機械学習を用いた旅行活動ツイート判定精度の実験的比較

○藤大友都　服部峻　砂山渡（滋賀県立大学）　高原まどか （龍谷大学）

概要　旅行活動を SNS に投稿した経験があるユーザは多い．このような旅行活動データの利活用の一つとして可視化が考えられ，新たな旅行先の決定や旅行需要の発見に寄与すると考える．しかし，SNS に蓄積されるデータは日々膨大になり続けており，旅行以外にも多種多様な情報が玉石混交である．そこで本稿では，SNS から旅行活動データをノイズなく網羅的に収集するため，ChatGPT など既存の様々な機械学習を用いて，旅行活動か否かの判定精度の比較実験を行う．

キーワード: 機械学習，分類，ChatGPT，SNS 分析，観光

1　はじめに

近年，旅行活動を SNS に投稿するユーザが増えている．日本交通公社の調査 [1] によると，実施した国内旅行の計画にあたり行った情報収集の媒体として,「SNS やブログ，動画サイト」を選択した割合が，若い世代になるにつれて高く，Z 世代では 3.5 から 5 割を占めている．また，旅行先を選択する際，SNS の情報を重視すると回答した層は全体の 2 割，自分自身の SNS に投稿することを意識して旅行先を選択した層が全体の 1 割を占めており，旅行活動において SNS は必要不可欠な存在となっている．

このような旅行活動データの利活用の一つとして，可視化が考えられる [2]．可視化を行うことで，ある観光エリアにおける各スポット毎の旅行活動の特性が見えるようになり，観光エリアにおいて体験できる活動が明確になる．可視化を通して，新たな旅行先の決定や旅行需要の発見に繋がり，旅行活動が活発になるのではないかと考えた．

しかし，SNS に蓄積されるデータは日々膨大になり続けており，旅行以外にも多種多様な情報が含まれている．旅行活動データを収集する際，収集の仕方を工夫しなければ，旅行と関係のない情報が混ざり，可視化の妨げになることが考えられる．例えば，彦根の旅行活動を可視化する場合，彦根城や四番町スクエアの観光が想定される．しかし，彦根駅前での街頭演説や，住民の日常生活の投稿など，旅行活動ではないものまで可視化されると，旅行活動が正しく認識できなくなる．

そこで本稿では，SNS から旅行活動データをノイズなく，かつ網羅的に収集するため，ChatGPT など既存の様々な機械学習を用いて，旅行活動か否かの判定精度の比較実験を行う．本稿では SNS のうち Twitter に着目し，ツイート本文のテキストから旅行活動か否かを判定する．

本稿の以降の構成は以下の通りである．まず 2 章で関連研究について，3 章で実験方法について詳述する．次に，4 章で実験結果をまとめ，5 章で今後の研究課題について述べる．

2　関連研究

本章では，関連研究について述べる．

2.1　Twitter ユーザの居住地推定に関する研究

西村ら [3] は，ツイートに出現する単語情報から，ユーザの居住地推定を行った．居住地が判明しているユーザのツイートから，地域ごとの特徴語リストを作成しておき，特徴語リストをもとに居住地推定を行うことで，精度が向上した．しかし，正しく居住地を推定できたユーザの正解率は平均して 0.43 であり，まだまだ改善が必要である．

2.2　ツイートを利用した災害情報採集手法に関する研究

藤田ら [4] は，Twitter 内に散らばる災害情報を代表語 1 単語（例えば，台風，地震など）で検索し，その後，自己教師あり分類器で災害情報を分類する手法を検討した．自己教師あり分類器を作製することで，代表語がツイートに含まれているか否かで分離するよりも，精度よく分類できることが確認できた．しかし，分類結果を可視化するにあたり，どの程度精度が高ければ有用なシステムとなるのか，明らかにされていない．

3　実験方法

我々はこれまでに，観光地における旅行活動ツイートの収集手法として,「地名」,「位置情報」,「特徴語」の 3 手法を組み合わせたフィルタリング手法により，旅行活動ツイートの網羅的な収集を検討した [2]．その結果，観光地「彦根」を旅行するツイートを 94 件収集した．しかし，ルールベースによる分類は精度が低いことが分かった．

そこで本稿では，既存の様々な機械学習を用いて，収集したツイートを「彦根の旅行活動」と「旅行活動でない」に分類する実験を行った．入力するデータは,「彦根の旅行活動ツイート 94 件」と「一般的な日本語ツイート 100 件」である．194 件のツイートのうち，学習データ: テストデータを 8:2 に分割して実験を行った．プログラミング言語は Python を使用した．

3.1　SVM

SVM (Support Vector Machine) は，教師あり機械学習モデルの一種で，異なるクラス間に境目となる直線や曲線を引くことにより，分類を実現する．主に分類のタスクに用いられる．入力にはベクトルが必要であり，本稿ではツイート（文書）のベクトル化手法として「TFIDF」と「Doc2Vec」の 2 種類を用いた．ライブラリは，TFIDF によるベクトル化と SVM の分類器作成には scikit-learn を，Doc2Vec によるベクトル化は gensim を用いた．尚，ハイパーパラメータはデフォルト値である．

3.2　fastText

fastText は，Word2Vec を基に作られた機械学習ライブラリである．fastText の入力フォーマットに合わせるため，形態素解析を先に行い，形態素間をスペー

スで区切る処理を行ってから入力した．尚，ハイパーパラメータは全てデフォルト値である．

3.3 ナイーブベイズ

ナイーブベイズは各特徴量が独立であると仮定し，出力の確率を計算するモデルである．各ツイートは，TFIDF によるベクトル化処理を行ってから入力した．分類器の作成は機械学習ライブラリ scikit-learn の GausianNB を用いて，ハイパーパラメータは全てデフォルト値とした．

3.4 BERT・RoBERTa

BERT (Bidirectional Encoder Representations from Transformers) は，文章を文頭と末尾の双方向から学習を行うことで，従来の機械学習から精度を向上したモデルである．事前学習モデルとして，cl-tohoku/bert-base-Japanese を用い，トークナイザには BertJapaneseTokenizer を使用した．学習データを使って転移学習を行い，テストデータで検証した．

RoBERTa (Robustly optimized BERT approach) は，従来の BERT を改良して，事前学習用データセットを大幅に増やし，事前学習の回数を増やした．また，バッチサイズを大きくし，長い文章を投入することで，BERT の精度を大きく上回っている．事前学習モデルは，rinna/japanese-roberta-base を，トークナイザには T5Tokenizer を使用した．学習データを使って転移学習を行い，テストデータで検証した．

3.5 ChatGPT

ChatGPT は，2022 年 11 月に OpenAI 社がリリースした大規模言語モデルの一種である．ユーザが入力した質問に対し，人間のような自然な対話形式で回答が出力されるシステムである．これを利用して，分類問題を解く．OpenAI の API を使用して，ChatGPT の gpt-3.5-turbo を用いた旅行活動ツイートの分類を行った．プロンプトは，旅行活動の定義のみを入力する簡易なものと，定義に加えて旅行活動ツイートの判断材料をいくつか追加した詳細なものの 2 種類で実験を行った．

まず，簡易プロンプトを以下に示す．

Q1: 今から，彦根の旅行活動ツイートか否かの分類を行っていただきます．旅行活動ツイートの定義は以下のとおりです．・観光地に移動し，観光地内の各スポットで何らかの活動を行うことを旅行とする

以下のように出力してください．

（ツイート本文）
判定: 彦根の旅行活動である or 旅行活動だが彦根ではない or 旅行活動ではない理由:（上記のように判断した理由）

例入力: 今日は彦根城に行きました．ひこにゃんが来てくれました．出力: 今日は彦根城に行きました．ひこにゃんが来てくれました．判定: 彦根の旅行活動である理由: ツイートには彦根城への訪問が明記されており，当日の行動であることが示されています．また，ひこにゃん（彦根城キャラクター）が登場したことから，彦根城での観光活動が行われたことが明確です．

準備が整い次第，分類してほしいツイートを入力します．

A1: 了解しました．準備が整い次第，分類してほしいツイートを入力してください．どのツイートでもお手伝いできます．

Q2: （ツイート 1 件）

次に，詳細プロンプトを以下に示す．

Q1: 今から，彦根の旅行活動ツイートか否かの分類を行っていただきます．旅行活動ツイートの定義は以下のとおりです．
・観光地に移動し，観光地内の各スポットで何らかの活動を行うことを旅行とする詳細は以下のとおりです．

① 観光地内スポットもしくは観光地名と，その場所での行動が含まれていること．行動が明記されていない場合でも，文脈から行動していると判断できる場合は旅行活動ツイートと判定する．

② 当日の投稿であること（”今日は◎◎行った!”は旅行活動だが，”先週の△△楽しかった！”や”来週△△行きたい!”は旅行活動ではない）日付情報が明記されていない場合は当日の行動とする．

③ 本人の活動であること（■■が今日◎◎行ったらしい．は旅行活動ではない）

④ I’m in ◎◎　の投稿は，その投稿が題材とする観光地に関する場所ならば旅行活動ツイート

⑤ 題材とする観光地以外にも複数の観光地と行動があった場合，題材とする観光地の旅行活動として採

⑥ 題材とする観光地に住んでいると考えられるユーザのツイートは旅行活動ツイートとしない（旅行活動ではないため）

以下のように出力してください．

（ツイート本文）
判定: 彦根の旅行活動である or 旅行活動だが彦根ではない or 旅行活動ではない理由:（上記のように判断した理由）

例入力: 今日は彦根城に行きました．ひこにゃんが来てくれました．出力: 今日は彦根城に行きました．ひこにゃんが来てくれました．判定: 彦根の旅行活動である理由: ツイートには彦根城への訪問が明記されており，当日の行動であることが示されています．また，ひこにゃん（彦根城のキャラクター）が登場したことから，彦根城での観光活動が行われたことが明確です．

準備が整い次第，分類してほしいツイートを入力します．

A1: 準備が整いましたので，分類してほしいツイートを入力してください．それに対して，指定された基準に基づいて判定と理由を提供します．ツイートをお待ちしています．

Table 1: Classification results of each machine learning.

Machine Learning Models	Accuracy	Recall of Travel Activities	Precision of Travel Activities	Recall of Japanese Tweets	Precision of Japanese Tweets
SVM (TFIDF)	0.749	0.520	0.970	0.986	0.679
SVM (Doc2Vec)	0.567	0.507	0.547	0.623	0.594
fastText	0.685	0.367	0.955	0.984	0.626
Naive Bayes (TFIDF)	0.731	0.853	0.663	0.636	0.836
BERT	0.493	0.440	0.469	0.542	0.512
RoBERTa	0.492	0.490	0.515	0.492	0.466
ChatGPT (Simple prompt)	0.655	0.309	0.935	0.980	0.601
ChatGPT (Detailed prompt)	0.747	0.479	1.000	1.000	0.671

Q2: （ツイート 1 件）

プロンプトは，Q1 (user)，A1 (assistant)，Q2 (user) の形式で入力した．回答の判定は,「彦根の旅行活動である」と出力した場合のみ，旅行活動と予測したものとし,「旅行活動だが彦根ではない」,「旅行活動ではない」，もしくは意図しない判定結果（例：彦根の旅行活動ではない）を出力した場合は，正しく判定できなかったとし，全て「彦根の旅行活動ではない」とした．

4　実験結果

実験結果を Table 1 に示す．

最終的に必要なのは彦根の旅行活動ツイートであり，旅行と関係のないツイートは極力拾わないようにしたい．このため，旅行活動ツイートの適合率・再現率を重視する．SVM は TFIDF でベクトル化を行うことで，適合率が高くなった．しかし，再現率が 0.520 と低いため，旅行活動ツイートの精度が高いが網羅的ではない．fastText は適合率が高いが，その分再現率が落ちている．ナイーブベイズは逆に，再現率が一番高い．網羅性を高めたい場合はナイーブベイズが有効であると考えられる．BERT・RoBERTa を比較すると，両者とも旅行活動ツイートの比較タスクでは精度が大きく変化しないことが分かった．最後に，ChatGPT は簡易なプロンプトよりも詳細にプロンプトを入力する方が精度が上がることが分かった．しかし，どちらのプロンプトの場合も，194 件中 20 件程度は判定結果が意図しない出力を返したため，再現率が低下した．

5　今後の研究計画

今後の研究計画として,「旅行活動データの可視化」「データ収集」「機械学習モデルの組み合わせ」を考えている．

5.1　旅行活動データの可視化ついて

今後，得られた旅行活動ツイートを用いて可視化することを検討している．我々は以前，収集した旅行活動ツイートを人工社会シミュレーションである artisoc[5] にエージェント化し，観光地の旅行活動の特徴が理解できるような可視化の手法について検討した [2]．しかしながら，テキストデータと artisoc の可視化結果を比較する評価実験を行ったところ，旅行需要の発見，旅行先決定支援，魅力的な可視化の観点ではテキストデータの方が評価が高いという結果となった．この結果を受けて，今後は実際の地図に重ねて表示することで可視化の視認性が向上するのではないかと考えた．具体的には，Google Maps や OpenStreetMap といった地図上に，ツイートに含まれる場所情報（彦根城や四番町スクエアなど）に合わせて表示する．

また，収集された旅行活動データに加え，あえて旅行活動に関係ないノイズを加えて可視化することを検討している．ツイートを機械学習により分類したとしても，少なからず旅行活動と関係ないノイズが含まれる可能性がある．例えば,「彦根城で紅葉の写真を撮った」であれば，彦根城での観光を示唆しているため旅行活動であると判断できる．しかし,「滋賀県立大学で授業を受ける」の場合，滋賀県立大学は彦根市にあるものの，授業を受けることは観光ではないため，旅行活動ではない．このようなノイズの割合が低い状態での可視化は旅行活動の方が大多数であるため，可視化時の影響は少ない．しかし，ノイズが増えるにつれて，旅行活動と関係ない活動により，彦根市など対象エリアの旅行活動が正しく認識できなくなることが考えられる．今後，含まれるノイズ割合と可視化の視認性との関係を実験することで，機械学習においてどこまで正解率や適合率を伸ばすべきかを検討する．

5.2　データ収集について

5.1 節の可視化を達成するためには，旅行活動ツイートに加え，ノイズとなる旅行活動ではないツイートが必要である．必要な旅行活動ツイートは「観光スポットを示す場所で，旅行活動を示す行動」であり，ノイズとなるツイートは「観光スポットを示す場所であるが，行動は旅行活動ではない」である．現在，旅行活動データは 2019 年 2 月頃のデータを 94 件収集しているが，今後は 2020 年から 2022 年の 10 月のツイートを 1000 件程度収集したいと考えている．

5.3　機械学習モデルの組み合わせについて

本稿では，各機械学習モデルをそれぞれ独立して比較を行ったが，複数の機械学習モデルを組み合わせることにより，再現率を改善できるのではないかと考えている．例えば，SVM (TFIDF) の旅行活動の再現率は 0.520 であるが，ChatGPT の再現率は 0.479 である．これらのモデルが旅行活動と判定した投稿を単純に足し合わせることにより，再現率の改善が見込まれる．ただし，SVM (TFIDF)，ChatGPT の両方が旅行活動であると判定した投稿が多いと，足し合わせても再現率は改善しないため，精査が必要である．また，再現率が

低く，適合率が高い ChatGPT は，旅行活動であるにも関わらず取りこぼした投稿が存在する．これを，再現率が高く，適合率が低いナイーブベイズに分類させることで，ChatGPT が取りこぼした旅行活動ツイートを拾うことができるのではないかと考えている．このように，異なる機械学習モデルを組み合わせることによる精度改善を検討する．

参考文献

1) 公益財団法人日本交通公社, “国内旅行における SNS・写真に対する意識／実態 〜JTBF 旅行実態調査トピックス〜,” https://www.jtb.or.jp/research/statistics-tourist-sns-pictures2022/（参照 2023-09-26）.
2) 藤大 友都, 服部 峻, 砂山 渡, “旅行活動ツイート可視化のためのフィルタリングと Artisoc エージェント化,” 第 15 回データ工学と情報マネジメントに関するフォーラム，5b-3-2 (2023).
3) 西村 駿人, 数原 良彦, 鷲崎 誠司, “地域特徴語選択を用いたマルチクラス分類による Twitter ユーザの居住地推定,” 電子情報通信学会 信学技報，Vol.112，No.367，NLC2012-37，pp.23–27 (2012).
4) 藤田 俊之，小林 亜樹，“マイクロブログにおける会話を利用した災害情報採集手法,” 電子情報通信学会論文誌 B，Vol.J106-B，No.1，pp.1–12 (2023).
5) artisoc4 - MAS コミュニティ - 構造計画研究所, https://mas.kke.co.jp/artisoc4/（参照 2023-10-03）.