テレビ字幕データを用いた感情分析による「ある日の日本の気分」推定に関する研究

イー フエイチー[1]　望月 源[2]
[1]東京外国語大学 言語文化学部 [2]東京外国語大学 大学院総合国際学研究院
{yee.hueichi.s0,motizuki}@tufs.ac.jp

概要

本研究は感情分析を用いて，とある日の日本社会全体がその日の出来事からどのような情緒的な雰囲気になっていたのか，いわば「日本の気分」を推定することを目的とする．テレビ字幕データの中から日々のニュース字幕を取り出し，日本語感情分析ライブラリ oseti により計測した各文の感情スコアを集計することで，ある日の感情を推測する．2019 年 10 月から 2020 年 12 月 31 日までの期間を対象に実際に日付ごとの感情分析によるスコアがどのように変化したのかを考察する．

1 はじめに

本研究は感情分析を用いて，とある日の日本社会全体がその日の出来事からどのような情緒的な雰囲気になっていたのか，いわば「日本の気分」を推定することを目的とする．過去の一連のニュースに関するテキストを分析することによって，過去の，とある期間の日本社会がどんな雰囲気なのか，あるいは現在の社会の雰囲気がどんななのか，具体的に世間の「雰囲気」を数値化する．数値化した結果から，とある期間の「日本の気分」を可視化する[1]．

この目的を達成するため，本研究では，テレビのニュース番組の字幕データを利用し，1 日単位での感情分析によるスコアを抽出する．1 年 2 ヶ月という比較的長い期間のデータを分析することによって，時間と共に日付ごとの感情分析によるスコアがどのように変化したのかを考察する．

2 調査方法

2.1 使用データについて

本研究では，東京外国語大学計算言語学研究室で収集整備している「日本語テレビ字幕コーパス」を用いる[2]．このコーパスは，東京都の地上波デジタルテレビ放送に付与されたクローズドキャプション(字幕)データを収集し，不要な情報を除去した後に，字幕表示のために断片化した文を 1 文にまとめたうえで字幕表示のタイミング情報を付与したものであり，以下の特徴がある．

- 2012 年 12 月から 2021 年 9 月の時点で 518,960 番組, 179,268,965 文，1,945,899,573 形態素．
- 各テキスト(番組)はテレビ局の番組情報に基づいて，アニメ／特撮，スポーツ，ドキュメンタリー／教養，ドラマ，ニュース／報道，バラエティ，映画，音楽，劇場／公演，趣味／教育, 情報／ワイドショー, 福祉, および，その他の 13 種類に分類されている．

本研究ではこのコーパスの中から，2019 年 10 月 1 日から 2020 年 12 月 31 日までの「ニュース/報道」の 16,362 テキストを字幕データとして使用する．

2.2 データ分析の方法

2.2.1 感情分析ツール

感情分析では，単語や用語ごとに「ポジティブ」「ネガティブ」の極性を定めた「感情極性辞書」を用いることが多い．代表的な日本語の感情極性辞書として「日本語評価極性辞書」がある．この日本語評価極性辞書は，「日本語評価極性辞書(用語編)」[3]と「日本語評価極性辞書(名詞編)」[4]に分類されている．用言編と名詞編の全表現には「ポジティブ」「ネガティブ」の評価極性が振り分けられており，用言編にはさらに「客観的」「主観的」の振り分けもある．本研究では，この日本語評価極性辞書をデータとして用いた日本語感情分析ライブラリ「oseti v.0.2」[5]を用いる．oseti は 1 文ごとの感情スコアを計算するツールとして GitHub にて公開されている．1 文を入力すると，文内の表現のスコアの平均値を

計算し，文全体としての感情スコアを-1 から 1 の範囲で出力する. ポジティブなスコアは肯定的な感情，ネガティブなスコアは否定的な感情，丁度ゼロとなるスコアはどちらでもないということを指す.

また，oseti は否定的な文脈が現れるとネガティブな極性を付ける．例えば，「お金」と「希望」は単独ではポジティブな極性である表現だが「お金も希望もない」という文ではポジティブな極性が取り消され，ネガティブな極性になるため，文全体のスコアはネガティブな値となる．本研究ではコーパス内の各文の感情スコア計算に oseti を用いる．

2.2.2 感情分析方法

とある日の「日本の気分」を表す数値の計算に当たって，まずは式 1 の「合計スコア」を導入する．合計スコアscr_{total}はその日のポジティブな文のスコアscr_pの総計とネガティブな文のスコアscr_nの総計を足し合わせたものであり，その日に放送された全部の文のスコアを足し合わせたものである．

$$scr_{total} = \sum_{p \in positive} scr_p + \sum_{n \in negative} scr_n \quad (式1)$$

ここで *positive* はポジティブな文の集合，*negative* はネガティブな文の集合をそれぞれ表す.

式 1 では，肯定的な感情を持つ文と否定的な感情を持つ文はスコアが打ち消し合う関係にある.

一方，肯定的な感情と否定的な感情がお互いに打ち消し合う関係にあるのかについては，疑問も生じる．例えば，ある日に，1 件の非常に悲しいニュース(スコアを 5.0 とする)と，5 つの少しだけ嬉しいニュース(各スコアを-1.0 とする)があったとする. 合計スコアの考え方では，5.0-(1.0x5)=0.0 となり，その日の感情は全体として「どちらでもない」ということになる．しかし，実際にその日に重大な出来事が発生しており，人々は起きた物事に強い感情を持っていたはずなので尺度としてしっくりこない部分もある．そこで，極性によらず社会の感情の強さを図る「感情の強度①」scr_{es1}を導入する(式 2).

$$scr_{es1} = \sum_{p \in positive} scr_p + \sum_{n \in negative} |scr_n| \quad (式2)$$

感情の強度①とは，その日のネガティブスコアを絶対値にし，ポジティブスコアに足し合わせたものである. 両方のスコアが同等に高い日があった場合，合計スコアではゼロに近くなるが，感情の強度①では反対に高くなる．評価極性に関係なく，その日の感情の動きの強さが分かる尺度となっている.

ただし，感情の強度①はデータの量に強く左右される．一般に，ニュースが多い日にはスコアを計算する文数も多くなるため，値が大きくなり，少ない日には値が小さくなる．そこで，データの量の影響を軽減する「感情の強度②」scr_{es2}を導入する(式 3).

$$scr_{es2} = \frac{\sum_{p \in positive} scr_p + \sum_{n \in negative} |scr_n|}{N} \quad (式3)$$

感情の強度②とは，感情の強度①を総文数 N で割ったものである．1 日のニュースの量に応じて感情の強度が調整されるため，感情の強度①に比べてデータ量の影響が少ない尺度となっている.

本研究では，1 日ごとのテレビ字幕データを以上の 3 つの計算方法で計算し，比較する.

3 感情データの可視化と分析

3.1 合計スコアのグラフ化

2019 年 10 月から 2020 年 12 月のデータについて，1 日ごとの感情を合計スコアで計算した結果を図 1 に示す．図 1 で，横軸は日付を示し，縦軸は合計スコアを示す.

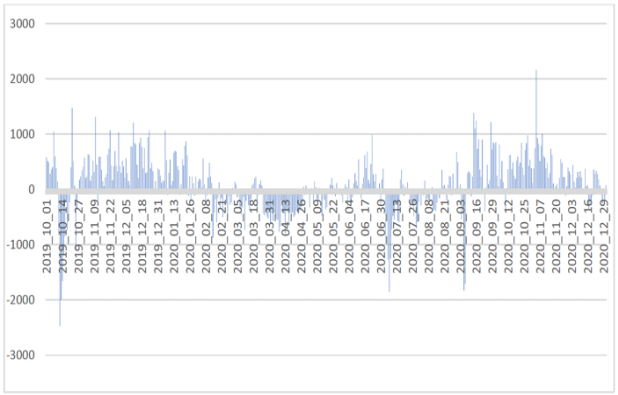

図 1 合計スコアの推移

期間全体の合計スコアの平均は 89.3 だった．合計スコアが最も高かった日は 2020 年 11 月 4 日で 2,161.8 だったが特に大きなニュースはなかった．一方, 最も低かった日は 2019 年 10 月 12 日で -2,471.0 だった．この日は，過去最高クラスの台風 19 号が伊豆半島に上陸した．より詳細に，期間中に多くの人に影響を与える出来事が起きた日と合計スコアの対

応を比較する．

2019 年 10 月 1 日には消費税率が引き上げられたが合計スコアは 578.3 であり，その後の日付のスコアと大きく変わらなかった．同月 12 日，25 日には台風が日本に上陸し，スコアが急減した．12 日は期間中最低の-2,471.0 で，25 日も前日までのポジティブスコアとは異なり急減して-1,083.3 だった．合計スコアからは，消費税率の話題は社会の感情を大きく左右することはなかったが，台風の話題は非常に強いネガティブな影響を及ぼしたと考えられる．

2020 年 2 月 3 日にクルーズ船ダイヤモンドプリンセスが日本に停泊し，13 日に新型コロナウィルスによる初めての死亡者が確認され，19 日にはクルーズ船からの下船が開始した．2 月 3 日と 19 日の合計スコアはそれぞれ 188.8 と 124.7 だった．新型ウイルス感染拡大のニュースに対し，社会は否定的な感情を抱くと予想されたものの，この両日の合計スコアはポジティブだった．しかし，合計スコアの全体的な傾向から見ると，その後の 2 月下旬ごろから数ヶ月ネガティブな期間が続く．2 月初頭の時点では，新型ウイルスに関する報道は他の話題と比べてまだ少なかったか，言葉の使い方が楽観的だったか，社会の新型ウイルスに対する印象が悪い方向に根付いてはいなかったのではないかと推測される．

2020 年 4 月 7 日に新型コロナウイルスの感染拡大により，第一回目の緊急事態宣言が発表された．5 月 4 日に緊急事態宣言の約 1 か月間の延長が発表され，5 月 25 日に全国で解除された．4 月 7 日の合計スコアは-549.057，5 月 4 日は 18.734，5 月 25 日は 116.262 だった．緊急事態宣言が発表された当初，合計スコアは全体的に低く，図 1 ではネガティブなスコアが連続した期間が多く，緊急事態宣言が発表された当日の合計スコアもネガティブだった．しかし，時間が経つにつれて合計スコアが上がり，5 月 4 日には合計スコアはややポジティブで緊急事態宣言が全面解除した 5 月 25 日になると合計スコアがさらにポジティブになった．

合計スコアでは，同じ評価極性の日が長い期間連続する傾向にある．2019 年 11 月 13 日から 2019 年 12 月 27 日までの 45 日間連続でポジティブスコアが続いた．その他，39 日間ポジティブスコアが続いた 2020 年 10 月 11 日から 2020 年 11 月 18 日までの期間と，33 日間連続でネガティブスコアが続いた 2020 年 3 月 26 日から 2020 年 4 月 27 日までの期間などが見られた．より大きな傾向としては，2019 年 10 月終わりから 2020 年 2 月までと，2020 年 9 月から 12 月はポジティブな傾向，2020 年 3 月から 6 月と 2020 年 7 月から 8 月にネガティブな傾向が見られた．

3.2 感情の強度①のグラフ化

同一期間の感情の強度①で計算した結果を図 2 に示す．図 3 に期間中の文の数を示す．

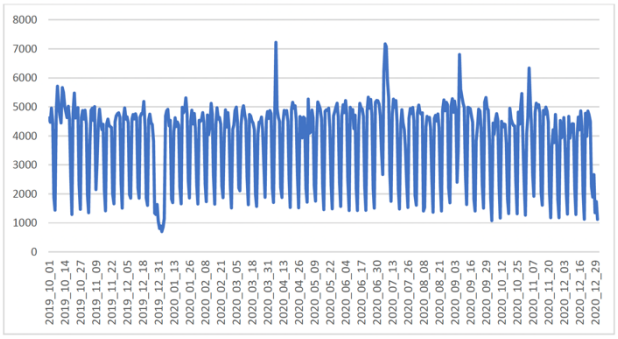

図 2 感情の強度①の推移

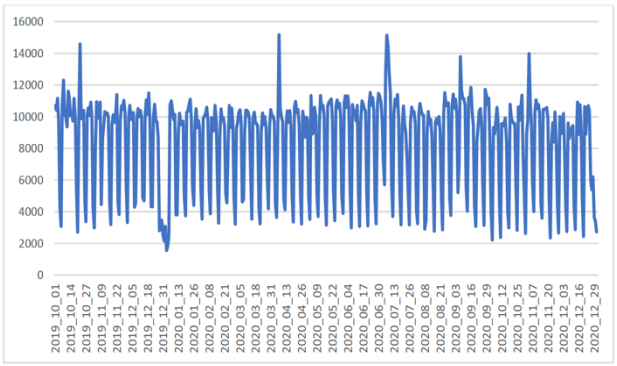

図 3 文の数

期間内で，感情の強度①は平均で 3,864.1 だった．値が最も高かった日は 2020 年 4 月 7 日で 7,220.5 だった．最も低かった日は 2020 年 1 月 3 日で 701.6 だった．同じ期間で文の数は平均で 8,407 文だった．文の数が最も高かった日は 2020 年 4 月 7 日で 15,167 文だった．文の数が最も低かった日は 2020 年 1 月 3 日で 1,561 文だった．どちらの場合も，値の最大，最小の日とニュース量の多少が完全に一致する．感情の強度①にはニュースの量が強く影響していることがわかる． 図 2 と図 3 では，形が一致するところが大きい．図 3 では文の数が非常に多かった日が 5 日のうち，4 日は感情の強度①も多かった．その 4 日は，緊急事態宣言が初めて宣言された 2020 年 4 月 7 日，豪雨により避難指示が出た 2020 年 7 月 7 日，台風 10 号が沖縄県に上陸した 2020 年 9 月 7 日と，アメリカ大統領選挙の結果が日本で発表された 2020

年 11 月 4 日である．どの日も報道されたニュースの量も多く，感情の強度①も文の数も多かった．

大きな傾向としては，感情の強度①はニュースの量に影響され易い．しかし，とある日に「ニュースの量が多かった」ということは，評価極性に関係なく，報道する価値があるニュースが多かったとも言える．そうなると，感情の強度①はどの日に重大出来事が起きたかの特定に役立つと考えられる．

3.3 感情の強度②のグラフ化

同一期間の感情の強度②で計算した結果を図 4 に示す．

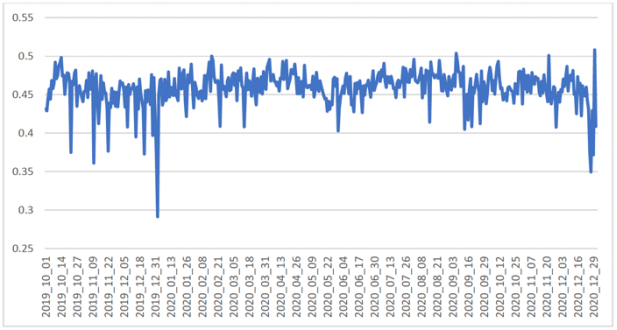

図 4 感情の強度②の推移

感情の強度②は平均で 0.459 だった．値が最も高かった日は 2020 年 4 月 7 日で，0.508 だった．最も低かった日は 2020 年 1 月 3 日で，0.291 だった．全数値の中で，感情の動きが弱かった下位 5%は 0.417 より低く，強かった上位 5%は 0.488 より高かった．感情が強かった日で，合計スコアがポジティブな日が 3 日，ネガティブな日は 23 日あった．一方，感情が弱かった日で，合計スコアがポジティブな日は 17 日，ネガティブな日は 7 日だった．感情の強度②が高く感情が強かった日は合計スコアがネガティブな傾向にあった．反対に，感情の強度②が低く感情が弱かった日は合計スコアがポジティブな傾向にあった．分析期間のデータを参考にする限り，否定的な感情は肯定的な感情よりも強い感情を呼び起こすと考えられる．

重大な出来事の発生と感情の強度②を照らし合わせると，感情が強かった日もあれば，弱かった日もある．新型ウイルスの初めての死亡者が発表された日の 2020 年 2 月 13 日の直後に感情が強く，14 日，16 日，17 日の感情の強度②が強かった．また，台風 19 号が上陸した 2019 年 10 月 12 日の辺りに 10 月 9 日，12 日，13 日と 14 日にも感情の強度②が強かった．しかし，台風 21 号が上陸した 2019 年 10 月 25 日は当日の感情の強度②が 0.478 でごく平均的な数値だった．同じ台風に関するニュースであっても、感情の強度②に必ず反映するとは限らない．

4 おわりに

本研究では，テレビのニュース番組の字幕データを利用し，1 日単位での感情分析によるスコアを抽出し，とある日の「日本の気分」の数値化を試みた．

3 つの数値化のうち，合計スコアでは，とある日に重大出来事が起きたかを特定するより，数日，数ヶ月などある程度長い期間での社会の雰囲気を調べることに向いていた．2020 年の新型コロナウイルス感染状況における報道の変容について，岸本らの分析がある[6]．岸本らの考察と図 1 の値が次のように非常に良く符号する．

「未知のウイルスが海外から日本へと徐々に身近に迫る様子が報道され，視聴者の関心は上昇傾向していた（図 1 で感染が始まった初期：2020 年 1 月 4～3 月 18 日が該当）．第一波に入ると，緊急事態宣言により生活環境は一変し，コロナ報道は活発化した．初の感染拡大や活発なコロナ報道により、視聴者の関心はピークに達していた（図 1 で第一波：3 月 19 日～6 月 21 日が該当）」（第一波が収束すると）「コロナ報道はやや落ち着き始めていた」「第二波に入ると，東京の感染者数報道が中心となるが、視聴者の関心は変化しなかった（図 1 で第二波：8 月 8 日～11 月 3 日が該当）」

感情の強度①はデータの量に左右されることが多いが，何らかの重大なできごとが発生した日を示す指標としての可能性がみられた．感情の強度②はデータ量の影響は受けにくいが，感情が強くかつ肯定的な日と，感情が弱くかつ否定的な日を検知する性能が限られており，重大な出来事の発生した日は特定しにくい．

とある日に重大出来事が起きたとして，その出来事についての報道が当日だけに限られることは少なく，その後も報道が続く．そのため，出来事が起きた当日以降でも感情の強さがピークアウトすることもある．今後より正確に，とある日の感情を掴むために，強い感情が連続する期間内で共通する単語を重大な出来事を表す話題として，話題ごとにニュースをまとめ，最初に報道された日を探し出すようなシステムの実現が考えられる．これにより，特定の日の重大な出来事とその時の「日本の気分」をより正確に検知できる可能性がある．

謝辞

本研究は JSPS 科研費 JP19H04224，JP20H00096 の助成を受けたものです.

参考文献

[1] イーフエイチー, テレビ字幕データを用いた感情分析による「ある日の日本の気分」推定に関する研究，卒業論文，東京外国語大学,3. 2022.
[2] Hajime Mochizuki and Kohji Shibano, Building Very Large Corpus Containing Useful Rich Materials for Language Learning from Closed Caption TV, World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, Volume 2014, No. 1, pp. 1381-1389, Association for the Advancement of Computing in Education (AACE), 10. 2014.
[3] 小林のぞみ、乾健太郎、松本裕治、立石健二、福島俊一、意見抽出のための評価表現の収集、自然言語処理、Vol.12、No.3、pp.203-222、2005.
[4] 東山昌彦、乾健太郎、松本裕治、述語の選択選好性に着目した名詞評価極性の獲得、言語処理学会第 14 回年次大会論文集、pp.584-587、2008.
[5] 池上 有希乃、 oseti v.0.2、 2019、https://github.com/ikegami-yukino/oseti（最終アクセス 2022.1.5）
[6] 岸本大輝、井原史渡、栗原聡、新型コロナウイルスの感染状況に対するテレビ報道の特徴と報道変容の分析、人工知能学会全国大会（第 35 回），pp.1-4, 2021.

A 付録

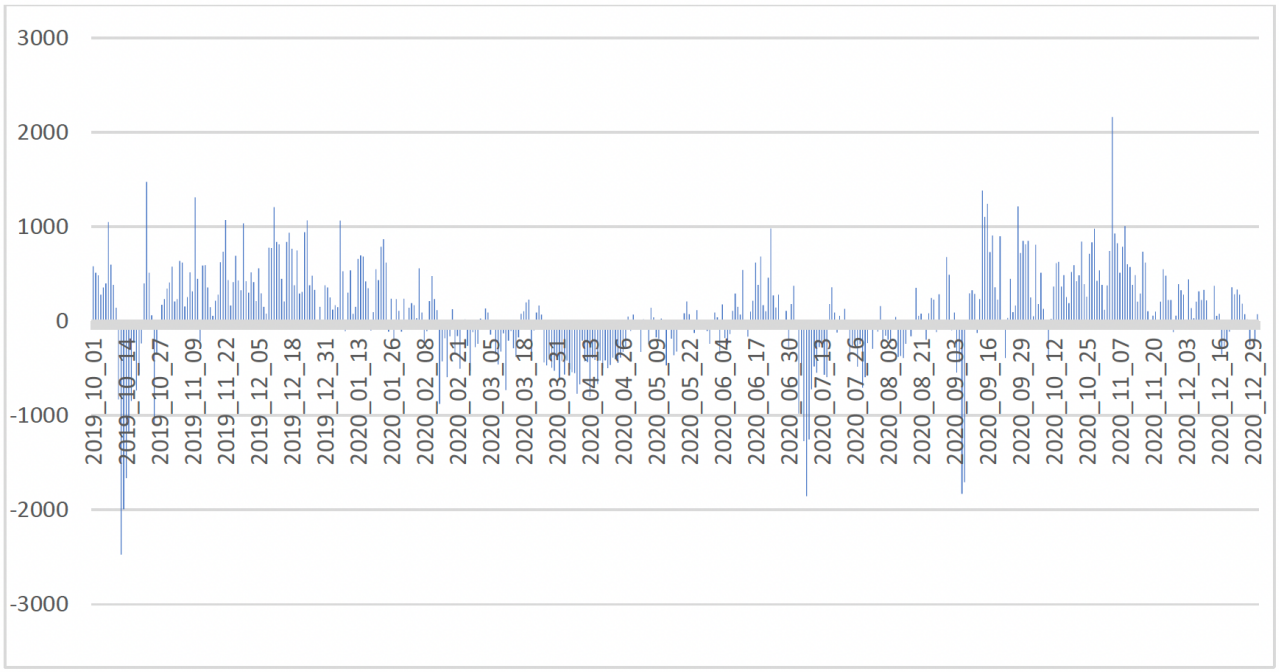

図 1 合計スコアの推移

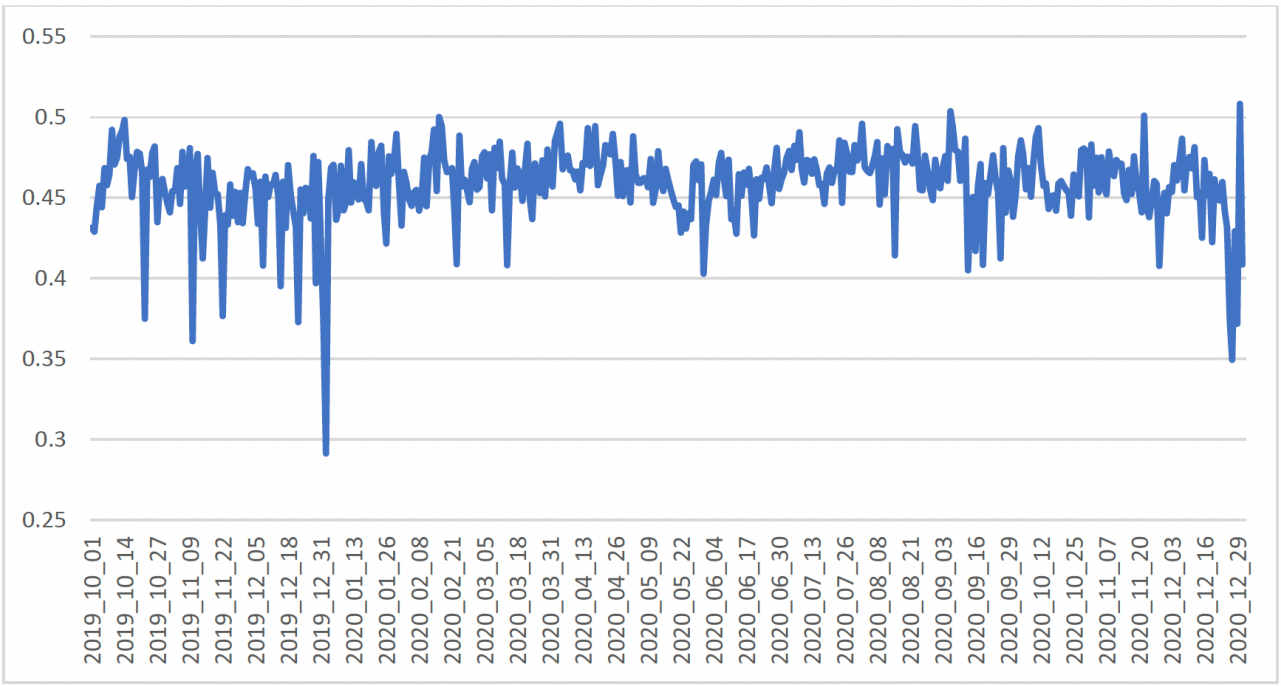

図 4 感情の強度②の推移