新型コロナウイルス感染症関連ツイートのトピックと感情の分析

小島章太郎[1]　内山清子[1]
[1]湘南工科大学 工学部 コンピュータ応用学科
18A6050@sit.shonan-it.ac.jp　uchiyama@sc.shonan-it.ac.jp

概要

本研究ではソーシャルネットワーキングサービス「Twitter」の投稿を分析し、新型コロナウイルス感染症に関するトピックと感情の分析を目的とする。その上で、日本語のツイートデータを対象とした、感情値の推移の分析とトピックモデルによる単語の分析とトピック割合推移の調査、深層学習による感情の分類モデルの構築を実施した。

1. はじめに

近年ではソーシャルネットワーキングサービス(SNS)に様々な文章が投稿されている。新型コロナウイルス感染症(COVID-19)は世界的に流行が拡大しており、世間に継続的な影響が発生している感染症である。COVID-19 に関する SNS 上の投稿を分析することで感染症に対する世間の意識変化や流行傾向などを分析できるのではないかと考えた。

本研究では SNS サービス「Twitter」の投稿を分析することで、COVID-19 に関する世間の感情や興味の変化について分析することを目標とした。感情分析については COVID-19 流行下において、日本の Twitter に対して 1 年以上にわたり適用している例が見られなかったことから、より大規模な感情分析を実現できるのではないかと考えた。

2. 関連研究

本研究では SNS 上の投稿の分析を行うため、インターネット上の投稿のトピック分析に関する先行研究を 2.1 節に記載する。また、COVID-19 に関する Twitter の分析の先行研究について 2.2 節に記載する。

2.1. インターネット投稿のトピック分析

時系列に沿って変化するインターネット上の投稿に対する分析として、佐藤らはレビューサイトの分析を行い、月ごとなど一定期間の TF-IDF 値を計算して季節を表す特徴語の抽出を実現している [1]。

大規模なイベントに対するトピックの分析について、北田らは東日本大震災を題材にトピックモデルの一種である LDA を用いて分析を行っている [2]。

ツイートに対する感情分析としては中村による「感情表現辞典」 [3]をもとに 10 軸に分類する手法が一般的となっている。山本らはこれを Twitter の投稿に適用し、感情分析を行っている [4]。同研究では Twitter 特有の単語表現や顔文字等の表現が感情表現辞典には含まれていないこと、ツイートという短文に対して 10 軸の感情軸を適用するのは過多であり、現実的ではないとされている。

2.2. COVID-19 に関する Twitter 分析

Lisa Singh らは 2020 年 1 月から 3 月にかけて世界各国の Twitter 上の COVID-19 に関する投稿を分析している [5]。一部の国については COVID-19 の症例件数とツイートの件数に一定の相関関係が見られたことから、SNS 上の投稿を分析することで感染拡大状況の推定・予測ができる可能性が示されている。

井原らは2020年1年間の日本語のツイートの分析を行い、感染拡大状況とツイート数の関係の精査、感情単語の時系列変化の検証等を行っている [6]。

機械学習を用いて COVID-19 関連ツイートの感情を分析する試みは英語圏では積極的に行われており Rustam らはツイートのネガティブ・ポジティブの二値分類を複数の機械学習モデルを用いて行い、精度を比較している [7]。

本研究では、新たに 2020 年 1 月から 2021 年 6 月にかけての日本語のツイートデータを用いて検証を行うことで長期間の分析を実現する。合わせてトピック分析を実施し、単語のトピック別の分類とトピックの出現推移の調査を実施する。また、先行事例では深層学習を用いて日本語の COVID-19 関連のツイートの感情分析を行っている例が見られなかったため、本研究では新たに日本語の COVID-19 関連ツイートに対する感情分析モデルの構築と比較を行う。

3. 使用データについて

本研究では、Twitter Academic Research API[i]を用いてツイートの取得を行った。検索キーワードとして「新型コロナウイルス」を使用し、2020 年 1 月 1 日から 2021 年 6 月 30 日を対象に、キーワードが含まれるツイート全件の取得を行った。取得したツイートは合計 6,549,683 件となった。このうち、メンション(@ツイート)については分析の対象外としているため、分析対象の件数は約 620 万件となった。

ツイート本文を処理する際、単語への分割が必要となるため形態素解析エンジンを用いて分かち書きを実施した。形態素解析エンジンは MeCab [8]を使用し、新語・造語に強みを持つ Neologd 辞書 [9]をシステム辞書として用いることで、Twitter 本文の分かち書きを実現した。解析にあたり、下処理として英数字の半角への統一、記号の除去、数字の除去等を行ったほか、Twitter 特有の処理として、エスケープ記号(&等)の復元、リンク・@メンション・ハッシュタグ記号の除去を実施している。

また、厚生労働省がオープンデータとして公開している COVID-19 新規陽性者数の推移データを比較対象として使用している [10]。はじめに、感染者数の増減をもとに期間分けを定義した。7 日間平均が増加し始めてから減少傾向に転じるまでの期間を増加期、減少傾向に入ってから減少が止まるまでの期間を減少期、2 週間以上感染者数が大きく変化しなかった期間を停滞期としている。期間分けの推移と感染者数の推移の対応を図 3.1 に示す。

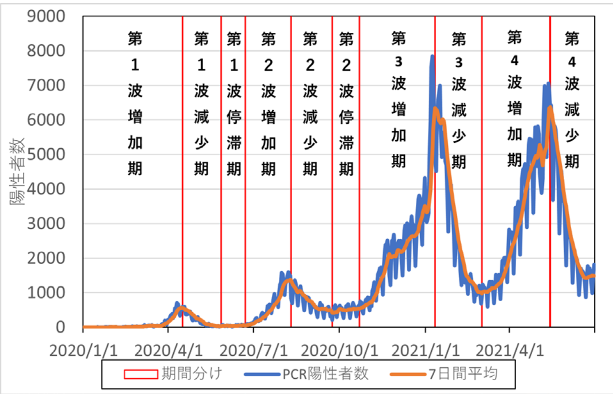

図 3.1 期間分けと感染者数の対応

次に、ツイート数と感染者数の比較と、単語の出現確率の感染者数との比較を試みた。ツイート数と感染者数の比較は、直接的な比例関係は見られなかった(図 3.2)。単語ごとの比較については、1 日毎にツイート本文を結合し 1 つの文書とし、その文書集合に対して TF-IDF 値の計算を行った。

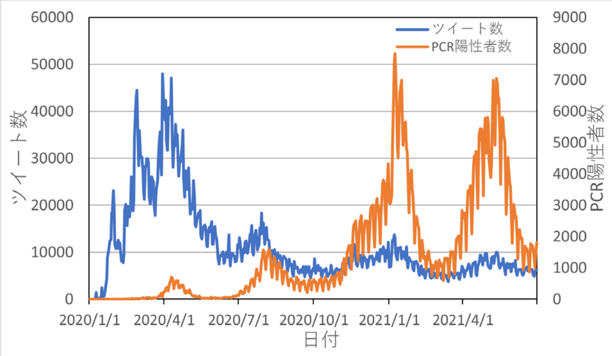

図 3.2 ツイート数と感染者数の関係

感染拡大初期と重なる第 1 波増加期については、「ステイホーム」や「ソーシャルディスタンス」、「テイクアウト」などいくつかの単語が感染者数との相関係数 0.8 以上の正の相関を示していた。第 1 波減少期では、「不要不急の外出」「臨時休業」「外出自粛」などが 0.8 以上の正の相関を示した他、「再開」などの単語が増加傾向となる負の相関を示していた。第 2 波以降では有意な相関を示す単語は確認されなかった。

4. 提案手法

本研究では、COVID-19 の状況下での Twitter の感情を分析するため、はじめに感情値の推移の分析を行った。続いてトピックの分析を行い、トピックの割合の推移について調査を行った。最後に深層学習を用いて感情分析を実施し、感情の分類モデルの作成を行った。

4.1. 日別平均感情値の分析

取得したツイートの本文に対し、感情分析ツール「MLAsk」 [11]を用いて 10 軸の感情分析を実施した。MLAsk は入力したテキストの形態素解析を行い、感情辞書と比較することで、感情表現辞典に基づきテキストを「好」「安」「哀」「厭」「怖」「怒」「恥」「昂」「驚」「喜」の 10 軸に分類することが可能である。また、単語の感情表現辞典との比較や、記号や顔文字の解析を行うことで、ネガティブ・ポ

i https://developer.twitter.com/en/products/twitter-api/academic-research

ジティブ、覚醒・非覚醒の 2 軸にテキストを分類することが可能である。

収集したツイートデータに対し、10 軸の感情分析を実施した。なお、MLAsk では感情表現辞典を元に定義された単語群を元に感情の分類を行うため、テキスト中の単語がいずれも感情表現辞典に記載されていない場合、無感情という判定となる。今回分析対象とした約 620 万件のツイート本文データのうち、ML-Ask にて感情が含まれていると判定されたツイートは約 115 万件(約 18.5%)であった。感情軸別の推移をプロットしたものを図 4.1 に示す。

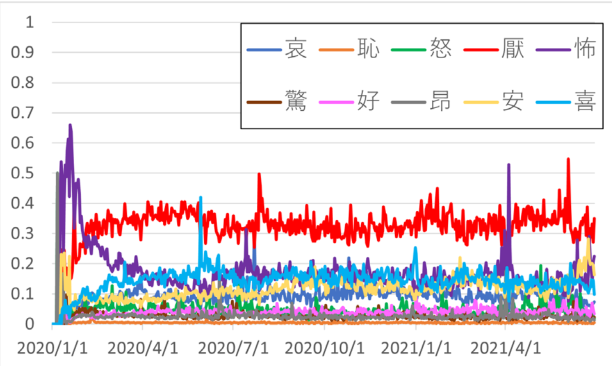

図 4.1 感情軸別の推移グラフ

感染拡大初期は「怖」という感情が半数以上を占めていたことが明らかとなった。全体的には 2020 年 2 月以降は「厭」という感情が 3 割～4 割で最多となっている。

第 1 回緊急事態宣言解除のタイミングや年末には「喜」という感情が瞬間的に増加するなど、特定の時点にいくつかの特徴が見られた。また、「恥」「昂」「好」「怒」などの感情はほぼ 1 割を下回った状態が続いており、Twitter の投稿ではあまり感情軸として確認することができなかった。

4. 2. LDA を用いたトピック分析と比較

続いて、トピックモデルの一種である潜在的ディレクリ配分法(Latent Dirichlet Allocation : LDA)を用いてトピック分析を行った。LDA では一つの文書に複数のトピックが存在していることが前提となっていることから、大規模で複数のトピックが混在している可能性が高い本研究では適していると考え、LDA を採用した。

はじめにトピック数の決定を行う必要があるため、分析トピック数を 2～21 まで 20 回変化させてトピックの分析を行い、指標である Perplexity と Coherence の値の確認を行った(図 4.2)。

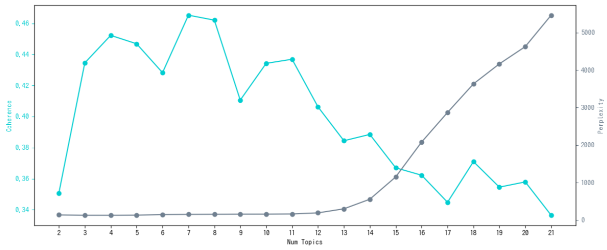

図 4.2 LDA トピック数別評価結果の推移

その結果、Perplexity が安定し Coherence の値が最も高かった 7 が適切なトピック数と判断し、トピック数として採用した。

実際にトピック数を 7 に設定し分類されたトピックごとの単語の例を図 4.3 に示す。

図 4.3 LDA トピック分類結果

LDA ではトピック数を設定すれば自動でクラスタリングが行われるものの、その結果については人間が手動で意味付けを推定する必要がある。今回の推定結果を表 4.1 に示す。

表 4.1 トピック項目の推定結果

トピック ID	トピック内容
Topic 0	感染者数への直接的な言及
Topic 1	社会的影響に関する言及
Topic 2	感染予防策などの言及
Topic 3	行政の動きに対する言及
Topic 4	検査・対策手法に関する言及
Topic 5	感染者数の報道への言及
Topic 6	行政アカウント等からの発信

各トピックの割合推移を分析した結果、感染拡大初期の 2020 年 1 月ごろは、緊急事態宣言の発令準備等に伴う Topic3:行政の動きに関する言及や、PCR 検査やワクチンなど Topic4:検査・対策手法に関する言及、マスクなど Topic2:感染予防策などの言及が多数を占めていた。その後、2020 年 4 月頃からは

Topic0:感染者数に関する言及が 25%程度を占め最多となっている。次いで Topic6:行政アカウント等からの発信が多かったものの、第 2 波が到来した 2020 年 7 月～8 月ごろは Topic5:感染者数の報道への言及が、年末には Topic1:社会的影響に関する言及が一時的に増加していた。なお、実際のトピック割合推移のグラフについて付録に示している。

4.3. 深層学習を用いた感情分析

MLAsk で解析することができるネガティブ・ポジティブの判定を教師データとして、深層学習を用いてツイートの感情属性の予測を行った。MLAsk ではラベル付けを行えなかった「ニュートラル」のツイートに関しては学習対象から除外している。ラベル付けを行うことができたネガティブ・ポジティブ各属性のツイート 2 万件ずつ、合計 4 万件のツイートを使用し、学習データと検証データを 7:3 に分割して検証を行った。予測に際しては日本語処理に対応した言語処理モデルの BERT など複数のモデルの比較を行い、精度の差を検証した。学習はすべて Tensorflow 2.5.0[ii]を用いて行った。実際に検証を行った汎用モデルとその正解率の一覧を表 4.2 に、BERT モデルとその正解率の一覧を表 4.3 に示す。Long Short Term Memory (LSTM)、Simple Recurrent Neural Network (RNN)、Global Average Pooling 1D については Tensorflow 標準のモデルを使用し、BERT については東北大学[12]、京都大学[13]、ホットリンク社[14]が配布しているモデルを使用し、ファインチューニングを実施した。

表 4.2 感情分析タスクのモデルと正解率一覧

モデル名称	エポック数	正解率
LSTM (64 層)	**6**	**94.23%**
SimpleRNN (64 層)	6	93.48%
Global Average Pooling 1D	20	93.17%

表 4.3 検証を行った BERT モデルと正解率一覧

モデル名称	エポック数	正解率
BERT(ホットリンク社 SNS コーパス)	**3**	**97.64%**
BERT(京都大 WWM 版)	3	95.18%
BERT(京都大 通常版)	3	94.98%
BERT(東北大 WWM 版)	3	93.27%

Tensorflow 標準の汎用モデルの中では LSTM が最も高い正解率を示し、続いて SimpleRNN、Global Average Pooling 1D の順となった。

BERT の中で最も正解率が高かったのはホットリンク社の SNS コーパスを用いた BERT モデルで、唯一 97%台となっている。その後に京都大のモデル、東北大のモデルが続いている。全体的に汎用モデルと比較し、BERT は高い正解率を示すことができた。

5. 考察

本研究では、ツイートデータに対する感情分析とトピック分析をもとに COVID-19 に関するツイートの分析を実施した。

日別感情値の推移については、緊急事態宣言の発令等に関する時期に感情の変化を見ることができた。一部の感情軸についてはツイートデータ上ではあまり確認できなかったことから、ツイートに対する感情軸については最適化の余地があると考えられる。

トピック分析については、LDA を用いることである程度適切なトピック分析を実施することができたと考えられる。2020 年 4 月頃から現在に至るまで、感染者数への言及が最も多くトピックとして現れており、感染者数が感染症の一つの指標として世間の関心が高い状態が続いていると考えられる。

日本語ツイートに対する深層学習を用いた感情分析については SNS コーパスによる BERT モデルがもっとも高い正解率を示した。同 SNS コーパスは本研究と同じくツイートを元に構築されており、約 8600 万ツイートをもとに構築された大規模なコーパスであることが理由と考えられる。また、どの BERT モデルも 90%を超える正解率が確認された。今回はネガティブ・ポジティブの二値分類だったため、比較的高い結果を出すことができたことが考えられる。

6. おわりに

本研究では、COVID-19 関連のツイートデータを収集し、その分析と感染者数との比較を行った結果、感染者数との有意義な相関関係は見られないことを示した。また、感情分析を実施し時期ごとの感情特性を明らかにした他、深層学習を用いてツイートデータの感情を高精度で分類することを示した。

今後は他の文脈におけるツイートデータの活用や、他のインターネット上のテキストデータへの応用を検討していきたい。

[ii] https://github.com/tensorflow/tensorflow/releases/tag/v2.5.0

参考文献

1. **佐藤裕次郎, 山西良典 , 西原陽子.** 宿泊施設のレビューの時系列分析による季節を表す特徴語の抽出. 人工知能学会, 2019. SIG-AM-21-19.
2. **北田剛士, ほか.** 東日本大震災時のツイートのトピック系列の可視化と分析. 2015 年度 人工知能学会全国大会 論文集, 2015. 2B3-NFC-02a-1.
3. **中村明.** 感情表現辞典. 東京堂出版, 1993. 978-4490103397.
4. **山本湧輝, 熊本忠彦 , 灘本明代.** ツイートの感情の関係に基づく Twitter 感情軸の決定. 第 7 回データ工学と情報マネジメントに関するフォーラム 論文集, 2015.
5. **Lisa Singh, ほか.** A first look at COVID-19 information and misinformation sharing on Twitter. National Institutes of Health, 2020.
6. **井原史渡, 岸本大輝 , 栗原聡.** 新型コロナウイルスに伴う Twitter の分析と感染状況との関連. 第 35 回人工知能学会全国大会論文集, 2021.
7. **Furqan Rustam, ほか.** A performance comparison of supervised. PLOS ONE, 2021.
8. **MeCab: Yet Another Part-of-Speech and Morphological Analyzer.** (オンライン) (引用日: 2021 年 5 月 21 日.) https://taku910.github.io/mecab/.
9. **佐藤敏紀.** Neologism dictionary based on the language resources on the Web for Mecab. (オンライン) 2015 年. (引用日: 2021 年 3 月 9 日.) https://github.com/neologd/mecab-ipadic-neologd.
10. 新型コロナウイルス感染症について ＞ オープンデータ ｜厚生労働省. (オンライン) 厚生労働省. (引用日: 2021 年 8 月 27 日.) https://www.mhlw.go.jp/stf/covid-19/open-data.html.
11. **PtaszynskiMichal.** ML-Ask: Affect Analysis System. Michal Ptaszynski / Research. (オンライン) (引用日: 2021 年 7 月 2 日.) http://arakilab.media.eng.hokudai.ac.jp/~ptaszynski/repository/mlask.htm.
12. **東北大学 乾研究室.** cl-tohoku/bert-base-japanese. Hugging Face. (オンライン) (引用日: 2021 年 10 月 29 日.) https://huggingface.co/cl-tohoku/bert-base-japanese.
13. BERT 日本語 Pretrained モデル **. 京都大学 黒橋・褚・村脇研究室.** (オンライン) 2020 年 11 月 21 日. (引用日: 2021 年 10 月 29 日.) https://nlp.ist.i.kyoto-u.ac.jp/?ku_bert_japanese.
14. **大規模日本語 SNS コーパスによる文分散表現モデルの公開 : hottoSNS-BERT の配布.** 株式会社ホットリンク公式ブログ. (オンライン) 株式会社ホットリンク, 2019 年 3 月 11 日. (引用日: 2021 年 10 月 29 日.) https://www.hottolink.co.jp/blog/20190311_101674/.

A 付録

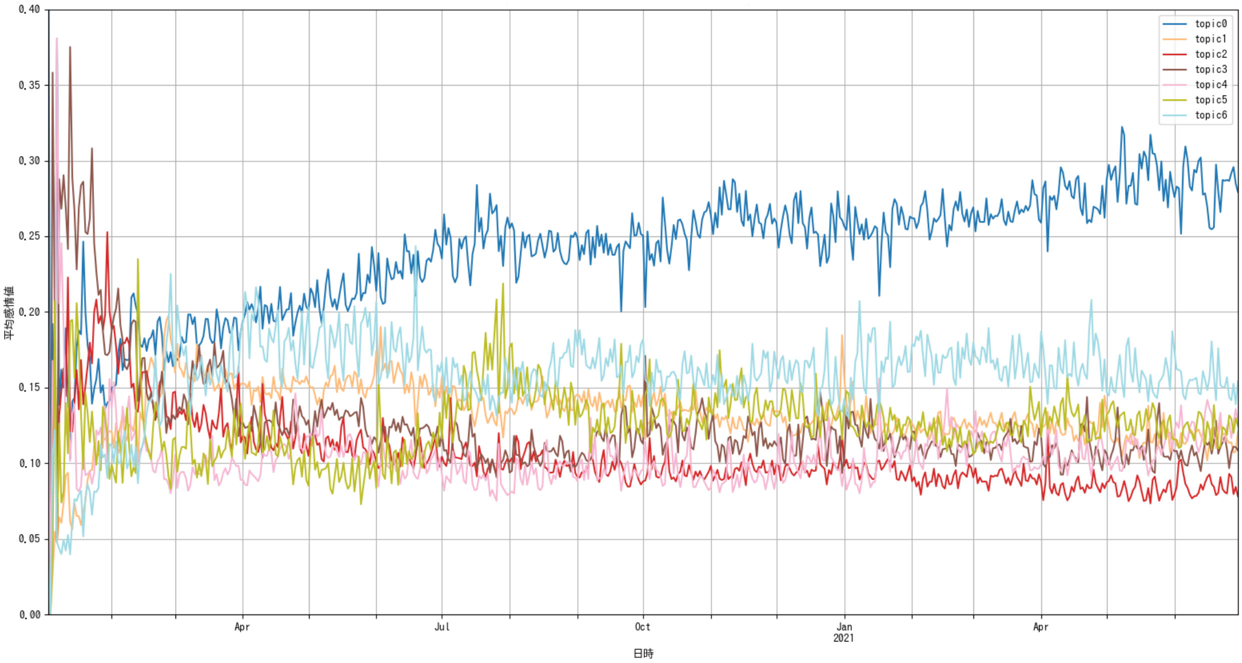

図 **A1** 新型コロナウイルス **LDA** トピック推移