

一般口演 | 知識工学

一般口演9

自然言語処理・テキストマイニング

2021年11月20日(土) 09:10 ~ 11:10 E会場 (2号館2階222+223)

[3-E-1-03] Fuzzy-C-Means法を用いた COVID-19に関するセンチメント分析 ー日本語と対象とした Twitterユーザーに関する検証ー

*上田 龍一郎¹、韓 豊¹、張 洪健¹、青木 智大²、小笠原 克彦²（1. 北海道大学保健科学院, 2. 北海道大学大学院保健科学研究所）

*Ryuichiro Ueda¹, Feng Han¹, Hongjian Zhāng¹, Tomohiro Aoki², Ogasawara Katsuhiko²（1. 北海道大学保健科学院, 2. 北海道大学大学院保健科学研究所）

キーワード：COVID-19, Sentiment analysis, Twitter

2020年から COVID-19による感染拡大に伴い諸外国ではロックダウンなどの政策が行われた。一方で SNSを用いた公衆衛生の情報収集を目的とした利用者が増加し、先行研究では SNSデータを利用したセンチメント分析は効果的な公衆衛生の政策を行うための世論の把握に役立つと指摘している。そこで本研究では日本で施行された緊急事態宣言による国民の感情の変化についてセンチメント分析を行うことで明らかにし、世論の動向を把握することを目的とする。本研究では NTTデータ社「なずきのおと」により日本における1回目の緊急事態宣言の前後約2週間である2020年3月23日から4月21日を対象に「コロナ」、「COVID-19」、「新型肺炎」のキーワード含む Tweetデータを対象とした4,894,009件を用いて日別ごとの処理を行った。センチメント分析には Tweetデータを単語ごとにベクトル化する処理を行い、感情辞書の単語にはクラスター分析の1つである Fuzzy-C-Means法を採用し、怒り、悲しみ、驚き、嫌悪、心配、信頼、喜びの7つの感情クラスターを設定した。次に感情クラスターの中心点と単語ベクトルで処理された Tweetがどの感情に一番近いかをベクトル距離の算出をした。分析の結果、対象とした期間内における Tweet全体による感情の割合は心配(21.5%)、信頼(17.7%)、嫌悪(15.9%)、喜び(13.7%)、悲しみ(11.0%)、怒り(10.3%)、驚き(9.8%)となり、時系列による感情分析では緊急事態宣言発令日で嫌悪の感情値の割合が86.6%であった。本研究の結果から、日本における緊急事態宣言による公衆衛生の介入は SNS上において一般市民の感情の変化に影響する可能性が高いことが示唆され、公衆衛生の介入に対しての一般市民への影響、世論の反応をタイムリーに把握するための資料提供となると考える。

Fuzzy-C-Means 法を用いた COVID-19 に関するセンチメント分析

-日本語を対象とした Twitter ユーザーに関する検証-

上田龍一郎^{*1}、韓豊^{*1}、張洪健^{*1}、青木智大^{*2}、小笠原克彦^{*2}

^{*1} 北海道大学大学院保健科学院、^{*2} 北海道大学大学院保健科学研究所

Sentiment Analysis on COVID-19 using the Fuzzy-C-Means Method -Analysis of Twitter users in Japanese-

Ryuichiro Ueda^{*1}, Feng Han^{*1}, Hongjian Zhang^{*1}, Tomohiro Aoki^{*2}, Ogasawara Katsuhiko^{*2}

^{*1} Graduate School Health Sciences, Hokkaido University, ^{*2} Faculty of Health Sciences, Hokkaido University

Abstract

With the spread of COVID-19, Study of Sentiment analysis to examine the public's concerns and feelings about COVID-19 are increasing. In this study, we conduct a sentiment analysis on Japan to identify changes and trends in sentiment toward COVID-19. We conducted a daily analysis using 4,894,009 tweets of Twitter data including the keywords "corona(コロナ)," "COVID-19," and "new pneumonia(新型肺炎)" for the period from March 23 to April 21, 2020, which is approximately two weeks before and after the first declaration of a state of emergency in Japan. We also processed Tweet data into vectors for each word, and employed the Fuzzy-C-Means method, a type of cluster analysis, for the words in the sentiment dictionary, and set up seven sentiment clusters (Negative: anger, sadness, surprise, disgust, Neutral: Anxiety, Positive: trust, and joy), and conducted sentiment analysis divided into Tweet groups and retweet groups. The results of the analysis showed a mixture of positive (joy) and negative (Anxiety, disgust) emotions. In addition, the retweet group tended to post significantly more negative content. It is thought that the increase in negative feelings associated with the spread of COVID-19 infection and the feelings of expectation for infection prevention measures due to the declaration of a state of emergency prompted the increase in positive values, and it is thought that topics related to COVID-19 tend to spread negative content.

Keywords: COVID-19, Sentiment analysis, Twitter

1. 緒論・目的

2019 年 12 月に中国湖北省で発生した新型コロナウイルス感染症(COVID-19)により諸外国ではロックダウンによる強い行動規制が行われ一般市民への精神的な悪影響が指摘された¹⁾。一方 COVID-19 の感染拡大以降テキスト情報から個人の抱く感情を分析するセンチメント分析を用いることにより、個人の心理学的変化や国民の懸念内容を反映させ公衆衛生、公共政策への情報提供が行われている。

本研究では SNS データを用いたセンチメント分析を、日本語ユーザーを対象に行い COVID-19 の感染拡大初期に対する一般市民の感情、および動向を検証する。

2. 方法

2.1 使用データ

本研究では日本における緊急事態宣言発令から前後約 2 週間の 2020 年 3 月 23 日 00:00:00 から 2020 年 4 月 21 日 23:59:59 を対象に「コロナ」、「COVID-19」、「新型肺炎」のキーワードを含む日本語データを対象とした 4,997,353 件のランダムサンプリングによる Twitter 投稿データを使用した。データの预处理としてテキストデータの処理に対応できない記号を含む Twitter データとハッシュタグ(例:#コロナ)、URL のみのテキストデータを削除し、計 4,965,100 件とした。

2.2 テキストデータ処理

センチメント分析を行うにあたり、初めにテキストデータの形態素解析を行った。形態素解析は語句の境界判別を行い、文章を構成する文法的に意味を持つ最小の単位を決定する処理であり、本研究では日本語形態素解析オープンソースの MeCab(ver0.996)と新語表現に対応した mecab-ipadic-NEologd²⁾を導入し形態素解析を行った。

次に解析処理後のテキストデータに Word2Vec を用いたベ

クトル化を行った。Word2Vec は単語の意味を低次元のベクトルにより表現するモデルであり、これにより自然言語処理における類似語の抽出や単語の意味上の計算が可能となる。

2.3 センチメント分析

本研究では JIWC-Dictionary(日本語感情表現辞書)³⁾から怒り、心配、嫌悪、悲しみ、驚き、信頼、喜びの 7 種類の感情を設定し、辞書に含まれる単語を用いて Fuzzy-C-Means(FCM)法によるクラスタリングを行うことで 7 つの感情クラスターを設定した(図 1)。FCM 法は主観的なあいまいさを扱うための枠組みであるファジィ理論に基づいた非階層型クラスタリングの 1 つであり、対象としたデータに対してそれぞれのクラスターへの所属度を表すパラメータを用いることにより各クラスターの対してどの程度属するかを算出する。これにより感情などの曖昧性を含む単語の分類に対して FCM 法を用いることで詳細な分類を可能にしている。

感情	辞書内に含まれる単語例
怒り (Anger)	怒った、怒り、怒る、悪い、嫌がらせ、イライラ、うるさい、ゴミ、暴言、煽り、理不尽な、騒音、迷惑、被害、虐待、裏切り…
驚き (Surprise)	いきなり、かかって、がん、サプライズ、びっくり、偶然、知った、知って、突然、解散、詐欺、驚いた、驚き、発見…
悲しみ (Sad)	悲しい、悲しがる、悲観、悲愁、哀感、哀傷、哀切、泣き、泣き叫ぶ、嘆き、涙、涙声、追悼、痛嘆、センチメンタル…
嫌悪 (Disgust)	嫌いな、嫌がらせ、嫌な、うるさい、テロ、犯罪、犯人、ひどい、悪、悪かった、批判、無い、無し、無視、嘘、汚い…
心配 (Anxiety)	不安、不安だ、不安な、不安に、病気、症状、このまま、この先、なかなか、今後、心配、状況、異動、確認、考えた、考える…
信頼 (Trust)	仲間、任せ、依頼、信用、信頼、頼り、頼んで、助け、助けて、守って、親友、親身に、関係、サポート、フォロー…
喜び (Joy)	遊び、遊んで、楽しい、楽しかった、出かけた、おいしい、食事、できた、できて、会って、会話、笑い、笑顔、好きな…

図 1 感情辞書単語例

センチメント分析全体の経路図を図 2 にまとめた。テキストデータのベクトル化により各 Tweet のベクトル座標を作成し、

FCM 法により生成した感情クラスターの重心ベクトルとの距離計算を行い、最もベクトル距離の小さい値をそのツイートの感情として決定した。また、本研究におけるセンチメント分析全般においてはプログラミング言語である Python(3.9.4)を用いて行った。

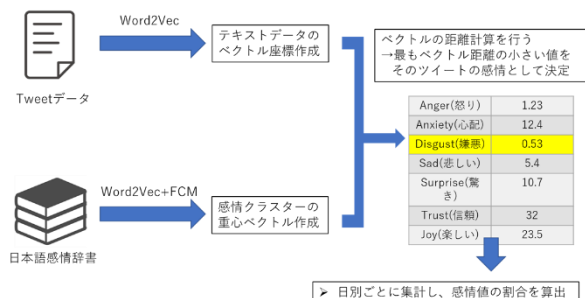


図2 センチメント分析全体の経路図

本研究におけるセンチメント分析を2つの方法により検討を行った。1つ目は緊急事態宣言発令日(2020年4月7日)からの前後約2週間を分け、感情値7種類の割合を日ごとに求め、緊急事態宣言発令前後における感情値の割合を算出した。

2つめは抽出したTwitterデータをユーザー本人が投稿したTweet群と、拡散を目的とした自身または他ユーザーの投稿を再度投稿するRetweet(RT)群に分け、感情値をPositive(信頼、喜び)とNegative(怒り、心配、嫌悪、悲しみ)に分類した日ごとの集計値を用いて群ごとの比較検討を行った。また、2つの検討方法ともに統計ソフト JMP(16.0)を用いて Mann-Whitney U test を行った。

3. 結果・考察

はじめに緊急事態宣言発令前後における感情値の変化に関して表1に示す。センチメント分析全体の結果 Positive が Tweet 群において前後の期間でそれぞれ41.7%,33.3%と高い割合である一方で Negative に該当する心配、嫌悪の感情が Tweet 群、RT 群の両方で高い割合を占めていた。対象期間においては COVID-19 の感染拡大初期における Negative な感情の上昇と同時に緊急事態宣言による感染拡大防止策への期待の感情が Positive の割合の上昇に影響していたことが考えられる。英語ユーザーを対象としたセンチメント分析の結果においても感染拡大初期において可能性のある政策への期待による Positive な感情と未知の感染症への恐怖に対する Negative な感情が同時に高値を示している⁴⁾。

表1 センチメント分析全体の結果

	2020年3月23日-2020年4月6日				2020年4月7日-2020年4月21日			
	Tweet群		RT群		Tweet群		RT群	
	合計	%	合計	%	合計	%	合計	%
怒り	61,335	9.5	210,385	11.3	68,189	10.4	322,849	15.9
心配	109,999	21.5	587,540	32.5	164,325	23.4	211,853	12.3
嫌悪	35,367	4.7	322,462	21.0	52,089	7.2	419,620	21.8
悲しい	38,737	5.2	106,351	6.4	33,939	4.7	150,847	8.1
驚き	67,789	4.7	111,044	8.6	44,320	5.7	257,841	13.6
信頼	87,551	12.7	218,199	13.9	119,582	17.0	182,596	10.2
喜び	272,879	41.7	174,382	9.8	217,074	33.3	330,374	18.2

次に Tweet 群と RT 群に分けた感情値の検討について日ごと感情値の割合を Mann-Whitney U test により検討した結

果を表2に示す。緊急事態宣言発令前後2週間において RT 群が Tweet 群よりも Negative な内容を投稿する傾向が有意にみられた(発令前: $r=0.29, p=0.02$, 発令後: $r=0.40, p=0.002$)。感染拡大初期の段階においては COVID-19 に関する内容は Negative な感情を含む投稿を拡散させる傾向が考えられ、先行研究では英語圏ユーザーを対象とした COVID-19 の感染拡大初期段階において RT による投稿は否定的な内容が多いことを指摘している⁵⁾。

表2 Tweet 群、RT 群による検討

	Tweet(n=60)		RT(n=60)		検定統計量 r	p
	中央値	四分位偏差	中央値	四分位偏差		
Negative_宣言前	0.040	0.036	0.066	0.065	0.29	0.02*
Negative_宣言後	0.038	0.045	0.063	0.049	0.40	0.002*

	Tweet(n=30)		RT(n=30)		検定統計量 r	p
	中央値	四分位偏差	中央値	四分位偏差		
Positive_宣言前	0.089	0.192	0.054	0.038	0.22	0.22
Positive_宣言後	0.099	0.164	0.108	0.088	0.18	0.34

本研究の限界点として分析対象として使用した Twitter ユーザーの年齢層には偏りがある。そのため本研究における結果は必ずしも国民全体を代表するものではない可能性が考えられる。また使用した Tweet データには位置情報は含まれていないため、都道府県によって母集団が不均一なサンプルである場合がある。次にセンチメント分析による感情は7種類のうちの1つに決定していることから複数の感情が考えられる場合や設定した感情に該当しない場合を考慮できていない。

4. 結論

日本における1回目の緊急事態宣言発令前後を対象としたセンチメント分析を行った結果、COVID-19 の話題においては Positive な感情と Negative な感情が同時に混在している一方、情報を拡散させる目的である RT には Negative な内容を含む投稿が多い傾向にあることがわかった。したがって今後感情値の結果を踏まえたコミュニケーションの分析を行うことで一般市民の懸念内容を調査する必要があると考える。

5. 参考文献

- 1) Soubhik Barari, Stefano Caria, Antonio Davola et al. Evaluating COVID-19 Public Health Messagin in Italy: Self-Reported Compliance and Growing Mental Health Concerns. medRxiv 2020.
- 2) Mecab-ipadic-NEologd : Neologism dictionary for MeCab, 2020. [https://github.com/neologd/mecab-ipadic-neologd (cited 2020-Aug-18)].
- 3) JIWC-Dictionary(日本語感情表現辞書) [https://github.com/sociocom/JIWC-Dictionary (cited 2021-Aug-18)]
- 4) Jia Xue, Junxiang Chen, Chen Chen, Chengda Zheng, Sijia Li, Tingshao Zhu. Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. PLoS ONE 2020.
- 5) Koyel Chakraborty, Surbhi Bhatia, Siddhartha Bhattacharyya, Jan Platos, Rajib Bag, Aboul Ella Hassanien. Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers: A study to show how popularity is affecting accuracy in social media. Applied Soft Computing 2020 ; Volume 97, Part A.