コロナ禍の状況を自由記述文で記録し分析する試み

那須川哲哉* 鈴木祥子* 村岡雅康* 平野真理**
*日本アイ・ビー・エム株式会社 東京基礎研究所 **東京家政大学 人文学部
*{nasukawa, e30126, mmuraoka}@jp.ibm.com **hirano-m@tokyo-kasei.ac.jp

1 はじめに

新型コロナウイルス感染症の流行による危機的状況（以降「コロナ禍」）が続く中，日常生活におけるコロナ禍の影響を記録し分析可能にすることで，例えば感染者数の増減傾向といった変化の方向性を捉えるなど，何らかの有用な知見の獲得に結びつけられる可能性があるのではないかと考えた．本稿では，テキストデータを中心としたコロナ禍に関するデータ収集の取り組みと，この取り組みから得られた知見を紹介する．

2 コロナ禍データの記録・収集

2.1 コロナ禍データ収集方針

大きな災害に関するテキストデータを収集し分析する取り組みとして，著者らは，東日本大震災及び熊本地震の際に，Twitter へ投稿された関連ツイートを収集し分析した[1,2]．東日本大震災の際には，「地震」「被災」「原発」というキーワードに加え，#jisin 及び #jishin というハッシュタグを対象としてデータを収集した．また，熊本地震の際には，Tweet Location の情報と被災地の地名のキーワードを手がかりにして，データを収集した．各々のケースで，毎日数十万件規模のデータが集まり，IBM® Content Analytics[3]などのテキストマイニングツールを用いることにより，被災地で不足している物資の情報などを捉えることができた．

今回のコロナ禍に関しては，震災と異なり，被災地や被災者が限定されない上，日々の生活全般にどのような影響が生じるかの予想が困難である．そのため，何らかのキーワードを設定してデータ収集を行うのは難しい．また，コロナ禍がどれだけの期間続くか分からず，長期的な取り組みを行なう必要があると考えた．そこで，コロナ禍で感じたことや気付いたことを多くの人に自由記述形式でテキスト化してもらい，それを集めて分析するというアプローチを取ることにし，入手可能なデータを出来るだけ多く収集することにした．

2.2 Slack チャンネルの作成とデータ収集

まずは，IBM 社内で利用している Slack[i]と呼ばれるコミュニケーションツール上に，コロナ禍に関する情報を投稿したり閲覧したりするためのチャンネルを作成した．その上で社内の参加者を募り，書き込みを依頼した．2020 年 4 月 21 日にこの取り組みを始めて以来，グループ会社や海外を含む多様な部署から参加者が集まり，2021 年 1 月 15 日時点の参加者は 348 名になっている．投稿データの言語は日本語で,「マスクを買いに行ったがどこにも売っていない」「久しぶりに電車に乗ったら車内広告が少なかった」といった体験談や「こんな報道があった」という URL 付きの報告など多様な内容が簡潔に表現されている．各投稿には投稿者の名前が表示され，閲覧者が絵文字リアクションという機能で各投稿に対し「残念」「確かに」といった反応を示せるようになっている．

この Slack チャンネルへの 2020 年 12 月末までの書き込み件数の状況を図 1 に示す．

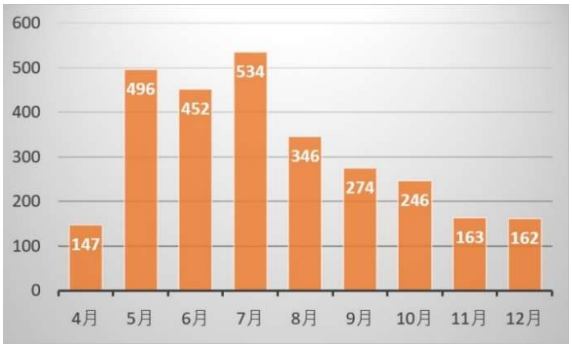

図 1: Slack チャンネルへの月別投稿データ件数

2020 年 12 月末までに合計 2,820 件の投稿が集まり，各投稿の平均文字数は，中に含まれる URL の文字数も含めて 141.1 文字となっている．

[i] https://slack.com/

2.3 コロナ日記の収集

IBM 社員とは異なる母集団のデータも集めて比較したいという考えから，下記 2 大学の学生にも協力を仰ぎ，データ収集を行なった．

滋賀大学データサイエンス学部において，2020 年 6 月 13 日に「データサイエンス実践論 A」という講義でテキストアナリティクスを紹介する機会を活かし，その受講生 41 名に依頼して，5 月 30 日から 12 日間，コロナ禍で感じたことや気付いたことなどを簡単な日記形式で記録・提供してもらった．Slack チャンネルへの投稿と異なり他人に読まれることは意識せずに書き溜めたデータをレポート的に提出してもらった．同日中でもトピックが異なる場合は，複数件のデータにするよう依頼し，また，受講課題として基本的に毎日書くようガイドした．その結果，536 件のデータが集まり，各データの平均文字数は 39.5 文字であった．日別の件数分布を図 2 に示す．

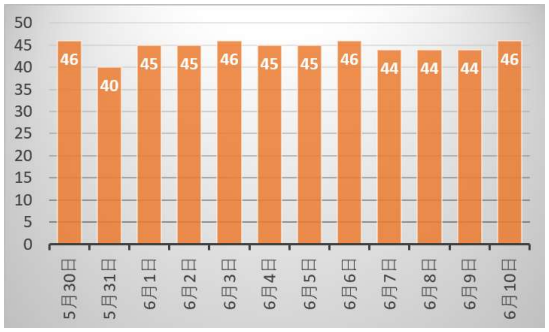

図 2: 滋賀大学コロナ日記の日別データ件数

同様に，首都圏にある A 大学に所属する学生 11 名に対しても依頼を行い，2020 年 5 月 14 日からの 2 ヶ月間にわたるコロナ日記を提供してもらった．その結果，7 月 19 日までの 521 件のデータが集まった．各データの平均文字数は 153.2 文字であった．日別の件数分布を図 3 に示す．

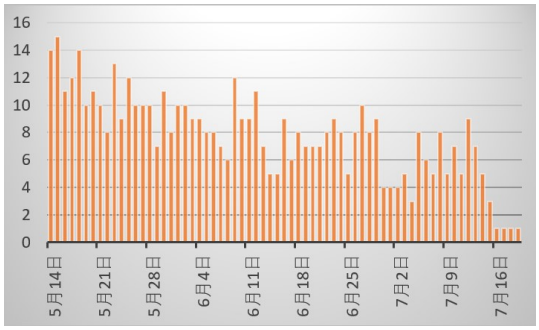

図 3: A 大学コロナ日記の日別データ件数

[ii] https://www.ibm.com/cloud/watson-discovery/

2.4 データ収集の難しさと方向性

この試みで，データ収集を長期的に続けている Slack チャンネルに関しては，開始後 2 ヶ月で参加者が 2 百名を超え，さらに趣旨説明のセミナーなどを行った結果，6 月末には 3 百名を超えた．その後も参加者数は増加しているが，投稿件数はなかなか増えない状況が続いている．原因の一つとして，記名式であり，多くの同僚に読まれることを意識して記入することに敷居の高さを感じる参加者が多いことが考えられる．

当初は，数十万〜数百万件規模の大量のデータを収集し，テキストマイニング技術[4,5]を用いて，個々のテキストに目を通すだけでは気付かないような知見の獲得を実現することを目指していた．現在では，コロナ禍の長期化が見えてきたこともあり，継続的に収集することにより，コロナ禍の継時的変化を捉えることを重要視するようになってきている．

3 コロナ禍データの分析と考察

3.1 記述内容の比較

前節で紹介した 3 種のデータを IBM Watson ® Discovery[ii] の Content Mining application というテキストマイニングツール(以下 WD と略記)に投入し，3 種をまとめた全データにおける出現確率よりも，各データにおける出現確率が高い表現を特徴的な表現として抽出した結果，下記の傾向が見られた．

- 滋賀大学データサイエンス学部のデータ 536 件に特徴的な表現は，「一日中家」「試験」「暑い」「講義」「彦根」「就活」「バイト」など
- A 大学のデータ 521 件に特徴的な表現は，「美味しい」「寝る」「食べる」「嬉しい」「頑張る」「ゆっくり」「辛い」「犬」「癒す」「姉」「飲む」「泣く」「イライラする」「レポート」など
- A 大学データの収集期間と同じ 2020 年 5 月 14 日から 7 月 19 日までに IBM の Slack チャンネルに投稿された 1,089 件のデータに特徴的なのは「記事」「エレベーター」「オフィス」「COVID」「ハンドソープ」「登校」など

コロナ日記においては，日々感じたことが独白的に記述されているのに対し，IBM のデータは Slack チャンネルで公開される性質上，新聞やネットの記

事で見つけた情報を共有するような内容が多く，業務や生活全般に関する広い話題が含まれている．

3.2 有益と感じられる情報

2020 年末までに IBM の Slack チャンネルに投稿されたコロナ禍情報は，合計 2,820 件という，全件に目を通して分析できる量であり，マイニングによって意外な気付きを見いだすことは難しいが，WD により，図 4 のように，「ワクチン」に関する投稿が次第に増えてきているといった知見が得られる．

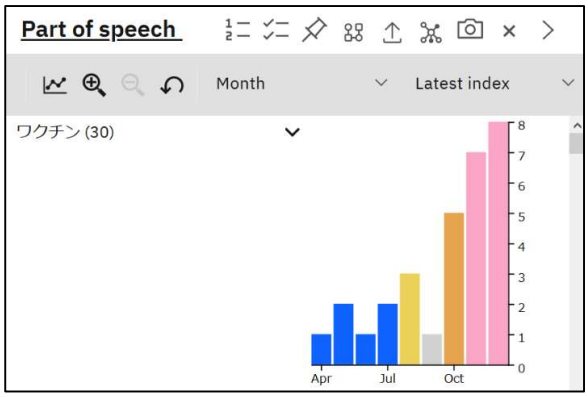

図 4: 増加傾向にある表現の WD による検出画面

2,820 件のうち最も多くのデータに出現する名詞が「人」であり，492 件(17.4%)に含まれている．次が「コロナ」の 407 件(14.4%)で，「マスク」の 380 件(13.5%)が続く．コロナ禍データの大半が「コロナ」という表現を含まないことから，コロナ禍に関するデータを SNS 等から収集するためには，コロナ禍に関する多様なトピックの表現をキーワードとして用いる必要があると考えられる．従って，この試みで収集しているコロナ禍データはコロナ禍関連表現を収集するために活用することもできる．例えば，「リモート手土産」「帰省暮」「幸先詣」といったコロナ禍で生まれてきた新しい表現が，「という言葉」という表現と共起しているのを見出すことができる．

日々この Slack チャンネルにアクセスしている中で，著者らが特に実利性が高いと感じたのが物品の販売に関する情報である．例えば，マスクに関しては，データ収集を開始した当初は，「買えずに困っている」という投稿が多かったが，2020 年 5 月に入ってから次第に「売っていた」「値下がりした」という投稿が見られるようになり，この情報を見て，自宅周辺で適切な値段になるのを待つような動きも見られた．また，「スーパーでバターが買えなくて困っている」という投稿に対し，「ドラッグストアでは買えた」という投稿があり，困っていた投稿者も買えるようになったことが報告されている．

参加者の多くがテレワークを続け外出を控えており，自ら目にできる情報が限られている中，このチャンネルの情報が個々人の視野を広げる効果を出していると感じられる．

3.3 筆者の性格特性の変化の分析

テキストデータから読み取れる情報は，記述された内容だけではない．Big Five Model[6,7]などで数値化できるようになった筆者の性格特性がテキストに表現されることが報告されている[8]．性格特性は，基本的には個人の特徴として一貫性があることが期待されるが，その判断基準が「人生を楽しんでいる」かどうかといった本人の感覚に依存していることからも，個人の状況に応じ，多少なりとも変動する．

著者らは日本語テキストから筆者の性格特性を推定するシステム（IBM Watson Personality Insights： 以下 PI と略記）を開発し[9]，これを用いて，2016 年に発生した熊本地震の被災者のツイートを分析した際には，big5_conscientiousness（誠実性・几帳面さ）の facet_cautiousness（注意深さ）の値が被災中に一時的に高まる現象を，また，突然入院した患者のツイートの分析では big5_agreeableness（協調性）の値が入院中に一時的に高まる[10]現象を見出した．

今回収集したコロナ禍データでも同様の性格特性の変化が見られないかを調査した．まずは，A 大学のコロナ日記を対象として，2020 年 5 月 14 日から 6 月 2 日までの 20 日間（新型コロナウイルスの PCR 検査新規陽性者数が減少傾向にある収束期），6 月 3 日から 6 月 22 日までの 20 日間（新規陽性数が少ない安定期），6 月 23 日から 7 月 13 日(一部 19 日)までの約 20 日間（新規陽性数が増加傾向にある再拡大期）に分け，日記を記述してくれた 11 名の学生の各期間のテキストから PI で推定される性格特性の変化を調査した．さらに，同期間に IBM の Slack チャンネルへの投稿が多かった 9 名のテキストから PI で推定される性格特性の変化も調査した．双方のデータにおいて変化量が比較的大きいのは，big5_conscientiousness の facet_dutifulness（忠実性）であるという結果が得られた．その変化状況を図 5 と図 6 に示す．これらの図からは，自粛を守る生活が続く一方で，そのリターンがないことで徐々に動機付けも下がり，安定期になって一部の学生や社員の忠実性が低下している様子が見て取れる．さらに

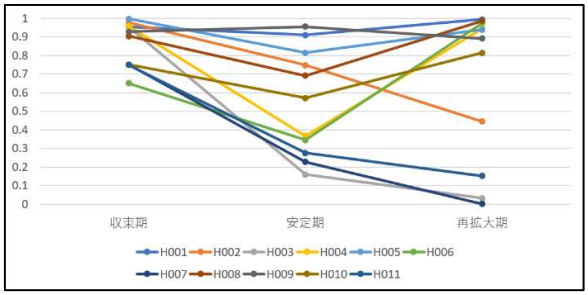

図 5: A 大学コロナ日記の筆者の dutifulness の変化

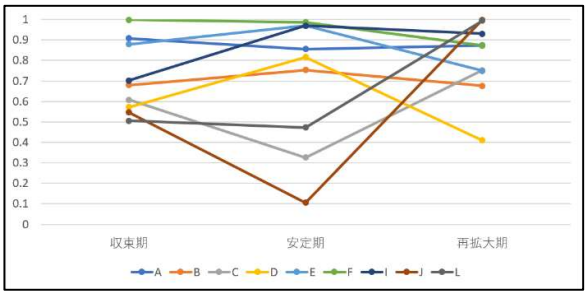

図 6: IBM の Slack への投稿者の dutifulness の変化

再拡大期に，忠実性の値が再び高まる（戻る）ケースとさらに下がるケースに分かれる様子が見られる．

とりわけ学生においてそのばらつきが顕著であり，このことは，オンライン授業導入当初に大学から学生に対して行われていた遠隔でのサポートが，安定期に入って手薄になったことを反映している可能性もある．こうした長期間にわたる非常事態において学生のドロップアウトを防ぐためには，むしろ安定期のサポートが重要であることがうかがえる．

3.4 感染者数増減傾向推定の可能性

この試みを始めるにあたり，変化のきっかけの一部でも捉えられないものかと期待していた感染者数の増減傾向は，コロナ禍が続く中で，非常に多くの要因に左右される難しい問題であることが分かってきた．Slack チャンネルに蓄積されたデータには，量的な少なさもあり，感染拡大や縮小傾向の特定のきっかけが示されていることは無さそうである．

それでも，今回収集しているデータに加え，日々報道される様々な情報や第一波・第二波を通して，ある仮説を考えるようになった．その仮説とは，危機感を訴える情報が減って安心感が広がると拡大傾向が始まり，拡大傾向に伴い危機感を訴える情報が増えると自粛による行動変容が進み，縮小傾向に転ずるというものである．この仮説が第三波に当てはまるなら，2020 年 12 月上旬には収束傾向に移行するのではないかという希望的観測をしていた．しかし，残念ながら 2021 年 1 月に入っても拡大傾向が止まらない状況である．

IBM の Slack データから，その背景を示唆しうる投稿を確認できた．2020 年 11 月 20 日に『テレビを見ていたら，「これ以上，何をすれば良いのでしょう？」という質問が寄せられていました.』という投稿があり，そこに 12 名が「同感です」というリアクションをしていた．このことから，多くの人が「これ以上何もできない」と考え，行動変容が起こる余地が減っていた可能性が考えられる．

こういったデータや図 5，6 のようなデータが示唆しうるコロナ疲れの状況を把握することで，新規陽性者数の増減傾向の変化につながる行動変容の余地の大小を多少なりとも捉えられる可能性はありそうだと考えられる．

4 おわりに

コロナ禍で自然言語処理の技術を活かそうという多くの取り組み[11-15]が存在する中，日常生活の状況を記録し役立てようとする試みを示した．社内外の多くの協力者に依存する取り組みであり，データ収集の難しさから，ビッグデータを集めることは出来ていないが，これまでに集まったデータの価値は大きいと感じている．

コロナ禍において，ずっと巣ごもり状態にある人がいれば，コロナ禍前と変わらぬ活動を続ける人もいて，人々の行動や感受性が多様化している．性格特性の変化にも大きな個人差が認められる．データの継時的変化に顕著な特徴が現れない性格特性も多く，それは，コロナ禍の持つストレス状況のあいまいさや，リスクの捉え方における個人差を反映しているとも推察される．こういった多様性を理解し寛容性を高めることにつながるなら大きな意義があり，その可能性がこのデータには含まれていると考える．

謝辞

本取り組みに協力してくださった多くの方々に感謝いたします．

参考文献

1. Murakami, Akiko, and Tetsuya Nasukawa. "Tweeting about the tsunami? Mining twitter for information on the Tohoku earthquake and tsunami." In Proceedings of the 21st International Conference on World Wide Web, pp. 709-710. 2012.
2. Murakami, Akiko, Tetsuya Nasukawa, Kenta Watanabe, and M. Hatayama. "Understanding requirements and issues in disaster area using geotemporal visualization of Twitter analysis." IBM Journal of Research and Development 64, no. 1/2 (2019): 10-1.
3. Zhu, Wei-Dong, Asako Iwai, Todd Leyba, Josemina Magdalen, Kristin McNeil, Tetsuya Nasukawa, Nitaben Nita Patel, and Kei Sugano. IBM content analytics version 2.2: Discovering actionable insight from your content. IBM Redbooks, 2011.
4. 那須川哲哉. テキストマイニングを使う技術/作る技術-基礎技術と適用事例から導く本質と活用法, 東京電機大学出版局. 2006.
5. 那須川哲哉, 吉田一星, 宅間大介, 鈴木祥子, 村岡雅康, 小比田涼介. テキストマイニングの基礎技術と応用, 岩波書店. 2020.
6. Goldberg, Lewis R. An alternative" description of personality": the big-five factor structure. Journal of personality and social psychology 59.6: 1216, 1990.
7. McCrae, R. R. and John, O.P. "An introduction to the five‐factor model and its applications." Journal of Personality, 60(2), 175-215, 1992.
8. Mairesse, F., Walker, M.A., Mehl, M.R., and Moore, R.K., Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. In Journal of Artificial intelligence Research, 30: 457-500, 2007.
9. 那須川哲哉, 上條浩一. "日本語における筆者の性格推定の取組み" 言語処理学会第 23 回年次大会発表論文集, pp.807-810. 2017.
10. 那須川哲哉, 上條浩一, 榎美紀, 鈴木祥子, 山下紗苗, 上泰, 権藤恭之, 北村英哉, 尾崎由佳. "テキストから推定される筆者の性格特性情報の活用の試みと考察" 言語処理学会第 26 回年次大会発表論文集, pp.1439-1442. 2020.
11. 荒牧英治, 若宮翔子, 矢田竣太郎, "COVID-19 に関する自然言語処理", 自然言語処理 2020 年 12 月, Journal of Natural Language Processing Volume 27 Number 4, pp.933-937, 2020.
12. 河原大輔, "オープンコラボレーションによる COVID-19 世界情報集約サイトの構築", 自然言語処理 2020 年 12 月, Journal of Natural Language Processing Volume 27 Number 4, pp.939-943, 2020.
13. Zhang, Edwin, Nikhil Gupta, Rodrigo Nogueira, Kyunghyun Cho, and Jimmy Lin. "Rapidly Deploying a Neural Search Engine for the COVID-19 Open Research Dataset." In Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. 2020.
14. Wolohan, J. T. "Estimating the effect of COVID-19 on mental health: Linguistic indicators of depression during a global pandemic." In Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. 2020.
15. Sun, Shuo, and Kevin Duh. "CLIRMatrix: A massively large collection of bilingual and multilingual datasets for Cross-Lingual Information Retrieval." In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 4160-4170. 2020.