Title	テキストマイニングによる「新型コロナ不満情報」の特徴抽出と特徴分析
Sub Title	
Author	成, 誠(Cheng, Cheng) 林, 高樹(Hayashi, Takaki)
Publisher	慶應義塾大学大学院経営管理研究科
Publication year	2021
Jtitle	
JaLC DOI	
Abstract	
Notes	修士学位論文. 2021年度経営学 第3849号
Genre	Thesis or Dissertation
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO40003001-00002021-3849

慶應義塾大学大学院経営管理研究科修士課程

学位論文（　　2021　年度）

論文題名

テキストマイニングによる「新型コロナ不満情報」の特徴抽出と特徴分析

主　査	林　高樹
副　査	中村　洋
副　査	大林　厚臣
副　査	

氏　名	成　誠

論 文 要 旨

所属ゼミ	林高樹研究会	氏名	成　誠

（論文題名）

テキストマイニングによる「新型コロナ不満情報」の特徴抽出と特徴分析

（内容の要旨）

【背景】

コロナ禍よって、私たちの生活様式や、社会の各システムは急激な変化を強いられている。その中で、私たちの間には新型コロナに関わる新たな不満や不安が生じている。そのような不満や不安を理解することで、行政の的確な対応や新ビジネスの創出へとつながることが期待される。

【目的】

本研究の目的は、テキストマイニングの手法により「新型コロナ不満情報」という文書データの特徴を抽出し、分類することで、「コロナ禍」に伴う新たな不満を把握し、企業に対してはビジネスチャンスの創出、または政府・自治体に対しては今後の対策改善に何らかのヒントを見つけ出す。

【研究方法】

2018 年に Google の Jacob Devlin らによって発表され、学術研究や実用化が進んでいる自然言語処理アプローチである BERT とその改良版の SentenceBERT を用いて、新型コロナ不満の口コミデータを分析しその特徴理解を試みる。BERTopic によるトピック分類と BERT による感情分析(極性分析)はほん研究の主な方向性である。

【結論と示唆】

「コロナ不満」のテキストデータに対するトピック分類を行うことで、「コロナ禍」において、人々の不満トピックが常に変化していることが明らかになった。一方、その中、「学力低下」、「子供の教育」などずっと存在している不満や不安の話題もある。感情分析（ポジネガ判定）では、人々の不満程度の数値化を実現したため、政府・自治体にとって、より早く「ターゲット」を見つけ出せる。その中、特に「女性」、「専業主婦」、「アルバイト・パート」という特徴を持っている人々の不満程度が高いことが明らかになった。

本研究のように、「コロナ不満」のテキストデータを取り上げ、「BERT」という最新な自然言語処理アプローチで不満特徴を抽出する方法自体は、企業または自治体にとって、参考になれる部分もあると考えられる。例えば、「コロナ不満」を収集する公式プラットフォームを構築することで、自治体が即時に住民の不満トピックと不満程度を把握することができ、早めに対応策を考えて、「With コロナ」住民の暮らしの改善に繋がると思う。

【新規性】

筆者が調べた限り、「コロナ不満」のテキストデータを取り上げ、「コロナ禍」において人々の不満意識や不安意識を分析する先行研究はまだ存在していない。本研究では、最新な自然言語処理アプローチ「BERT」を活用して、トピック分類と感情分析により、最新なデータである「コロナ不満」を分析する。従来の自然言語処理アプローチに比べて、精度の高い且つ意味のある結果を目指す。

目次

1. 序論 5
  1.1 研究背景 5
    1.1.1 「コロナ禍」の振り返り 5
    1.1.2 「コロナ禍」による人々の不満や不安意識 5
  1.2 自然言語処理(NLP) 7
    1.2.1 従来の自然言語処理(NLP)アプローチ 7
    1.2.2 BERT 7
    1.2.3 Sentence BERT と BERTopic 8
2. 研究目的 9
3. 先行研究 9
  3.1 日本国内の先行研究 9
    3.1.1 テキストデータによる感情分析に関する研究 9
    3.1.2 「BERT」に関する研究 10
  3.2 海外の先行研究 11
    3.2.1 テキストデータによる感情分析に関する研究 11
    3.2.2 「BERT」に関する応用研究 11
  3.3 先行研究の課題 12
4. 研究方法とデータ概要 12
  4.1 データ概要 12
  4.2 データ特徴 13
  4.3 データ前処理 14
  4.4 研究方法 15
    4.4.1 BERTopic によるトピック分類 15
    4.4.2 BERT による感情分析(ポジネガ分析) 16
    4.4.3 日本語 BERT 訓練済みモデル 17
5. 分析結果 17
  5.1 コロナ不満のトピック分類 17
  5.2 コロナ不満の不満程度数値化（ポジネガ判定） 22
6. 結論と考察 27
  6.1 トピック分類の結論 27
  6.2 感情分析の結論 27
  6.3 考察 28
7. 限界・課題 28
8. 謝辞 29
9. 参考文献 30
付録 32

1. BERT によるトピック分類のプログラム(SentenceBERT のみ) 32
2. BERT によるトピック分類のプログラム(BERTopic) 36
3. BERT による感情分析（ポジネガ判定）のプログラム 37

1. 序論

1.1 研究背景

1.1.1 「コロナ禍」の振り返り

2019 年 12 月末、中国・武漢で発生した新型コロナウイルス感染症（COVID-19）は、瞬く間に全世界に広がって、2020 年 3 月に WHO は世界的流行（パンデミック）を表明し、諸外国ではこの事態を対応するため、外出禁止令やロックダウンなど強制力のある措置がとられ、人々の日常生活や経済活動は大きな打撃を受けた。

日本でも、2020 年 1 月中旬、クルーズ船ダイヤモンド・プリンセス号での旅客が新型コロナウイルスの感染を確認した以来、国内の感染状況は急激に拡大した。2020 年 4 月 7 日、ついに政府は東京など 7 都府県に 1 回目の緊急事態宣言を発令し、16 日に全国に対象を拡大した。5 月 25 日まで続けた緊急事態宣言期間の中、事実上の外出自粛、在宅勤務、休校またはオンライン授業や飲食店の時短営業など感染防止策の実施は余儀なくなれた。これを受けて、日本国内でも、国民の生活様式、各業界の実態と経済活動には大きいな変化が起こった。その後、感染状況の緩和とリバウンドが続けて、2021 年 12 月までには、合計 4 回の緊急事態宣言が発令された。この原稿を書いている今でも、オミクロン株の感染拡大が始まって、日本国内の新型コロナの完全収束がまだ不透明な状態になるため、「コロナ禍」が予想以上に長引いた。

人々の生活様式、各業界の実態、国家の経済状況が急激に変わっている「コロナ禍」の時代において、政府にとって、次回の緊急事態を備えるため、如何に対応策を万全にする能力と如何に住民の日常生活と暮らしを改善する能力が問われる。一方で、企業にとっては、「コロナ禍」からチャンスを掴み、新たなビジネスの種を見つける能力も問われる。

1.1.2 「コロナ禍」による人々の不満や不安意識

「コロナ禍」による生活様式の変化や未来の景気状況が不透明になっている中、人々の生活は急激な変化を強いられ、それに伴う不満意識や不安意識が新たに生まれる。特に、2020 年 3 月からの「With コロナ」の暮らしにおいて、人々の不満が多いと言われる。2020 年 5 月から、電通デジタルはの専門チームから 2 回実施した「With コロナ社会における、不満意識調査」の結果によると、不満レベル高と不満レベル中の回答割合が約 85%に達して、コロナ禍で変化した暮らしに不満を抱いており、かなり人が潜在的な不満を抱いていると思われる。また、具体的な不満を見ると、2 回の調査ともに「人・友人と会えない」や「外食ができない、飲食店の営業時間が短い」や「外出できない」など様々な不満が集計されたが、1 回目は特に日常生活への変化に対する不満が多く、2 回目では人生や思い出作りに影響を及ぼす新不満が多く出てきた。

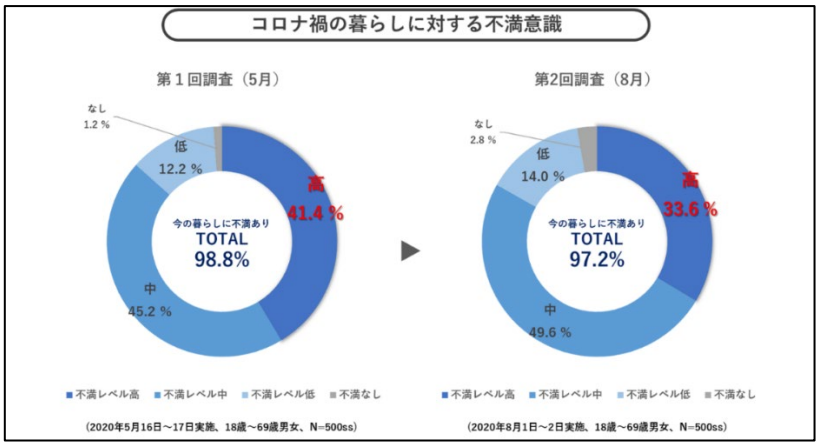


図１－１　「With コロナ社会における、不満意識調査」結果

（出所：「Fu-man insight lab」の調査より、https:// www.dentsudigital.co.jp/release/2020/0908-000610）

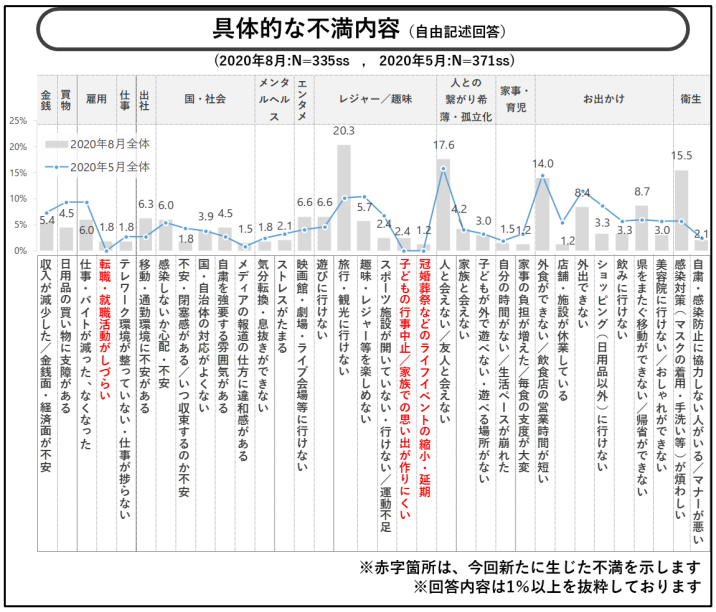


図１－２　「With コロナ社会における、不満意識調査」結果（具体的な不満）

（出所：「Fu-man insight lab」の調査より、https:// www.dentsudigital.co.jp/release/2020/0908-000610）

実際、政府・自治体では積極的に住民の「コロナ不満」を収集している。例えば、東京都の都政情報サイトから見ると、2020 年 4 月から毎月に新型コロナウイルス感染症対策に対する不満や懸念していることをアンケート設問の形式で集まっている。また、神奈川県もコロナ禍において県民の生活クオリティを向上するため、「新型コロナウイルス感染拡大による県民の不満や不安調査」をテーマとして、2 年連続に県民アンケートを実施した。

一方、With コロナの生活が続く中、人々の「不満意識」への分析は新しい価値の種の発見につながると考え、多くの企業でも「コロナ不満」に対する調査に乗り出した。上述の電通デジタルによる調査以外、調査会社クロス・マーケティングが新型コロナウイルスが消費者の行動や意識に与える影響の把握を目的として、全国 47 都道府県に在住する 20～69 歳の男女 2, 500 人を対象に「新型コロナウイルス生活影響度調査」を計 3 回実施した。また、株式会社 Insight Tech はと情

報・システム研究機構・国立情報学研究所（以下 NII）は 2021 年 3 月 23 日より新たに、Insight Tech が運営するサービス「不満買取センター」において収集した「新型コロナ不満アンケート」データも提供し始めた。

今後、これらの「不満」データをどのような手法で分析して、そして、得られた結果を如何に活用できるかはますます重要であろう。

上述の「コロナ不満」に関する分析調査のほとんどは、アンケート調査回答者の自由回答に基づいた集計結果である。しかし、回答者が書いた具体的な不満に対して、テキストマイニングで何らかの特徴を抽出する研究は、筆者が調べた限り、まだ存在してない。したがって、本研究では、上述の「不満買取センター」から収集された「コロナ不満」のデータを取り上げ、特に「具体的な不満」というテキストデータの項目を取り上げて、最新の自然言語処理アプローチで分析を行う。新生活様式に伴う不満と新たに生じた不満を把握することで、ニューノーマルに定着していくと考えられる価値意識について考察する。

1.2　自然言語処理(NLP)

1.2.1　従来の自然言語処理(NLP)アプローチ

自然言語処理は人間だけが処理できる記号の集合体を、何らかの手法でコンピューターに処理させ、数値など意図するものを出力させる手法である。

従来の自然言語処理アプローチでは、いくつかの限界とデメリットがある。単語の出現位置、出現頻度や品詞だけで考えられているため、結構語句の意味の理解につらがらないという限界がある。また、ルールベースの処理しかできないというデメリットもある。さらに、BERT が発表される前に存在していた一般的な自然言語処理アプローチの場合、文章の前後のコンテンツではなく、単一方向からしか学習することはできない。例えば、ELMo[Peters,　(2018)]方法または OpenAI GPT[Radford,　(2018)]方法のいずれも左から右の方向にしか学習せず、つまり、文章タスクや Q&A などの前後の文脈が大事なものでは有効ではない。

1.2.2　BERT

BERT とは「Bidirectional Encoder Representations from Transformers」を指し、2018 年 10 月 11 日に Google が発表された最新な自然言語処理モデルである。

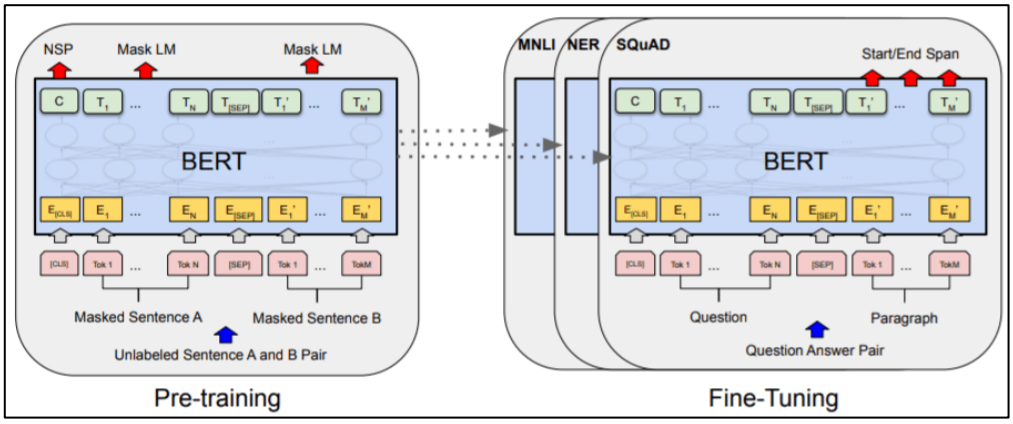


図 1－3　BERT 事前学習とファインチューニングの仕組み

（出所：Jacob Devlin. BERT: Pre-training of deep bidirectional transformers for language understanding(2019)）

BERT の仕組みは、既に入力されたシーケンスを元に別のシーケンスを予測する事前学習モデルであり、入力されたラベル（名前）が付与されていない分散表現を Transformer が処理を行い、学習することである。実際には、Transformer が Masked Language Model と Next Sentence Prediction という 2 つの手法を同時進行で進めていき、学習する。

BERT は双方向のテキスト処理を遂行できるのため、文章の文脈をよく理解できる特徴がある。また、BERT には、汎用性が高いという特徴もある。これまでのタスク処理モデルの場合、特定のタスクだけに対応していたが、BERT であれば、モデルの構造を修正することなく、さまざまなタスクに応用させることが可能になる。最後に、BERT は「ラベルが付与されていないデータセット」でも処理することができるメリットがあるため、少ないデータでも始められる手法として応用に際しては大量の計算用テキストデータを用意することが不要になる。

1.2.3 Sentence BERT と BERTopic

Sentence BERT は BERT をファインチューニングして良質な文章ベクトルを生成する手法である。その仕組みは、文字通り BERT をベースに構築されており、そして、事前学習された BERT モデルにプーリング層を追加することで高精度の文埋め込みの精度を獲得する。

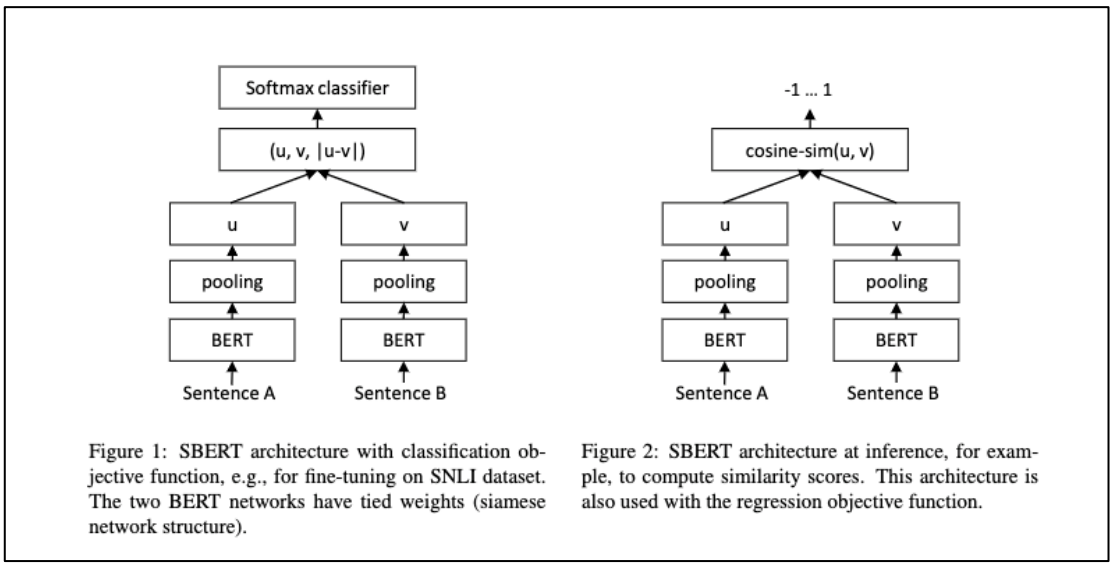


Figure 1: SBERT architecture with classification objective function, e.g., for fine-tuning on SNLI dataset. The two BERT networks have tied weights (siamese network structure).

Figure 2: SBERT architecture at inference, for example, to compute similarity scores. This architecture is also used with the regression objective function.

図 1－4　Sentence-BERT によるクラス分類と類似度の計算

（出所: https://www.vareal.co.jp/column/sentence-bert%E8%AB%96%E6%96%87-%E5%92%8C%E8%A8%B3/）

Sentence BERT による高精度の文章ベクトル化によって、チャットボット、問い合わせサイトへの質疑応答システム、類似文章検索と分類などのことが可能になる。本研究では、「コロナ不満」のテキストデータに対して、何らかの特徴を見つけ出すことが目的の一つのため、Sentence BERT という技術が適切だと思われる。

2. 研究目的

本研究の目的は、テキストマイニングの手法により「新型コロナ不満アンケートデータ」という文書データの特徴を抽出し、分類することで、「コロナ禍」に伴う新たな不満を把握し、企業に対してはビジネスチャンスの創出、または政府・自治体に対しては今後の対策改善に参考価値を見つけ出す。具体的に、以下の 3 つの目標達成を目指す。

① BERT モデルを用いて、テキストデータから「コロナ不満」のトピックを抽出・分類し、人々がコロナ禍において、どんな不満を持っているのかを確認することである。

② クロス集計と BERT モデルによるポジネガ判定を用いて、それぞれのトピックの不満において、不満の特徴を見つけ出し、どんな人（性別・職種・年収など）、どこで、どんな不満程度になるのかを確認すること。

③ 「コロナ不満」の特徴に合わせることで、With コロナ・After コロナのの生活をサポートするため、ビジネスに対する示唆だけではなく、今後、今回のような緊急事態が発生する場合、どう対応するのかを自治体に向けて存在している課題と改善案を提示すること。特に、「コロナ不満」トピック分類の結果とポジネガ判定の結果に合わせて、不満の出現頻度と不満程度という 2 つの軸を設定することで、両者がともに高い不満が最初に対応すべき不満と特定する。

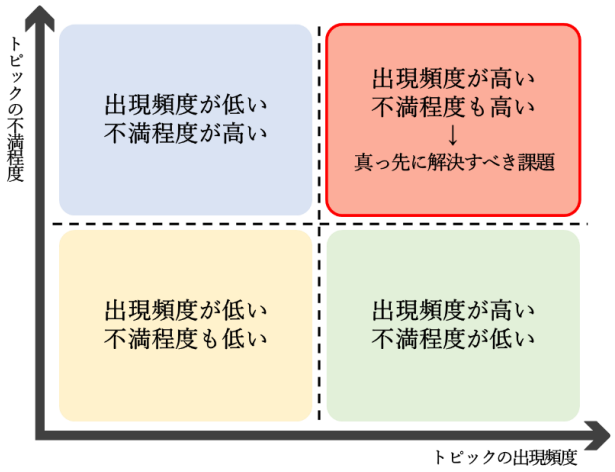


図 2－1　トピック不満程度とトピック出現頻度の 2 軸

3. 先行研究

3.1　日本国内の先行研究

3.1.1　テキストデータによる感情分析に関する研究

本研究を遂行するにあたり、日本国内における「感情分析」をテーマとする先行研究について調査を行った。特に、テキストデータから人々の感情を分析する先行研究はいくつか存在する。（四方田, 2021）はコロナ禍において、SNS で投稿された「休校に関する不満」を分析対象とし、KH CODER による計量テキスト分析と感情極性辞書に通じて、投稿者の社会的な関心の内容を分析した。感情分析の結果、休校に対して、投稿者の社会的な関心は、「子供への影響」・「保護者への影響」・「社会への影響」と「防疫政策」に分けられ、その中特に、学業と運動不足への不安と家庭での育児負担増大などのことに対する関心が強いという示唆が得られた。（渡邊, 2017）は「マイナンバー」をキーワードとした SNS のつぶやきデータを分析対象とし、Twitter を用いた政府広

報に関する感情分析を行った。感情分析の手法として、KH CODER による共起ネットワークの構築とポジネガ判定を行った。感情分析の結果、マイナンバーカードに関する認知度がまだまだ発展途上と、政府のマイナンバーに関する宣伝と広報がまだまだ足りなく、効果もイマイチという結論も明らかになった。また、感情に対するクロス集計分析から、利便性を高めるサービスの拡大とネガティブ感情を正の関係性を示す場合もあることを示唆した。（花井, 小口, 2008）は CMC 利用者の感情表現を読み取るために、E メール文章のデータを用いて、単語を抽出して、「分類語彙表」に基づいて、「安心・焦燥・満足」や「苦悩・悲哀」などの感情表現を分類した。具体的な分析手法はコレスポンデンス分析が挙げられる。

一方、本研究と同じく、「不満調査データセット」を利用し、感情分析を行う先行研究もいくつか存在する。（松本, 三道, 2019）は「不満データ」の中に女性の投稿に目をむき、職種別・年代別に分けて比較分析することで、ビジネス特にマーケティングに役立つ情報を抽出した。研究手法は KH CODER によるコレスポンデンス分析を採用した。その結果、30・40 代女性と専業主婦が子供に関する不満の声が多く、商品数量よりも価格に対する不満が多いという示唆が得られた。（王, 参道, 2019）は同じく「不満データ」を用いて、投稿者の「鉄道」に対する不満を収集し、コーディングルールを作成しながら、ベイズ学習に基づいて、感情抽出とともに投稿者の不明な属性を推定した。その結果、青少年の女性がダイヤに対する不満が圧倒的に多く、元々の欠損データを推定することで、テキストマイニング全体の結果がより簡潔、明確になったという示唆が得られた。

3.1.2 「BERT」に関する研究

2018 年、最新な自然言語処理モデルとなる「BERT」が誕生して以来、日本国内でも「BERT」を用いて分析を行う研究が増えてきている。その技術が特に、テキスト分類や文章の文脈解析などの領域で活用されている。本研究ではトピック分類とレビュー分析を行うにあたり、「BERT」を採用する。ここでは、参考となる日本国内の「BERT」による文書分類または「BERT」による口コミ分析の先行研究事例をいくつか紹介する。

（鈴木, 2021）らは「BERT」を用いてステークホルダーの活動報告を SDGs に写像する文書分類器を構築し、SDGs の各項目の連環関係を可視化した。日本語の語彙を理解するため、論文では東北大学乾・鈴木研究室が公開した訓練済み日本語 BERT モデルを使用し、分類制度の評価指標については、Precision, recall, fl-score を設定した。日本政府、自治体などが公開した SDGs 関連の文書を分析することで、BERT ベースの文書分類器の制度が 0.91－0.96 になり、高い分類機能が示され、将来、BERT ベースの文書分類器が世界規模の SDGs 目標達成のためのマッチング支援に十分期待できると示唆した。（泉, 2020）らは災害対策構築の効率化に向けて、SNS 投稿の中の災害情報を収集し、「BERT」を用いて文書分類をすることで、災害発生の位置情報を抽出して、災害時の投稿分類における有効性を示唆した。「BERT」モデルとして京都大学の黒橋・村脇研究室から公開された日本語訓練済みモデルを使用した。その結果、BERT による大雨・台風時の投稿の文書分類には再現率が 0.760、真陰性率が 0.858 に達するため、大雨・台風時に投稿された情報が正確に分類され、かつ必要な情報と不必要な情報も分離することができると示唆した。今後、災害文書分類モデルに合わせることで、投稿者の位置情報を抽出できるアルゴリズムを構築すれば、災害発生後の対応がより早めに行うことができるとした。

3.2 海外の先行研究

3.2.1 テキストデータによる感情分析に関する研究

同じく、海外での「感情分析」をテーマとする先行研究について調査を行った。(Maria, 2017)らはさまざまな言語で人々の意見を表現するテキストスニペットから感情を検出するための高速で柔軟な一般的な方法論を提案した。ギリシャ語と英語の両方の言語でのオンラインユーザーレビューを含む 4 つのデータセットでの実験を通じて、テキスト文書がベクトルで表され、極性分類モデルのトレーニングに使用された機械学習アプローチの分析精度が一番高いと報告した。(Mohamed M.Mostafa, 2013）は Twitter から 3516 件のランダムなテキストサンプルを抽出して、テキストマイニングの手法に通じて、Nokia、T-Mobile、IBM、KLM、DHL などの有名ブランドに対する消費者の感情を評価した。具体的に、既知の方向性を持つ約 6800 のシード形容詞を含む専門家が事前定義した辞書を使用して、テキストのポジティブ性とネガティブ性を評価した。その結果、いくつかの有名なブランドに対する一般的にポジティブな消費者感情が示されること以外、テキストデータに対する定性的および定量的方法論の両方を使用することにより、この研究は国際的なブランドに対する態度に関する議論に幅と深さが追加されたことも示唆した。

一方、(Yabing Zhao, 2018）らは、全体的な感情分析ではなく、「顧客満足度」の評価と予測に向けて、旅行会社のサイトからの 127, 629 件のレビューのサンプルを使用して、オンラインテキストレビューの技術的属性とレビューコミュニティへの顧客の関与性を測った。主観性、多様性、長さなど 7 つの変量を設置し、テキストデータから顧客満足度の回帰モデルを構築する結果、オンラインレビューの読みやすさと長さは、顧客満足度と負の関係性が示され、一方で、オンラインレビューの感情の極性は、顧客満足度と正の関係性があるという仮説の妥当性を明らかにされた。

3.2.2 「BERT」に関する応用研究

海外では、国内より早く「BERT」モデルによる分析が始まっているため、「BERT」関連の先行研究が多く存在している。筆者はテキスト分類とレビュー分析において、BERT モデルを用いた先行研究を調査した。

(Hu, 2019）らは潜在的な顧客の購買意欲に関する重要な情報を見つけ出すため、従来の単語・トークンベースでの自然言語処理(NLP)と異なり、文脈重視の「BERT」を用いた顧客レビュー読解（RRC）を行った。その結果、「BERT」モデルを導入することで、RRC 感情分析の精度とパフォーマンスが向上した。さらに、この研究で提出されたポストトレーニングは他のレビューベースのタスクに適用し、マーケティングの意思決定をサポートするのに非常に効果的であることを示唆した。(Jieh-Sheng Lee, 2019）らは事前にトレーニングされた「BERT」モデルを微調整し、それを特許分類に適用することに焦点を当て、200 万件を超える特許の大規模データセットを効率的に分類をした。この BERT ベースのアプローチは、単語の埋め込みを使用した CNN 方式という従来の自然言語処理のアプローチを上回ることを示唆した。(Santiago González, 2020）はさまざまの機械学習シナリオにおける「BERT」モデルと古典的な NLP アプローチのパフォーマンスを比較することで、「BERT」の優れた性能とその柔軟性を明らかにした。また、機械学習アルゴリズムに供給される従来の TF-IDF ボキャブラリーに対する BERT 動作の実験を行うことで、「BERT」は従来の NLP 問題をスムーズに解決できることを示した。

3.3 先行研究の課題

上述の先行研究を参考にし、筆者はその課題を以下のように整理した。

まず、「感情分析」を中心とする先行研究から見ると、「不満」をテーマとして取り上げ分析する先行研究は多いが、その「不満」は概ねマーケティング分野につながるケースが多い。つまり、顧客が商品またはサービスに対する不満をテキストマイニングのアプローチによる分析が主流となる一方、「コロナ不満」のような人々の生活に関する不満を解析する先行研究が少ない。「コロナ不満」のデータセットは公開されて日が浅いため、日本国内においても、海外においても、「コロナ不満」をテーマとする先行研究は、筆者の調べた限り存在しなかった。また、感情分析の研究手法については、テキストからキーワード（トークン）を抽出し、単語ベースでコレスポンデンス分析やベイズ学習を行うケースが多いのが現状である。しかし、この方法ではテキストのニュアンスや文脈に対する解析の精度がある程度下がることから、長い文書データをより正確に分析するため、もっと新しい、最先端のテキストマイニング技術が必要となる。2018 年の「BERT」モデルの誕生はこの課題解決に一つの可能性を示したが、筆者の調べた限り、今のところ、「BERT」技術を応用し、感情分析を行う先行研究は存在していない。

一方、「BERT」を中心とする先行研究は、研究テーマは主に文章の分類やテキストデータの予測などの領域に集中している。しかし、「BERT」モデルには最初に Google が開発した技術であるため、英語に対する適応性が高く、日本語を「BERT」に適用するためには日本語訓練済みモデルを BERT に埋め込み必要性がある。そのため、現時点において、「BERT」の研究対象は大多数が英語で書かれたテキストであり、日本語のテキストデータに対する応用まだまだ少ないのが現状である。また、文章の文脈をよく理解できる機能を持っているため、単語ベースの言語極性辞書に頼らず、「BERT」だけを利用して、テキストデータのポジネガ性を判定することができる。筆者の調べた限り、「BERT」モデルをテキストマイニング手法とし、日本語文書データの「ポジティブ感情・ネガティブ感情」を判定する先行研究または感情数値化に関する先行研究は稀少である。

本論文は上述の先行研究の課題を捉えたうえ、テキストマイニングにおける最新な文書分類技術「BERT」モデルを用いて、社会的に重要な日本語テキストである「新型コロナ不満アンケートデータ」を取り上げ、トピック分類とポジネガ判定を試み、両者の結果に踏まえて、ビジネスへの考察だけではなく、不満解消に向けて、自治体に向け何らかの提言をすることになった。

4. 研究方法とデータ概要

4.1 データ概要

本研究で使うデータセットは株式会社 Insight Tech が運営する「不満買取センター」上で実施した、新型コロナウイルスについてのアンケートに関するデータである。このデータは、国立情報学研究所(NII)の運営するデータリポジトリ(IDR)から入手した。

(https://www.nii.ac.jp/dsc/idr/fuman/fuman_covid19.html)

アンケートは時期を変えて 4 回実施されており、今回の分析に使用したデータは 4 回すべての結果を含んでいる。4 回のアンケートは、新型コロナウイルスの感染拡大～緊急事態宣言解除後の 2020 年 3 月～6 月に実施されたものである。その中に、第 2 回の有効回答件数が 4060 件に達していることを除いて、他の 3 回の有効回答件数は 3000 件ぐらいの水準である。

アンケートの中身は、回答者の性別、年齢や職業などの基本情報以外に、アンケート設問に対する回答が記録されている。毎回アンケート調査で出された設問が少し変化があるが、「コロナに対する懸念していることや不満はありますか」という問題が必ず含まれ、必須項目として、全ての回答者がこの問題について記入している。本研究では、主に回答者の基本情報と「コロナ不満」のアンケート回答に注目する。

表 4.1　新型コロナ不満アンケートデータの共通情報

カラム名	説明	備考
id	回答 ID	同一アンケート内でのみ有効な ID
user_id	ユーザ ID	アンケートデータ内で共通となる ID で、同じユーザが複数のアンケートで回答している場合は同じユーザ ID が付与
created_at	回答日時	
gender	性別	以下のいずれかより選択：男、女
prefecture	居住都道府県	47 都道府県より選択
age	年齢	
job	職種	以下のいずれかより選択：専業主婦、アルバイト、会社員（事務系）、会社員（技術系）、経営者・役員、学生、無職など

（出所：株式会社 Insight Tech が公表した不満データセットをもとに筆者作成）

表 4.2　新型コロナ不満アンケートデータの一例

ID	性別	年齢	職業	子供	コロナに対する懸念していることや不満
2	男	36	会社員（技術系）	1 人	致死率もあまり高くないのに、世界が過敏になりすぎてる気がする。
14	女	40	アルバイト	1 人	報道が過剰で不安を煽ってると思う。
88	女	36	自由業	1 人	妊娠中&花粉症なので、とにかくマスク不足を何とかしてほしい。

（出所：株式会社 Insight Tech が公表した不満データセットをもとに筆者作成）

4.2　データ特徴

TF-IDF 法でサンプルデータの中の頻出単語を抽出することが可能であり、その集計を見ると、大まかに不満回答の傾向が分かる。

表 4.3　TF-IDF による頻出単語

	複合語	スコア		複合語	スコア
1	感染者	21170.25	7	感染者数	1216.99
2	コロナウイルス	10006.74	8	感染拡大	1156.04
3	高齢者	5918.51	9	新型コロナ	1123.13
4	新型コロナウイルス	3132.28	10	マスク不足	1059.09
5	コロナウィルス	1500.93	11	花粉症	963.76
6	緊急事態	1472.50	12	感染症	927.97

合計 13049 件のサンプルデータに対して、不満投稿者の基本情報を集計した結果、女性回答者が全体の約 71%に占めており、「専業主婦」が職業となる回答者の割合も約 24%に達して、次に多いのは約 23%の「会社員（事務系と技術系）」である。そのため、本研究で得られた結果は、女性且つ専業主婦の不満の声をより反映されるとみられる。一方、年齢層から見ると、ほとんどの回答者が 30 歳から 50 歳までの中年層であり、子供・学生・高齢者からの回答が少ない。そのため、20 代以下の若年層と高齢者の不満の特徴を反映されにくい可能性もある。また、回答者の年収から見ると、年収が 800 万円以下の回答者が約 88%に占めており、年収 800 万円以上の富裕層からの不満の回答が比較的に少ない。したがって、結果を解読する際に、その 3 点に注意しなければならない。

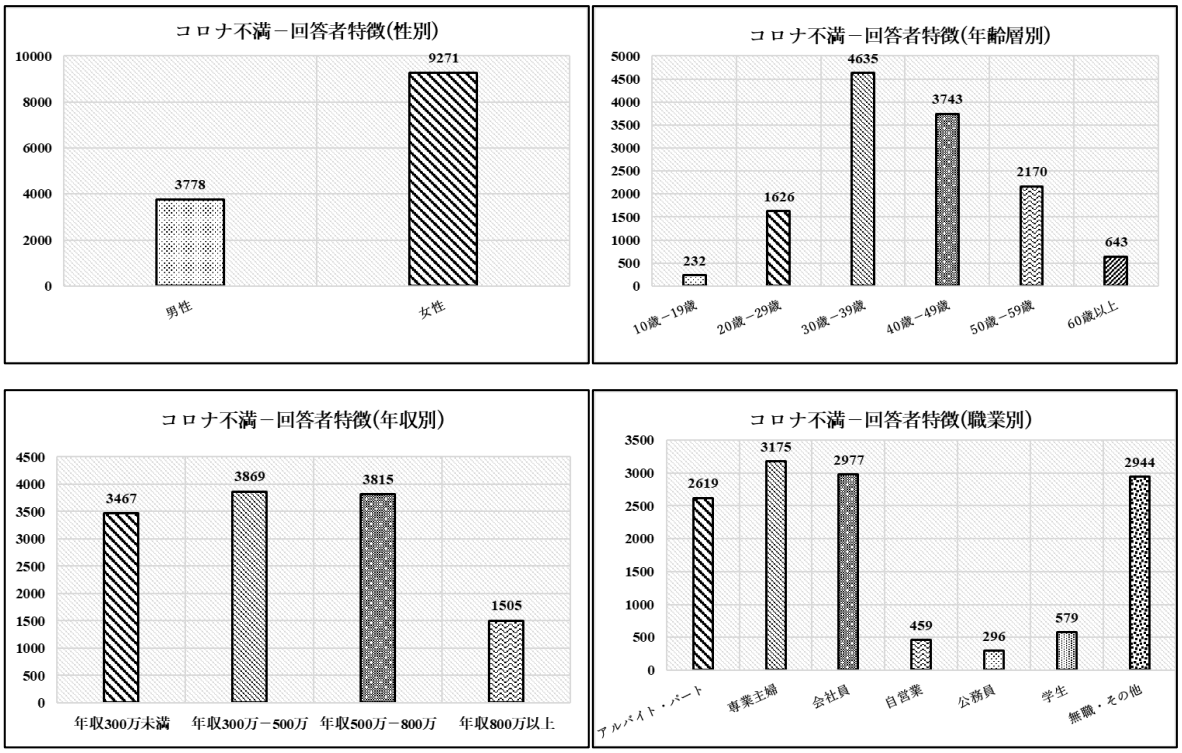


図 4－1　コロナ不満－回答者の基本情報集計(性別、年齢、年収、職業)

4.3　データ前処理

テキストデータに対して、下記のように前処理を行い、有効回答件数は合計 13049 件である。

① サンプルデータである「コロナ不満調査アンケート」には、「コロナに対する何か不満や不安はありますか」という設問だけではなく、「コロナによる困ったこと」や「コロナはいつ終息する」などの合計 14 個の設問と回答が入っている。本研究の研究対象はただ「コロナ不満」であるため、回答者の基本情報と設問 14 の「コロナに対する何か不満や不安はありますか」というテキストデータを取り上げる。

② 「コロナ不満」の回答の中には、回答者の感情を示す絵文字や特殊記号が入っているが、今回取り上げた東北大学が開発した日本語学習済みモデルでは、絵文字と特殊記号への学習をしていないため、アンケートデータの中のすべての絵文字や特殊記号を削除した。

③ 「コロナ不満」の回答の中には、1 つの回答の中、箇条書きで多数の不満や不安を投稿する

回答者が存在する。その複数の不満はそれぞれ異なる主旨を持っているため、「BERT」による文脈理解の正確性を考え、ここでは、箇条書きを含む合計 23 件のデータを全て分離して、1 つのデータには 1 個不満を持っている構造を維持する。

4.4　研究方法

本研究の分析には、Google 社が機械学習の研究を目的として開発した Google Colaboratory を使用する。BERT によるトピック分類とポジネガ判定の部分では、Python をプログラミング言語として使用する。また、BERT の出力結果に対して統計分析を行う際に、統計ソフト「SPSS27.0」を使用する。

4.4.1　BERTopic によるトピック分類

「言葉の意味」を AI にて理解するため、トピックモデルという自然言語処理の分野で用いられる解析手法が応用できる。トピックモデルでは、文章が複数の潜在的なトピックからなり、それらは確率的に生成されると仮定し、単語がそのトピックの確率分布に従って出現すると捉える。本研究で取り上げたデータ「コロナ不満」について書いた回答の場合、そこには「健康」についてだけではなく、「子供の教育」や「経済」などの話題についても書かれていることが想像される。これをトピックモデルで解析すると、得られた各トピックの出現しやすさ（確率分布）と文章中に出現する単語の特定のトピックにおける出現しやすさ（確率分布）の値に基づいて、テキスト分類ができる。

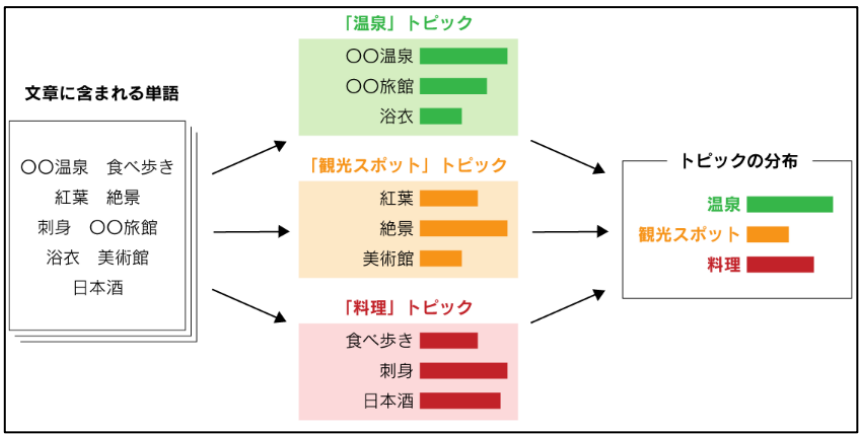


図 4－2　トピックモデルによるテキスト分類（イメージ）

（出所：SODA トピックモデルとは？ https://www.nico-soda.jp/blog/post/20210121_000096.html）

日本語のテキストデータに対するトピック分類の研究はいくつか存在する。（松河, 2017）らは LDA によるトピックモデルの分析に基づいて、授業評価アンケートの自由記述にトピック分類を行い、全体的に人間の感覚に合った分類を示した。ただし、LDA によるトピックモデルの分析では、R などのソフトウェアによる解析に時間がかかることと、トピックに対応したラベルの作成（トピック命名）に負担がかかるという課題が残された。

一方、BERTopic は BERT の一つの拡張モデルとして、1 章で紹介された Sentence BERT 技術をベースに、さらに BERT 埋め込みモデルとクラスベースの TF-IDF を活用して、トピックの説明に

重要な単語を残しながら、簡単に解釈できるトピックを可能にする高密度クラスターを作成するトピックモデリング手法である。また、生成されたトピックを簡単に解釈して視覚化することもできる。従来の LDA や NMF 手法に比べて、理解しやすい上に作業負担が軽いというメリットがある。また、（Abeer Abuzayed, 2021）らは BERTopic によるトピック分類の結果を、LDA および NMF 手法による分類結果と比較したが、BERTopic によって生成された全体的な結果は、NMF および LDA より良い結果を示した。

そのため、本研究では、BERTopic を用いて、下記の分析を行う。

① 「コロナ不満」を月ごとに（3 月－6 月）トピック分類をする

② 「コロナ不満」を回答者の年代別・職業別・年収別・居住地別でトピック分類をする

また、大量なテキストデータがあった場合、何百ものトピックが生成され、意味の近いトピックが混在する可能性があるため、本研究では最小類似度が 90%に超える重複のトピックを削除する。その後、抽出されたすべてのトピックに対して、最も頻繁に出現する上位 10 件のトピックに絞って、各トピックの意味を見て、ラベルを付け（トピック命名）と不満分野の特定をする。

4.4.2 BERT による感情分析(ポジネガ分析)

本研究で扱われる感情分析では、全体的にテキストの感情を解読することではなく、最も一般的なポジティブ・ネガティブ判定の 2 値分類とテキストのポジネガ得点から回答者の不満程度を読み取る分析である。具体的には、「コロナ不満」の回答の中、感情を示す単語ごとにポジティブなら加点・ネガティブなら減点し、総合点によってテキスト全体を判定する。そして、その総合得点が大きければ大きいほど（1 に近い）、テキストのポジティブ性が強く、総合得点が小さければ小さいほど（-1 に近い）、テキストのネガティブ性が強く、つまり、不満の程度が高いと読み取ることができる。

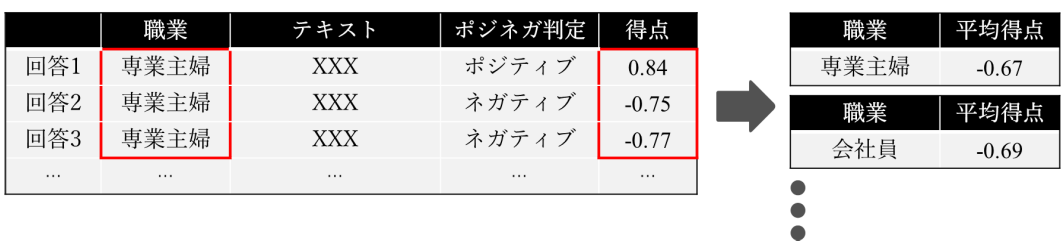
本研究では、下記の 2 つの方向で「コロナ不満」の不満程度を計算する。

① 月ごと（3 月－6 月）の回答者全体の不満程度を算出する

② 回答者の基本情報に基づき、グループ別（年代別・職業別・年収別・居住地別）の不満程度を算出する

不満程度の計算は感情分析から得られた各テキスト（不満の回答）のポジネガ得点に基づき、平均値を取る方法である。ポジネガ得点は-1 から 1 までの範囲である。グループ別で分析する場合、回答者の基本情報に合わせて、グループに分けて、ポジネガ得点の平均値を計算する。例えば、職業別で分析する場合、「専業主婦」や「会社員」などの基準でグループ内のポジネガ得点の平均値を算出する。そうすることによって、年代別や職業別など、グループ別の不満程度を比較できる。

図 4－3　不満程度の集計（イメージ）



4.4.3　日本語 BERT 訓練済みモデル

BERT は元々英語に基づいて作られたため、日本語のテキストデータを分析する場合には、その精度を上げるため、日本語の訓練済みモデルが必要である。現時点で多く使われているのは、東北大学・乾研究室で開発した日本語訓練済みモデルである。このモデルは日本語版 Wikipedia をコーパスに用いて訓練して、汎用言語モデル BERT に適応したものであり、現在は HuggingFace 製の BERT において、日本語訓練済みモデルとして追加されている。（五井野, 2021）らは、この訓練済みモデルを使って、医療文書から病名や症状などの固有表現を抽出して、その精度の高さも明らかにした。一方、（芝山, 2021）らは、東北大学で開発したモデルを含む 3 つの日本語訓練済みモデルを比較した結果、東北大学・乾研究室のモデルの精度が一番高く、汎用性も比較的に高いことを示した。そこで、本研究でも東北大学・乾研究室の BERT 日本語訓練済みモデルを使用する。

5. 分析結果

5.1　コロナ不満のトピック分類

図 5－1 は「BERTopic」によるトピック分類の出力結果の一部である。図 5－1 の示した通り、出現頻度が上位になるトピックが抽出され、「Count」列の数値はこのトピックに当てはまる回答の件数で、「Name」列はこのトピックの特徴である。その中で、注意を要するのは、うまくトピック抽出ができていないことを示すは「Topic＝-1」と出力される回答である。月ごとのトピック分類から見ると、約 2 割から 3 割の回答がうまく分類できていないため、以下の分析ではこの回答「Topic＝-1」を除くことにする。

	Topic	Count	Name
1	-1	769	-1_また_マスク_トイレットペーパー_そして
2	1	187	1_ただ_と思う_1日でも早い終息を_政府の初動が遅いのが国民にしわ寄せがきている
3	2	170	2_また_学校が休みになり_大人も不必要な外出をやめないと広がる一方_学校を休校にしたのはよ...
4	3	157	3_また_怖い_と思う_あと
5	4	151	4_マスク_トイレットペーパー_マスク不足_マスクだけでなく
6	5	149	5_経済がストップして_景気が悪くなる_経済の停滞_物流も止まることを懸念している
7	6	117	6_デマが多い_また_トイレットペーパー_あきらかにデマなのに買い占めが起こった
8	7	65	7_2回かかる事もあるらしいし_人として思いやりを持ってひとりひとりが意識を変えて生活して欲...
9	8	47	8_うがい_手洗い_whoの症状のない人のマスクは必要ないってことで_対抗できるワクチンなど...
10	9	43	9_20_職託職員公務員なので_打撃を受けてるのは勤め人よりも事業主なのに_非正規の人には保...
11	10	35	10_あります_pcr検査を渋る理由が色々取りざたされていて興味深い_過度の検査は混乱を招く...
12	11	34	11_3月の公演は中止になっている_ライブの延期や中止になっていて心配_好きなアーティストの...
13	12	31	12_コロナ_あとマスク騒動だが_年度末に被っているので過剰な自粛要請のせいで経済が止まるの...

	Topic	Count	Name
1	-1	1084	-1_また_そして_あと_しかし
2	1	335	1_マスク_怖い_マスク不足_また
3	2	262	2_また_毎日_無症状_マスコミは
4	3	225	3_通信費_光熱費_また_経済破綻
5	4	171	4_早く終息して欲しい_いつ終息するのか_みんなが_不安
6	5	171	5_4年後の自分の就職に響くのではないかということ_死だと考えています_満足に買い物できない...
7	6	125	6_早くワクチンを開発してほしい_わからない_また_1度かかって治っても
8	7	79	7_医療崩壊_医療崩壊が心配_1人1人が絶対に病気にならないように気をつけていかねばならない...
9	8	67	8_また_毎日ヒヤヒヤしながら乗っているので頭がおかしくなりそう_外出_政府のゆるめの対応が...
10	9	48	9_3月末に通院し_早く特効薬作ってほしい_治療方法がいつ確立されるか_批判ばかりする人のせ...
12	10	41	10_辛い_gwまでの娯楽が奪われることが確定している今_本当に苦痛_本当に苦痛_ネットのあ...
11	11	41	11_妊娠中なので_現在妊娠中なので_4月中に出産予定_感染したら子どもに影響があるのか不安...
13	12	29	12_ある程度は合わせてほしい_悪いことが起こらないか心配_最悪_母に会いに行きたいし_検査...

図 5－1　BERTopic によるトピック抽出の出力結果（一部）

表 5.1 から表 5.4 までにはトピック分類の集計結果となる。分析を遂行するにあたり、筆者は頻繁に出現している上位 10 個のトピックに属する各不満回答の内容を見て、トピックの命名（ラベル付け）を行い、「コロナ不満」の分野を特定する。3 月から 6 月のトピック抽出の結果を比較して見ると、月ごとにトピックの内容が大幅に変化していることが明らかになった。

表 5.1　BERTopic によるトピック抽出の集計結果(3 月)

Topic	件数	トピック概要	ラベル	分野
1	800	マスクと消毒液不足 トイレットペーパー不足	物資の不足	日常生活
2	297	騒ぎすぎ マスコミ 流行を喜んでいるようにしか思えない	マスコミ	情報
3	159	花粉症 うがい、手洗い	コロナ対策	コロナ対策
4	145	マスク不足、消毒液不足 買い占めをやめてほしい	物資の不足 デマ	日常生活 情報
5	130	早く収束してほしい 正しい情報だけを流してほしい 子どもが保育園に 4 月入園予定	コロナ対策 情報信憑性 子供の教育	コロナ対策 情報 教育
6	130	早く薬の開発をしてほしい マスコミが感染者をコロナを撒き散らした悪者のように報じるのはやめてほしい 新生児に対する処置	治療方法 マスコミ コロナ対策	コロナ対策 情報
7	93	オリンピック 景気が悪くなる	イベント中止 景気悪化	経済
8	87	コロナのせいでいろんなイベントが中止	イベント中止	娯楽
9	78	学校が休みになり 学生が外で遊びまくっている	休校問題	教育
10	75	政府の初動が遅い	政府の対応	政府

3 月のトピック分類の結果から見ると、出現頻度がトップになるトピックは「マスク不足」や「トイレットペーパー不足」など日常生活またはコロナ感染を防ぐために必須な物資に関するトピックである。これらの商品が不足すると、コロナ禍における日常生活に大きな支障が出るのため、不満の分野が「日常生活」と特定する。次に多いのは、マスコミが騒ぎすぎなどといった「情報」と特定することができる不満であった。また、新学期に迎えて子供の教育問題を心配する声も上位にランクインしており、トピック内容が「倒産」、「景気」に言及する「経済」に対して不満や不安がある回答もかなり多い。

表 5.2　BERTopic によるトピック抽出の集計結果(4 月)

Topic	件数	トピック概要	ラベル	分野
1	1460	マスク不足	物資の不足	日常生活
2	204	マスクが買えないこと	物資の不足	日常生活
		経済破綻で無茶苦茶な世界になってしまうこと	景気悪化	経済
		給与が下がること	収入減少	仕事
3	140	光熱費と通信費	生活コスト	日常生活
4	104	早くワクチンを開発してほしい	ワクチン開発	コロナ対策
		治療法や薬が早く見つかってほしい	治療方法	
5	64	医療崩壊	医療体制	医療
		政府の対応が遅すぎる	政府の対応	政府
		仕事の減少	収入減少	仕事
6	58	地方への感染拡大が心配	コロナ感染	コロナ感染
		後から体に弊害が出たりしないか心配		
		感染したらどうなるのかという不安		
7	46	あいまいな対策が続いていること	コロナ対策	コロナ対策
		子どもの学校が休校になり	子供の教育	教育
		自宅にいるのも飽きるのでこの自粛ムードがいつまでも続く事を懸念している	自粛疲れ	娯楽
8	28	仕事がなくなるのではないか	収入減少	仕事
		雇用が維持されるのか		
9	20	マスクに対してみんな神経質になりすぎで	周りの行動	日常生活
10	16	だれかに移してしまうのではないか	コロナ感染	コロナ感染
		子どもが罹患してしまうことが心配である		

4 月に入ると、不満回答のトピックが徐々に変化していることが確認できる。最も多い話題は変わらず「マスク不足」という問題だが、「景気悪化」・「収入減少」・「仕事減少」など景気の悪化によって、自身の仕事や収入に悪影響が与えられたことに訴える不満が多くなってきた。それに伴い、「光熱費」や「通信費」など生活コストに気になるトピックも上位にランクインした。その他、「ワクチン開発」を強く望んでいる声と「政府の対応が遅すぎる」と批判するトピックに言及した内容も頻繁に出現している。

表 5.3　BERTopic によるトピック抽出の集計結果(5 月)

Topic	件数	トピック概要	ラベル	分野
1	1152	ワクチンの開発 治療薬の開発	ワクチン開発 治療方法	コロナ対策
2	264	失業者の増加 不況になって倒産 犯罪の増加	雇用 倒産 犯罪率	経済
3	152	日本の対応が遅すぎるすべてにおいて	政府の対応	政府
4	114	学力低下が心配です	学力低下	教育
5	99	国の対応が遅い 政府の存在意義が分からなくなり	政府の対応	政府
6	86	政府の要請に対して過剰な反応を示すお店 生活リズムが狂う	生活リズム	日常生活
7	78	第 2 波が来る	感染拡大	コロナ感染
8	76	娯楽など自粛ができない人々の行いによってなかなか封じ込めには時間がかかり終息できない	周りの行動	日常生活
9	71	1 度陽性の人が陰性となり再び陽性となる １度感染して治っても再度かかってしまう	感染拡大	コロナ感染
10	37	会社が倒産しないか 経済が悪くなり航空会社が倒産するのではないのかと思う	倒産 景気悪化	仕事 経済

5 月に入ると、「マスク不足」など日用品の不足を訴える不満がほぼなくなった。代わりに、「不況による失業問題」・「不況による治安悪化と犯罪増加」など「経済」に関する不満や不安が上位にランクインした。また、「日本政府の対応が遅すぎる」と批判する声も先月 4 月の結果により多いである。その中、一番不満や心配の声が上がったのは、「ワクチン開発」や「治療薬開発」にめぐるコロナ感染を防ぐための対策になる。さらに、「子供の学力」など子供の教育問題に心配する声が 3 月、4 月に引き続き上位にランクインした。

表 5.4　BERTopic によるトピック抽出の集計結果(6 月)

Topic	件数	トピック概要	ラベル	分野
1	781	いつまでこのような生活が続くのか いつワクチンができるのか	自粛疲れ ワクチン開発	日常生活 コロナ対策
2	261	経済の悪化 景気悪化	景気悪化	経済
3	236	また感染が広がること 感染が多い地域からも人が来るようになり不安	感染拡大	コロナ感染
4	155	医療崩壊 経済の停滞 治安の悪化 犯罪の多発	医療体制 景気悪化 治安と犯罪	医療 経済 治安
5	144	第 2 波が来て 秋冬に大きな第 2 波がくる	感染拡大	コロナ感染
6	144	緊急事態宣言が解除されてから感染再拡大	政府の対応	政府
7	129	学校が再開されたが 小学生の子どもの学習の遅れ	子供の教育	教育
8	128	政府の対応の遅さ	政府の対応	政府
9	113	早くワクチン 2022 年頃になって少しでも安心出来るといい	ワクチン開発	コロナ対策
10	102	新型コロナウイルスの収束が見えない	感染拡大	コロナ感染

6 月のトピック分類の結果から見ると、不満や不安の声が最も多いのは、「コロナ禍が長引くことによって、今の生活に疲弊感が出る」など自粛生活またはコロナ禍の生活に疲れが出ると訴える不満であった。次に多いのは、先月に引き続き「経済悪化」に関する内容であった。また、「子供の教育への心配」と「政府への批判」の声も引き続き上位にランクインした。さらに、緊急事態宣言の解除にあたって、コロナ感染の第 2 波が来るという心配な声も上位にランクインした。

5.2　コロナ不満の不満程度数値化（ポジネガ判定）

まず、月別ポジネガ得点の推移と回答全体の平均ポジネガ得点は表 5.5 に示す。ポジネガ得点の全体平均は-0.5658 であり、4 月の第 2 回調査と 6 月の第 4 回調査における回答のポジネガ得点の平均値は全体平均より下回る状態である。5 月の第 3 回調査のポジネガ得点は全体平均よりやや上回るが、かなり接近した値である。一方、ポジネガ得点の平均値が一番低いなのは、3 月の第 1 回調査の回答であった。この結果により、「コロナ禍」が発生した後、最初の 3 月に回答者の不満程度が比較的に低い水準になったものの、4 月に入ると不満程度が一気に上昇した。さらに、5 月の不満程度がある程度緩和されたが、6 月に入るとまた上昇する傾向が明らかにした。

表 5.5　全体・月別ポジネガ得点（不満程度）

調査回	回答件数	平均ポジネガ得点
第 1 回調査－3 月	2996	-0.5166
第 2 回調査－4 月	4060	-0.5971
第 3 回調査－5 月	2996	-0.5520
第 4 回調査－6 月	2997	-0.5863
全体	13049	-0.5658

3 月と 4 月のポジネガ得点の差が大きい理由を探してみたところ、それはネガティブ判定の件数の差であった。3 月の第 1 回調査において、約 19%の回答がポジティブと判定され、その得点が正の値のため、全体平均値を押し上げた。一方、4 月の第 2 回調査の中、3 月より少なく約 15%の回答がポジティブと判定されたため、全体の平均ポジネガ得点が相対的に低くなった。

表 5.6　全体・月別ポジネガ得点（不満程度）

調査回	ポジティブ件数	平均ポジ得点	ネガティブ件数	平均ネガ得点
第 1 回調査－3 月	556	0.8368	2440	-0.8250
第 2 回調査－4 月	589	0.8190	3471	-0.8374

続いて、回答者の基本情報に基づいて、グループ別でポジネガ判定をする結果は表 5.7－表 5.12 に示す。男女別、職業別、年収別、年齢層別と子供を持っているか否かという合計 5 つの標準でグループ分けをした。

男女別のポジネガ得点から見ると、比較的に男性のほうのネガティブ性が低く、女性の平均ポジネガ得点が-0.5938 であり、サンプル全体の得点-0.5658 よりもかなり下回るため、全体的に女性の不満程度が高いという傾向が読み取れる。

表 5.7　男女別ポジネガ得点（不満程度）

性別	回答件数	平均ポジネガ得点
男性	3778	-0.4970
女性	9271	-0.5938

男性と女性のポジネガ得点の差が大きい理由を探すため、そのポジネガの判定件数と得点が表 5.4 に示す。男性と女性のネガティブ判定の割合はそれぞれ 82%と 84%と、それなりの差が見られていないが、女性のネガティブ得点が-0.8509 男性より大幅に下回るので、平均値を押し下げたと考えられる。

表 5.8　全体・月別ポジネガ得点（不満程度）

性別	ポジティブ件数	平均ポジ得点	ネガティブ件数	平均ネガ得点
男性	682	0.8372	3096	-0.7909
女性	1446	0.7975	7825	-0.8509

職業別のポジネガ得点から見ると、全体平均値-0.5658 によりも得点が下回る職種は公務員、アルバイト・パート、専業主婦（主夫）と自営業であった。その中、専業主婦（主夫）の平均得点が-0.6126 に達しており、全職種の回答の中にネガティブ性が一番高い。それに伴い、専業主婦（主夫）による回答件数も 3175 件の多い水準であるため、「コロナ禍」において、回答者の中において、専業主婦（主夫）らの不満程度が一番高いと言える。続いて平均得点が低いなのは、自営業と公務員による回答であった。ただし、自営業者と公務員による「コロナ不満」の投稿件数がそれぞれ 459 件と 296 件で、それほど多くはないため、必ずしも自営業と公務員の不満程度が高いということが言い難い。次に多いのは、回答件数 2619 件、平均得点が-0.5742 のアルバイト・パート職であった。全体の平均得点にかなり接近している数値であったが、回答件数の多さを考えて、ここではアルバイト・パート職によるコロナ不満程度が高いことが読み取れる。

一方、職種が会社員、役員、自由業と学生による回答のポジネガ得点が全体平均値-0.5658 により上回る。その中、会社員による不満投稿の平均得点が-0.5594 であり、全体平均に上回るが、かなり接近している数値であった。そのため、会社員によるコロナ不満程度が中等と読み取れる。そして、学生による投稿が 579 件と比較的に多い水準であったが、そのポジネガ得点が僅か-0.4652 であり、全職種の中で一番低い値になる。そのため、回答者の中、学生のコロナ不満程度が比較的に低いことが明らかになった。最後、無職・その他のポジネガ得点を見ると、その値も-0.4724 であり、全職種の中で二番目の低い値であるため、まだ仕事をしていない人または仕事を持っていない人のコロナ不満程度が低いことが読み取れる。

表 5.9　職業別ポジネガ得点（不満程度）

職業	回答件数	平均ポジネガ得点
会社員	4515	-0.5594
経営者・役員	101	-0.4811
公務員	296	-0.5823
アルバイト・パート	2619	-0.5742
専業主婦（主夫）	3175	-0.6126
自由業	334	-0.5516
自営業	459	-0.5993
学生	579	-0.4652
無職・その他	971	-0.4724

年収別のポジネガ得点から見ると、300 万円以上の回答者に対して、年収が上がると平均ポジネガ得点の数値が下がるという傾向がある。特に、年収が 300 万円－500 万円の回答者の得点が-0.6197 と一番低い水準になる。また、年収が 300 万円以下の回答者による不満投稿も、ネガティブ性が強いことが明らかになった。表 5.5 の職業別平均得点と一緒に見ると、その傾向になり理由がある程度分かる。一般的に、収入が少ないアルバイト・パート職と専業主婦のポジネガ得点が低く、それが年収別の得点傾向と一致している。そのため、全体的に年収が低い人のコロナ不満程度が低く、年収が高い人の不満程度または不満への関心が少ないということが考えられる。

表 5.10　年収別ポジネガ得点（不満程度）

年収	回答件数	平均ポジネガ得点
300 万円以下	3467	-0.5893
300 万円－500 万円	3869	-0.6197
500 万円－800 万円	3815	-0.5183
800 万円以上	1898	-0.5083

回答者年代別の結果を見ると、中年層の回答者と高齢者が若年層よりポジネガ得点の平均値が低い。全体的に年齢が上がるとともに、不満投稿のネガティブ性が強くなるという傾向がある。つまり、年齢が上がると、コロナ不満程度が強くなると読み取れる。特に、60 歳以上の回答者の不満投稿が少ないものの、ネガティブ性が一番強く、全体平均-0.5658 に下回る-0.5864 の数値であった。次に多いのは、40 歳－49 歳までの中年層からの回答になる。一方、29 歳以下の若年層の回答から見ると、平均ポジネガ得点がそれぞれ-0.4755 と-0.5163 であるため、若年層のコロナ不満程度が比較的に低くことが明らかになった。

表 5.11　年齢層別ポジネガ得点（不満程度）

年齢	回答件数	平均ポジネガ得点
19 歳以下	232	-0.4755
20 歳－29 歳	1626	-0.5163
30 歳－39 歳	4635	-0.5695
40 歳－49 歳	3743	-0.5829
50 歳－59 歳	2170	-0.5689
60 歳以上	643	-0.5864

最後に、サンプルデータ全体の女性の割合と専業主婦の割合が高いため、女性が関心を持つ家庭と子供について、子供の数の違い基づいて、グループ別でポジネガ得点を比較した。その結果から見ると、予想と異なりそれなり傾向のある得点分布には得られなかった。子供を持っていない回答者の不満投稿の平均ポジネガ得点がやや低いが、子供人数 1 人－3 人までの回答者の平均値と接近しているため、子供人数だけでコロナ不満程度を判断することが言い難いである。一方、不満投稿の件数自体が少ないが、4 人以上子供を持っている回答者のポジネガ得点-0.6073 であるため、ある程度、子供人数が 4 人以上の回答者のコロナ不満程度が高いと考えられる。

表 5.12　家庭人数（子供を持っている人数）ポジネガ得点（不満程度）

子供人数	回答件数	平均ポジネガ得点
0 人	5990	-0.5710
1 人	2726	-0.5604
2 人	3339	-0.5567
3 人	881	-0.5760
4 人以上	113	-0.6073

各グループ間のポジネガ得点の差を確認するため、本研究は男女別・職業別・年齢層別と年収別のポジネガ得点に対して、SPSS による分散分析と T 検定を行い、グループ間の有意差を確認した。

独立サンプルの検定

		等分散性のための Levene の検定		2 つの母平均の差の検定						
									差の 95% 信頼区間	
		F 値	有意確率	t 値	自由度	有意確率 (両側)	平均値の差	差の標準誤差	下限	上限
ポジネガ得点	等分散を仮定する	67.165	.000	15.388	14493	.000	.1706589314	.0110900911	.1489209368	.1923969259
	等分散を仮定しない			15.094	6394.467	.000	.1706589314	.0113064312	.1484945380	.1928233247

図 5－2　男女別ポジネガ得点・T 検定

T 検定の結果によると、男女別ポジネガ得点の F 値の有意確率が 0.000 であるため、1%の有意水準のもと、男性のポジネガ得点と女性のポジネガ得点の分散が異なると確認できる。T 値の有意確率も 0.000 であるため、1%の有意水準のもと、男性のポジネガ得点と女性のポジネガ得点の平均値は差異があり、つまり、性別によって、人々のコロナ不満に対するネガティブ具合も異なると言える。

分散分析

ポジネガ得点

	平方和	自由度	平均平方	F 値	有意確率
グループ間	8.803	4	2.201	5.784	.000
グループ内	4316.011	11342	.381		
合計	4324.814	11346			

図 5－3　職業別ポジネガ得点・分散分析

多重比較

従属変数:　ポジネガ得点

最小有意差

(I) 職業	(J) 職業	平均値の差 (I-J)	標準誤差	有意確率	95% 信頼区間 下限	95% 信頼区間 上限
1	2	-.022041466	.0151518751	.146	-.051741765	.0076588330
	3	-.037897546*	.0142875868	.008	-.065903690	-.009891402
	4	.0208518538	.0302213684	.490	-.038387262	.0800909691
	5	-.114850353*	.0272306348	.000	-.168227113	-.061473594
2	1	.0220414659	.0151518751	.146	-.007658833	.0517417647
	3	-.015856080	.0162834373	.330	-.047774437	.0160622767
	4	.0428933196	.0312145207	.169	-.018292546	.1040791854
	5	-.092808887*	.0283288310	.001	-.148338302	-.037279473
3	1	.037897546*	.0142875868	.008	.0098914017	.0659036902
	2	.0158560801	.0162834373	.330	-.016062277	.0477744369
	4	.0587493997	.0308042549	.057	-.001632274	.1191310735
	5	-.076952807*	.0278761275	.006	-.131594844	-.022310770
4	1	-.020851854	.0302213684	.490	-.080090969	.0383872616
	2	-.042893320	.0312145207	.169	-.104079185	.0182925461
	3	-.058749400	.0308042549	.057	-.119131073	.0016322741
	5	-.135702207*	.0385522302	.000	-.211271254	-.060133160
5	1	.114850353*	.0272306348	.000	.0614735937	.1682271128
	2	.092808887*	.0283288310	.001	.0372794732	.1483383017
	3	.076952807*	.0278761275	.006	.0223107703	.1315948444
	4	.135702207*	.0385522302	.000	.0601331600	.2112712541

*. 平均値の差は 0.05 水準で有意です。

図 5－4　職業別ポジネガ得点・分散分析（多重比較）

分散分析の結果によると、職業別のポジネガ得点の F 値の有意確率が 0.000 であるため、1%の有意水準のもと、職業間のポジネガ得点の差異が存在していると確認できる。図 5－4 の多重比較

の結果を具体的に見ると、会社員(職業 1)のコロナ不満程度は専業主婦(職業 3)と学生(職業 5)の不満程度に比べて、有意確率が 0.000 であるのため、差異が存在していると言える。他にも、職業別のグループ間の差異が確認できる。特に、学生(職業 5)はほかのすべての職業に比べて、コロナ不満程度が異なる。

しかしながら、年齢層別と年収別の分散分析の結果では、有意確率がそれぞれ 0.05 に上回るのため、年齢層別と年収別のコロナ不満程度に差が明らかではないことも確認できた。そのため、得点平均値に基づいて、コロナ不満程度を判断する際に、年齢層別と年収別という基準ではなく、男女別と職業別をもっと注目すべきかもしれない。

6. 結論と考察

本節では、本研究の分析結果から、どんな知見または意味が得られるのか、その考察について述べる。

6.1 トピック分類の結論

「コロナ不満」のテキストデータに対するトピック分類を行うことで、「コロナ禍」において、人々の不満トピックが常に変化していることが明らかになった。一方、その中、ずっと存在している不満や不安の話題もある。例えば、3 月から 6 月まで、「学力低下」、「子供の教育」など「教育」分野の不満を訴える回答者がずっと存在しているため、このような固定的な不満に目をむき、新商品・新サービスを開発するチャンスがある。政府・自治体にとっては、「コロナ不満」のトピック分類の結果を用いて、最新な「不満」トレンドを把握でき、頻繁に出現しているトピックに対して、早めに制度設計を万全にすることが可能となる。

6.2 感情分析の結論

感情分析（ポジネガ判定）では、人々の不満程度の数値化を実現したため、政府・自治体にとって、より早く「ターゲット」を見つけ出せる。つまり、不満程度の高い特定なグループに対して、彼らの「コロナ禍」における生活を支援・サポートできるような商品・サービスを開発したり、彼らの「コロナ禍」における不満を解消でき、今後もその不満発生を防げるような対策を完備することができる。

表 6.1　ポジネガ得点のクロス集計(職業別・性別)

性別/職業	会社員	アルバイト・パート	専業主婦(主夫)	自営業	学生
男性	-0.5304	-0.5591	-0.8500	-0.6244	-0.4680
女性	-0.5689	-0.5846	-0.6124	-0.5495	-0.4543

また、5 章の分散分析と T 検定の結果にみれば、男女別と職業別のグループ間にポジネガ得点の差があったため、クロス集計表を作成し、特定のグループのネガティブ性を確認する。結果、性別に問わず、アルバイト・パート職のコロナ不満程度がともに高い一方、女性・専業主婦のグループと男性・自営業のグループのコロナ不満程度が高いことが分かる。そのため、政府・自治体が With コロナの生活支援策を考える時、アルバイト・パート職に対するサポートを優先すべきかもしれない。さらに、女性の専業主婦に対する支援策や男性の自営業者に対する支援策も早め

に解決すべき課題とみられる。このような分析に基づくことで、支援対象者のターゲットを明確にすることが可能になり、自治体の支援策の改善に役立つことが期待される。

6.3 考察

本研究で試みた「新型コロナ不満アンケートデータ」のテキストデータを用いて、「BERT」などの最新な自然言語処理アプローチで不満特徴を抽出する方法は、企業または自治体にとり、有益な情報を提供することが期待される。政府・自治体では、この「コロナ不満」の研究方法に基づいて、住民から日常的な不満を収集するプラットフォームを構築し、定期的にその不満をテキストマイニングに通じて特徴抽出をすることができる。例えば、公式アプリ「Cocoa」または「LINE」に通じて、「コロナ不満」を収集するアンケート調査を定期的に発表することができる。そうすることで、自治体が即時に住民の不満トピックと不満程度を把握することができ、早めに対応策を考えて、「With コロナ」住民の暮らしの改善に繋がることが期待される。一方、企業は、「コロナ不満」を収集し分析することで、新ビジネス創出に役立つばかりでなく、社内でもこのような「不満」を分析するプラットフォームを構築することで、タイムリーに「With コロナ」の社員が仕事や日常生活に対する不満を把握でき、早めの対策を取ることで、社員の OL 改善やモチベーション向上につなげることが可能となる。

7. 限界・課題

本研究の課題と限界について、以下の 3 点が挙げられる

① サンプルデータの限界性について

4 章のデータ概要でも紹介した通り、本研究で取り上げた「コロナ不満」のデータは、株式会社 Insight Tech が運営する「不満買取センター」上で実施したアンケート調査である。アプリ利用者はこの「コロナ不満」に関するアンケート調査を回答することで、一定なポイントを得られ、アプリ内のネットショップで何らかの商品を交換することができる。そのため、政府・自治体によるアンケート調査に比べて、回答者の代表性、つまり、回答者のコロナ不満に対する答えが一般人に代表するかどうか問題である。

同様に、すべてのアンケートは携帯アプリまたはウェブサイトで行ったため、基本的に回答者が IT リテラシーが高い人と思われ、高齢者など比較的に IT リテラシーの低い人の「コロナ不満」の声を反映していない可能性が高い。また、回答者の女性の割合と、専業主婦の割合が高めたため、得られた不満トピックの特徴や不満程度の特徴もそれを反映して偏っている可能性がある。

② 「BERT」によるトピック分類の課題について

本研究を遂行するあたり、「BERT」技術を用いたが、出力結果から見るといくつかの課題が残される。まず、トピック分類の結果から見ると、毎月のトピック分類の中、Topic が-1、つまり、うまく分類できていない、有意ではない回答が約 2 割－3 割まで存在している。分類の精度を上げるため、事前学習モデルを変えるなどの工夫を図ることで、分類できないトピックを出来る限りに減少させる必要がある。また、分類されたトピックの中でも、複数の内容を含んでいるトピックもある。例えば、4 月のトピック分類の結果の中、「4 年後の自分の就職に響くのではないか」・「満足に買い物できないので困る」・「持病がある人の方が重症化する」という関係性の弱い内容が同一のトピックに出現した。このような場合、トピックにラベル付けることは難しい。

③ 「BERT」によるポジネガ判定（感情分析）の課題について

本研究は、「コロナ不満」のテキストデータに対して、「BERT」による感情分析を採用して、各回答のネガティブ性を数値化した。しかし、分析結果から見ると、元々「不満や不安」を表すネガティブなデータだが、BERT の感情分析により、「ポジティブ」と判定された場合もある。その理由は、言葉の曖昧性と考えられる。例えば、3 月のアンケート調査の一部「ポジティブ」と判定された回答例を表 7.1 に取り上げる。

表 7.1　ポジティブと判定された不満回答の一例

ID	Text	判定	スコア
129	4 月から子供が受験生になるので色々、テストとかあるのにそれの妨げになるかも。早く、薬や予防接種とか、出来上がって欲しいです。	ポジティブ	-0.7967
841	家族が持病があるから、みんなで収束の方向に行くように協力してほしい。政府には、過剰なくらいに対策をお願いしたいです。	ポジティブ	-0.8905
2849	自分さえよければという考えでなく、譲り合う気持ちを持つようになって欲しい。	ポジティブ	-0.8581

アンケートデータの中では、自身の不満を表現する回答だけではなく、「コロナ禍」における自身の希望または今後に向けてどう対応すべきかという内容も多かった。したがって、その内容がポジティブかまたはネガティブかの明確な傾向が見られず、「BERT」のポジネガ判定に影響を及ぼしたと考えられる。例えば、表 7.1 で示したように、不満回答の中で「〇〇すべき」や「〇〇にしてほしい」などアドバイス的な意見が「ポジティブ」と判定されやすい傾向がある。今のところ、「BERT」による感情分析は、「ポジティブ」と「ネガティブ」の二値分類を実現することができるが、自然言語の曖昧性がかなり強く、「ポジティブか」・「ネガティブか」の 2 つの極性だけで言葉の意味を完全に判断することが難しいと考えられる。今後、感情分析において、このような課題の解消を図るようなアプローチが出ることが期待される。

8. 謝辞

本研究では、 国立情報学研究所の IDR データセット提供サービスにより株式会社 Insight Tech が運営する Web サービス「不満買取センター」から提供を受けた「新型コロナ不満アンケートデータ」を利用しました。提供いただき誠にありがとうございました。

また、本論文の作成にあたり、多くの方々にご指導ご鞭撻を賜りました。指導教員である林高樹教授には、本研究を進めるにあたり様々なアドバイスとご指導を賜りました。特に BERT による分析の部分で、適切なご助言を賜りました。副査の中村洋教授と大林厚臣教授も分析手法の改善や修論の示唆など、細部にわたるご指導をいただきました。そして、ゼミの同級生には常に意義のある議論を頂き、分析手法への理解が深まれ、精神的にも支えられました。

最後に、皆様には、本研究の遂行にあたり多大なご助言、ご協力頂きました。ここに誠意の意を表します。

9. 参考文献

[1] 泉翔太, 堀太成, 山根達郎, 全邦釘, 藤森祥文, 森脇亮(2020). Deep Learning を用いたマイクロブログ投稿文の災害情報分類, J-Stage, AI・データサイエンス論文集 2020 年 1 巻 J1 号, 398-405.

[2] 渡邊真治(2017). マイナンバー普及への感情要因の影響に関する分析, 2017 年春季全国研究発表大会, B2-2

[3] 四方田健二(2021). 新型コロナウイルス感染拡大に伴う休校に対する社会的関心: Twitter 投稿内容の計量テキスト分析と感情分析, 名古屋学院大学教職センター年報第 5 号(2021), 49-61.

[4] 藤井義久(2020). 中学生における欲求不満とキレやすさとの関係, 学校メンタルヘルス 10 巻(2007), 398-405.

[5] 富井久義(2020). 新型コロナウイルス感染症は遺児世帯の生活にどのような影響を及ぼしたか―遺児世帯の家計と教育・進路選択への影響, 社会情報研究・第 2 巻 2 号, 398-405.

[6] 中澤政孝, 亀井且有, 前田陽一郎, クーパー・エリック(2020). BERT を用いた単文の感情極性推定手法の提案とその有効性, 日本知能情報ファジィ学会, 第 36 回ファジィシステムシンポジウム, TA2-3.

[7] 伏木田稚子, 北村智, 山内祐平(2012). テキストマイニングによる学部ゼミナールの魅力・不満の検討, 日本教育工学会論文誌(36), 165-168.

[8] 松河秀哉, 大山牧子, 根岸千悠, 新居佳子, 岩﨑千晶, 堀田博史(2017). トピックモデルを用いた授業評価アンケートの自由記述の分析, 日本教育工学会論文誌(41), 233-244.

[9] 鈴木かの子, 松井孝典, 川久保俊, 増原直樹, 岩見麻子, 町村尚(2021). BERT モデルを用いた SDGs に関するマルチラベル文書分類器の構築とマッチングシステムの開発, 人工知能学会全国大会論文集, 第 35 回, 4H3-GS-11d-01.

[10] 王 帥, 三道 弘明(2019). テキストマイニングによる不満から読み取る鉄道に関する問題－ベイズ学習による投稿者の不明な属性の推定, NII-IDR ユーザフォーラム 2019, P13.

[11] 松本佳依, 三道弘明(2019). 「あなたの不満買い取ります」から読み取る女性の声－テキストマイニングによる抽出, NII-IDR ユーザフォーラム 2019, P12.

[12] 堀部めぐみ, 笹岡沙也加, 長沼美紗, 長谷川栞, 原英彰, 中村光浩(2018). テキストマイニングによる産後うつについて母親が思うことの分析 ―ソーシャルメディアにおける発言の内容から, 看護科学研究, 2018 年 16 巻 2 号, 53-63.

[13] 花井 友美, 小口 孝司(2008). E メールの交換過程における感情表現の出現パターン:テキスト・マイニングを用いた分析, 社会心理研究, 2008 年 24 巻 2 号, 131-139.

[14] Abeer Abuzayed, Hend Al-Khalifa (2021). BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique, Procedia Computer Science, Volume 189, 2021, 191-194.

[15] Joseph W. Newman, Robert A. Westbrook (1978). An Analysis of Shopper Dissatisfaction for Major Household Appliances, JMR, Journal of Marketing Research (pre-1986), Chicago Vol.15, Iss.000003.

[16] Mohamed M.Mostafa (2013). More than words: Social networks' text mining for consumer brand sentiments, Expert Systems with Applications, Volume 40, Issue 10, August 2013, 4241-4251.

[17] Maria Giatsogloua, Manolis G. Vozalis, Konstantinos Diamantaras, Athena Vakali, George Sarigiannidis, Konstantinos Ch. Chatzisavvas (2017). Sentiment analysis leveraging emotions and word embeddings,

Expert Systems with Applications, Volume 69, March 2017, 214-224.
[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Association for Computational Linguistics, Volume 1, 2019, 4171–4186.
[19] Guang Qiu, Xiaofei He, Feng Zhang, Yuan Shi, Jiajun Bu, Chun Chen (2010). DASA: Dissatisfaction-oriented Advertising based on Sentiment Analysis, Expert Systems with Applications, Volume 37, Issue 9, September 2010, 6182-6191.
[20] Avishek Garain, Sainik Kumar Mahata (2019). Sentiment Analysis at SEPLN (TASS)-2019: Sentiment Analysis at Tweet level using Deep Learning, Cornell University, Computation and Language, arXiv:1908.00321.
[21] Santiago González-Carvajal, Eduardo C. Garrido-Merchán (2021). Comparing BERT against traditional machine learning text classification, Cornell University, Computation and Language, arXiv:2005.13012.
[22] Jieh-Sheng Lee, Jieh Hsiang (2019). PatentBERT: Patent Classification with Fine-Tuning a pre-trained BERT Model, Cornell University, Computation and Language, arXiv:1906.02124.
[23] Zhengjie Gao, Ao Feng, Xinyu Song, Xi Wu (2019). Target-Dependent Sentiment Classification With BERT, IEEE Access, Volume:7, 11 October 2019, 154290 – 154299.
[24] Hu Xu, Bing Liu, Lei Shu, Philip S. Yu (2019). BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis, Cornell University, Computation and Language, arXiv:1904.0223.
[25] 石田基広(2017), R によるテキストマイニング入門(第 2 版), 森北出版.
[26] 小林雄一郎(2017), R によるやさしいテキストマイニング, オーム社.
[27] 株式会社 Insight Tech :「不満買取センター」不満調査データセット, 国立情報学研究所情報学研究データリポジトリ
[28] 株式会社 Insight Tech :「不満買取センター」新型コロナ不満アンケートデータ, 国立情報学研究所情報学研究データリポジトリ

付録

1. BERT によるトピック分類のプログラム(SentenceBERT のみ)

```
#Google Drive にマウント
from google.colab import drive
drive.mount('/content/drive')
```

```
#MeCab インストール
!apt install mecab libmecab-dev mecab-ipadic-utf8
!pip install mecab-python3

!apt install git make curl xz-utils file
!git clone --depth 1 https://github.com/neologd/mecab-ipadic-
neologd.git
!echo yes | mecab-ipadic-neologd/bin/install-mecab-ipadic-neologd -n -a
!ln -s /etc/mecabrc /usr/local/etc/mecabrc

!pip install -q transformers
!pip install fugashi
!pip install ipadic
```

```
#BERTopic インストール
%%capture
!pip install bertopic[all]
!pip install bertopic[visualization]
!pip install flair
!pip install umap-learn
```

```
!pip install unidic-lite
```

```
#ファイル読み込み
import csv
import pandas as pd
import numpy as np
#3 月から 6 月までの不満テキストデータ
Fuman_data_03 = pd.read_csv('drive/My Drive/Fuman_Data/fuman_test03_BER
T.csv',encoding = 'UTF-8')
Fuman_data_04 = pd.read_csv('drive/My Drive/Fuman_Data/fuman_test04_BER
T.csv',encoding = 'UTF-8')
```

```
Fuman_data_05 = pd.read_csv('drive/My Drive/Fuman_Data/fuman_test05_BER
T.csv',encoding = 'UTF-8')
Fuman_data_06 = pd.read_csv('drive/My Drive/Fuman_Data/fuman_test06_BER
T.csv',encoding = 'UTF-8')
#データ配列転換
fuman03 = Fuman_data_03.loc[:,"Text"]
fuman04 = Fuman_data_04.loc[:,"Text"]
fuman05 = Fuman_data_05.loc[:,"Text"]
fuman06 = Fuman_data_06.loc[:,"Text"]
```

```
import pandas as pd
import numpy as np
#Transformer 導入
from bertopic import BERTopic
import transformers
from transformers import BertJapaneseTokenizer, BertModel
# Load sentence transformer model
from sentence_transformers import util, SentenceTransformer
sentence_model = SentenceTransformer("stsb-xlm-r-multilingual")

# Create documents embeddings
embeddings = sentence_model.encode(fuman03, show_progress_bar=False)
```

```
#UMAP と HDBSCAN 導入
import umap
import hdbscan
# Define UMAP model to reduce embeddings dimension
umap_model = umap.UMAP(n_neighbors=15,
                       n_components=10,
                       min_dist=0.0,
                       metric='cosine',
                       low_memory=False).fit_transform(embeddings)

# Define HDBSCAN model to perform documents clustering
cluster = hdbscan.HDBSCAN(min_cluster_size=10,
                                 min_samples=1,
                                 metric='euclidean',
                                 cluster_selection_method='eom',
```

```
                                    prediction_data=True).fit(umap_model)
```

```
#可視化
import matplotlib.pyplot as plt

# Prepare data
umap_data = umap.UMAP(n_neighbors=15, n_components=2, min_dist=0.0, met
ric='cosine').fit_transform(embeddings)
result = pd.DataFrame(umap_data, columns=['x', 'y'])
result['labels'] =cluster.labels_

# Visualize clusters
fig, ax = plt.subplots(figsize=(20, 10))
outliers = result.loc[result.labels == -1, :]
clustered = result.loc[result.labels != -1, :]
plt.scatter(outliers.x, outliers.y, color='#BDBDBD', s=2)
plt.scatter(clustered.x, clustered.y, c=clustered.labels, s=2, cmap='hs
v_r')
plt.colorbar()
```

```
#c-TF-IDF
fuman03_df = pd.DataFrame(fuman03, columns=["Text"])
fuman03_df['Topic'] = cluster.labels_
fuman03_df['Doc_ID'] = range(len(fuman03_df))
fuman03_df_per_topic = fuman03_df.groupby(['Topic'], as_index = False).
agg({'Text': ' '.join})
#frequency of each word
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer

def c_tf_idf(documents, m, ngram_range=(1, 1)):
    count = CountVectorizer(ngram_range=ngram_range).fit(documents)
    t = count.transform(documents).toarray()
    w = t.sum(axis=1)
    tf = np.divide(t.T, w)
    sum_t = t.sum(axis=0)
    idf = np.log(np.divide(m, sum_t)).reshape(-1, 1)
```

```
    tf_idf = np.multiply(tf, idf)

    return tf_idf, count

tf_idf, count = c_tf_idf(fuman03_df_per_topic.Text.values, m=len(fuman0
3))
#Topic Representation
def extract_top_n_words_per_topic(tf_idf, count, fuman03_df_per_topic,
n=20):
    words = count.get_feature_names()
    labels = list(fuman03_df_per_topic.Topic)
    tf_idf_transposed = tf_idf.T
    indices = tf_idf_transposed.argsort()[:, -n:]
    top_n_words = {label: [(words[j], tf_idf_transposed[i][j]) for j in
 indices[i]][::-1] for i, label in enumerate(labels)}
    return top_n_words

def extract_topic_sizes(df):
    topic_sizes = (df.groupby(['Topic'])
                    .Text
                    .count()
                    .reset_index()
                    .rename({"Topic": "Topic", "Text": "Size"}, axis='
columns')
                    .sort_values("Size", ascending=False))
    return topic_sizes

top_n_words = extract_top_n_words_per_topic(tf_idf, count, fuman03_df_p
er_topic, n=20)
topic_sizes = extract_topic_sizes(fuman03_df); topic_sizes.head(10)
```

```
#Topic Reduction
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
for i in range(20):
    # Calculate cosine similarity
    similarities = cosine_similarity(tf_idf.T)
    np.fill_diagonal(similarities, 0)
```

```
    # Extract label to merge into and from where
    topic_sizes = fuman03_df.groupby(['Topic']).count().sort_values("Te
xt", ascending=False).reset_index()
    topic_to_merge = topic_sizes.iloc[-1].Topic
    topic_to_merge_into = np.argmax(similarities[topic_to_merge + 1]) -
 1

    # Adjust topics
    fuman03_df.loc[fuman03_df.Topic == topic_to_merge, "Topic"] = topic
_to_merge_into
    old_topics = fuman03_df.sort_values("Topic").Topic.unique()
    map_topics = {old_topic: index - 1 for index, old_topic in enumerat
e(old_topics)}
    fuman03_df.Topic = fuman03_df.Topic.map(map_topics)
    fuman03_df_per_topic = fuman03_df.groupby(['Topic'], as_index = Fal
se).agg({'Text': ' '.join})

    # Calculate new topic words
    m = len(fuman03)
    tf_idf, count = c_tf_idf(fuman03_df_per_topic.Text.values, m)
    top_n_words = extract_top_n_words_per_topic(tf_idf, count, fuman03_
df_per_topic, n=20)

topic_sizes = extract_topic_sizes(fuman03_df); topic_sizes.head(10)
```

2. BERT によるトピック分類のプログラム(BERTopic)

```
#BERT・埋め込みモデル・東北大学・3 月

#Sentence Transformer
from bertopic import BERTopic
import transformers
import MeCab
from transformers import BertJapaneseTokenizer, BertModel
from flair.embeddings import TransformerDocumentEmbeddings

touhoku = TransformerDocumentEmbeddings('cl-tohoku/bert-base-japanese')
Topic_model3 = BERTopic(calculate_probabilities=True,verbose=True,nr_to
pics="auto",embedding_model=touhoku)
```

```
topics_03, probs_03 = Topic_model3.fit_transform(fuman03)
```

```
#BERT・埋め込みモデル・東北大学・3月：結果出力

Topic_model3.get_topic_info()

#BERT・埋め込みモデル・東北大学・4月

Topic_model = BERTopic(calculate_probabilities=True,verbose=True,nr_top
ics="auto",embedding_model=touhoku)
topics_04, probs_04 = Topic_model.fit_transform(fuman04)

#BERT・埋め込みモデル・東北大学・4月：結果出力

Topic_model.get_topic_info()

#BERT・埋め込みモデル・東北大学・5月

Topic_model5 = BERTopic(calculate_probabilities=True,verbose=True,nr_to
pics="auto",embedding_model=touhoku)
topics_05, probs_05 = Topic_model5.fit_transform(fuman05)

#BERT・埋め込みモデル・東北大学・5月：結果出力

Topic_model5.get_topic_info()

#BERT・埋め込みモデル・東北大学・6月

Topic_model6 = BERTopic(calculate_probabilities=True,verbose=True,nr_to
pics="auto",embedding_model=touhoku)
topics_06, probs_06 = Topic_model6.fit_transform(fuman06)

#BERT・埋め込みモデル・東北大学・6月：結果出力

Topic_model6.get_topic_info()
```

3. BERT による感情分析（ポジネガ判定）のプログラム

```
!apt install mecab libmecab-dev mecab-ipadic-utf8
!pip install mecab-python3

!apt install git make curl xz-utils file
!git clone --depth 1 https://github.com/neologd/mecab-ipadic-
neologd.git
!echo yes | mecab-ipadic-neologd/bin/install-mecab-ipadic-neologd -n -a
!ln -s /etc/mecabrc /usr/local/etc/mecabrc
```

```
!pip install -q transformers
!pip install fugashi
!pip install ipadic
```

```
from transformers import AutoTokenizer, AutoModelForSequenceClassificat
ion
from transformers import pipeline

tokenizer = AutoTokenizer.from_pretrained("daigo/bert-base-japanese-
sentiment")

model = AutoModelForSequenceClassification.from_pretrained("daigo/bert-
base-japanese-sentiment")
```

```
import pandas as pd #データ分析
import numpy as np #科学計算

data = pd.read_csv('/content/drive/My Drive/Fuman_Data/fuman_test03_BER
T.csv',encoding = 'UTF-8')
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
filename = input('file path input->')

with open(filename) as f:
  sentence = f.readlines()
  #print(doc)

sentiment_analyzer = pipeline("sentiment-analysis",model="daigo/bert-
base-japanese-sentiment",tokenizer="daigo/bert-base-japanese-
sentiment")

result = list(map(sentiment_analyzer, sentence))

for i, res in enumerate(result):
  print(sentence[i], result[i])
```