた理由は，放射線と被ばくという 2 つの語それぞれに特徴的な単語が存在したからだと考えられる。放射線という語に関しては，がんに対する放射線治療による副作用の影響を懸念する意見が多く，被ばくという語に関しては，一般撮影や CT 検査等による医療被ばくの影響を懸念する意見が見受けられた。また，E や F グループでは電磁波や内部被ばく等に

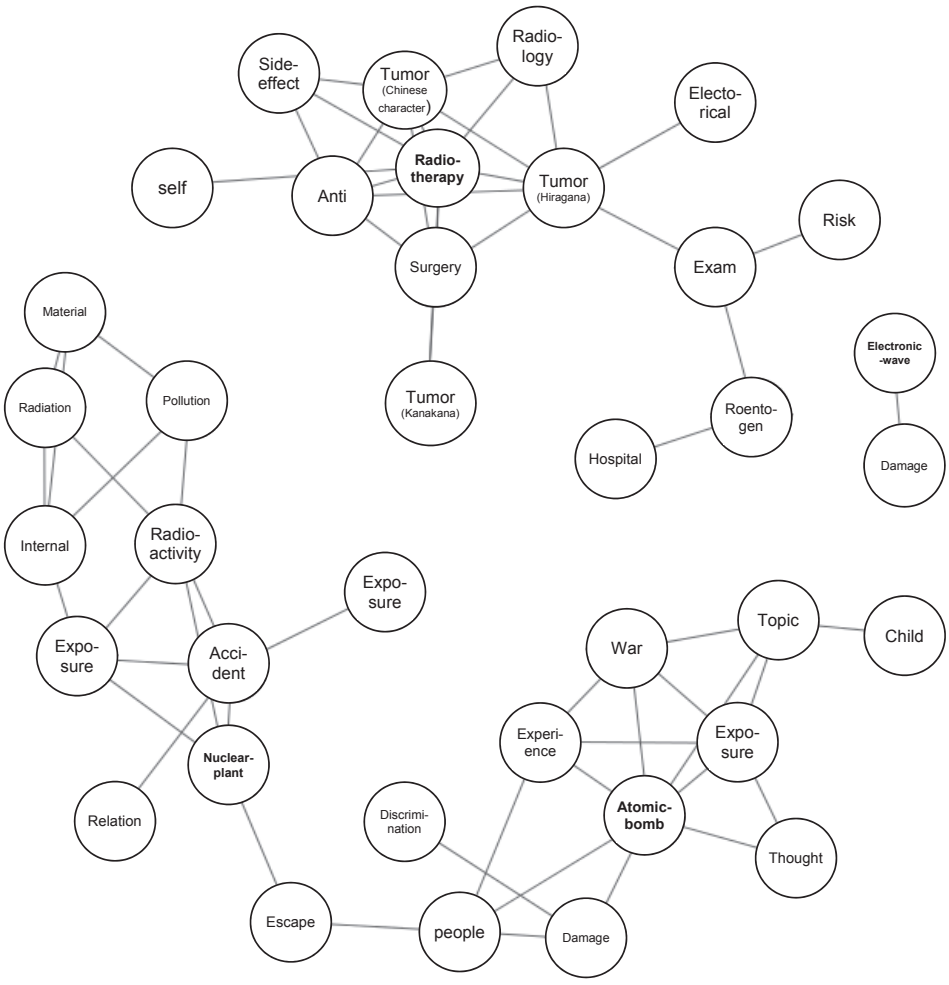

Fig. 4　Co-occurrence network of the data set.

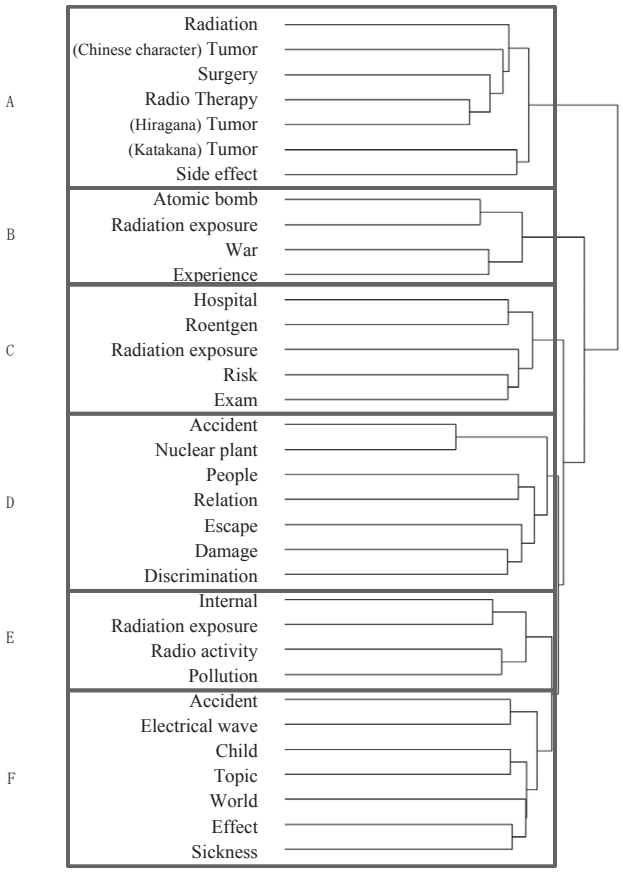

Fig. 5 Hierarchical clustering of the data set.

Table 1 List of existing words by hierarchical clustering.

Cluster category	Extraction words	Number of appearances
A	Radiation	910
	Tumor (Chinese character)	112
	Surgery	57
	Radio Therapy	242
	Tumor (Hiragana)	152
	Tumor (Katakana)	60
	Side effect	47
B	Atomic bomb	151
	Radiation exposure	552
	War	62
	Experience	48
C	Hospital	54
	Roentgen	45
	Radiation exposure	41
	Risk	40
	Exam	60
D	Accident	138
	Nuclear plant	132
	People	64
	Relation	41
	Escape	37
	Damage	61
	Discrimination	45
E	Internal	73
	Radiation exposure	425
	Radio activity	97
	Pollution	60
F	Accident	41
	Electrical wave	37
	Child	56
	Topic	115
	World	38
	Effect	56
	Sickness	51

よる身体影響を不安に思う意見も存在することが把握できた。医療被ばくに対するクラスターが 2 群あることから医療被ばくに対する懸念が他の群より詳細に抽出されており，市民からの関心度合いが高いと考えられる。

iii　推測した文章群ごとの割合と単語出現頻度

4 つのカテゴリの健康影響に対する不安意見の割合と不安意見に使われる頻出用語を **Table 2** に示す。今回のデータセットからは，医療に関する不安が最も多いことがわかる。時間経過により原発による影響が薄れ，比較的日常に近い医療に関する被ばくに対して不安を抱く市民が多いと考えられる。

Table 2 The major keywords as for health effects in each category.

Medical 640/ 1,528 (about 42 %)
X-ray, Cancer, Radiotherapy, Side effect, Risk, CT, etc…
Nuclear plant 224 / 1,528 (about 15 %)
Internal exposure, Pollution, Radioactivity, Escape, etc…
Atomic bomb 306 / 1,528 (about 20 %)
Child , War, Genetic.
Electromagnetic wave 48 / 1,528 (about 3 %)
Electromagnetic wave, MRI

3.　本研究の限界および今後の課題

本研究は，Twitter 上のツイートのテキストデータを使用しており，ツイートの検索単語，ツイートの煩雑な文章の影響を受けることが避けられない。「放射線」,「被ばく」という検索ワード以外の単語を含めると本研究の結果と差異が生じることや，ツイートは書籍や論文のように文章の構成が整っていないことが知られており，不安文章の分類時に係り受けや文脈を考慮できていないという限界があった。

また，今後は，データ量を増加するほか，地理情報，性別等のユーザーの情報を活用して分析を行うことで地域ごとや性別ごと，年齢別などによる分析も有効と考えられる。

課題としては，不安意見をどのように定義付けするこ

とで高い抽出精度を向上ができるか，各カテゴリ内の不安意見を詳細に分析できる手法の開発が挙げられる。前者に関しては，今回は放射線や被ばくに対しての不安意見だと認めることができたらタグ付けを行うというルールを設けていたが，観察者間でばらつきが見受けられた。精緻なルールを設ける必要がある。後者に関しては，今回は各カテゴリ内でどのような単語がよく使われているかを把握できたが，更にデータを増やし詳細に不安意見を分析することができれば市民のニーズを正確かつ幅広く知ることができる。

V 結　　論

本研究では，Twitter を用いることで市民の放射線による健康不安を効率的に集め，共起ネットワークと階層的クラスタリング分析により不安の傾向や新たなニーズも把握できることを示した。本研究で用いたデータでは，放射線による不安意見は「医療」と「原子力発電」と「原子力爆弾」と「電磁波」に大別され，「医療」に関する不安文章が最も多いことが明らかになった。各カテゴリの主要な不安ニーズを把握し，重点的に対策を行うことで，被ばくの不安解消を効果的に行うことができると考えられた。

謝　　辞

九州大学システム情報科学研究院所属の内田誠一様および同研究室の皆様には，本研究におけるテキストデータの処理方法について多くの御助言ご指導を頂きました。深謝申し上げます。

利益相反の開示

開示すべき利益相反状態はない。

参 考 文 献

1) A. FURUNO, H. TAKASHITA, H. TOKUNAGA and H. HORIKOSHI; Risk communication practice after the Tokyo Electric Power Company's Fukushima Daiichi Nuclear Power Station Accident ―awareness of Fukushima residents in internal dosimetry―. *JAEA-Review*, **022** (2014).
2) L. UKKOLA, H. OIKARINEN, A. HAAPEA and O. TERVONEN; Patient information regarding medical radiation exposure is inadequate: patients' experience in a university hospital. *Eur. Radiol.*, **23** (4), 114–119 (2017).
3) L. UKKOLA, H. OIKARINEN, A. HAAPEA and O. TERVONEN; Information about radiation dose and risks in connection with radiological examinations: what patients would like to know. *Eul. Radiol.*, **26** (2), 436–443 (2016).
4) S. RAJIB, Y. TAKEUCHI and S. MATSUURA; Risk communication, learning from megadisasters: lessons from the Great East Japan Earthquake, 356–367 (2012).
5) T. MOTOYOSHI and Y. YOSHIDA; Radiation risk communication after the Great East Japan Earthquake. Kansai University (2014).
6) F. TORIUMI; Big data collection on Twitter, *Jpn. J. Organi. Sci.*, **48** (4), 47–59 (2015).
7) N. KAJI and N. YOSHINAGA; Super-information society based on big data ―information technologies of searching the whole, from platform technologies to applications―: 6. natural language processing for leveraging social big data. *Jpn. J. Inf. Process.*, **56** (10), 982–989 (2015).
8) T. KAWASHIMA; Consumer needs extraction method from Twitter, *DEIM Forum*, **B5-1** (2016).
9) A. MIURA, F. TORIUMI, M. KOMORI, N. MATSUMURA and K. HIRAISHI; Relationship between emotion and diffusion of disaster information on social media: case study on 2011 Tohoku Earthquake. *Jpn. J. Artif. Intell.*, **31** (1), NFC-A_1-9 (2016).
10) N. KOBAYASHI, K. INUI, Y. MATSUMOTO and K. TATEISHI; Collecting evaluative expressions for opinion extraction, *J. Nat. Lang. Process.* **12** (3), 203–222 (2005).
11) M. HIGASHIYAMA, K. INUI and Y. MATSUMOTO; Learning sentiment of nouns from selectional preferences of verbs and adjectives, *Proc. 14th Annu. Meeting. Assoc. Nat. Lang. Process.*, 584–587 (2008).
12) K. HIGUCHI; A two-step approach to quantitative content analysis: KH Coder tutorial using Anne of Green Gables (Part I), *Ritsumeikan Soc. Sci. Rev.*, **52** (3), 77–91 (2016).
13) H. IGARASHI, M. FUKUSHI and S. HOSHINO; Analysis of human error among radiological technologists with using of the questionnaire and text mining. *Jpn. J. Health Sci.*, **13** (2), 59–70 (2010).
14) W. KASHINO, S. TACHIBANA, T. HIRAMOTO and Y. SEKI; Shiminn iken no shuushuushisutemu de erareta tui-to karano「hoikuenn」「kyouiku」ni kansuru ikenn chuushutsu, *Assoc. Nat. Lang. Process.*, 533–536 (2017) (in Japanese).

峰松　優（みねまつ　ゆう）
2018 年九州大学医学部保健学科放射線技術科学専攻卒業。同年九州大学大学院医学系学府保健学専攻医用量子線科学分野修士課程入学。2020 年同大学院修了予定。同年（株）フィリップス・ジャパン入社予定。