Twitter データを用いたヘビーユーザ特定

小川 歩　　鈴木 愛海　　櫻井 義尚
明治大学 総合数理学部
ev201070@meiji.ac.jp　ev201013@meiji.ac.jp　sakuraiy@meiji.ac.jp

概要

SNS が提供するデジタル活動データは，ユーザ行動や意見の収集に貢献している．特にヘビーユーザからのフィードバックは，企業にとって重要な意見であり，直接的な収益に結びつく．しかし，手動でのヘビーユーザ特定は時間と金銭がかかる．そこで， 本研究では機械学習を使って自動的にヘビーユーザを特定し，ChatGPT と比較する実験を行う．研究の成果が企業やプロバイダーにヘビーユーザ特性の洞察を提供し，マーケティング戦略の最適化に寄与することが期待される．

1 はじめに

1.1 研究目的

近年， Twitter（現 X）や Instagram 等のソーシャル・ネットワーキング・サービス(SNS)が広く普及している．これによって，消費者による商品やサービスに対する意見が SNS 上に多く投稿され，マーケティング上有益な情報が蓄積されている．その中でも，サービスに多くお金を費やすユーザ(以下ヘビーユーザと記載する)は消費において主要な影響力を持つため，彼らのニーズに答えることは企業にとって収益向上への重要な一環である．しかし，そのユーザがヘビーユーザか否かを人手により判断することは定常性に欠ける上にコストがかかるというデメリットがある．そこで本研究では，ヘビーユーザ評価指標を基に，機械学習を用いたヘビーユーザ特定モデルを作成し，ユーザの意見が多く蓄積されている SNS の 1 つである Twitter を対象に，東京ディズニーリゾートのファンの中からヘビーユーザを特定しその精度を検証する．ヘビーユーザの特定をし，その声を聞くことでサービスや製品の改善及び顧客満足度を向上させ，長期的な顧客ロイヤルティを築くことを目的とする．

1.2 先行研究

1.2.1 Twitter データを用いた性別予測[1]

SNS 上の文章分析には文脈に応じた分散表現を獲得することができる BERT (Bidirectional Encoder Representations from Transformers)がしばしば利用され，テキスト分類や固有表現抽出において高い精度を発揮するとの報告がなされている． Burghoorn らは Twitter 上から収集したテキストデータに性別ラベルを付与したデータセットを元に， BERT モデルを構築し，それを用いたユーザの性別予測の研究を行った．その結果，少量のデータを用いた性別予測において BERT モデルの有効性を示している．

1.2.2 深層学習を用い興味傾向推定[2]

SNS 上の情報を用いて，ユーザの属性を推定する研究は以前より盛んに行われている．石川らは Twitter ユーザの「いいね」の傾向を基に，ユーザの興味傾向推定を行った． 学習には，8 層の隠れ層を持つディープニューラルネットワークが用いられ，その結果として 8 割を超える精度での予測が可能であると結論付けている．この結果から見て取れるように，機械学習はユーザ属性の特定分野において非常に有望であり，高い精度での予測が期待できる．

1.3 本研究の立ち位置

前章に示す通り，BERT モデルを用いた Twitter テキスト分析は様々な手法で行われている．しかしながら，サービスや商品の消費量に着目しコンテンツにおけるヘビーユーザを特定するシステムの構築を行っている研究は存在しない点で，本研究は新規性を具備している．

2 データ

研究の出発点として，本研究ではディズニーパークのヘビーユーザを明確に特定するため，SNS 上でのツイートデータを採用した．これにより，ディズニーパークに深い関心を寄せるユーザのアクティビティを捉え，独自の条件に基づいてデータセットを構築した．データセットの構成に先立ち，我々がどのようにヘビーユーザを定義し，収集条件を設定したかを詳細に説明する．

2.1 データセット

2.1.1 データセットの構成

データの構造は表 1 に示された通りである．本研究では，主にヘビーユーザと NOT ヘビーユーザ（ヘビーユーザではないユーザ）の 2 つに分類した．ヘビーユーザ群には，インフルエンサー，フォロワーが 1000 人以上のユーザ，フォロワーが 1000 人以下のユーザの 3 つのサブグループがあり，NOT ヘビーユーザ群には，ファンと NOT ファンの 2 つのサブグループが存在する．これにより，データセット内のユーザを詳細かつ多面的に分類し，分

析の幅を広げた． インフルエンサー, ファン, 及び NOT ファンの定義については付録に記載する．

表 1 データセットのユーザ構成

ヘビーユーザ			NOT ヘビーユーザ	
インフルエンサー	フォロワー 1000 人以上	フォロワー 1000 人以下	ファン	NOT ファン
40	120	340	300	200

2.1. Twitter API

Twitter API を活用し，ユーザ 1,000 人から個別にユーザごとに 100 件ずつ，計 100,000 ツイートを収集した．これにより，各ユーザの投稿パターンや傾向を詳細に分析し，データセットの多様性を確保した．そのうちランダムに抽出した 7,000 件を学習データとし，3,000 件をテストデータとして使用した．

2.2 ヘビーユーザの定義

本研究において，「ヘビーユーザ」は東京ディズニーリゾートを頻繁に訪れ，訪れた際に商品またはサービスにお金を費やしているユーザ，または東京ディズニーリゾート関連施設に多くお金を費やしているユーザを指す． アノテーション時に自ら基準を設け，採用したヘビーユーザの判断基準は以下の通りである．ただし，どれか 1 つだけの条件を満たすだけでは不十分であり，複数項目（特に 1 番を重要視）を満たすユーザをヘビーユーザとしてカウントした．

1. 訪問頻度：
・ 1 か月あたりの入園回数が多い（注：イベントやハロウィン期間などの限定的に多い場合は除外）
2. 居住地の熱意：
・ 居住地が地方であっても月 1 回以上の頻繁な訪問がある場合は高評価
3. 購買活動：
・ ディズニーパークのショップでの購買活動が見られ，話題の商品にも関心を示す
4. 宿泊施設利用：
・ ディズニーホテルを利用しているかどうか（オフィシャルホテルかどうかも判断材料）
5. 飲食利用：
・ ディズニーパークやディズニーホテルのレストランを利用しているかどうか
6. エンターテインメントへの熱意：
・ ショー，グリーティング，アトラクションに対する熱意が見られる（例：同じアトラクションへの執着，ディズニープライオリティパス購入の有無）
7. 専門的知識：
・ ディズニーパークに対する特別な知識を有する（例：ディズニーパークの植物に詳しい，パレードに使われる機材に理解がある）

これらの基準を総合的に評価し，複数の条件を満たす者をヘビーユーザとして本研究の対象とした．

2.3 ツイート収集

2.3.1 ツイート収集条件

論文のデータ収集条件では，ツイート最低収集条件を以下のように設定した．対象はディズニーパークのヘビーユーザとし，これに該当するユーザを特定するための条件を定めた．

· 総ツイート数:
ユーザの総ツイート数が 500 以上であることを条件とした．これにより，積極的にツイートを行っているアクティブなユーザを対象とした．
· ディズニーパーク訪問頻度:
ユーザが月に 1 回以上ディズニーパークに行っていることが条件である．[3]の p. 26 の「ヘビーユーザ」の定義を参考に，この条件を選定した．この条件では，ユーザが頻繁にディズニーパークに訪れているヘビーユーザを特定した．訪問回数は，ツイート文に明確に言及されていなくても，写真や実況の投稿などから判断した．
· ディズニーパーク専念:
ディズニーパーク以外のディズニー関連イベントや映画への投稿が少ないことを条件とし，ディズニーパークのヘビーユーザに限定した．これにより，ディズニーパークに特化したユーザを対象とした．

これらの条件に基づき，データセットを構築した．ユーザの総ツイート数とディズニーパークへの頻繁な訪問が特定の基準を満たす場合に，ヘビーユーザとしての認定を行った．

2.3.2 ツイート収集方法

研究のツイート収集において，以下の仮説を立てて検討した．

仮説①: ヘビーユーザ同士の相互フォローの可能性
ディズニーパークのヘビーユーザ同士は，共通の興味を有しており，互いにフォローし合っている可能性が高いとの仮定から，同じ趣味を共有するユーザ同士の繋がりを把握することが重要である．

仮説②: ディズニーファン特有の呼称の使用

ディズニーファンが特有の呼称を使用するユーザには，ヘビーユーザが多く含まれる可能性があるとの仮定から，特定のコミュニティに焦点を当ててツイート収集することが有益である．

仮説③：ヘビーユーザの有益な意見の発信
ヘビーユーザが有益な情報を提供している可能性が高いとの仮定から，マーケティングの観点から価値のある情報を抽出する手法に注力することが重要である．

これらの仮定に基づき，効果的なツイート収集方法を選定した．

(1)　ワード検索：
・　ハッシュタグ：#TDL_now，#TDS_now，#インパ などのディズニーパーク関連のハッシュタグを対象にした．
・　ファン特有の呼称：ディズニーリゾートラインに対するリゾラ，スプーキー“Boo!”パレードに対するスキブ，ミート・ミッキーに対するミトミなど，ディズニーパークファンが使用する．独自の呼称を抽出する．
(2)　ヘビーユーザの RT(リツイート)：
ディズニーパークに関連するツイートをリツイートするユーザから収集を行い，ヘビーユーザの行動を追跡する．
(3)　おすすめユーザ：
ユーザが活発にディズニーパークに関連するツイートを行っている場合，そのユーザがおすすめとして表示される可能性が高いため，おすすめ欄からのツイートも収集対象とする．
(4)　引用 RT(リツイート)：
ヘビーユーザが発信する意見や情報が他のユーザによって引用リツイートされた場合，有益な情報が含まれている可能性が高いため，引用 RT からのツイートも収集する．

仮定_1 に基づき，ヘビーユーザ同士の相互フォローの可能性を検証するために，(2)ヘビーユーザの RT(リツイート)と(3)おすすめユーザの手法を主要な収集手法として採用した．これにより，同じ趣味を共有するユーザ同士の繋がりを効果的に把握し，ヘビーユーザの同定に重点を置いている．
仮定_2 に基づき，ディズニーパークファンが使用する特有のワードを抽出するために，(1)ワード検索の手法を導入した．これにより，ディズニーパークファンが活発に交流する場でのヘビーユーザを捉え，網羅的なデータ収集を実現した．
最後に，仮定_3 に基づき，ディズニーパークファンが有益な情報を発信している可能性が高い(4)引用 RT(リツイート)の手法を採用した．引用 RT(リツイート)には意見が多く投稿され，その中から有益な情報を提供しているユーザを参照することで，マーケティングの観点から価値のあるデータを収集できると期待している．

2．4 データの前処理

実験を行うにあたり，データセットに前処理を施し，2 種類のデータを構築する．一方のデータには，絵文字を削除する処理を行い，もう一方には施さずそのまま学習を行った．絵文字は文章学習精度に大きく寄与することからこの 2 種類のデータを用いることとする．url の削除，改行記号(n/)の削除，記号の半角変換，メンションの削除については，両方のデータに共通して施す．

3 ヘビーユーザ特定の提案手法

3.1 BERT モデル

文章分類には， 東北大学自然言語処理研究グループにより 2023 年 6 月 22 日に公開された，訓練済み日本語 BERT モデル(bert-base-japanese-char-v3)を使用する．この BERT モデルは 12 レイヤー，768 次元の隠れ層，12 個のアテンションヘッドを持つ．入力テキストは，Unidic2．1．2 辞書を基に単語レベルのトークン化をし，後に WordPiece サブワードによるトークン化がなされる．トレーニングには CC-100 データセットの日本語部分と Wikipedia Cirrussearch ダンプファイル(2023 年 1 月 2 日時点)から生成されたテキストコーパスを使用，テキスト分割には fugashi 及び mecab-ipadic-NEologd が使用される． 実験で使用する際のハイパーパラメータについては，データローダのバッジサイズを 8，モデル学習は 10epoch まで繰り返し行った．

3.2 モデル精度の評価手法

本研究では，100 件のテキストデータを入力した際，対応するユーザ ID に対してヘビーユーザ，NOT ヘビーユーザのラベルを正確に分類できた場合を正解とする．正解ラベルが 2 種類であることから，2 値分類の予測に広く用いられる，Accuracy，Precision， Recall， F1-score をベースに評価を行う．

4 実験

4.1 実験の流れ

本研究は，ヘビーユーザデータセットの作成，ヘビーユーザ特定モデルの作成，モデルの精度検証，ChatGPT による比較検証の 4 段階により構成される．

4. 1. 1 ChatGPT を用いた比較検証

ChatGPT で以下のプロンプトを実行．OpenAI の gpt-4-1106-preview を使用して，各テキストデータのラベルを予測し，Accuracy， Precision，Recall， F1-score を算出.

・このデータセット内の body と label を用いて，ラベル予測してください。その時，F1score， Precision， Recall， Accuracy を計算してください。

---

5 実験結果

表 2 実験結果

	F1 score	Precision	Recall	Accuracy
BERT 絵文字有り	0.58	0.57	0.60	0.57
BERT 絵文字無し	0.56	0.71	0.63	0.59
ChatGPT 絵文字有り	0.53	0.57	0.60	0.57
ChatGPT 絵文字無し	0.52	0.57	0.61	0.57

5.1 考察

BERT モデルの実験結果によれば，「絵文字有り」および「絵文字無し」のデータセットでの性能は類似しているが僅かに「絵文字無し」のモデル精度が上回っている．特に「絵文字無し」の Precision が 0．71 と高いことが注目される．これはデータセットの特性がモデルにとって識別しやすいことを示唆している．F1 スコアも両データセットで均衡した性能を示し， BERT モデルが異なるテキストデータに対して一定の汎用性を持つことを示唆している．一方で，ChatGPT では絵文字有りモデルと絵文字無しモデルで精度に大きな差は見られなかった．

BERT モデルと ChatGPT の比較では， BERT が両データセットで高い性能を示している．「絵文字有り」では，両モデルの精度に大きな違いが見られないが，「絵文字無し」では，全体的に BERT モデルの精度が高く，特に Precision の差が顕著であった． 総じて，「絵文字無し」のデータセットにおいてはヘビーユーザ特定モデルが高い Precision を示し，モデルの陽性予測が高い信頼性を持っていることが示唆された．一方で，絵文字の使用がモデルの性能に影響を与えており，絵文字の適切な処理方法の探索が必要であると考えられる．ディズニー特有の言葉や略称の取り扱いにも改善の余地があり，これらの課題に対処することでモデルの性能向上が期待される．

5.2 今後の課題

今後の改善点として，データセットを見直すことによる予測精度の向上が期待される．今回行った検証で，ディズニーファン特有の略称及び呼称の学習が本データセットでは網羅できなかったことから，それらの呼称をまとめたコーパスを作成することによる精度向上見込みがある．また，本研究の課題として，テキストのみに着目しヘビーユーザ特定を行っている点が挙げられる．東京ディズニーリゾートファンの傾向として，キャラクターの写真を熱心に撮影するという行動が多く見られた．また，写真はユーザがディズニーパークを訪れているかどうかを判断する重要な材料にもなり得る．此度の検証では，一切の画像情報を削除しているため，本研究と画像分析を組み合わせることにより更に高精度なヘビーユーザ特定が実現すると予想される．

6 おわりに

ツイートを使用し，東京ディズニーリゾート利用者のヘビーユーザを機械学習により特定する試みを行った．その結果，すべての精度指標において約 60%の精度が得られ，特に絵文字を削除したデータセットを用いたモデルの Precision が 70%を超える高い性能を示した．一方で，コーパスの作成や画像情報の考慮により，更なる予測精度向上が期待される．ヘビーユーザ特定モデルは有用性を持ちつつも，改善の余地があることが示唆され，今後の研究ではこれらの課題に取り組み，モデルの性能向上を図るべきである．また，自然言語処理分野におけるモデル選択の重要性が明らかにされた． BERT モデルは絵文字を含まないテキストデータにおいて高い分類性能を示し，複雑なテキスト構造を理解し特徴を捉える能力を持っている．一方で，ChatGPT は一般的なテキスト分類タスクにおいて精度向上の伸び代に限界がある．これらの知見は，特定のタスクやデータセットに最適なモデルを選択する指針となりうる．ヘビーユーザには固有の特徴があり，これが新しいマーケティング戦略の契機となると予想される．今後は，ヘビーユーザの意見を更に深く分析し，マーケティング戦略の洗練に活かしていく必要がある．

謝辞

本研究は JSPS 科研費 20K11960 の助成を受けたものです.

参考文献

1. Burghoorn, Maaike et al. ″Gender prediction using limited Twitter Data.″ ArXiv abs/2010.02005 (2020): n. pag.

2. 石川理一朗, 山本雄平, 佐野睦夫. ″深層学習を用いた Twitter 利用者のトピックに対する興味傾向推定.″ 2021 年度 情報処理学会関西支部 支部大会 講演論文集, 2021.

3. MRI リサーチアソシエイツ株式会社『3 万人データから解く新たな消費者像「仮説検証」から「仮説発見」へテーマパーク編』(オンライン) https://dl19w3jlhkm4w.cloudfront.net/2018/3万人データから解く新たな消費者像_テーマパーク編 181220_2.pdf(アクセス日：2023 年 1 月 10 日)

付録

・「インフルエンサー」とは
3, 000 人以上のフォロワーを有し, 東京ディズニーリゾートに関するツイートを頻繁に行うユーザを指す. これには YouTube チャンネル, 個人ブログ, Twitter タイムラインでディズニーリゾート情報を発信するユーザが含まれる.

・「ファン」とは
関心を持ち日常的に関連ツイートを行うが, ディズニーリゾートでの支出が確認できないユーザを指す.

・「NOT ファン」とは
ディズニーリゾートに関するツイートを 1 つ以上行っているが, パークに関する言及が恒常的でないユーザを指す.