6V-05

潜在的ディリクレ配分法を用いた 一人称代名詞による SNS の投稿内容への影響の分析

松澤 拓海† 川又 泰介† 松田 源立†

成蹊大学 理工学部 情報科学科†

1. はじめに

現代において SNS は人々の生活から切り離せない存在として多くの人が利用している。SNS にはそれぞれ特有の特徴があり、用途に応じて複数の SNS を使い分けているユーザも多い。その中でも Twitter は自分の好きなことや思っていることをつぶやくようにして書き込むスタイルとなっており、投稿内容（ツイート）にユーザのその場の気分といった感覚が反映されやすい SNS である。本研究では、そのようなツイートに反映されるユーザの感覚を分析することを目的とし、他言語には珍しい日本語独自の特徴として、多様な一人称代名詞の存在に着目して分析を行い、各一人称代名詞が含まれたツイートごとに特徴を抽出することを目指す。

2. 先行研究

研究を進めるにあたって、研究の方向性決めやできることを探るために参考にした論文を 3 点示す。

2.1 新型コロナウイルスに伴う Twitter の分析と感染状況との関連性[1]

[1]では新型コロナウイルスに対する動向を Twitter のツイートを活用して時系列を追って分析している。[1]では新型コロナウイルスに対するツイートを集める際にツイート本文中に「新型肺炎」「コロナ」「ウイルス」「ウィルス」「武漢」というキーワードが含まれていることを条件としている。本研究ではこの手法を参考にし、キーワードを「私」「自分」などの一人称代名詞に変更し、ツイートを収集した。また、[1]では集めたツイートを形態素解析することで単語ごとに分割して分析を進めており、その際に使用していたストップワード群を本研究でも利用した。

2.2 大学生の一人称の使用についての研究[2]

[2]では大学生の一人称代名詞について、話す場合と SNS 上で使用する場合とに分け、どの一人称を使用するのか大学生の男女別に分析されている。また、その一人称をいつから使い始めたのかを知るために、アンケートを実施して調査している。本研究では、一人称を分析する際に調査項目として扱われていた一人称の一部を今回の分析対象として利用する。

2.3 2019 年大阪ダブル選挙における Twitter 分析 [3]

[3]では、2019 年 4 月 7 日に投票が行われた大阪ダブル選挙における Twitter の分析をしている。分析する手法として潜在的ディリクレ配分法を用いて 10 分類のトピックを作成している。本研究では、[3]のトピック数やツイート数の関係を参考にトピック数とツイート数を決定した。

3. 研究手法

3.1 ツイートの収集

11 月 4 日～12 月 26 日までの期間に投稿された日本語のツイートから特定の一人称代名詞を含むツイートを収集した。収集対象は[2]に挙げられた「私」「自分」「俺」「僕」「うち」「あたし」「わたくし」の 7 種類の一人称を含むツイートとした。ツイートの収集方法として Google Spreadsheet のアドオンの Twitter Archiver[4]を活用して、上記各一人称のいずれかを含むツイートを収集した。この際にリツイートとリプライは除外した。

3.2 形態素解析

収集したツイートを読み込み、日本語形態素解析システムの Mecab[5]を用いて名詞の単語リストを作成した。この際に記号や絵文字などのデータのノイズが多種類あったので、この時にストップワードとして単語リストから除外した。また、使用頻度が極端に少ない単語を除去するため、頻出単語のうち上位 5000 語までを利用した。

3.3 分析方法

各一人称を含むツイート群から 3000 件ずつランダムに抽出し、合計 21000 件のデータから潜在的ディリクレ配分法（以下 LDA）を用いて一人称共通のトピックを抽出した。トピック数は一人称の種類と同様の 7 とした。scikit-learn の LatentDirichletAllocation の fit 関数を LDA のツールとして利用した[6]。doc_topic_prior は 0.15 に設定した。次に、上記で抽出した一人称共通のトピックと各一人称を含むツイートとの関連の強さを調査する為に、transform 関数で各ツイートの各トピックとの関連の強さを計算した。更に各一人称を含む 3000 件のツイート群ごとに、その平均値を求めた。

4. 結果

4.1 全ての一人称に関する共通トピック

表 1 は、いずれかの一人称を含むツイート群から LDA により抽出された一人称共通の 7 つのトピックを示す。各トピックにおいて影響力が強い単語を 20 個表示している。まず、一部に重複する単語も存在するが、概ね異なる単語群の影響力が強いことから、異なる 7 種類のトピックが抽出されたことが分かる。さらに、表 1 の左欄に各トピックの解釈を与えた。このトピックの解釈を行う際には、トピックに含まれている単語とともに、各トピックと関連の強いツイートも参考に解釈を行った。ここでは、今後の考察で重要性が高い「仕事」「娯楽」及び

The Analysis of the Effects of First Personal Pronouns in Posted Contents of SNS by Latent Dirichlet Allocation
†Takumi Matsuzawa, Taisuke Kawamata, Yoshitatsu Matsuda Faculty of Science and Technology, Department of Computer and Information Science, Seikei University

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.

表 1　一人称共通の各トピックの上位単語と解釈

トピック 1（身体）	心 達 数 普通 猫 性 的 金 バカ 頭 体 今 可能 年 未来 者 事 気 期待 英語
トピック 2（仕事）	日 目 今日 年 今年 月 円 仕事 気 参加 分 万 前 疲労 夜 番 店 来年 誕生 昨日
トピック 3（娯楽）	方 全部 曲 彼女 言葉 過去 次 手 応援 顔 大切 最高 感情 音 個 天才 嘘 意味 料理 自身
トピック 4（学生）	君 笑 子 友達 前 方 力 愛 話 嫌 誰 感謝 質問 万 箱 絶対 車 勉強 相手 今
トピック 5（恋愛）	夢 男 女 声 的 名前 方 定期 絵 本 電話 今 最近 先 話 風 点 気 馬 競馬
トピック 6（親子）	様 今 明日 者 娘 動画 親 今日 子 頃 色 息子 雪 子供 歳 派 妹 服 配信 気
トピック 7（大人）	事 誰 世界 今 家 有馬 記念 無理 大 位 日本 墓 度 家族 今日 部屋 酒 風呂 年 所

「学生」の 3 種類のトピックについて解釈の根拠を述べる。仕事トピックに関しては、「仕事」「疲労」「夜」など会社員の生活に関係しそうな単語が多く、また「店」などは営利組織であり、このトピックを代表するツイートの中にも企業名などが含まれたツイートが見られた為、仕事に関するトピックと解釈した。次に娯楽トピックについては、「曲」「応援」「音」「料理」など趣味や娯楽に関する単語が多く、「最高」や「感情」などといった単語も見られている。加えて、このトピックを代表するツイートには娯楽に関するものが多く見られた為、娯楽に関するトピックと解釈した。学生トピックに関しては「友達」や「勉強」など、学校を連想させる単語が多い。また、集めたツイートを調査した結果、質問箱という SNS で流行している匿名質問サービスが多く見られており、このトピック構成している「質問」や「箱」という言葉は質問箱に関係したツイートが大きく影響していると考えられる。このサービスを使う年代は学生が多いと考えられるので、学生に関するトピックだと解釈した。

4.2 一人称によるトピックの強さ

表 2 は、表 1 の各トピックと各一人称のツイートとの関連性を算出し、その平均値をリスト化したものである。リスト中に赤と青で記されている数値があるが、こちらは各一人称でトピックに対して関連が強く出ている一番高い数値を赤で、次に高い数値を青で記している。この結果から、ほとんどの一人称では学生トピックに対して 1 番関連性が強いといえる。2 番目に関連性が強いのは、娯楽トピックと仕事トピックのどちらかだと分かった。このうち娯楽トピックは「俺」「あたし」「わたくし」といった比較的日常的な一人称との関連が強く仕事トピックは「私」「自分」「僕」「うち」といった比較的堅い一人称との関連が強かった。また「僕」という一人称は特異的に親子トピックとの関連性が強かった。

各トピックに対しての関連の強さで最大と最小の差が最も大きい（トピックの偏りが大きい）のは「私」であった。以下、「自分」「僕」「わたくし」「うち」「あたし」の順であり、差が最も小さいのは「俺」であった。

表 2　各一人称を含むツイート群と各トピックの関連の強さの平均値

	身体	仕事	娯楽	学生	恋愛	親子	大人
私	0.124	0.145	0.137	0.181	0.136	0.142	0.136
自分	0.142	0.156	0.138	0.170	0.127	0.143	0.131
俺	0.125	0.138	0.152	0.153	0.145	0.147	0.140
僕	0.129	0.155	0.129	0.140	0.137	0.172	0.138
うち	0.127	0.158	0.150	0.158	0.122	0.132	0.153
あたし	0.130	0.146	0.151	0.158	0.136	0.134	0.146
わたくし	0.124	0.139	0.148	0.161	0.143	0.139	0.147

4.3 考察

学生トピックが最も支配的であった結果から、一人称を使用する Twitter の利用者の多くは若者であることが示唆される。しかし、2 番目に関連が強かったトピックが仕事トピックと娯楽トピックに関しては、一人称によって大きな差があり、「俺」「あたし」「わたくし」といった日常的な一人称が、感覚的な部分が強い娯楽トピックと強い関連があることが観察されたのは興味深い現象である。また、トピックの最大値と最小値の差に関しても、「私」「自分」「僕」は大きく、「俺」「あたし」「わたくし」は小さいという現象が観察され、一人称によってツイートの特徴に大きな差があることが示唆されている。

5. まとめ

本研究では 7 種類の一人称を対象として、Twitter のツイートを集め、LDA による分析を行い、一人称によってツイートのトピックに大きな違いがあることを示した。今後は一人称の種類を増やして同様の調査を行うことを予定している。また、今回使用した一人称の「うち」は文章中で「～うち」などで使われている可能性があるため、より性能の高い辞書等を用いることでデータの精度を向上させたいと考えている。

謝辞

本研究は JSPS 科研費 JP21K12036 の助成を受けたものである。

参考文献

[1]井原史渡, 岸本大輝, 栗原聡 “新型コロナウイルスに伴う Twitter の分析と感染状況との関連性”, 第 35 回人工知能学会全国大会論文集 (2021).
[2]野原　加奈子, 松田　勇一, “大学生の一人称の使用についての研究”, 宇都宮共和大学　都市経済研究年報, 15(0), 91-122 (2015).
[3]梅原　英一, 小川　祐樹, 平川　敦貴, “2019 年大阪ダブル選挙における Twitter 分析”, 第 34 回人工知能学会全国大会論文集 (2020).
[4]https://digitalinspiration.com/product/twitter-archiver/ (2021 年 10 月 26 日アクセス)
[5]https://pypi.org/project/mecab-python3/ (2021 年 10 月 1 日アクセス)
[6]https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html, (2021 年 12 月 20 日アクセス)

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.