A Preliminary Analysis of Newspaper Editorials on COVID-19 Using Natural Language Processing Technologies : Differences among Newspapers and Further Research

自然言語処理技術を用いた新型コロナウイルスに関する新聞社説の予備的分析
―新聞社ごとの違いと研究の展望―

A Preliminary Analysis of Newspaper Editorials on COVID-19 Using Natural Language Processing Technologies
-Differences among Newspapers and Further Research-

星野 雄介（武蔵野大学 経営学部 准教授）[1]

Abstract

This paper aims to analyze newspaper editorials on COVID-19 using natural language processing techniques, especially topic modeling, sentiment analysis, and sentence similarity, and to identify the effectiveness and limitations of the techniques and the prospects for research. The results of editorials related to COVID-19 from January 2020 to October 2021 shows that (1) the issues related to COVID-19 are diverse, and the reaction of newspapers varied. The results suggest that the tone of the newspapers are not necessarily in line with existing “liberal-conservative” axis. (2) However, some problems with natural language processing are stressed, such as problems derived from fine-tuning in topic modeling and similarity analysis, probabilistic methods in topic modeling, and accuracy problems in sentiment analysis. (3) However, quantifying using natural language processing technologies showed the possibilities to analyze media trends in more detail by incorporating external variables such as the number of COVID-19 cases or severe cases, and by comparing newspaper editorials from years other than 2020-2021.

キーワード：新型コロナウイルス、社説、自然言語処理、トピック・モデリング、感情分析、極性分析、文書類似度、コサイン類似度
Keywords: COVID-19, Newspaper Editorial, Natural Language Processing, Topic Modeling, Sentiment Analysis, Polarity Analysis, Document Similarity, Cosine Similarity

１．はじめに

本論文の目的は、新型コロナウイルス（以下、COVID-19）に関する新聞社説を、自然言語処理技術、特にトピック・モデリング・感情分析・文書類似度分析を用いて分析し、手法の有効性と限界、研究の展望を明らかにすることである。具体的には、COVID-19 に関する社説には、①どのようなトピックが含まれているのか。②COVID-19 自体、そして政府に対してどれくらいポジティブなのか。③どのような場合に内容が近寄るのか、について明らかにしたうえで、研究の限界と展望について考察していく。

COVID-19 は 2019 年 11 月に確認され、2020 年に入ると、世界的に拡大を始めた。2020 年 3 月には WHO が、COVID-19 の感染拡大がパンデミックであるとの認識を示し、各国に対して対策強化を求めた。その後、COVID-19 は世界中で何度も流行と終息を繰り返している。2022 年 2 月 4 日時点では、世界の COVID-19 感染者は 3 億 8803 万人、死者は 571 万人であり、日本でも感染者数 3019 万人、死者 1 万 9054 人に達している[2]。

収束しない COVID-19 に対して、様々な対応がとられている。第 1 が入院やワクチンといった医療対応である。第 2 が、活動制限であり、日本においては 4 回の緊急事態宣言、地域ごとにまん延防止等重点措置が実施された。休校やテレワーク、オンライン講義、飲食店の営業時間短縮によって人々の活動に制約がかかった。第 3 が経済政策であり、各種の給付金や経済キャンペーンなどがあげられる。2020 年前半の定額給付金や、旅行や観光などの需要を喚起する目的の「Go To キャンペーン」、事業者に対する各種給付金などが代表的である。

このような状況下で、様々な世論が生じている。古くから存続するメディアである新聞・雑誌・ラジオ・テレビに加え、2010 年代に流行したソーシャル・メディアにおいても、様々な立場の人々が様々な意見を述べており、それは世

論を形作っている。日本における COVID-19 に関する新聞報道について、Parvin et al. (2020) らは 2020 年 1 月から 3 月の英字新聞である *China Daily* と *Japan Times* のディスコースを比較し、3 ヶ月の間に「健康と対策」から「経済・政治・社会福祉」へと焦点が移ったことを明らかにした。ソーシャル・メディアの分析として、鳥海（2020）は、2020 年 1 月 17 日から 4 月 30 日の間の Twitter でのツイートを分析し、COVID-19 は 2 月末に一般化したことや、「気のゆるみ」が 3 月の連休中に観測されたことを明らかにしている。

そのような世論を生み出すメディアにおいて、新聞は重要な位置を占める。Scheufele と Tewksbury（2007）によると、新聞は 3 つの方法で人々の認識や意思決定に影響を与える。第 1 の方法がフレーミング（framing）であり、人々が物事を理解する枠組みを形作ること、第 2 がアジェンダ・セッティング（agenda setting）であり、人々がどのような点に注目すべきかを、報道量や頻度によって指示することである。第 3 がプライミング（priming）であり、特定の論点がメディアによって強調されることによって、有権者が政権や政治的リーダーを評価する際の基準としてのその論点が機能する、という考えである。そして、そのような新聞の中でも、新聞社の主張、すなわち社論が強く反映されているのが、社説である。社説はその新聞社の論説委員会で議論し、論説委員会が執筆するといわれており、社論の中核をなす（高木 2004）。それゆえ、既存研究でも社説を世論の典型事例として取り上げられてきた。田中と藤井（2015）は 1950 年代からの公共事業に対する新聞論調の変遷を分析している。伊藤（2012）も、福島第一原発事故後の新聞社の論調を、社説を通して分析している。COVID-19 に関して、新聞社説を分析した研究としては、星野と平尾（2021）が挙げられる。星野と平尾（2021）によると COVID-19 に関する新聞社説のテキストマイニングによって、第 1 に感染者数の増減にかかわらず、感染や医療に関する論点が減少し、社会や経済に関する論点が高止まりしていること、第 2 に、「朝日新聞」と「読売新聞・毎日新聞・日本経済新聞」というクラスタに分類でき、クラスタ間で重視する論点の相違がみられた。

従来、新聞社説に関する研究手法として、内容分析やテキストマイニングが用いられることが多かった。内容分析は、「社会調査によって得られる質的データを分析する方法」（樋口 2006, p.3）と定義される。データの各部分あるいは

全体に、特定のコードを割り当てることで、分析していく方法であり、19 世紀から 20 世紀初頭、新聞の印刷部数が増大したことを背景に広まったといわれている。質的研究の一種であるが、付与されたコードは量的にも分析される。テキストマイニングとは「テキスト（文章）をマイニング（情報発掘）することであり、定性的な特徴を持つテキストを定量的に分析すること」（小木 2015, p.31）と定義でき、多様な研究が発表されている。

他方で、近年、自然言語処理という技術が普及始めている。北中（2020）に依拠しながら、これを簡単に整理していこう。自然言語とは我々が日常的に用いている言語であり、プログラミング言語や論理式などの人工言語・形式言語などと対比できる。自然言語処理とは、このような自然言語を処理する技術の総称である。主な処理内容としては文書分類・機械翻訳・文書要約・質問応答・対話などがあげられる。2022 年時点で実装されているサービスとしては、Google の Google アシスタント、Apple の siri、Amazon の Alexa、Microsoft の cortana などがあげられる。また、自然言語処理を得意とするベンチャー企業が多様なサービスを提供している。

これらの自然言語処理技術には様々なものがある。代表的な技術として、文書を構成するテーマを確率論的に明らかにする「トピック・モデリング」、文書の感情（ネガティブかポジティブ。あるいは喜怒哀楽のような感情）を明らかにする感情分析、文書の類似度を明らかにする「文書類似度分析」、単語をベクトル化して単語の持つ文脈を明らかにする「Word2Vec」などがあげられる。

これらの自然言語処理は近年、人文学や社会科学にも普及を始めている。筆者が専門とする経営学分野でも、経営情報を扱う経営情報学会や、多様なバックグラウンドを持つ人々が参加する研究イノベーション学会で論文が発表されるのみならず、2021 年には、日本における経営学分野の中心的学会誌である『組織科学』においても新たな分析方法として自然言語処理が紹介された（例えば、村瀬他 2021; 山口　2021）。しかしながら、COVID-19 の新聞分析においては、いまだ自然言語処理技術が用いられていないようである。そこで本論文では、各新聞社の社論を自然言語処理技術を用いて分析したうえで、発見事実と研究の展望を考察する。

本論文の構成は以下のとおりである。次節において、トピック・モデリング

を用いて新聞社説を構成するトピックの時系列変化や新聞社ごとのトピックの偏りを明らかにしていく。第 3 節において、感情分析を用いて社説の論調を分析する。この説では社説を構成する文書を 1 文ごとに分割して分析しており、内容分析との関連がみられる。第 4 節において、同じ日の各新聞社の社説間の類似度を測定することで、どういう場合に論点が一致するのか、を明らかにしていく。第 5 節は、得られた結果を整理し、今後の研究の展望を議論する。

２．トピック・モデリングを用いた分析

第 2－4 節では共通して、①その節で用いられる手法の紹介、②データの範囲・取得方法・データ概要、③分析方法、④結果、⑤小括という構成をとる。

①　トピック・モデリング

本節では、「トピック・モデリング」を用いた新聞社説の分析を行う。トピック・モデリングとは、確率論的アルゴリズムに基づいて話題やテーマといったトピックを抽出する手法である。トピック・モデリングの手法としては、潜在的意味解析（Latent Semantic Index：LSI)や LSI を確率生成モデルとした確率的潜在意味解析（Probabilistic Latent Semantic Indexing：PLSI)、潜在的ディリクレ配分法（Latent Dirichlet Allocation：LDA）が代表的である。本論文では LDA を用いたトピック・モデリングを行う。LDA とは、「文書は複数のトピックから構成されるということを前提としたモデルである。単語は潜在的にトピックを持ち、同じトピックを持つ語同士は共起しやすい、という考え方に基づき、単語のトピックを確率的に求める手法」（土村 2017, p.178）である。

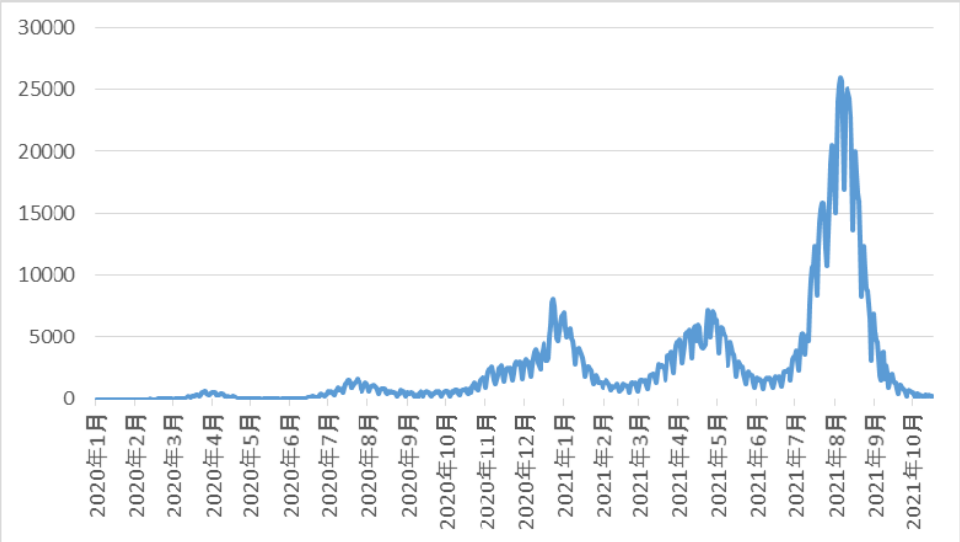
②　データ

本論文では、朝日新聞・毎日新聞・読売新聞という日本を代表する全国紙と、同じく日本を代表する経済専門紙である日本経済新聞（以下、日経新聞）の、合計 4 紙の新聞社説をデータとする。もちろん、日本においては、これらの新聞以外にも全国紙・地方紙・専門紙が数多く発行されているものの、本論文で選択された 4 紙は、一般的に日本において主要な新聞とみなされているから、

十分な分析が行えると判断される。

分析期間は、2020 年 1 月 1 日から 2022 年 10 月 31 日とした。COVID-19 は 2019 年 11 月に発生しているものの、日本で初の COVID-19 の感染者が発見されたのは 2020 年 1 月 15 日であり、開始時点として 2020 年 1 月を設定することに一定の妥当性があると思われる。また、終わりを 2022 年 10 月末としたのは、図 1 のように 2021 年 7-9 月のいわゆる第 5 波の流行も終息したタイミングであるためである。

図 1　日本の COVID-19 感染者数推移（日次）



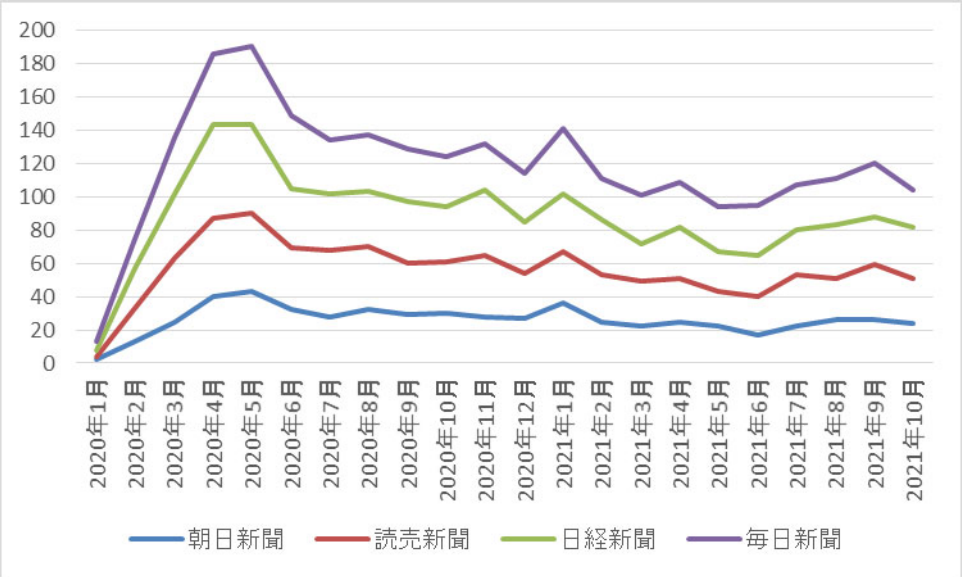
出所：厚生労働省特設サイト [3]。

データは、星野・平尾（2021）と同様に、朝日新聞・毎日新聞・読売新聞・日経新聞の各紙のオンラインデータベースである「聞蔵 II ビジュアル」（朝日新聞）、「毎索」（毎日新聞）、「ヨミダス歴史館」（読売新聞）、「日経テレコン２１」（日経新聞）から、「新型肺炎」「コロナ」の 2 種類のキーワードで検索した。この検索方法では、記事タイトルのみならず、本文のなかに 1 回でも「新型肺炎」「コロナ」という単語が含まれる記事が検索される。そのため、COVID-19 を中心論点としていない記事も多く抽出されるが、COVID-19 は我々の社会の多方面に影響を与えていることから、むしろより現実を説明できると思われる。検索された記事の中から社説のみを取得した。その後、タイトルと本文を接合、

すなわち、タイトルを第 1 文とみなし、分析を進めていった。

以上の手順で取得された社説は、全 2612 本であった。内訳は、朝日新聞 574 本、毎日新聞 667 本、読売新聞 709 本、日経新聞 662 本であり、月ごとの社説本数の推移は図 2 のようになっている。一見して明らかなように、学校が休校となった 2020 年 3 月や、緊急事態宣言が発出された 2020 年 4-5 月に社説本数が増加している。その後、2021 年 1 月の 2 度目の緊急事態宣言発令時に一時的に本数が上昇するものの、2021 年 6 月までは緩やかに減少し、第 5 波と呼ばれる、2021 年 7-9 月に再び上昇している。

図 2　新聞社説の推移



注：積み上げ式の折れ線グラフである。

③ 分析方法

分析は LDA を用いている。LDA を問わず機械学習においては、各種のパラメターのファインチューニングが極めて重要である。そこで、分析するにあたり以下のように各種パラメターを調整した。

ストップワード

ストップワードとは、自然言語処理にあたり処理対象外とする単語のことで

ある。通常は、あまりに一般的すぎ、かつ出現回数が多いため、文脈を捉えるのに不都合な単語、例えば「する」「ある」「ら」などをストップワード対象とする。日本語のストップワードとしては、國府他が作成したものなどがある（國府他 2013）が、本論文では、日本語のストップワードとしては最もポピュラーな Slothlib の日本語ストップワードを用いている[4]。さらに、Slothlib 日本語ストップワードには登録されていないものの、分析を通して不要だと思われる単語（「期」「か月」「年度」など）を追加でストップワードとして登録している。

品詞

本論文では分析に当たり、品詞を名詞のみに限定した。既存研究（例えば、吉田 2020）と同様に、社説が取り扱っている内容に焦点を当てるためである。

単語の出現回数によるスクリーニング

本論文では分析に当たり、50%以上の社説で出現する単語は除外した。あまりに頻繁に出現する単語を分析に含めることで、むしろトピックの広がりがわかりにくくなると考えられるためである。さらに、50 回以下しか出現しない単語も除外している。これら 2 つの処理は、ストップワードのような辞書作成したのではなく、プログラム上で実現した。ただし、これらの処理が持つ問題については、第 5 節で議論する。

形態素分析

日本語は英語などと異なり、「スペース」で単語ごとに区切られているわけではない。そのため、文字が連続している文書から文法や品詞情報に基づき単語単位で切り出す必要がある。このことを形態素分析という。日本語の形態素分析ツールとしては、「MeCab」「ChaSen」「JUMAN」などが代表的である（伝 2009）。また、プログラミング言語 Python においては、MeCab の辞書である ipadic を利用した「Janome」[5]や、株式会社リクルートの AI 研究機関と国立国語研究所との共同研究によって生み出された「GiNZA」[6]があげられる。本論文では、次に見る最新の辞書を使用するために、MeCab[7] を用いている。

辞書

MeCab で形態素分析を行う際、専用の辞書を参照する。通常の MeCab では ipadic という MeCab に同梱された辞書を参照しているが、この ipadic に含まれている単語は多くなく、また、新しい単語に対応していない。COVID-19 に関しては、「三密」や「コロナ禍」といった言葉が生み出されているが、ipadic ではこれらに対応することができないのである。そこで、本論文では、この問題に対応するために、「mecab-ipadic-NEologd」[8] という辞書を用いる。mecab-ipadic-NEologd は、ipadic では正しく分割できない固有表現などの語を 300 万語以上採録しており、週に 2 回自動的に更新される。そのため、2020 年以降に普及した単語についても分析可能となる。分析は 2022 年 2 月上旬であり、COVID-19 に関連した単語も含まれている。

トピック数

トピック・モデリングにおいて、文書にどれだけのトピックが含まれているのかは、事前に決定する必要がある。その際は、Perplexty と呼ばれる指標に着目することが多い。そこで本論文では以上の条件をもとに、予備的に分析し、Perplexity の観点からトピック数を 40 とした。

ライブラリ

LDA の実行には「Gensim」[9] を用いた。Gensim は大規模なコーパスを用いたトピック・モデリングや文書インデックス作成、類似度分析のための Python ライブラリであり、トピック・モデリングでは頻繁に用いられていることから（例えば、de Oliveira Capela& Ramirez-Marquez 2019)、本論文でもこれを採用する。

④ 結果

トピックと構成単語

LDA を実行しトピックと構成単語が得られた。その構成単語から、筆者自身の判断でトピック名をつけた。構成単語のすべてが 1 つのトピックを、必ずしも示しているわけではないが、一定の範囲で文脈を想定できると思われる。そ

の結果が表 1 である。

表 1　トピック・トピック名・構成単語

トピック番号	トピック名	構成単語
1	地域と学校	学校, 地域, コロナ, 支援, 子ども, 中国, 日本, 米, 問題, バイデン
2	東アジア情勢	日本, 中国, ロシア, 首相, 官僚, 予算, 保有, 要求, 強化, 会議
3	政治と五輪	選手, 国民, 五輪, 事件, 議員, 首相, 説明, 日本, 東京五輪, 自民党
4	五輪開催	大会, 子ども, 五輪, 選手, 組織, 開催, パラリンピック, 委, コロナ, 支援
5	デジタル	デジタル, 庁, 社会, 課題, デジタル化, 国民, 行政, 企業, データ, 問題
6	企業経営	企業, 制度, 問題, コロナ禍, 国会, 議論, 雇用, 首相, 対応, 人材
7	国際的対応	感染, 中国, WHO, 調査, 対応, 米, 日本, 協力, ウイルス, 現地
8	日米中	中国, 日本, 首相, 米国, 政権, 世界, 総裁選, 米, 経済, 安定
9	社会保障	女性, 企業, 大学, コロナ禍, 日本, 制度, 社会保障, 雇用, 支援, 仕事
10	与党	首相, 給付, コロナ禍, 国民, 政治, 危機, 自民党, 感染, 問題, 選挙
11	ワクチン接種（運用面）	接種, ワクチン, 自治体, 感染, 高齢者, 支援, 影響, コロナ, コロナ禍, 子ども
12	景気回復	企業, 回復, 製造業, 感染, 改善, 景気, 需要, リスク, 3月, 経済
13	若年層の生活	子ども, 文化, 日本, 中国, 社会, 地域, 企業, 若者, 感染, コロナ禍
14	総裁選	首相, 政権, 問題, 安倍, 国民, 政治, 自民党, 総裁選, 情報, 自民党総裁
15	米国金融	経済, 米, 米国, ドル, 成長, 政権, 日本, 株価, コロナ禍, 政策
16	政策対応	コロナ, 感染, 問題, 経済, 議論, 対応, 政策, 政権, 予算, 首相
17	学校	学校, 負担, 感染, 国民, 子ども, 判断, 問題, 制度, 課題, 自治体
18	経済活動	企業, 引き上げ, 投資, 経営, 賃金, 価格, 支援, 制度, コロナ禍, 成長
19	自治体による医療対応	自治体, 患者, 施設, 価格, 接種, 利用, 確保, 高齢者, 医療, 企業
20	経済縮小	コロナ, 感染, 企業, 日本, 経済, 縮小, 対応, 回復, 需要, 米

トピック番号	トピック名	構成単語
21	米国経済	米国, 米, 工場, 企業, 世界, バイデン, 日本, 生産, 巨大, 社会
22	経済回復	経済, 感染, 日本, 回復, 拡大, コロナ, ワクチン, 中国, 対応, 状況
23	ワクチン接種（国）	接種, ワクチン, 企業, 米, 日本, 世界, 支援, 中国, 経済, 証明
24	世帯支援	子ども, 経済, 中国, 検討, 支援, 国民, 世帯, 首相, 案, 香港
25	自治体による支援	自治体, 支援, 地域, 地方, コロナ, 対応, 制度, 首相, コロナ禍, 住民
26	地域社会	影響, 子ども, 予算, コロナ禍, 自治体, 社会, 経済, 支援, 今後, 対応
27	衆院選	首相, 国民, 政権, 衆院選, 政策, 選挙, 政治, コロナ, ワクチン, 野党
28	休校と五輪	選手, 休校, 北朝鮮, 感染, 影響, 地域, 学校, 対応, 参加, 五輪
29	宣言と効果	国民, 宣言, 延長, 感染, 対応, 支援, 首相, 効果, 企業, 感染対策
30	子どもへの対応	コロナ, 感染, 自治体, 子ども, 医療, 対応, 避難, 学校, 入院, 家庭
31	地域支援	自民, 需要, 首相, 地域, 支援, 生産, 感染, 供給, 経済, 衆院選
32	ワクチン開発	ワクチン, 接種, 開発, 日本, 体制, コロナ, 経済, 米, 感染拡大, 供給
33	緊急事態宣言	感染, 宣言, 緊急事態宣言, 解除, 発令, 事業者, 防止, 措置, 協力, 首相
34	五輪と感染対応	選手, 大会, 五輪, 開催, 競技, 観客, 無観客, 会場, 感染, IOC
35	米中対立	中国, 国民, 暴力, 投票, 市民, 党, 米国, 批判, 野党, 事態
36	首相動向	首相, 日本, 地域, 問題, 米国, 政権, 合意, 企業, 国民, 目標
37	地域医療体制	接種, 感染, ワクチン, 医療, 体制, 病床, コロナ, 重症, 患者, 自治体
38	感染拡大抑止	飲食店, 自治体, 要請, 再編, 金融機関, 政策, 業者, コロナ, オンライン, 知事
39	新卒採用	学生, 企業, 情報, 採用, 外国人, 国会, 調査, オンライン, 不安, 対応
40	国民生活	国民, コロナ, 五輪, 予算, 首相, 経済, 社会, 開催, コロナ禍, 政策

例えば、トピック 5 は、「デジタル, 庁, 社会, 課題, デジタル化, 国民, 行政, 企業, データ, 問題」という単語で構成されていることから、デジタル庁の発足と社会のデジタル化を示していると思われる。もちろん、COVID-19 に関連したトピックも多くみられる。例えば、トピック 4「大会, 子ども, 五輪, 選手, 組織, 開催, パラリンピック, 委, コロナ, 支援」とあり、コロナ禍での五輪開催だと思われる。他にもワクチン接種に関するトピック（トピック 11、23)、人々の生活にかかわるトピック（トピック 13、28、30、39、40）などがみられる。

この 40 のトピックは、さらにいくつかのカテゴリに集約することができよう。筆者の判断で 40 のトピックを「コロナ対応」「医療」「生活」「経済」「国際」「政治」「五輪」の 7 つに分類した結果が、表 2 である。国内での COVID-19 の感染拡大に対する影響や対策を多角的に検討しているといえよう。

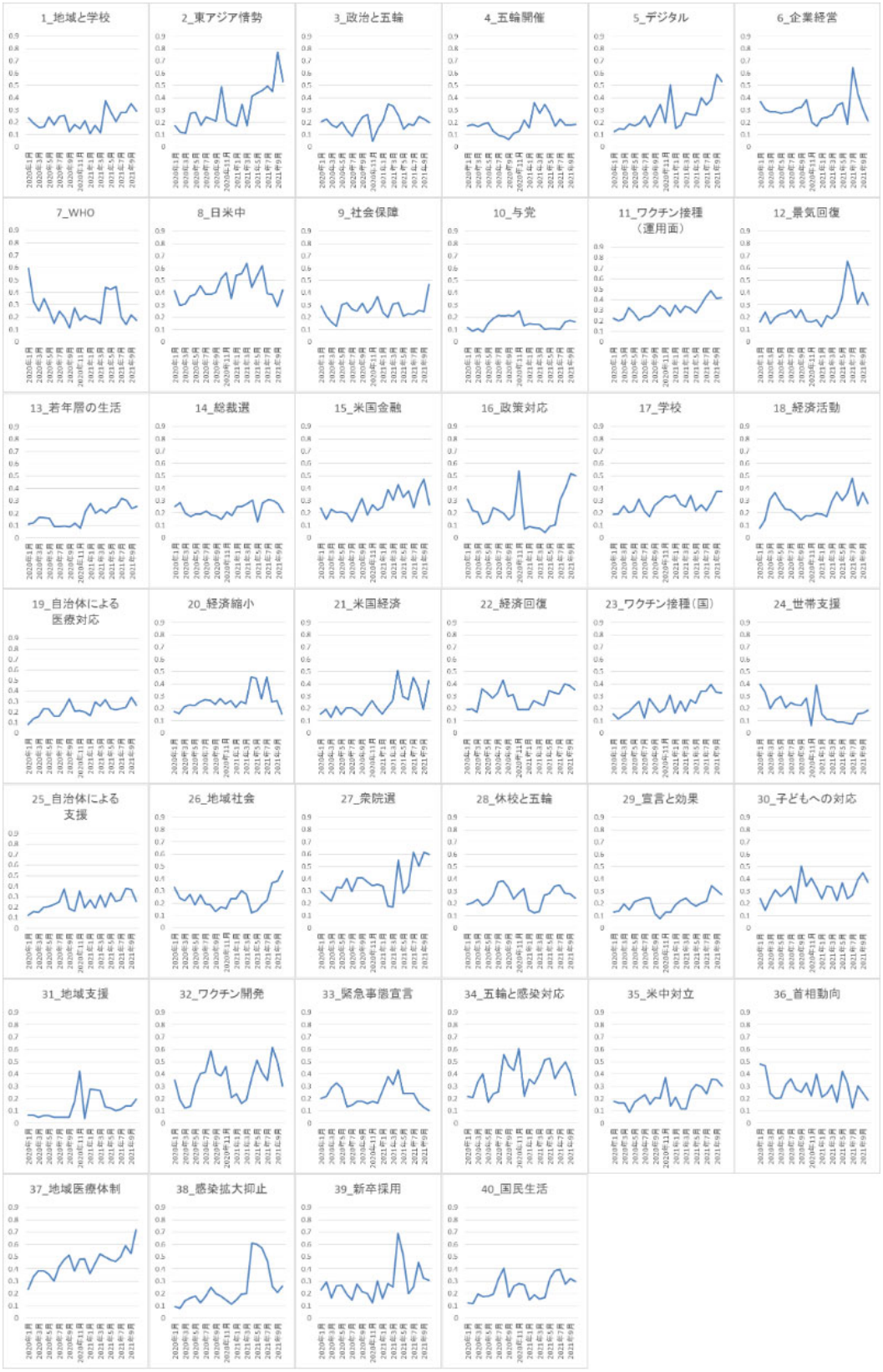
表 2　トピックのカテゴリ

カテゴリ	トピック
【コロナ対応】	9_社会保障・16_政策対応・29_緊急事態宣言と効果・33_緊急事態宣言・38_感染拡大抑止
【医療】	11_ワクチン接種（運用面）・19_自治体による医療対応・23_ワクチン接種（国）・32_ワクチン開発・37_地域医療体制
【生活】	1_地域と学校・13_若年層の生活・17_学校・24_世帯支援・25_自治体による支援・26_地域社会・28_休校と五輪・30_子どもへの対応・31_地域支援・39_新卒採用・40_国民生活
【経済】	6_企業経営・12_景気回復・15_米国金融・18_経済活動・20_経済縮小・21_米国経済・22_経済回復
【政治】	3_政治と五輪・5_デジタル・10_与党・14_総裁選・27_衆院選・36_首相動向
【五輪】	4_五輪開催・34_五輪と感染対応
【国際】	2_東アジア情勢・7_国際的対応・8_日米中・35_米中対立

トピックの時系列推移

40 のトピックの出現率を月単位の時系列で整理することで、各トピックが、どのような変遷をたどったかが明らかになる（図 3)。例えば、トピック 4「五輪開催」について、2020 年から 2021 年へと 1 年延期された五輪については、やはり延期が決まった 2020 年序盤と、東京五輪組織委員長の森会長による失言と辞任、後任人事が発生した 2021 年 2-4 月に出現頻度が増加している。また、トピック 5「デジタル」では、2020 年 9 月に発足が決定し、2021 年 9 月に設置された「デジタル庁」に関する議論だと考えることができる。また、トピック 13「若年層の生活」という論点では、2020 年よりも 2021 年のほうが、全般的に出現率が高くなっている。これは、コロナ禍が長引き、子どもを含めた若年層への影響が大きくなってきており、それに対応することの必要性を示唆していると思われる。

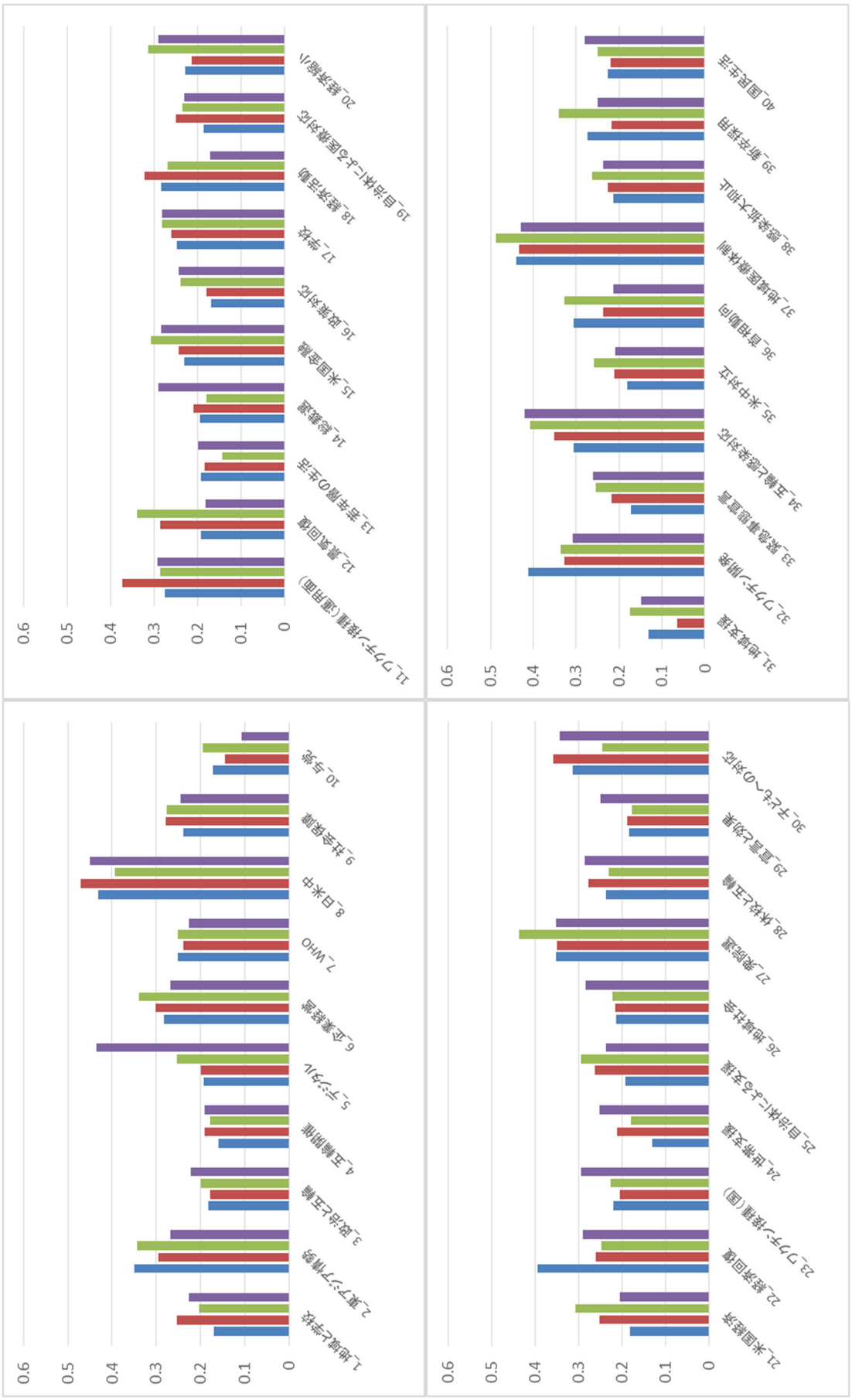
図３　トピックの時系列推移



新聞社ごとのトピックの偏り

最後に、トピックごとの単語の出現率を新聞社ごとのトピックの偏りを整理した（図 4)。一見して明らかなように、4 紙ともトピック 8「日米中」、トピック 37「地域医療体制」について比較的頻繁に言及されていることが分かる。次いで、トピック 27「衆院選」、トピック 34 五輪と感染対策、トピック 32「ワクチン開発」などに言及されている。逆にトピック 31「地域支援」、トピック 13「若年層の生活」、トピック 24「世帯支援」などについては、相対的に言及の頻度が少ないといえる。新聞社ごとでは、毎日新聞はトピック 5「デジタル」、朝日新聞はトピック 22「経済回復」について、読売新聞はトピック 11「ワクチン接種（運用面)」について、他紙よりも言及頻度が高い。本論文のサンプルのうち、唯一の経済専門誌である日経新聞は、トピック 12「景気回復」、トピック 27「衆院選」、トピック 35「米中対立」、トピック 39「新卒採用」などに着目していることが明らかとなった。

図４　新聞社ごとのトピック



⑤ 小括

本節では、トピック・モデリングという文書中に含まれるトピックを明らかにする自然言語処理を用いて分析を行った。分析の結果、COVID-19 に対する影響を多面的に検討していること、一部のトピックについては、時系列変化の背景も比較的容易に考察できること、そして、新聞社ごとに着目するトピックが異なっていることも明らかとなった。従来、朝日新聞と毎日新聞はリベラルあるいは左派とみなされてきた（小林＆竹本 2016）。しかしながら、本結果をみると、同じリベラル派であっても着目する論点が異なっていることが分かる。

３．感情分析

①感情分析について

本節では COVID-19 に関する新聞社説を、感情分析という自然言語処理の手法によって分析していく。自然言語処理における感情分析とは、もともとはテキストに対してポジティブやネガティブといった評価極性を判定する極性分析が主流であった。例えば Koyano et al.（2021）はオンラインミーティング上での発言の感情をポジティブ－ネガティブに分析している。他にも評価極性の代わりに、感情を Ekman の 6 種類の基本感情などに分類する方法も開発され、実際に研究に用いられている（難波&福田 2021）。例えば、ML-ASK というプログラムでは、「哀・恥・怒・厭・怖・驚・好・昂・安・喜」という 10 感情で文書を分析することができる。鳥海（2020）は、この ML-ASK を改良した pymlask[10] を用いて、COVID-19 に関するツイートの感情を分析している。このような感情分析は、経営学においてはトピック・モデリングとならんで活用されていると指摘されている（村瀬他 2021）。本節では感情分析のうち、ポジティブ－ネガティブを測定する極性分析を用いて、新聞社説に含まれる COVID-19 と政治に対する新聞社の感情を分析している。

③ データ

本節で用いるデータは、第2節と同様に、朝日新聞・読売新聞・毎日新聞・日本経済新聞の4紙において2020年1月1日から2021年10月31日までに発表された、

2612本の社説である。社説タイトルを第 1 文とみなしたことも、第2節と共通している。ただし、分析に当たり、第2節では1つの社説全文を分析対象としていたが、本節の感情分析では1文ごとに分割して分析していく。1文ごとに分割する理由は2つある。第1が、内容分析の手法と関連させるためである。内容分析では「line-by-line」と言われるように、1文ずつどのような内容なのかをコーディングしていく。そのため、1文ずつ分割することは、内容分析とのつながりを考慮するのであれば、妥当だと思われる。第2に、分析に用いるライブラリの都合である。本節では、以下に述べるように2つの感情分析ライブラリを用いるが、そのうちの1つである「daigo/bert-base-japanese-sentiment」[11]は、分析できる文字数が500文字に制限されている。他方で、新聞社説は平均して972文字であることから、社説を分割する必要があった。機械的に社説を分割するよりも、1文ごとに分割したほうが、前後の文脈をすべて無視できるという点で望ましいと判断した。なお、2文単位、3文単位で分割・分析しても結果に大きな違いが生じなかった。1文ごとに分割した結果、7万106文を取得した。

続いて、文単位で COVID-19 関係、政治関係の 2 種類にコーディングしていく。しかしながら、7 万文を超えていることから、目視でコーディングしていくのは困難を伴う。そこで、本節では「コロナ」「肺炎」という 2 つの単語のうちいずれかが含まれている文を「COVID-19 関連文」、「首相」「総理」「内閣」「政権」「与党」「自民」「自公」「政府」のうちいずれかが含まれている単語を「政府関係文」とみなした。「COVID-19 関連文」は全 6202 文、「政府関連文」は全 9180 文取得できた。時系列推移と新聞ごとの文数は図 5 と図 6 となっている。

図 5　「COVID-19 関連文」「政府関連文」出現数（時系列推移）

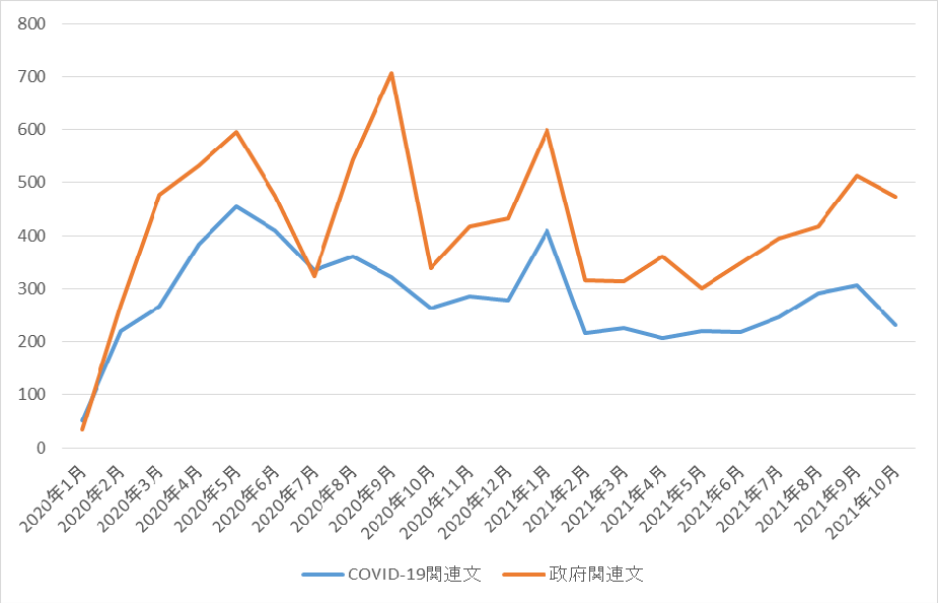
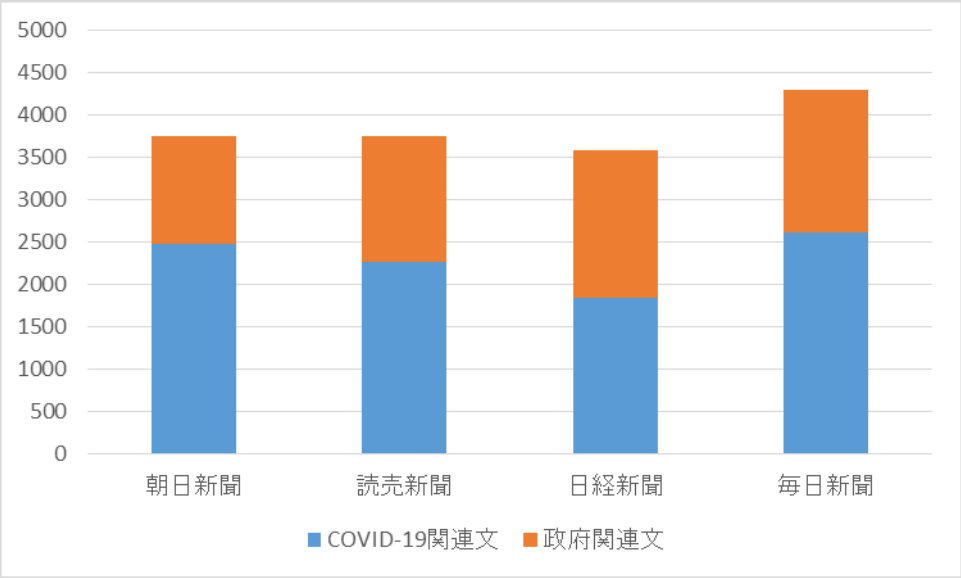


図 6　「COVID-19 関連文」「政府関連文」出現数（新聞社ごと）



④ 分析方法

本節では感情分析のために 2 つの Python ライブラリを用いる。1 つ目は、池上有希乃氏が作成した「oseti」[12] という感情分析ライブラリである。このライブラリは、東北大学の乾・鈴木研究室が作成した「日本語評価極性辞書」（小林

他 2005; 東山他 2008）を用いて、文中に含まれる単語の極性を判別する。例えば、oseti では文書中に、ネガティブな単語が 2 つ、ポジティブな単語が 3 つあった場合、「ポジティブ：3、ネガティブ：2、トータル：+1」というように分析される。また、oseti は否定形にも対応しており、例えば「悲しくない」という文は、「悲しい（ネガティブ）」＋「ない（否定形）」と判断され最終的にポジティブと判定される。しかしながら、元の辞書の都合上、ポジティブ・ネガティブの度合いを分析することはできず、また、複雑な文に対応するのが難しいものの、既存研究（例えば、Koyano et al. 2021; Adach & Negishi 2020）でも用いられており、一定の妥当性があると思われる。本節では、ポジティブ、ネガティブ、ポジティブ－ネガティブ差分（ポジティブ単語数－ネガティブの単語数）の 3 つの数値を用いる。

2 つ目のライブラリは ydaigo 氏が開発した「daigo/bert-base-japanese-sentiment」（以下、daigo/bert）である。このライブラリは東北大学の鈴木・乾研究室が作ったBERTの事前学習モデル[13]をもとに作られた感情分析用のライブラリである。BERTとは、2018年にGoogleが発表した自然言語処理の手法であり、「Bidirectional Encoder Representations from Transformers」の頭文字をとったものである。BERTは当時、様々なタスクにおいて最高スコアを獲得していたが、その理由の一つはTransformerというアーキテクチャを用いて、文脈を読むことが可能になったためと言われている。BERTのモデルを作るためには本来ならテキストを収集して学習させる必要がある。しかしながら、東北大学の乾・鈴木研究室は日本語のWikipediaを学習させた事前学習モデルを公開している。daigo/bertはその事前学習モデルを極性分析用にチューニングしたものである。上述のosetiと異なり、このライブラリで文書を判別させると、ポジティブ度合・ネガティブ度合を算出することが可能である。ただし、ネガティブ度は1から引くことで算出している。

④ 結果

oseti の分析結果

まず、全文を対象に oseti を用いてポジティブ・ネガティブを分析した。一文ごとのポジティブ－ネガティブ差分の月次平均を整理したものが、図 7 である。

図 7　oseti 分析結果（時系列）

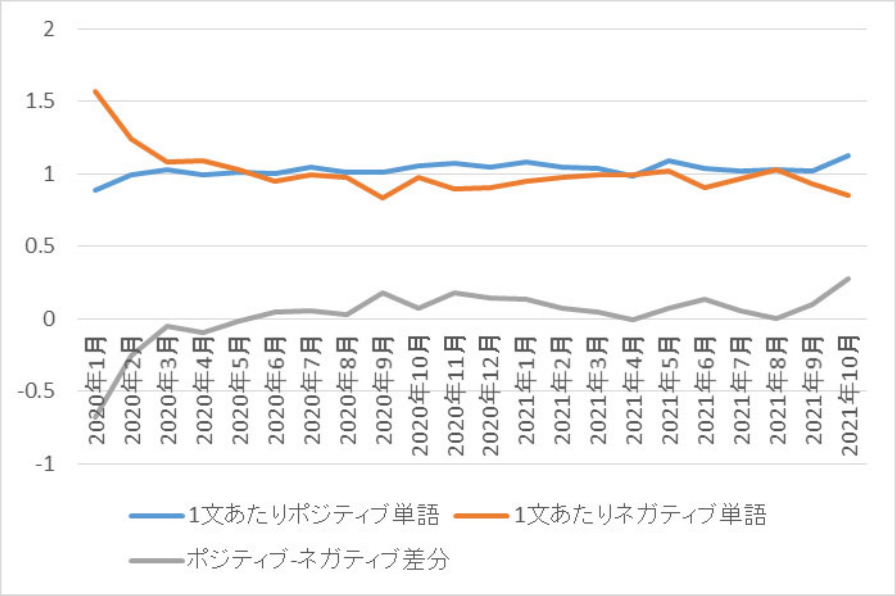
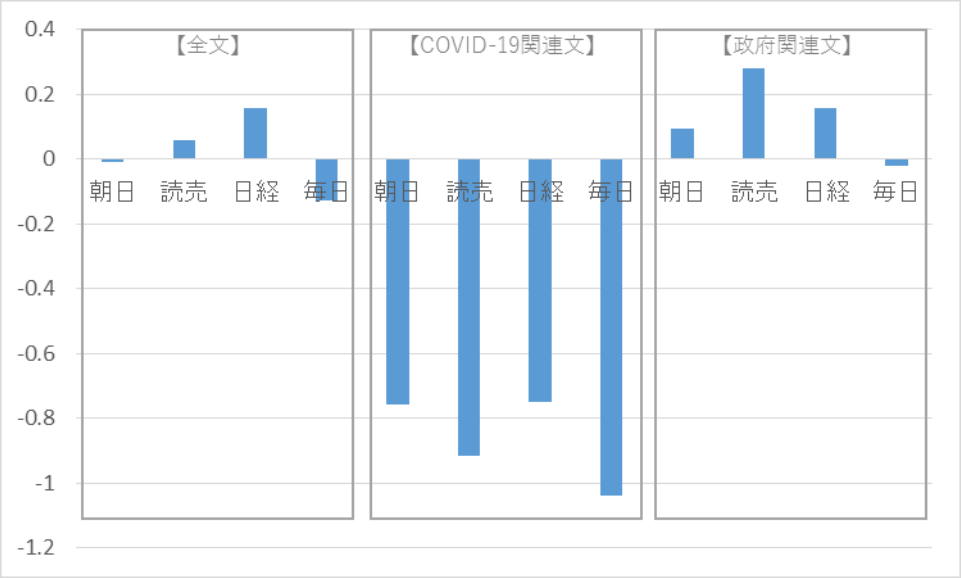


図 7 より、2020 年 1-4 月ではポジティブ－ネガティブ差分（灰色線）がゼロを下回っていた。ポジティブ単語数（青線）・ネガティブ単語数（橙線）平均を見ると、この時期では否定的な単語が多くみられた一方、ポジティブな単語が少なかったことを示している。他方で 2020 年 5 月以降はおおむね 1 文に含まれるポジティブな単語数は、ネガティブな単語数を上回ることで、全体としてポジティブになっていることが分かる。特に、2021 年 5－6 月、9－10 月といった、比較的、新規感染者数が落ち着いていた時期は、ネガティブな単語が減少していた。

続いて、新聞ごとの①全文、②COVID-19 関連文、③政府関連文のポジティブ－ネガティブ差分を整理したものが、図 8 である。

図 8　oseti 分析結果（新聞社ごと）

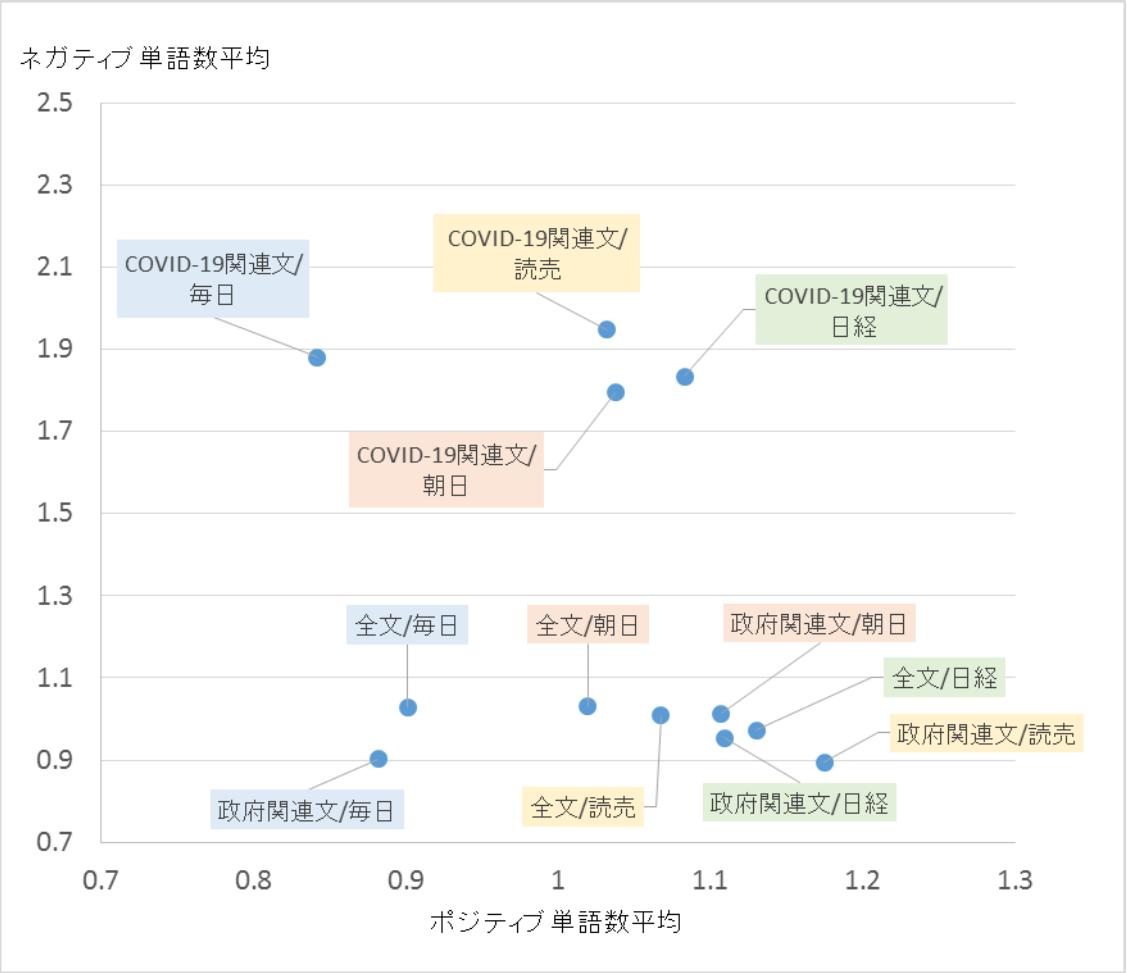


興味深いことに朝日新聞と毎日新聞は、全文でも否定的である。少なくともこの COVID-19 に関して両紙は否定的な論調を持っているといえる。他方で読売新聞と日経新聞は全体としてポジティブである。このような「朝日・毎日」と「読売・日経」といった分類は、既存研究との整合性が高い（竹川 2012; 小林&竹本, 2016; 星野&平尾, 2020）。COVID-19 関連文について、毎日新聞は最もネガティブであり、政府関連文でもネガティブである。毎日新聞と同じように分類される朝日新聞では、COVID-19 関連文はややネガティブであるが、政権に対してはポジティブな論調を示している。他方で読売新聞は COVID-19 関連文においてはネガティブであるものの、政権に対してはポジティブな論調をとっている。既存研究においては党派性の観点から新聞を分類することが多かった。しかしながら、党派性を構成する個別の論点については、新聞間で異なっているということを示唆している。

この論調の違いの原因は、ポジティブな単語の増減なのか、ネガティブな単語の増減なのかは、図 8 からは不明である。そこで、最後に、ポジティブ単語数平均を横軸に、ネガティブ単語数平均を縦軸に置き、新聞ごとの①全文、②COVID-19 関連文、③政治関連文の観点から散布図を作成した（図 9）。この図によって、全文－COVID-19 関連文、全文－政権関連文での論調の詳細な比較が

可能となる。

図 9　ポジティブ単語とネガティブ単語数



注：見出しの色は、新聞ごとにそろえてある

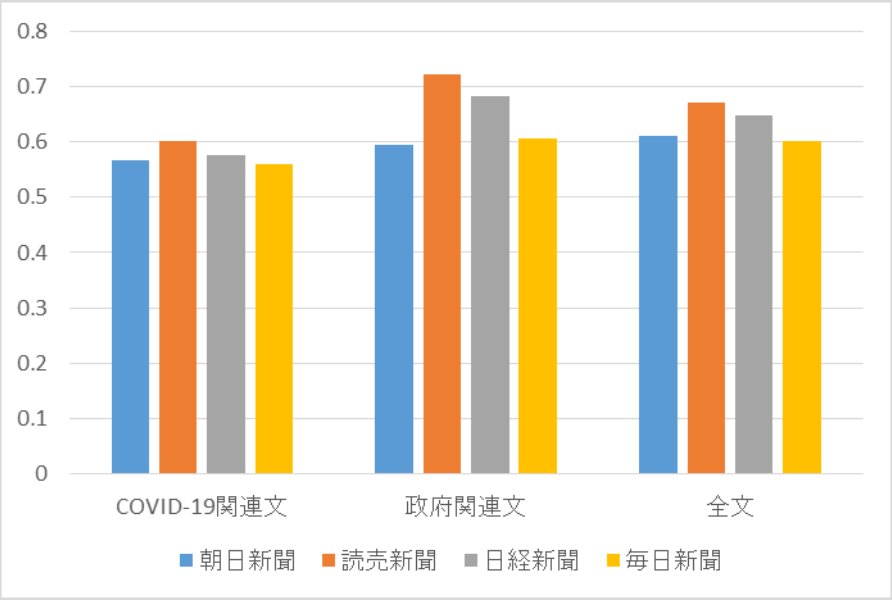
この図から、4 紙ともに、全体と比較したとき COVID-19 関連文では、ポジティブな単語がやや減り、ネガティブな単語が増えることが分かった。しかし、ネガティブ単語の増え幅に新聞間で相違があることから、図 8 のようになったといえる。他方で、政府に対しては状況が異なっている。朝日新聞はネガティブな単語はそれほど変わらず、ポジティブな単語が増加しているが、毎日新聞と日経新聞ではポジティブ・ネガティブ両方の単語がやや減少している。読売新聞と日経新聞はネガティブな単語がやや減少していることは毎日新聞と同様だが、ポジティブな単語が増加している。

daigo/bert の分析結果

続いて、daigo/bert を用いた分析結果を紹介する。図 10 は①全文、②COVID-19 関連文、③政府関連文についての全新聞での時系列での分析結果であり、図 11 は新聞社ごとの分析結果である。

図 10　daigo/bert の分析結果（時系列）

図 11　daigo/bert の分析結果（新聞ごと）



分析の結果、全体に対して COVID-19 関連文はややネガティブであった。政権に対しては時期ごとに変化するものの、ならしてみればポジティブになることが分かった。新聞社ごとに見ても、朝日新聞と毎日新聞が相対的にネガティブであり、読売新聞と日経新聞がポジティブであった。oseti という辞書ベース、daigo/bert という機械学習ベースの 2 種類の感情分析において、おおむね一致した結果が出たことから、本節の発見は、一定の頑健性を持つといえる。

⑤ 小括

本節では、oseti という辞書ベースの感情分析器、daigo/bert という機械学習ベースの感情分析器を用いて新聞間の COVID-19 に関する社説の感情を分析してきた。分析の結果、COVID-19 は相対的にネガティブに捉えられ、政府に対してはポジティブにとらえられていることが分かった。また、新聞社間では比較的ネガティブな論調を持つ朝日新聞・毎日新聞、相対的にポジティブな読売新聞・日経新聞と分類できた。

４．社説間の類似度

① 文書類似度

前節でも見た通り、新聞社には党派性があること、すなわち社論に遠近があると考えられる。しかし、そのような社論を数字で測定する試みとしては、星野と平尾（2020、2021）のようにKH Coderというテキストマイニング・ソフトウェアを用いた分析などがあるものの、決して多くはない。そこで、本節ではコサイン類似度という類似度指標を用いて、社説間の遠近を数量化していく。

コサイン類似度とは、ベクトル空間モデルにおいて、文書間の類似度を計算する手法である。具体的には、各文書に出現する単語と出現頻度をベクトルとみなし、それらのベクトル間の角度を計算する。もし、コサイン類似度が 1 なら文書は非常に似ており、ゼロなら全く似ていないと判断できる。

文書間の類似度をコサイン類似度によって計算する場合、単語の重要度に注目することが多い。単純な単語出現回数のみで分析した場合、「する」「ある」「ない」といった意味がないにも関わらず多く出現する単語が重要視される一方で、限られた文書にしか出現しないが文書を特徴づける単語の重要度が下がってしまう。この問題点を解決するために「tf-idf」（term frequency – inverse document frequency）という尺度が用いられる。tfは「ある単語の一つの文書における出現頻度」であり、idfは「ある単語が複数の文書のなかで、いくつの文書に出現しているか」を計算した指標である。「tf-idf」はtfとidfを掛け合わせたものであり、自然言語処理においては一般的に用いられている（難波 2020）。本論文では、tf-idfを用いてコサイン類似度を計算することで、新聞社説間の類似度を測定する。

② データ

データは前節までと同様に2020年1月1日から2021年10月31日までの朝日新聞・読売新聞・毎日新聞・日本経済新聞の社説のうち、「コロナ」「肺炎」が含まれているものをデータとした。社説の本数は合計で2612本であり、社説のタイトルも第 1 文とみなしている。ただし、社説単位での類似度を測定することが目的であることから、前節と異なり、1文単位で分割せず、全体のままを用いることとした。

分析では、ある 1 日の社説を新聞社間で総当たりで類似度を計算していく。例えば、2020 年 2 月 15 日は、4 紙において、1 本ずつ「コロナ」を含む社説があった。この場合は、「朝日-日経」「朝日-毎日」「朝日-読売」「日経-毎日」「日経-読売」「毎日-読売」の 6 通りのコサイン類似度を計算する。この考え方によると、1 社しか COVID-19 社説を発表していない日は類似度計算ができないためデータから除かれる。例えば 2020 年 1 月 24 日、2020 年 1 月 29 日は毎日新聞が 1 本ずつ社説を発表しているのみであったため、分析から除外された。このような日は 33 日分、43 社説であった。日付と社説本数が一致していないのは、1 社のみが COVID-19 関連社説を 2 本発表した日があったためである。

続いて、同じ日に同じ新聞で、COVID-19 社説が 2 本発表される場合の対応について説明する。どちらの社説と近いかは事前に不明である以上、すべての組み合わせを分析する必要がある。例として、2021 年 1 月 5 日は朝日新聞・毎日新聞・読売新聞は 1 本ずつの COVID-19 社説を発表していたが、日経新聞は 2 本発表していた。日経新聞の COVID-19 社説をそれぞれ日経 A、日経 B としたとき、以下の組み合わせが生じる。もし全 4 紙が 2 本ずつ発表していた場合、16 の組み合わせが生じることになる。このような処理を行った結果、全 8155 組の組み合わせデータを取得した。

（組み合わせ例）
日経 A-朝日、日経 A-毎日、日経 A-読売
日経 B-朝日、日経 B-毎日、日経 B-読売
朝日-毎日、朝日-読売、毎日-読売

③ 分析方法

コサイン類似度および tf-idf の計算には、Python の機械学習ライブラリとして一般的な scikit-learn[14] を用いた。品詞については、第 3 節の分析とそろえるためにも、名詞のみに限定した。

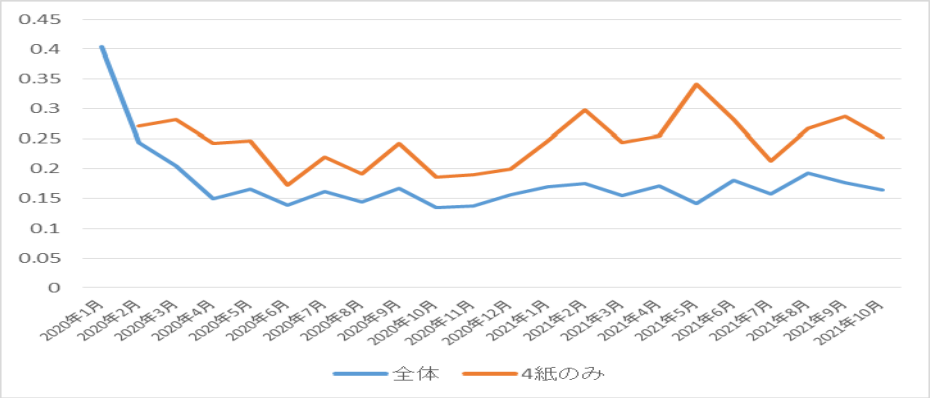
得られたコサイン類似度を日次と新聞間で整理していく。ただし、前者においては注意が必要である。なぜなら、1 日で新聞 A と新聞 B で COVID-19 社説が 2 本ずつあった場合、計 4 つの組み合わせ、すなわち 4 つのコサイン類似度

が取得されているためである。この場合、新聞社間の類似性を測定するためには 4 つ類似度の平均値か最大値のどちらかが望ましいだろう。本論文では、最大値を採用することとした。

④ 結果

図 11 は全組み合わせの月次の類似度の推移を示している。一見して明らかなように、日本における COVID-19 の感染拡大初期である 2020 年 1-3 月は、比較的高いコサイン類似度を示している（青線）。これは、COVID-19 という未知の外生的ショックに対して、比較的同じような反応をしたことを示している。しかしながら、そのような反応は 2020 年 4 月からは落ち着き、おおむねフラットとなっている。また、図 12 には 4 紙すべてが COVID-19 社説を発表した日のみを抽出して時系列で整理したコサイン類似度も示されている（橙線）。4 紙すべてのほうが、全体よりもコサイン類似度が高い。これは 4 紙すべてが注目するほどの重要な何かが起こったことを示している。また、2021 年 6 月、同年 2 月、同年 9 月に類似度がスパイク状になっているのも、同様に、4 紙すべてが特に注目する何かが発生したと考えられる。

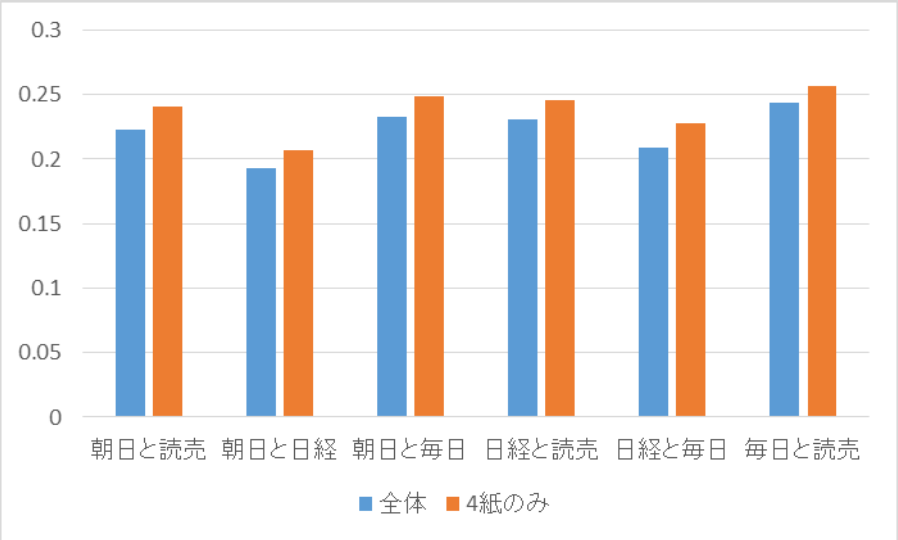
図 11　コサイン類似度の時系列推移



もとのデータを見てみると、2021 年 6 月は党首討論や G7 サミット、2 月は森五輪会長の辞任・後任人事、緊急事態宣言の延長、コロナ関連法といったテーマで、2021 年 9 月は菅総理の退陣と岸田新総裁、デルタ株の感染拡大と緊急事態宣言、パラリンピック閉幕というテーマで 4 紙が社説を発表している。

続いて、新聞社ごとの類似度を整理したものが、図 12 である。朝日新聞は読売新聞よりも毎日新聞のほうがコサイン類似度が高い。これは既存研究とおおむね一致している。ただし、経済専門紙である日経新聞と朝日新聞・毎日新聞とは類似度が相対的に低いというのも、容易に理解可能である。また、全組み合わせと比べ、4 紙すべてが社説を発表した場合のほうが、コサイン類似度が高くなるのは図 11 と共通であった。

図 12　新聞社間のコサイン類似度



4 紙すべてのコサイン類似度平均が最も高かった 3 つの日の社説を整理する。上でも見たように、森五輪会長の辞任（2021 年 2 月 13 日）、菅総理の辞任（2021 年 9 月 4 日）と自民党新総裁に岸田氏が選ばれた（2021 年 9 月 30 日）のタイミングで、高い類似性を示している。次点では、バイデン米大統領就任（2020 年 11 月 20 日）、の 2021 年度予算案（2020 年 12 月 20 日）、菅自民党新総裁就任（2020 年 9 月 15 日）というタイミングが高い類似度を示していた。

2021 年 2 月 13 日　コサイン類似度：0.5797

- 朝日　森会長辞任　目を覆うばかりの混迷
- 日経　森五輪会長の辞任を旧弊改める契機に
- 毎日　森会長辞任と後継人事　旧弊を改めていく契機に
- 読売　五輪会長辞任　森氏の決断遅れが混乱広げた

2021 年 9 月 4 日　　コサイン類似度： 0.56908

- 朝日　菅首相１年で退陣へ　対コロナ国民の信失った末に
- 日経　危機下で指導力発揮できる体制を
- 毎日　菅首相が辞意表明　独善と楽観が招いた末路
- 読売　菅首相退陣へコロナ克服に強力な体制作れ

2021 年 9 月 30 日　コサイン類似度： 0.558667

- 朝日　自民新総裁に岸田氏　国民の信を取り戻せるか
- 日経　国民の声に耳を傾ける岸田政権に
- 毎日　自民新総裁に岸田氏　「安倍・菅」路線から脱却を
- 読売　岸田自民新総裁　政策を肉付けし安定政権築け

COVID-19 の文脈で最も類似度が高かったのは以下のとおりである。順位は全体での順位を示している。COVID-19 の文脈では、緊急事態宣言の解除について、各紙が似た論点を発表していることが分かる。

7 位　　2021 年 3 月 19 日　コサイン類似度： 0.5399

- 朝日　展望みえぬ宣言解除　再拡大阻止に全力をあげよ
- 日経　緊急事態の全面解除後も万全の対策を
- 毎日　４都県の宣言解除へ　懸念を拭う対策が必要だ
- 読売　緊急事態解除　リバウンド回避へ警戒怠るな

10 位　2020 年 5 月 15 日　コサイン類似度： 0.52855

- 朝日　宣言一部解除、再流行への備え怠るな
- 日経　再流行警戒しながら慎重に経済再開を
- 毎日　緊急事態の一部解除　感染拡大引き続き警戒を
- 読売　緊急事態の解除　油断せず段階的に活動再開を

11 位　2021 年 2 月 27 日　コサイン類似度： 0.5261

- 朝日　６府県先行解除　再拡大防止を最優先に
- 日経　宣言解除後の感染再拡大防げ
- 毎日　首都圏以外で宣言解除　緩み招きかねぬ首相対応
- 読売　６府県宣言解除　段階的緩和で感染再拡大防げ

最後に、個別の組み合わせ時に、どのようなテーマの時に高い類似度を示したのか。各組み合わせで最も高い類似度を示した日付・コサイン類似度・社説タイトルを以下に整理した。朝日新聞と毎日新聞では、2021 年 9 月の岸田自民党新総裁について、高い類似度を示している。しかしながら、それ以外の組み合わせでは最低賃金（朝日－日経)、香港（朝日－読売、日経－毎日)、日銀政策（日経－読売)、COVID-19 国内初の死者（毎日－読売）となっており、一致する論点に散らばりがあることが分かった。

【朝日新聞－日経新聞】2020 年 7 月 24 日　コサイン類似度：0.6366

- 朝日新聞：最低賃金　引き上げの歩み継続を
- 日経新聞：最低賃金を無理なく上げる基盤づくりを

【朝日新聞－毎日新聞】2021 年 9 月 30 日　コサイン類似度：0.6843

- 朝日新聞：自民新総裁に岸田氏　国民の信を取り戻せるか
- 毎日新聞：自民新総裁に岸田氏「安倍・菅」路線から脱却を

【朝日新聞－読売新聞】2020 年 8 月 4 日　コサイン類似度：0.6410

- 朝日新聞：香港の選挙、崩れていく自由の基盤
- 読売新聞：香港議会選延期　民主主義の形骸化を懸念する

【日経新聞－毎日新聞】2020 年 5 月 29 日　コサイン類似度：0.6369

- 日経新聞：香港経済支える「一国二制度」の重大危機
- 毎日新聞：香港の国家安全法制　政治の自由奪う禁じ手だ

【日経新聞－読売新聞】2021 年 3 月 20 日　コサイン類似度：0.6369

- 日経新聞：日銀は柔軟でわかりやすい政策運営を
- 読売新聞：日銀政策見直し　副作用抑えて経済の下支えを

【毎日新聞－読売新聞】2020 年 2 月 15 日　コサイン類似度：0.7006

- 毎日新聞：新型肺炎国内初の死者　感染拡大の防止に全力を
- 読売新聞：新型肺炎　国内感染踏まえた医療体制を

⑤ 小括

本節ではコサイン類似度を用いて、COVID-19 に関する新聞社説の類似度を分析してきた。分析の結果、第 1 に 2020 年 1-3 月といった感染序盤において高い類似度を示していた。これは、未知の外生的ショックに対する反応だと考えられる。第 2 に、4 紙すべてが「コロナ」「肺炎」の含まれる社説を発表するときには、相対的に高い類似性を示していた。これは、4 紙が重視するほどの事態が発生したためだと思われる。第 3 に、個別の新聞間では、最も高い類似度を示す社説の内容に、散らばりが見られた。これは細かな論点に対しては、反応の度合いが新聞によって多様である可能性、あるいは、単に分析サンプルが少なかったために発生した偶然という 2 つの解釈が可能である。

５．おわりに

①要約と考察

本論文ではここまで、近年の自然言語処理技術の広まりに対応して、トピック・モデリング、感情分析、文書間類似度という手法を通じて、COVID-19 に関連した新聞社説を分析してきた。

第 2 節において、トピック・モデリングを用いた分析の結果、新聞社は COVID-19 に対する影響を多面的に検討していること、一部のトピックについては、時系列変化の背景も比較的容易に考察できること、そして、新聞社ごとに着目するトピックが異なっていることも明らかとなった。従来、朝日新聞と毎日新聞はリベラルあるいは左派とみなされてきた（小林&竹本 2016）。しかしながら、本結果をみると、同じ政治的立場であっても着目する論点が異なっていることが分かる。

続いて、第 3 節において、2 種類の感情分析ライブラリを用いた分析では、COVID-19 は相対的にネガティブに捉えられ、政権についてはポジティブにとら

えられていることが分かった。また、比較的ネガティブな論調を示すグループ（朝日新聞・毎日新聞）、相対的にポジティブな論調を示すグループ（読売新聞・日経新聞）に分類できた。

最後に、第 4 節では、コサイン類似度を用いて、COVID-19 に関する新聞社説の類似度を分析した結果、第 1 に 2020 年 1-3 月といった感染序盤において高い類似度を示すが、その後低下すること、第 2 に 4 紙すべてが社説を発表するときには、相対的に高い類似性を示すこと、第 3 に個別の新聞間では、最も高い類似度を示す社説の内容に、散らばりが見られたことが明らかとなった。

以上の発見事実は、新聞社間の共通性よりも多様性を示唆ししている。従来は、政治的論点に対する社論の観点から、朝日新聞・毎日新聞をリベラル派、読売新聞・産経新聞を保守派とみなす研究が多かった。既存研究では新聞社の党派性を論じる際には、教科書問題（竹川 2012）や国会議論（畑中他 2009）、歴史観（崔 2021）といった問題から切り込むことが多かったためだと思われる。しかしながら本論文での分析結果は、個別の論点ではむしろ多様性を示し、リベラル・保守といった観点では分析しきれないことが示唆できる。新聞は政治のみならず、経済・生活・教育といった多様な側面に言及している。そのため、それらの多様な側面においては、必ずしも政治的立場とは一貫せず、主張の異なる新聞と論調が一致することもありうるのである。

②今後の展望

本論文には様々な課題が残されている。それを技術的課題、研究方向性の 2 つの観点から整理していこう。

技術的課題

トピック・モデリングついては以下の問題がある。本論文では「mecab-ipadic-NEologd」という最新単語も含まれる辞書を用いている。他方で、日本のテキストマイニングを用いた研究では、KH Coderが代表的なソフトウエアである。このKH Coderはipadicを用いていることから、分析に連続性がないともいえる。次に、50%以上の文書で出現する単語を分析範囲から除外したが、この「50%」という根拠は筆者らの判断による。品詞についても既存研究

を参考に名詞のみに制限していたが、より広い文脈を考えるのであれば、ほかの品詞も加えたほうがいいかもしれない。さらに、そもそもトピック・モデリングという手法は確率論的な手法であることから、同じパラメターで分析したとしても、同じ結果にならないという問題点がある。

感情分析については以下の問題がある。第 1 が、分析精度の問題である。感情分析を行った研究者は実感しているだろうが、2022 年現在、人間の読解力を超えた感情分析を行うプログラムは開発されていないようである。そのため、本論文に限らず感情分析の結果には一定の限界が存在している。第 2 が、そもそもどの感情分析ライブラリを用いるかという問題がある。本論文では oseti と daigo/bert を用いているが、鳥海（2020）が用いた pymlask や、マイクロソフト社が提供する「Microsoft Asure TextAnalytics」[15] といったプログラムだと別の結果が出る可能性が高い。

コサイン類似度分析にも上述の品詞の問題がある。

研究方向性

本論文は、COVID-19 に関する新聞社説を 3 つの自然言語処理技術を用いて分析してきた。分析結果からは一定の示唆を得られたものの、今後は理論的貢献を目指すべきである。今後の研究方向性としては、第 1 に 3 つの結果の関係を分析することであろう。トピック・モデリングで得られたトピック対して新聞社ごとの論調が異なるのか、文書類似度が高い新聞社説で論調が異なるのか、といった研究方向である。第 2 に本論文で分析された数字以外の数字、例えば感染者数などを用いることで、新聞社説の論調は何が影響しているかを明らかにしていく方向性である。第 3 が、ほかの年の社説と比較することである。新聞の論調は 2020 年以降のコロナ禍固有の論調かもしれない。そのため、コロナ禍前・収束後の社説の論調と比較すると、外生的ショックに対して、新聞社がどのように反応したか明らかになる。また、同じ外生的ショックとして 2008 年のリーマンショック、2011 年の東日本大震災時の社説も比較対象となるだろう。

本論文には以上の課題がある。しかしながら、本論文は自然言語処理技術を用いた新聞社説研究の可能性を開くとともに、その限界と今後の展望を明らかにしており、一定の貢献をなしえたと考える。

謝辞

本論文は JSPA 科研費 20H01542、20H01540、21K01663 の助成を受けたものである。

注釈（各種ウェブサイトについては、2022 年 2 月 1 日最終閲覧）

1　Yusuke Hoshino (Associate Professor of Faculty of Business Administration in Musashino University, Japan)

2 NHK 特設サイト（https://www3.nhk.or.jp/news/special/coronavirus/data-widget/）

3 厚生労働省特設サイト（https://www.mhlw.go.jp/stf/covid-19/open-data.html）

4 Slothlibストップワード
（http://svn.sourceforge.jp/svnroot/slothlib/CSharp/Version1/SlothLib/NLP/Filter/StopWord/word/Japanese.txt）

5 Janome（https://github.com/mocobeta/janome）

6 GiNZA（https://github.com/megagonlabs/ginza）

7 Mecab（https://github.com/taku910/mecab）

8 mecab-ipadic-neologd（https://github.com/neologd/mecab-ipadic-neologd）

9 Gensim（https://github.com/RaRe-Technologies/gensim）

10 pymlask（https://github.com/ikegami-yukino/pymlask）

11 daigo/bert-base-japanese-sentiment
（https://huggingface.co/daigo/bert-base-japanese-sentiment）

12 oseti（https://github.com/ikegami-yukino/oseti）

13 bert-japaanese (https://github.com/cl-tohoku/bert-japanese)

14 scikit-learn（https://github.com/scikit-learn/scikit-learn）

15 Microsoft Asure TextAnalytics
（https://azure.microsoft.com/ja-jp/services/cognitive-services/text-analytics/）

参考資料

【英語文献】

Adachi, Y., & Negishi, T. (2020, August). Development and evaluation of a real-time analysis method for free-description questionnaire responses.

In *2020 15th International Conference on Computer Science & Education (ICCSE)* (pp. 78-82). IEEE.

de Oliveira Capela, F., & Ramirez-Marquez, J. E. (2019). Detecting urban identity perception via newspaper topic modeling. *Cities*, *93*, 72-83.

Koyano, Y., Oguchi, T., Akagaki, S., Mitani, H., Yasunaga, T., & Tobita, H. (2021, November). Development and Evaluation of Online Meeting System to Promote Effective Communication. In *Proceedings of the Future Technologies Conference* (pp. 712-725). Springer, Cham.

Parvin, G., Rahman, M., Ahsan, R., & Abedin, M. Media Discourse About the Pandemic Novel Coronavirus (COVID-19) in East Asia: The Case of China and Japan, Available at SSRN: https://ssrn.com/abstract=3603875.

Scheufele, D. A., & Tewksbury, D. (2007). Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of communication*, *57*(1), 9-20.

【邦文文献】

伊藤宏. (2012). 福島第一原発事故以降の原子力報道:事故後３ヶ月間の新聞社説の論調から見えてくること. *プール学院大学研究紀要, 52*, 199-212.

小木しのぶ. (2015). テキストマイニングの技術と動向. *計算機統計学*, *28*(1), 31-40.

北中英明. (2020). テキストマイニングによる文献研究―営業研究分野への適用事例―. *拓殖大学経営経理研究*, *117*, 27-44.

國府久嗣, 山崎治子, & 野坂政司. (2013). 内容推測に適したキーワード抽出のための日本語ストップワード. *日本感性工学会論文誌*, *12*(4), 511-518.

小林哲郎, & 竹本圭祐. (2016). 新聞読者は極性化しているか. *日本世論調査協会報「よろん」*, *117*, 22-26.

小林のぞみ, 乾健太郎, 松本裕治, 立石健二, & 福島俊一. (2005). 意見抽出のための評価表現の収集. *自然言語処理*, *12*(3), 203-222.

高木佐知子. (2004). メディア・ディスコースのイデオロギー表出ストラテジー--イラク戦争関連の社説における" We-group" の考察. *大阪府立大学言語*

文化研究, *3*, 9-19.

竹川俊一. (2012). 社説と報道によるフレーミング分析: 2001 年歴史教科書問題に関する朝日と読売を事例に. *マス・コミュニケーション研究, 80*, 211-229.

崔昌幸. (2021). [論説] 批判的言説分析による 「自虐史観」 の研究--『朝日新聞』 と 『読売新聞』 における社説を題材として--. *社会システム研究*, *24*, 183-202.

土村成美. (2017). トピックモデルによるアガサ・クリスティ作品の計量的分析. *じんもんこん 2017 論文集*, 177-182.

伝康晴. (2009). 多様な目的に適した形態素解析システム用電子化辞書（< 特集 > 日本語コーパス）. *人工知能*, *24*(5), 640-646.

鳥海不二夫. (2020). COVID-19 下の情報拡散. *第 11 回横幹連合コンファレンス予稿集*, A-4-6.

難波英嗣. (2020). テキスト間の類似度の測定. *情報の科学と技術*, *70*(7), 373-375.

難波英嗣, & 福田悟志. (2021). ネットからの不安感の情報抽出. *感性工学*, *19*(4), 163-170.

畑中允宏, 村田真樹, & 掛谷英紀. (2009). 新聞社説・国会議事録に基づく言論のイデオロギー別分類. *言語処理学会第 15 回年次大会発表論文集*, 408-411.

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*, 584-587.

樋口耕一. (2006). 内容分析から計量テキスト分析へ-継承と発展をめざして. *大阪大学大学院人間科学研究科紀要*, *32*, 1-27.

星野雄介, & 平尾毅(2020). バブル経済崩壊後のイノベーションに関する新聞報道: 新聞社説のテキストマイニングを通じて. *武蔵野大学経営研究所紀要, 2*, 71-94.

星野雄介, &平尾毅. (2021). 新型コロナウイルス（COVID-19）に関する新聞社説の論調: 時系列の変化と新聞社ごとの特徴. *武蔵野大学経営研究所紀要*, *3*, 72-92.

細貝亮. (2010). メディアが内閣支持に与える影響力とその時間的変化: 新聞社説の内容分析を媒介にして. *マス・コミュニケーション研究*, *77*, 225-242.

村瀬俊朗, 王ヘキサン, & 鈴木宏治. (2021). アンケート調査を越えて―自然言語処理や機械学習を用いたログデータの活用を模索する―. *組織科学*, *55*(1), 16-30.

山口真一. (2021). 組織・消費者間コミュニケーション研究の現在: テキスト分析, 行動データ分析, アンケート調査分析. *組織科学*, *55*(1), 4-15.

吉田政之. (2020). リスク情報開示におけるリスクの種類とその変遷―トピックモデルを用いて―. *原価計算研究*, *44*(1), 116-128.