旅先における失敗リスクを把握可能にするための機械学習を用いた失敗談ツイート抽出方法の構築と静岡県内観光地での適用

吉田伊武貴　東京都立大学都市環境学部都市環境学科自然・文化ツーリズムコース
倉田陽平　東京都立大学院都市環境科学研究科観光科学域

キーワード：ジオタグ付きツイート、失敗談、ツイート分類

Ⅰ．はじめに

旅行は日常生活と異なる見知らぬ土地での多額の消費活動である.見慣れぬ旅行先での思わぬ失敗やトラブルは珍しくない．「「旅のトラブル」に関する意識調査」によると「旅行先でトラブルに遭遇した」という質問に「はい」と回答した人は 46%だった(LINE トラベル jp 2019).ましてコロナ禍の旅行では訪問先の感染対策の状況など感染リスク回避の注意が必要となっている.よって観光情報には旅先の魅力だけでなく旅先のリスク回避に役立つ情報も望まれる.観光者の多くは事前に雑誌等を駆使し観光地の情報を収集する．近年のインターネットや SNS の普及により，Web 上には観光関連情報が豊富に存在しWebやSNSに関連した観光研究も進められている．Web 上のテキストデータに対して行われた研究には,石川ほか(2016)は大量の旅ブログ文書からの機械学習を用いた旅行ノウハウ情報の抽出,鈴木ほか(2019)は大量の位置情報付きツイートから迷いやすい箇所の抽出,斎藤ほか(2016)は位置情報付きツイートに現れる感情の分布を地図化する手法をそれぞれ提案している.

一方で SNS 上には各地での失敗談や経験談を述べた情報が日々投稿されていることが推測される.日本語での失敗談には「・・しすぎた」「・・してしまった」などある特定の単語等が使用されることが考えられる.本研究はそれらに注目し,日本語の特有表現を足掛かりに機械学習を用いて大量の位置情報付きツイートから各地の失敗談や経験談を述べたツイートの抽出を試みる.さらに抽出したツイートを地図上にプロットすることで旅行先の潜在的なリスクの可視化や旅行者のリスク回避となるか考察を行う．様々な種類の失敗談ツイートを横断的に抽出・可視化し，その観光情報としての価値を論じるものとして本研究は管見の限り先駆的である.

Ⅱ．研究目的

本研究では大量のツイートから機械学習により失敗談を抽出し,地図上に可視化することで,旅行者が事前に旅行先での失敗や様々なリスクなどネガティブな情報を効率的に入手できる可能性を創出することや旅先の潜在的リスクの可視化・リスク回避の一助となる情報を体系的に導出することを目的とする.

Ⅲ．研究方法

2016～2017 年に投稿された Twitter の位置情報付きツイート約 1143 万件を用いて,失敗談抽出フィルタリングを機械学習により構築する.機械学習のモデル及び分類器作成に使用するデータとして人手により失敗談・非失敗談とするツイートをサンプルとしてそれぞれ約 1 万件ずつ収集する.機械学習のモデルおよび分類器は,教師あり学習と深層学習および転移学習の 2 パターンで作成し評価指標により比

較を行う.より性能の良い機械学習のモデル及び分類器で全てのツイートデータから失敗談の抽出を行う.抽出後,位置情報をもとに QGIS を使用して地図上にプロットする.各地の失敗談分布を観察し,各地の失敗談が旅行の参考になるか考察を行う.

3.1 本研究における「失敗談」

図 1 のように忌避されるような失敗経験やネガティブな経験を失敗談とする.そのため実際には失敗していないネガティブな経験（図 1 中の混雑や電波に関連したツイートなど）も失敗談として取り扱う.一方で寝坊や遅刻・食べ過ぎのような原因が個人に帰するのみのものは失敗談には含めない.

・バイクの駐車場無い
・電波弱くて携帯つながりにくい
・目当てのプリン売り切れてた～
・金閣寺人多すぎて吐きそう
・Suica 使えない…
・来たけど工事中で見れなかった

図 1　失敗談ツイートの例

3.2 ツイートデータの前処理

対象とするツイートデータは 2016 年～2017 年にかけて日本国内で一般人に投稿されたものとする．日本語以外の言語や天気・地震等の自動投稿のツイートを除去した．html タグや URL や - （）@ユーザー名などの不要な文字列の除去や単語の表記揺れ(リンゴ・りんご・林檎など)や全角・半角文字の混在に対して変換し単語の統一も同時に行った.さらに機械学習用訓練データ･学習成果の評価データ用に人手で集めるサンプルツイートは季節性の偏りを防ぐため四季ごとに抜粋した.

3.3 本研究で扱う機械学習

Python を使用言語とし,Google 社が提供しているクラウド上で実行できるサービス Google Colaboratory を利用する．機械学習ライブラリには scikit-learn と Keras Bert を使用する.本研究は 1）教師あり学習,2）深層学習および転移学習,の二つの手法を用いる.1）には scikit-learn から SVM のアルゴリズムを採用する.SVM は主に分類問題に利用され,データを決定境界で分類するときにデータと決定境界の距離が最大になるように学習する.2)には Keras Bert から BERT(Devlin *et al.*2018)を利用する.BERT は深層学習された事前学習済み言語モデルであり,転移学習することで任意のタスクに応用できる.評価指標には正解率,精度,再現率, F 値を用いる.また過学習になっているか調べるため学習曲線も出力する.

Ⅳ．結果と考察

4.1 指標比較

図 2 より双方とも過学習の傾向は見られなかった.表 1 よりいずれの指標においても BERT モデルが良い結果を得た.よって BERT モデルを採用した.

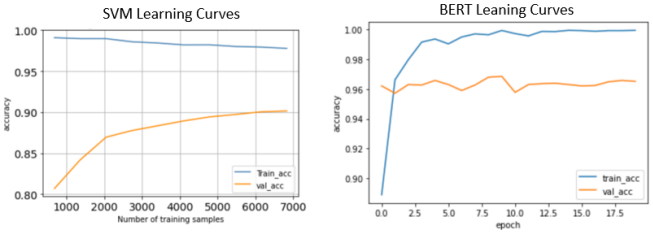

図 2　学習曲線

表 1　分類タスクの評価比較

	正解率	精度	再現率	F 値
BoW+TF-IDF・SVM	0.75	0.68	0.93	0.79
BERT	0.96	0.97	0.96	0.96

4.2 失敗談抽出

総ツイート数 11,427,750 件のうち,本研究にて提案した失敗談抽出フィルタリングを使用した結果,2,434,578 件（全ツイートの約２割）が抽出された.

さらに失敗談に用いられやすい言語表現を

把握するため,機械学習および転移学習によって抽出された失敗談ツイートから各単語の出現頻度を集計した.集計結果から助詞や固有名詞などを除外し,出現頻度の上位の単語を図３にまとめた.これらの言語特徴が含まれていると失敗談として分類されやすいと推測される.また言語表現ではないため本研究の分析のうえで除外された文章に添えられる顔文字や絵文字なども失敗談において多く確認された.

・「れば良かっ」	・「したほうが」	・「やらかし」	・「やってしまっ」
・「やっちゃっ」	・「ちゃっ」	・「じゃっ」	・「orz」
・「しくっ」	・「のに」	・「すぎる」	・「・・すぎ」
・「最悪」	・「全然」	・「悲報」	・「無理」
・「死んだ」	・「終わってる」	・「泣いた」	・「やばい」
・「すごい」	・「こんなに」	・「ここまで」	・「・・そう」
・「きつい」	・「つらい」	・「がっかり」	・「不味い」
・「思ったより」	・「期待より」	・「面倒」	・「不便」
・「残念」	・「したかっ」	・「あんまり」	・「微妙」
・「キレそう」	・「ふざけ」	・「不快」	・「腹立っ」
・「意味わから」	・「邪魔」	・「ぼったくり」	・「ひどい」
・「だるい」	・「アホ」	・「バカ」	・「ムカつく」
・「台無し」	・「うるさい」	・「嫌」	・「・・・ない」
・「困っ」	・「できな」	・「れない」	・「れなかっ」
・「・・にく」	・「間に合わず」	・「諦め」	・「せっかく」
・「間違っ」	・「うっかり」	・「忘れ」	・「イマイチ」
・「しか無い」	・「、、、」	・「。。。」	・「…」

図３　失敗談特有の言語表現

4.3 地図化

静岡県内の主要観光地から「日本平」や「三保の松原」がある静岡市,「三嶋大社」や「三島スカイウォーク」がある三島市,熱海市の失敗談分布から失敗談が旅行者の参考になるか考察する.その際,全ての失敗談ツイートから一部抽出し考察を行う.

4.4 考察

いずれの観光地でも混雑や行列に関するツイートが見られ,どこでそのツイートが投稿されたか確認できた.また改装工事や休業日に訪れてしまう失敗談も各地で見られた.これは観光客の確認不足と,工事や休業日などの情報が適切に観光客に届いていない発信不足の可能性も考えられる.また観光地によって特徴的な失敗談が確認された.紙面の都合上,ここでは「熱海」「日本平・三保の松原」における失敗談ツイートを地図化した結果(図 4 と図 5)のみを紹介する.熱海で確認された失敗談は,硫黄臭に言及したものがあり温泉街ならではの失敗談と考えられる.さらに熱海では大規模な花火大会が開催されることなどから夏の時期には混雑に関連した失敗談が多く「渋滞」「駐車場満車」の語句が散見された.静岡市に位置する日本平や三保の松原では曇天や霧によって富士山の眺望が見られない旨の失敗談が数多く確認された.この「富士山が見えない」を述べる失敗談は県内のほかの観光地においても非常に多く確認でき,静岡県の観光地ならではの失敗談であると思われる.失敗談をミクロスケールで観察することにより各観光地において観光者が遭遇しやすいトラブルに違いがあることや地域ならではの旅先での失敗経験を読み取ることができた.

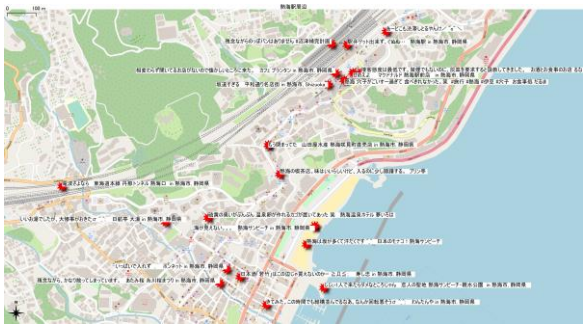

図４　抽出された失敗談分布「熱海」

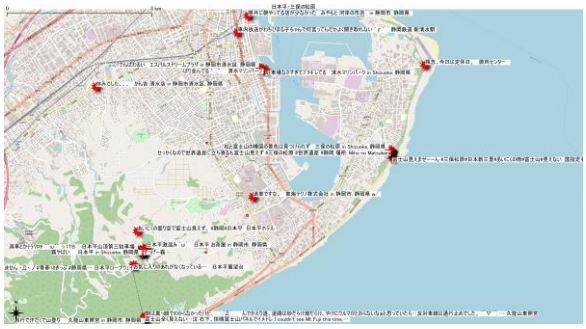

図５　抽出された失敗談分布「日本平・三保の松原」

Ⅴ．まとめと今後の課題

本研究では大量の位置情報付きツイートから失敗談を機械学習で抽出する手法を提案した.結果,失敗談分布を可視化することで各地の失敗経験を把握できた.

一方,抽出された失敗談には,観光者にとって,参考にならないものも多かった.役に立つ失敗

談と役に立たないものの判別も課題として挙げられる.それらツイートは今後除去するか,あるいは各地域における位置情報付きツイートに占める失敗談ツイートの割合を母比率の検定を利用して有意差を見るなどしてリスク評価を行う必要性も考えられる.

参考文献

LINE トラベル jp（2019）「旅のトラブル」に関する意識調査. https://www.travel.co.jp/guide/article/39747/ (2021 年 1 月 2 日閲覧)

石川綾美・難波英嗣・石野亜耶・竹澤寿幸(2016) 旅行口コミサイトからの旅行ノウハウ情報の自動抽出. DEIM FORUM 2016 第 8 回データ工学と情報マネジメントに関するフォーラム（第 14 回日本データベース学会年次大会

斎藤一・横川祥司(2016) 感情語辞書と位置情報付きツイート分析に基づいたアプリケーション「EmoNavi」の観光利用の検討. 北海道情報大学紀要,vol28-1,p103-110

鈴木亮平・廣田雅春・荒木徹也・遠藤雅樹・石川博(2019) 位置情報付きツイートを用いた観光地周辺の迷いやすいスポットの発見.データベースシステム研究会

Devlin,Jacob,Chang,Ming-Wei,Lee,Kenton,Toutanova, Kristina(2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Association for Computational Linguistics : Human Language Technologies", Volume.1,pp.4171-4186