BERTopicを利用したTwitter（現X）における　くまモン・ファンダムの言論分析

畠山　真一

1　はじめに

本研究では、畠山（2023）で提示したデータを再分析し、くまモンに関するTwitter（現X）への投稿をもとに、コアなファン層の特質について明らかにすることを目的とする。本研究で対象とするデータは、2021年9月29日～2022年1月5日の期間に収集されたものであるため、現在Xと呼ばれているSNSプラットフォームを旧来のTwitterと呼ぶ。

本研究では、投稿のトピックを深層学習を応用したBERTopicと呼ばれる手法を用いて分析し、その結果としてくまモン・ファンの成長過程とくまモンを実体化の間に関連性があることを明らかにする。

2　使用したデータについて

本分析のベースとなるデータは、Twitter社から提供されているAPIを利用して、「くまモン」を本文に含む投稿を収集したものであり、その基本的な情報は次のとおりである。

- 収集した期間：2021年9月29日～2022年1月5日
- 本文に「くまモン」を含む投稿のすべてから、次の条件に当てはまるものを削除した。
  - 「送料無料」という広告目的のキーワードを含む投稿を削除した(1)。
  - 同一内容の複数投稿については、1つのみを残した。
  - 投稿から絵文字、顔文字、URLを削除した。
  - neologdnモジュールを利用して正規化した(2)。

投稿そのものに加えて、上述のデータには、次のような情報が含まれている。

・ユーザ名、ユーザのプロフィール欄、投稿日時、返信（reply）ならば誰宛か、RT数など

以後、上述のデータをくまモンデータと呼ぶ。

3　くまモンデータの基本情報

くまモンデータの基本的な情報は次のとおりである。

・投稿の総数は、66，214である。
・くまモンデータを構成するユーザの総数は、27，935ある。

当該期間（2021年9月21日～2022年1月5日）の投稿数を横軸に、その数の投稿を行ったユーザの人数を縦軸にとったヒストグラムを作ると図1のようになる（縦軸は対数）。図1が示すように、当該期間に1～2個の投稿をするユーザが圧倒的に多いものの、この期間に403ポストしたユーザもおり、ある程度ロングテール的な様相を見せている。

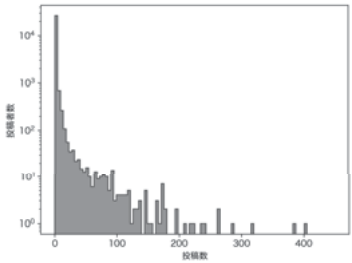

図1　投稿数とユーザ数の連関

続いて、基本的な形態素解析結果について述べる。本研究では、次のようなセッティングで投稿内容の形態素解析をおこなった。

・形態素解析器として、mecabを使用した。
・システム辞書は、mecab-ipadic-NEologdnを使用するとともに（佐藤・橋本・奥村 2017）、独自のくまモンデータに含まれる独自の表現（たとえば、「おはくま」など）を含んだ個人辞書を使用した。
・形式名詞など、一般的なstop-wordについては除外した。

このようなセッティングのもとでくまモンデータに含まれる投稿内容に対して形態素解析を行った結果、投稿に出現する名詞、動詞、形容詞、間投詞の上位10語として表1が得られた。

表１　単語の出現数

単語	出現数
くまモン	66619
さん	8363
今日	6928
くま	5053
熊本	5052
おはくま	4399
モンちゃん	3919
笑	2893
ありがとう	2875
ちゃん	2758

4　BERTopic を用いた投稿からのトピック抽出

4．1　コア・ユーザとライト・ユーザ

前節で述べたように、くまモンデータは、極めて活発にポストを行うユーザと1、2度の言及に留まるユーザが併存しているのが特徴である。ここで、全投稿数をおおまかに3分割し、投稿数とユーザ数の関係を見てみると、次の表2が得られた（カッコ内が当該ユーザの投稿数である）。

表２　投稿数と投稿者数

	投稿数 1 回	投稿数 2 回以上 12 回未満	投稿数 12 回以上
ユーザ数	21298	6087	550
投稿数	21298 投稿	18700 投稿	26216 投稿

このように、約2％にすぎない５５０ユーザが、全投稿のおよそ40％弱をポストしている。これらの活発なユーザを、これ以降、コア・ユーザと呼ぶことにしよう。

反対に、圧倒的多数を占める（76％程度）当該期間に「くまモン」という語を含む投稿を1回のみ行った投稿者をライト・ユーザと呼ぶことにしよう。

これ以降、コア・ユーザとライト・ユーザの投稿からBERTopic を用いてトピックを抽出し、比較分析を行う。

4．2　BERTopic とは

前節で述べたコア・ユーザは、どのような投稿を行っているのだろうか。

次節ではこの点を、いわゆる深層学習にもとづく自然言語処理をリードしたと言える言語モデルの一種であるBERT（Bidirectional Encoder Representations from Transformers）をトピック分析に応用したBERTopic を利用した分析を提示するが（Devlin etal 2019; Grootendorse 2022）、本節では、分析ツールであるBERTopic の基本的なフレームワークを提示する。

BERTopic は、おおまかに言えば、次の3段階のステップを踏むことで文章のトピックを抽出する。

1. 文章を埋め込みベクトルに変換し（sentence-BERTが利用される）、類似度を計算できるようにする。
2. 得られたベクトルの次元削減を行った上で（UMAPを使用する）、クラスタリングを行い（類似した内容を持つ文章をグループ化する）、それによって得られたクラスタを一つのトピックとみなす。なお、クラスタリング手法に関しては、HDBSCAN、k-Meansなどが選択可能である。
3. c-TF-IDF cもしくはBM25に基づき、各クラスタの特徴語を抽出する。これが、各トピック（先の各クラスタ）を代表する主要語である。

BERTopicは従来のLDA（Latent Dirichlet Allocation）に代表されるようなトピック分析（Beil el al. 2003）に比べ、語彙や文の意味を分散表現として利用しながら、意味の類似性に基づくクラスタリングを行っており、LDAに比べて高い精度でトピック抽出ができることが報告されている（川原，2023）。

BERTopicは極めて多彩なパラメータを持つが、本研究では、次のようなセッティングでトピック抽出を行った。

1. sentence-BERTの日本語モデルとして、sentence-bert-base-ja-mean-tokens-v2を利用した[3]。
2. 次元削減には、UMAPを、クラスタリングには、HDBSCANを使用した（ハイパーパラメータについてはデフォルト値を利用した）。
3. 特徴語検出には、文章長の影響を受けにくいとされるBM25を利用した。

4．3　コア投稿者言説からのトピック抽出

前節で述べたようなセッティングのもとで、コア・ユーザの投稿からなる文書集合に対して、BERTopicを用いてトピック抽出したところ、約半数（１３，０７５投稿）の投稿が分類不能となったものの、18個のトピックが得られた。表３は、当該トピックに分類される投稿数の上位５グループを表にまとめた

ものである。なお、表3における「ラベル」名は、本稿の筆者がトピック主要語をもとに解釈した結果であり、BERTopic からは得られない要素である。

紙幅の制約のため、トピックへの貢献度について Topic 0 のみグラフ化すると、図2のようになる。

表3・図2から理解されるように、分類可能であった文書集合の半数が、くまモンを擬人化した「くまモンへの手紙」であることは特筆すべき現象と思われる。

同様に Topic 1 は、くまモン体操を Topic 2 はくまモンスクエアでの出演に関するクラスタであり、コア・ユーザが単なる図像ではなく、「元気よく動き回り現実世界に身体を持つ」くまモンについて数多く言及していることが理解される。

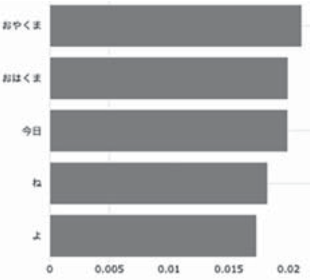

図2　トピックへの貢献度

表3　BERTopic によるトピック抽出（コア・ユーザ）

Topic	分類された投稿数	トピック主要語	ラベル
0	12688	おやくま・おはくま	くまモンへの個人的な手紙
1	76	体操・ダンス	くまモン体操
2	70	11 月・スクエア	くまモン出演情報
3	59	ぼっち・ジョージ	クリスマス
4	29	合唱・絵描き歌	くまモンの絵描き歌

4. 4　ライト投稿者言説からのトピック抽出

前節と同様の手法でライト・ユーザの投稿から BERTopic を用いてトピック抽出したところ、分類不可能な投稿がコア・ユーザの場合と同じくおよそ半数あった。抽出されたトピックは、１４４個あり（有意味なクラスタが１４４個あったと言い換えられる）、コア・ユーザによる投稿集合と異なり、多様なトピックが出現していることが理解される。次の表4は、当該トピックに分類される投稿数の上位5グループを表にまとめたものである。

これらのトピック集合は、前節のコア・ユーザのものと大きく

表4　BERTopic によるトピック抽出）（ライト・ユーザ）

Topic	分類された投稿数	トピック主要語	ラベル
0	3658	かわいい・好き	くまモンへの愛着
1	552	熊本城・熊本	熊本県の象徴
2	541	グッズ・買った	モノとしてのくまモン
3	454	美味しい・クレープ	パッケージアイコンとしてのくまモン
4	153	歌・fns 歌謡祭	TV の中のくまモン

異なっている。それは、身体性の顕現性に関する差である。ライト・ユーザのくまモンへの愛着を示すTopic 0 に含まれる投稿を見てみると、身体性が感じられないものも多い（感じられるものももちろんあり、コア・ユーザと同様に「くまモンへの手紙」と解釈できるものも数多くある）。

- マスクしてサムズアップしてるくまモン良すぎる
- ひこニャン、くまモンは双璧
- くまモン笑笑たしかに強そう

また、Topic 1 から Topic 3 までは、身体を持たないアイコンとしてのくまモンがトピックとなっている投稿と解釈できる。ただし、Topic 4 については、TVで見るという制約はあるものの、身体を持ち動き回るくまモンをトピックとする投稿集合と解釈できる。

このように、Topic 4 という例外はあるものの、ライト・ユーザによる投稿に出現するくまモンは身体性があまり感じられず、一種の「アイコン」にとどまっていると解釈するのが妥当であろう。

5　コア・ユーザとライト・ユーザの投稿集合の差異について

先述のように、コアユーザは「くまモンへの手紙」という性質を持つ投稿が圧倒的に多く、身体を持ち現実世界に定位するくまモンへの愛着を表現するという傾向を持っている。コア・ユーザも最初の段階ではライト・ユーザであったであろうということを考え合わせると、アイコンとしてのくまモンを愛好するライトなファンから成長した結果、コア・ユーザは、くまモンに身体性を見出すようになったと分析するのが妥当であろう。

ここで問題となるのが、ファンの「成長」を支える環境をどのように構築するかという点である。これに関しては、佐藤（2018）の議論されているファンダムへのアプローチが有効であると考えている。

6　おわりに

本レポートでは、Twitter における「くまモン」を含む投稿を BERTopic によって分析することを通じて、くまモン・ファンの成長過程を明らかにした。

注

(1) 畠山(2023)では、この商用tweetの削除が充分でなかったため、本稿では徹底した。

(2) neologdnは、https://pypi.org/project/neologdn/からダウンロードできる。

(3) https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

参考文献

Biel, D.M., Ng, A.Y., and Jordan, M.I.(2003). " Latent Dirichlet allocation." Journal of Machine Learning Research, 3, pp. 993-1022.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171 – 4186, Minneapolis, Minnesota. Association for Computational 6 Linguistics.

Grootendorst, M. (2022). "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." arXiv preprint arXiv:2203.05794.

畠山真二(2023)「Twitterにおけるくまモンの評判分析―ファンコミュニティの視点から―」尚絅語文第12号, 7-16

堀田治(2015)「超高関与消費のマーケットインパクト」『AD STUDIES』, 51, 15-20

川原一修(2023)「トピックモデルによる市場変動要因の抽出」言語処理学会第29年次大会(NLP 2023), pp.2190-2194, 言語処理学会

佐藤一誠(2015)『トピックモデルによる統計的潜在意味解析』コロナ社

佐藤尚之(2018)『ファン・ベース』ちくま新書

佐藤敏紀・橋本泰一・奥村学(2017)「単語分かち書き辞書mecab-ipadic-NEologdの実装と情報検索における効果的な使用法の検討」言語処理学会第23回年次大会(NLP 2017), pp.875-878, 言語処理学会