失敗談ツイートのテキスト分析による北海道における観光失敗リスクの把握

吉田伊武貴，倉田陽平
東京都立大学大学院都市環境科学研究科観光科学域
yoshida-ibuki@ed.tmu.ac.jp

概要：北海道内で観光客が発信したと思われる失敗談がつづられた大量のツイートを抽出し、テキスト分析を行った。KHcoder を用いた計量テキスト分析と教師なし学習によるテキストのクラスタ分析により北海道における観光中の失敗談の特徴を抽出した。これにより，北海道観光に典型的なリスクを明らかにした。

Keywords：ジオタグツイート,失敗談,計量テキスト分析,BERT

1．はじめに

旅行は日常生活とは離れた場所での消費行動である.そのため旅行者は慣れない土地での失敗やトラブルを起こしてしまうことは珍しくない.そこで筆者ら(2021)は日本国内の位置情報が付与された Twitter のツイートデータから機械学習を用いて失敗談を述べているツイートの自動抽出を試み,地図上での可視化までを行った.その結果,失敗談分布の考察から各観光地で特徴的な観光失敗経験があることが示唆された.一方で抽出された失敗談ツイートに対する十分なテキスト分析には至らなかった.本研究では対象地を北海道の各観光地として失敗談ツイートに対するテキスト分析を行うことでツイートから定量的に観光失敗に関する情報抽出を目指す.

2．関連研究

実社会の観測のために Twitter のツイートに対してテキスト分析を行う研究は数多く進められている.Twitter は気軽に投稿できる点から日々膨大なツイートが投稿されている.Twitter を活用した研究では大規模なデータを取り扱うことが多い.そのため分析ではビックデータに対する効率的な手法が求められる.河合らの研究(2020)では神奈川県湘南地域 5 市に関するツイートデータに対して計量テキスト分析用のフリーソフトウェア KHcoder を用いて分析を試みている.河合らはツイートデータに対する共起ネットワーク図と各地域及び季節と単語の出現頻度の対応分析の手法を用いている.また和田(2019)は大規模なツイートデータを単語分散表現ベクトル Word2vec により 300 次元の単語ベクトルに変換した後,次元圧縮法により 3 次元に圧縮しクラスタ分析を行っている.本研究ではこれらの分析手法を取り入れツイートデータから北海道観光に典型的なリスクの情報抽出を試みる.

3．研究目的

観光失敗談の地域性を分析するにあたり,観光が主要産業の都道府県である北海道を対象地とした.本研究では北海道内のジオタグが付与された失敗談ツイートから、テキスト分析を用いて北海道における失敗談の特徴を把握すること,さらに GIS を用いたクラスタ分析により失敗談が集中している各地域の失敗談ツイートコーパスにテキスト分析を行うことで北海道の各地域の観光失敗の特徴や差異の把握を目的とする.

4.研究方法

まず 2016～2017 年に投稿された Twitter の位置情報付きツイート約 1143 万件に対して観光資源のポイントデータ（後述）から半径 500m以内に投稿されたものを抽出し，それらのツイートに対して筆者ほか（2021）によって構築された

失敗談フィルタを適用し失敗談ツイートを抽出する.次に北海道内の失敗談の特徴語を特定するため失敗談コーパスを全国の失敗談と北海道の失敗談の 2 グループに分けそれぞれのグループにおける TF-IDF 値を算出する.そして TF-IDF 値をそれぞれ比較し北海道の TF-IDF 値が高い単語上位 50 語を北海道失敗談特徴語として求める.そして計量テキスト分析ができる KHcoder を用いて共起ネットワーク図を作成し失敗談における単語の共起関係の考察を試みる.次に北海道内において失敗談が集中している地域を確認するため北海道内の失敗談ツイートをポイントデータとして GIS ソフトである ArcGIS Pro のクラスタ分析機能を使用してクラスタ分析を試みる.クラスタが確認できた市町村を失敗談集中地域とし,各地域と単語の出現頻度をもとにした対応分析を行うことで地域の失敗リスクの特徴の把握を試みる.また計量テキスト分析では分析しきれない北海道の失敗談特徴語を,和田(2019)のベクトルのクラスタ分析を参考に分析を行う. 本研究では和田(2019)が行っていた Word2vec による単語ベクトルではなく Devlin ら[2018]によって提案された自然言語処理のディープラーニングモデル BERT によって生成する文章ベクトルを用いてクラスタ分析を行う.文章ベクトルによる分析のメリットとしては文章の類似性によるクラスタとなるため単語のベクトルより直感的でわかりやすい点にある.ベクトルの可視化には和田(2019)の研究と同様に Google 社がオープンソースとして提供しているディープラーニングフレームワークの TensorFlow の可視化ツール TensorBoard のスタンドアロン版のウェブ UI である Embedding Projector を利用する.BERT により生成する文章ベクトルから特徴語を含んだ文書のベクトル群を抽出し次元圧縮によるクラスタリングを行い特徴語に関するクラスタ抽出を試みる.

5. 結果と考察

5.1 失敗談ツイートの抽出

観光資源のポイントデータは昭文社の販売する Mapple POI データ全国版 2014 年版を利用した.観光資源のポイントデータの 500m圏のツイート数は 7248202 件であった。それらのツイートに対して吉田ら（2021）によって構築された失敗談フィルタを適用した.抽出された失敗談ツイート数は 851885 件で観光資源周辺 500m 圏内の全ツイートの約 11%であった.

5.2 TF-IDF 値による北海道の失敗談特徴語

全国の失敗談の TF-IDF 値より北海道失敗談の TF-IDF 値が高い上位 50 の単語(動詞・形容詞・形容動詞)の結果をまとめたものが表である.「寒い」「滑る」「凍る」は北海道の寒冷な気候による観光失敗が想像できる.一方で「辛い」という単語は「つらい」のか「からい」のか表から読み取ることはできない.そのため次の共起ネットワーク図から「辛い」の共起関係をもとにどちらの使い方か読み取る.

表 1　北海道失敗談の特徴語

北海道				
寒い	寝る	走る	頼む	諦める
混む	疲れる	持つ	置く	登る
飲む	降る	強い	帰れる	作る
過ぎる	涼しい	遅れる	寄る	冷たい
入る	食う	もらう	閉まる	広い
着く	きつい	空く	出す	晴れる
無い	待つ	向かう	凍る	飲める
悪い	すごい	起きる	にくい	酔う
辛い	美味い	飛ぶ	滑る	来れる
入れる	戻る	おいしい	着る	曇る

5.3 共起ネットワーク

図 4 は北海道の失敗談ツイートによって作成された共起ネットワーク図である.「辛い」はカレーやラーメンにつながっていることから「からい」という意味であることが推測される.ほかにも「忘れる」は「買う」と「写真」が共起関係となっており「買い忘れ」「撮り忘れ」が推察される.しかし「忘れる」「買う」といった単語は 4.2 中の表に含まれていないため北海道特有の観光失敗ではない一般的な観光失敗と考えられる.

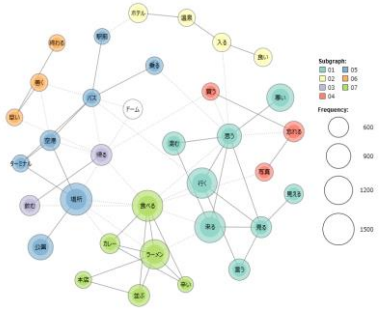

図 1　共起ネットワーク

5.4 季節と抽出語との対応分析

対応分析では原点付近に特徴がない単語が集まり原点から離れているほど特徴的である.四季は 4～6 月を春,7~9 月を夏,10～11 月を秋,北海道は 3 月にも降雪が確認される地域が多いことから 12～3 月を冬とした.四季と単語(名詞・動詞・形容詞・形容動詞)の対応分析の結果が図である.「寒い」「辛い」は秋に多く出現することがわかった.

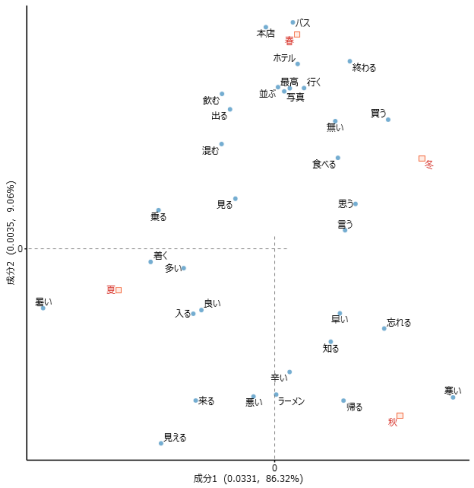

図 2　季節と抽出語との対応分析

5.5 GIS による北海道の失敗談クラスタ分析

ここまでの分析では北海道の失敗談を全道的にとらえてきた.しかし観光失敗は同じ道内でも地域ごとに特徴が異なる可能性が筆者ら(2021)によって示唆されている.よって地域による特徴の違いを把握するため失敗談が集中している地域を求め地域ごとにコーパスを分けて分析を進める.Arc GIS Pro のクラスタ分析機能を使用する.1 クラスタあたりの最小フィーチャ数を 1000,検索範囲を 1ｋｍとした.その結果が図 3 である.クラスタが確認された市町村は札幌市,函館市,千歳市,苫小牧市,小樽市,旭川市,帯広市,釧路市,根室市,北見市,稚内市であった.以降はこれらの地域を失敗談集中地域として各地域と単語の出現頻度による対応分析を行い各地の観光失敗リスクの特徴や違いの把握を試みる.

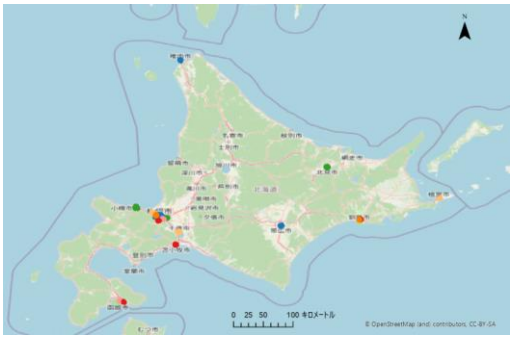

図 3　北海道内の失敗談クラスタ

5.6 各地域における抽出語の対応分析

GIS のクラスタ分析により観光失敗集中地域とした各地域と単語の出現頻度の対応分析を行った結果が図である.「辛い」は札幌における観光失敗の特徴的な単語であることが推察される.また「見える」は函館・稚内・根室の近くに位置している.「見える」は動詞の原形として抽出されており失敗談ツイートを確認すると「見えない」の意味で使用されていることが多かった.これらの地域では函館山や夜景,岬や半島など観光目的で眺望が期待出来るため「眺望を期待したが見えなかった」という観光失敗のリスクがあることが推測される.実際にこれらの地域で「見えない」を含んだ失敗談ツイートには函館山の展望台からの夜景や青森県が見えないといったものが確認できた.また「雪で見えない」の意味でつかわれているものも確認できた.

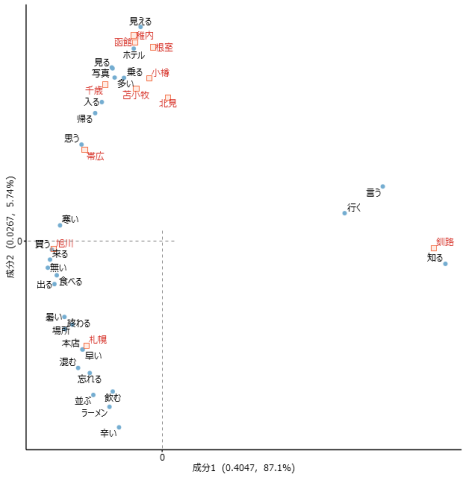

図 4　地域と抽出後との対応分析

5.7 文章ベクトルのクラスタ分析

ここまで対応分析と共起ネットワークによる分析を行ってきたが,5.2 で北海道の気候に関連していると考えられる「滑る」「凍る」「曇る」といった特徴語は抽出されなかった.それらの特徴語の分析のため,ベクトルのクラスタ分析を行う. Embedding Projector により次元圧縮ア

ルゴリズムである PCA,t-SNE によりベクトルを 3 次元に圧縮し可視化を行う.UI 上の検索窓にキーワード（ターゲット単語）を入力することでベクトル上意味の近い文章が選定される.学習回数は約 2000 回とした.図 5 は失敗談特徴語「滑る」から「滑」をターゲット単語に設定した場合のクラスタ分析の結果である.BERT は単語の意味も考慮できるモデルであることから似たような意味を持つツイートがクラスタを形成する.クラスタに含まれる「滑る」失敗談は雪や路面の凍結に起因するものであることが文面から推察できる.図 6 はターゲット単語を「凍」にした場合の結果である.「給油口のフタが開かない」など失敗談が確認された.図 7 はターゲット単語を「曇」にした場合の結果である.「曇り」により星空や景色などの眺望が遮られてしまう失敗談が確認できた.また寒さについて言及する失敗談ツイートもクラスタ内で確認された.

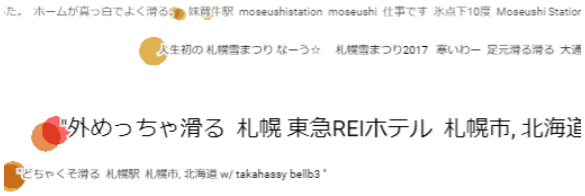

図 5「滑る」の失敗談クラスタ

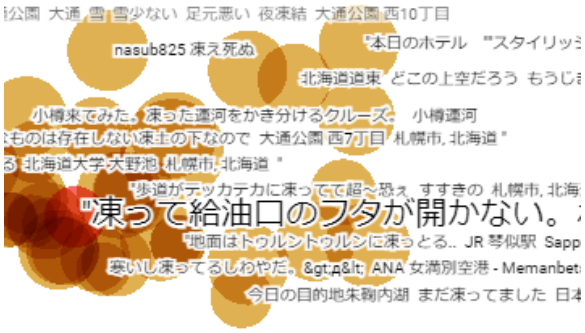

図 6「凍る」の失敗談クラスタ

図 7「曇る」の失敗談クラスタ

6.まとめ

本研究では北海道の観光失敗リスクについて観光失敗談ツイートを用いた計量テキスト分析及び文章のベクトルのクラスタ分析を行った.まず北海道全体の観光失敗リスクの把握のため TF-IDF 値を用いた失敗談特徴語の抽出および共起ネットワーク図を行った.次に失敗談のポイントデータから失敗談の集中している地域を選定し地域ごとの観光失敗の特徴の把握をテキスト分析により試みた.その結果,北海道では「寒い」「曇る」「凍る」「滑る」といった特徴語が確認でき,地域の対応分析により札幌では「辛い（からい）」,函館や根室では眺望に関連した観光失敗のリスクがある可能性が示唆された.また共起ネットワークと対応分析では分析できなかった特徴語については文章ベクトルを用いたクラスタリング分析を行った.その結果,北海道の観光失敗リスクを象徴するような失敗談クラスタが確認できた. 筆者ら(2021)では目視で失敗談の傾向を把握していたが, 本研究でのテキスト分析手法により失敗ツイートから旅行者の失敗傾向をシステマティックに把握できるようになった.今後の課題としては季節だけでなく投稿時間帯ごとの分析が必要である.また抽出された観光失敗談が観光者にとって有用かどうかの検証の必要性も挙げられる.

謝辞

本研究は科研費（21K12484）の助成を受けたものです.

参考文献

[1] 吉田伊武貴, 倉田陽平(2021)：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.第 7 回とうかい観光情報学研究会, pp.1-4, 2021 年 3 月, オンライン発表.

[2]Devlin,Jacob,Chang,MingWei,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

[4] 川合康央,池辺正典,池田岳史,益岡了(2020): Twitter 分析による都市キーワードの抽出. 日本デザイン学会研究発表大会概要集,67 巻, 日本デザイン学会 第 67 回春季研究発表大会,

[5] 和田伸一郎(2019)：インタラクティヴなデータ・ヴィジュアライゼーション・ツールを用いた Twitter データのクラスタ分析.人工知能学会全国大会論文集,JSAI2019 巻, 第 33 回 (2019)