E-032

Yahoo!知恵袋テキストデータの感情分析

Sentiment analysis of Yahoo! Chiebukuro text data

森藤 名都† 吉田 稔† 松本 和幸†
Natsu Morifuji Minoru Yoshida Kazuyuki Matsumoto

1. はじめに

近年、SNS における意見交換はますます活発になり、人々の感情や社会に対する反応がオンライン上に大量に蓄積されている。中でも、Yahoo!知恵袋[1]は日本最大級の Q&A プラットフォームであり、約 2600 万人の登録ユーザーと 1 億 6000 万件以上の質問を有している。ユーザーは単に情報を求めるだけでなく、自身の感情や意見を表明する場としても利用している。これらの膨大なテキストデータを分析することは、社会のトレンドや問題に対する人々の感情的な反応を理解する上で非常に有益である。 本研究では、Yahoo!知恵袋の質問テキストデータに対して感情分析を行い、ユーザーの感情傾向を明らかにすることを目的とする。具体的には、テキストを Negative, Positive, Neutral の 3 つの感情に分類し、主要な質問カテゴリ（政治、経済、スポーツ）ごとの感情分布の特徴を明らかにする。さらに、月ごとの感情割合の変動を観測し、その変動が特定の社会的な出来事や話題とどのように関連しているかを、頻出単語の分析を通じて考察する。

2. 関連研究

テキストデータからの感情分析に関する研究は数多く行われている。例えば、水門ら[2]は、コロナ禍における「COVID」や「コロナ」といったキーワードを含む Twitter 投稿を収集し、感情分析を行った。そして、その結果得られた感情値と GPS 記録に基づく人出データとを比較し、ネガティブな感情の高まりが約 1 ヶ月後の人出減少に繋がることを示した。この研究は SNS データが経済的・社会的活動の変化を予測する上で有効である可能性を示唆している。

水門らの研究が特定のトピック（COVID-19）に限定し、辞書ベースの感情分析手法を用いていたのに対し、本研究は Yahoo!知恵袋という異なるプラットフォームを対象とし、より広範なトピック（政治、経済、スポーツ）を扱う。また、感情分類には文脈理解に優れた深層学習モデルである BERT をファインチューニングして使用し、よりニュアンスを含んだ感情分析を目指す点で新規性がある。

3. 提案手法

本章では、本研究における感情分析の処理フローの概要を示す。データ収集、前処理、代表文抽出、BERT を用いた感情分類、月次感情割合の算出と可視化、そして変動要因の考察というステップで分析を進めた。

3.1 データ収集と前処理

分析対象として、国立情報学研究所から配布されている Yahoo!知恵袋データセット[3]のうち、2018 年 4 月から 2021 年 3 月までの Yahoo!知恵袋に投稿された質問テキストデータを使用した。カテゴリは「政治」（約 5.6 万件）、「経済」（約 4.2 万件）、「スポーツ」（約 7.5 万件）の 3 つに焦点を当てた。収集したテキストデータに対して、句読点の正規化、URL や不要な記号の除去、Janome[4]を用いた形態素解析によるトークン化、日付情報のパースといった標準的な前処理を施した。

3.2 代表文抽出

Q&A サイトの質問文はしばしば長文であり、複数のトピックや感情が混在することがある。そこで、投稿者の主たる感情が表れている可能性の高い文を「代表文」として抽出した。具体的には、各文に対して TF-IDF スコアと日本語評価極性辞書（名詞編）[5]に基づく感情スコアを算出し、これらの合計値が最も高い文を代表文とした。この処理により、分析対象をテキスト中の最も感情的に重要な部分に絞り込み、ノイズを低減することを目的とした。

3.3 BERT を用いた感情分類モデル

感情分類には、東北大学が公開している事前学習済み日本語 BERT モデル「cl-tohoku/bert-base-japanese-whole-word-masking」[6] を利用した。このモデルを、TIS 株式会社が公開している日本語ネガポジ判定データセット「chABSA-dataset」[7] を用いてファインチューニングし、テキストを Positive, Neutral, Negative の 3 つの感情ラベルに分類するモデルを構築した。構築したモデルのテストデータに対する性能は、Accuracy 90%、F1 スコア 0.79 であった。

3.4 感情変動要因の分析

Yahoo!知恵袋の各カテゴリについて、月ごとの感情割合（Positive と Negative の総数に対する Negative の割合）の変化率を算出し、その変化率が $\mu \pm 2\sigma$ （ここで、μ は平均、σ は標準偏差である）の範囲外にある月を特異点として識別した。これらの特異点を含む計 5 ヶ月間（該当月とその前後 2 ヶ月）の頻出単語を調査し、感情変化の要因を考察した。

4. 実験と考察

4.1 カテゴリ別感情分布

構築した感情分類モデルを用いて Yahoo!知恵袋の各カテゴリの質問テキスト（代表文）を分類した結果を表 1 に示す。 「政治」および「経済」カテゴリでは、Neutral（政治: 46.0%, 経済: 43.0%）および Negative（政治: 34.2%, 経済: 31.5%）と分類された投稿の割合が高かった。一方、「スポーツ」カテゴリでは、Neutral（49.6%）に加えて Positive（33.1%）と分類された投稿の割合が他の 2 カテゴリに比べて高かった。この結果から、政治や経済に関するトピックでは問題提起や懸念、批判的な意見が、スポーツに関するトピックでは応援や称賛、期待といった肯定的な感情が表出されやすい傾向が示唆される。

† 徳島大学大学院 Tokushima University

表 1 各カテゴリごとの感情分類結果

カテゴリ	Negative (%)	Neutral (%)	Positive (%)
政治	14,316(34.2)	19,286(46.0)	8,307(19.8)
経済	17,779(31.5)	24,259(43.0)	14,389(25.5)
スポーツ	12,982(17.2)	37,110(49.6)	24,766(33.1)

4.2 感情割合の変動と要因

月次の感情割合の変動を分析し、特に変化が大きかった時期とその要因として考えられる事象を以下の表にまとめる。「政治」カテゴリでは 2018 年 7 月にネガティブ感情の割合が顕著に増加した。この時期の頻出語には「死刑」「麻原」「災害」などが含まれており（表 2 参照）、オウム真理教元幹部の死刑執行や西日本豪雨といった社会的に影響の大きな出来事に対するユーザーの反応が反映されたものと推察される。

表 2 政治カテゴリの頻出単語

2018-6	2018-7	2018-8	2018-9	2018-10
人	人	日本	人	人
日本	日本	人	日本	日本
安倍	安倍	安倍	安倍	韓国
問題	死刑	問題	韓国	安倍
国	問題	韓国	問題	安田
韓国	国	国	日本人	問題
金	麻原	日本人	国	日本人
日本人	災害	アメリカ	石破	国
北朝鮮	執行	女性	自民党	国民
加計	自民党	自民党	税	消費

「経済」カテゴリでは 2020 年 2 月にネガティブ感情が増加した。頻出語には「税」「費」「会社」「所得」などが見られ（表 3 参照）、COVID-19 感染拡大初期における経済的な不安感が背景にある可能性が考えられる。

表 3 経済カテゴリの頻出単語

2019-12	2020-1	2020-2	2020-3	2020-4
円	万	円	万	万
万	円	万	円	円
保険	保険	保険	保険	保険
カード	費	申告	カード	金
年金	年金	年金	申告	申告
額	カード	カード	株	コロナ
人	年	確定	人	額
月	人	額	年金	カード
所得	会社	税	口座	所得
金額	申告	費	税	税

「スポーツ」カテゴリでは 2020 年 9 月にポジティブ感情が増加した。頻出語には「巨人」「優勝」「試合」「野球」などがあり（表 4 参照）、特定プロ野球チーム（読売ジャイアンツ）の快進撃と優勝への期待感が感情を高めた要因と推察される。

表 4 スポーツカテゴリの頻出単語

2020-8	2020-9	2020-10	2020-11	2020-12
選手	選手	選手	選手	選手
巨人	人	人	人	人
人	試合	巨人	日本	野球
試合	野球	野球	巨人	巨人
野球	巨人	日本	野球	投手
回	投手	試合	シリーズ	回
投手	回	プロ	プロ	プロ
筋	プロ	阪神	回	日本
監督	誰	優勝	戦	優勝
誰	優勝	好き	県	試合

これらの結果は、Yahoo!知恵袋上のユーザーの感情が、現実社会の出来事と連動して変動することを示唆している。ただし、経済トレンドのように、キーワードのみからの要因解釈が難しいケースも見られた。

5. おわりに

本研究では、Yahoo!知恵袋の質問テキストデータに対し、BERT ベースの深層学習モデルを用いた感情分析を行った。その結果、カテゴリごとの感情傾向の違いや、月次の感情割合の変動が実際の社会的な出来事と関連している可能性が示された。今後の課題として、以下の 3 点が挙げられる。第一に、感情ラベルの粒度を細分化したり、新たなスコアリング指標を導入したりするなど、感情分析手法自体の高度化である。第二に、質問だけでなく回答の感情も考慮し、Q&A コミュニティ全体の感情動態を捉えることである。第三に、質問文の曖昧な表現に対応し、単なる感情だけでなく投稿者の「意図」の分類へと分析を深化させることである。

謝辞

本研究は JSPS 科研費 JP24K15193 の助成を受けたものです。

参考文献

[1] Yahoo!知恵袋, https://chiebukuro.yahoo.co.jp/
[2] 水門善之, 田邉洋人, 和泉潔. コロナ禍における Twitter 情報を用いた感情値の計測と人出の関係, 2022 年度人工知能学会第 36 回全国大会発表論文集
[3] 国立情報学研究所情報学研究データリポジトリ「Yahoo!知恵袋データ（第 3 版）」, https://www.nii.ac.jp/dsc/idr/yahoo/chiebkr3/Y_chiebukuro.html
[4] 内田知子 (2023), Janome:Python による日本語形態素解析エンジン, https://github.com/mocobeta/janome
[5] 東北大学乾・岡崎研究室(2008-12) 日本語評価極性辞書（名詞編）, https://www.cl.ecei.tohoku.ac.jp/Open Resources-Japanese Sentiment Polarity Dictionary.html
[6] 日本語事前学習モデル cl-tohoku/bert-base-japanese-whole-word-masking, https://huggingface.co/cl-tohoku/bert-base-japanese-char-whole-word-masking
[7] TIS 株式会社ネガポジ判定データセット chABSA-dataset, https://github.com/chakki-works/chABSA-dataset