TopiCLEAR：文埋め込み表現のクラスタリングに基づくトピック抽出技術

藤田葵 [1]　山本泰智 [1]　中山悠理 [1]　小林亮太 [1,2]
[1] 東京大学大学院 新領域創成科学研究科
[2] 東京大学 数理・情報教育研究センター
{7213097981,r-koba}@edu.k.u-tokyo.ac.jp

概要

ソーシャルメディア上の短文テキストから主要な話題（トピック）を自動的に抽出することは，社会の動向を把握する上で重要な分析技術の 1 つである．しかし，従来手法であるトピックモデルは，短文における語彙共起の乏しさにより性能が低下することが知られている．本研究では，Sentence-BERT (SBERT) により得られる文埋め込みをクラスタリングすることでトピック抽出を行う手法 TopiCLEAR を提案する．4 つのデータセットを用いた評価により，提案手法 TopiCLEAR が人手でアノテートされたトピックと類似したトピックを抽出できることを示した．また，TweetTopic データセットに適用することで，TopiCLEAR が抽出したトピックの解釈性を検討した．

1　はじめに

近年，X (旧 Twitter)，Facebook，Reddit などのソーシャルメディアの普及により，人々はニュースや事件に対する意見や感想をリアルタイムに発信できるようになった．ソーシャルメディアの投稿（ポストやツイートなど）を分析することで，社会問題や政治・時事的話題に関する人々の関心の内容やその時間推移を把握することが可能となっている [1, 2]．特に，ソーシャルメディア上から大量に得られる短文テキストから主要な話題（トピック）を自動的に抽出することは，データ駆動的に社会の動向を把握する上で重要な分析技術の 1 つである．

Latent Dirichlet Allocation (LDA) [3] に代表されるトピックモデルは，単語分布に基づき文書の潜在構造を推定することによってトピックを抽出する手法である．しかし，短文テキストでは共起統計が乏しいことに加え，つづりの不統一やくだけた表現も多いため，単語頻度に基づく手法の性能低下が指摘されている [4]．近年，事前学習言語モデルの登場により，テキスト分析の方法論は大きく進化した．特に，Sentence-BERT (SBERT) [5] は，文の意味的類似性を定量化できる点で画期的である．SBERT は文章を高次元ベクトルに変換し，コサイン類似度に基づいて比較することで，質問応答や検索といったタスクにおいて高い性能を示している．このように文の意味情報を捉える応用事例における有用性は，SBERT がトピック抽出の課題に対しても有効である可能性を示唆している．

本研究では，SBERT により得た埋め込みベクトルをクラスタリングすることによってトピック抽出を行う手法 TopiCLEAR を提案する．提案手法 TopiCLEAR および実験結果の詳細は文献 [6] を参照されたい．ソースコードは GitHub にて公開している[1)]．本研究の貢献は以下の 3 点である．

1. 文埋め込みのクラスタリングに基づくトピック抽出手法 TopiCLEAR を提案する．
2. TopiCLEAR が抽出したトピックとアノテーションに基づき人手で整理されたトピックとの比較評価を行った．
3. TweetTopic データセットを用い，抽出されたトピックの解釈性を定性的に検討した．

2　提案手法 TopiCLEAR

本研究では，多数の文書に対して，テキスト情報だけから K 個のトピックに分類する手法 TopiCLEAR を提案する．提案手法は，

1. 文書のベクトルへの埋め込み．
2. 埋め込みベクトルのクラスタリング．

の 2 ステップに基づく．以下では，各ステップにつ

1）https://github.com/aoi8716/TopiCLEAR

いて説明する．手法の詳細については 文献 [6] を参照されたい．

2.1 文書のベクトルへの埋め込み

まず，Sentence-BERT (SBERT) [5] を用いて文書をベクトルに変換する．文書全体の内容を表現した 1 つのベクトルを得ることで，単語数が異なる文書群を同じベクトル空間に埋め込むことができる．SBERT は，大規模データなどから事前学習されたモデルであり，文書埋め込みの方法としてよく知られている．本研究では Hugging Face の `all-MiniLM-L6-v2` [2] を用いて，文書を 384 次元のベクトルに変換した．

2.2 埋め込みベクトルのクラスタリング

得られた埋め込みベクトルをクラスタリングすることによって K 個のトピックに分類した．以下では，本研究で用いたクラスタリング手法を説明する．

前処理を行うことにより，予備的なトピックの分類結果を得た．具体的には，得られた埋め込みベクトル群に主成分分析 (PCA) を適用することにより，384 次元を 64 次元に削減し，その後，混合ガウスモデルを用いてクラスタリングを行った．

次に，Adaptive Dimension Reduction (ADR) [7] を適用することによって，前処理で得られた分類結果を洗練させる．ADR は，手順 1) 分離精度を向上させるための次元削減，手順 2) 混合ガウス分布に基づくクラスタリング，の 2 つの手順を繰り返す．手順 1) では，線形判別分析を 現在の分類結果 (トピックラベル) と次元削減されたベクトル (64 次元) に適用することによって，文書ベクトルの次元を $K-1$ 次元に削減する．ただし，K はトピック数であり，線形判別分析で得られる最大の次元が $K-1$ 次元であることが知られている．手順 2) では，得られた $K-1$ 次元のベクトルを，混合ガウス分布を用いてクラスタリングを行う．得られた分類結果がこれまで得られていたものと同じであれば，この結果を最終的な分類結果とした．そうでない場合は手順 1) に戻った．また，最大繰り返し回数 (10 回) に到達した場合にも得られた結果を最終的な分類結果とした．

2) https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

表 1 データセットの統計量

データセット	トピック数	平均語数	文書数
20News	20	135.2	18806
AgNewsTitle	4	5.2	119794
Reddit	5	34.6	37993
TweetTopic	6	12.3	6997

3 実験設定

3.1 データセット

本研究では，テキストデータと人手によって付与されたトピックラベルを含む 4 種類のデータセットを用いた．各データセットの統計量を表 1 に示し，概要を以下にまとめる．20News [8] はニュース記事からなるデータセットであり，ニュース記事を医学，銃問題，中東政治など，20 種類のトピックに分類された結果が付与されている．AgNewsTitle [9] はニュース記事のタイトルを対象としたデータセットであり，世界，科学・技術，スポーツ，ビジネスの 4 種類のトピックに分類されている．Reddit [10] は Reddit 上の投稿から構成され，subreddit に基づいて，パソコン，ニュース，映画，アメフト，人間関係の 5 種類のトピックに分類されている．TweetTopic [11] は Twitter 上の投稿を対象としたデータセットであり，5 名のアノテータによって，ビジネス，ポップカルチャーなど，6 種類のトピックに分類されている．

3.2 比較手法

本研究では，提案手法 TopiCLEAR を，以下の 7 つの既存手法と比較した．トピックモデルとして，LDA[3]，BTM[12]，ProdLDA[13] を採用した．次に，埋め込み情報を利用するトピックモデルとして，ETM[14]，CTM[15] を用いた．最後に，言語モデルに基づく手法として，BERTopic[16]，生成 AI (Google Gemini: Gemini 2.5 Flash) を採用した．Bag-of-Words に基づく手法 (LDA，BTM，ProdLDA，ETM，CTM) では，ストップワード除去や低頻度語の除外などの前処理を施した．なお，生成 AI はトピック抽出を直接行うことができなかったため，あらかじめトピック名（医学，銃問題など）を与えたゼロショット分類を行った．この比較は公正とは言えず，生成 AI に有利なものとなっているが，参考結果として示

すことにした．

3.3 評価指標

評価指標として，クラスタリング性能を測る Adjusted Rand Index（ARI）を採用し，人手による分類ラベル（トピック）との類似度を評価した．ARI は 2 つの分類結果が類似しているほど大きな値をとる指標であり，両者が完全に一致する場合に最大値 1 をとる．また，ランダムな分類と同程度の類似度の場合には 0 となる．

トピックモデルの先行研究では，coherence 指標 (UCI，NPMI など) を用いた評価が広く行われてきた．しかし，coherence 指標はラベルノイズに対する頑健性を持たないことが指摘されている [6]．そのため，本研究では coherence 指標による評価は行わなかった．

4 実験結果

本研究では，人手による分類結果との類似性を評価する定量的評価と，実データセットに適用した結果を議論する定性的評価の 2 つの観点から提案手法の解釈性を調べ，既存手法との比較を行った．

4.1 定量的評価

4 つのデータセットに対し，提案手法 TopiCLEAR を含む 8 つの手法でトピック抽出を行った．そして，Adjusted Rand Index（ARI）を用いて，人手による分類結果と各トピックモデルが得た分類結果との類似性を評価した (表 2)．

提案手法 TopiCLEAR は，全てのデータセットにおいて最も高い ARI を示した．この結果は，TopiCLEAR が抽出したトピックが，人手により抽出されたトピックと最も類似していることを示唆している．標準的手法である LDA は，20News では人手によるトピックと比較的類似した結果 (ARI: 0.215) を示した一方，短文データ（AgNewsTitle）や口語的なデータ（Reddit，TweetTopic）では，ARI が 0 に近い値をとった．

4.2 定性的評価

ここでは，代表的な既存手法である LDA と提案手法 TopiCLEAR を用いて，TweetTopic データセットから抽出されたトピックを比較する．まず，両手法が抽出したトピック間の類似性をコサイン類似度によって評価した．両手法が抽出したトピック同士の類似度は低く（最大でもコサイン類似度 0.14），異なる傾向のトピックが抽出されていることが確認された．

表 2 Adjusted Rand Index（ARI）による抽出されたトピックの評価. 1 番高いスコアを**太字**, 2 番目に高いスコアをイタリック体で示した.

手法	20News	AgNewsTitle	Reddit	TweetTopic
LDA	0.215	0.035	0.139	0.024
BTM	*0.217*	*0.354*	*0.163*	0.086
ProdLDA	0.165	0.212	0.077	0.105
ETM	0.066	0.131	0.125	0.008
CTM	0.179	0.242	0.152	0.125
BERTopic	0.027	0.005	0.096	0.026
生成 AI	0.092	0.042	0.058	*0.167*
TopiCLEAR	**0.446**	**0.529**	**0.418**	**0.307**

次に，両手法が抽出したトピックと人手によりアノテーションされたトピックを比較した（図 1）．TopiCLEAR により抽出された Topic 1 および Topic 3 は，人手による “sports & gaming” に対応しており，スポーツ関連の投稿が多く含まれていた．また，Topic 5 は “pop culture” に対応し，音楽関連の投稿が多く見られた．Topic 6 では複数の話題が混在していたものの，家族への挨拶など個人的なメッセージが中心であり，トピック内容の解釈は可能であった．一方，LDA が抽出した全てのトピックは，人手による複数のトピックが混在しており，トピックの内容を解釈することは困難であった．

5 おわりに

本研究では，SBERT による文書埋め込みをクラスタリングすることでトピックを抽出する手法 TopiCLEAR を提案した．定量的評価の結果，提案手法 TopiCLEAR は，生成 AI を含む 7 つの既存手法と比較して，人手によって抽出されたトピックとより類似したトピックを抽出できることを示した．さらに，TweetTopic データセットを用いた定性的分析により，TopiCLEAR によって解釈しやすいトピックが得られることを確認した．以上の結果から，TopiCLEAR は，大量のテキストデータからトピックを抽出する作業を効率化する手法として有用と考えられる．今後は，日本語 Twitter データなど多様な言語へ適用範囲を拡張しつつ，TopiCLEAR の有用性をさらに検証していきたい．

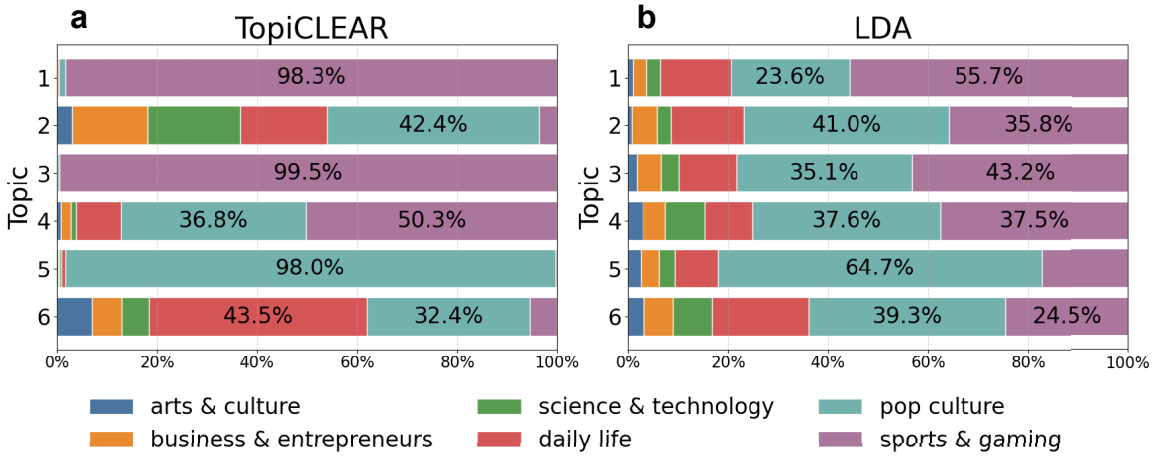

図 1　人間が抽出 (アノテーション) したトピックの構成比率. a. 提案手法 TopiCLEAR, b. 既存手法 LDA.

参考文献

[1] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Computational social science. **Science**, Vol. 323, No. 5915, pp. 721–723, 2009.

[2] Ryota Kobayashi, Yuka Takedomi, Yuri Nakayama, Towa Suda, Takeaki Uno, Takako Hashimoto, Masashi Toyoda, Naoki Yoshinaga, Masaru Kitsuregawa, and Luis E. C. Rocha. Evolution of public opinion on covid-19 vaccination in japan: Large-scale Twitter data analysis. **Journal of Medical Internet Research**, Vol. 24, No. 12, p. e41928, 2022.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. **Journal of Machine Learning Research**, Vol. 3, pp. 993–1022, 2003.

[4] Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. Short text topic modeling techniques, applications, and performance: A survey. **IEEE Transactions on Knowledge and Data Engineering**, Vol. 34, No. 3, pp. 1427–1445, 2022.

[5] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing**, pp. 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics.

[6] Aoi Fujita, Taichi Yamamoto, Yuri Nakayama, and Ryota Kobayashi. Topiclear: Topic extraction by clustering embeddings with adaptive dimensional reduction. **arXiv: 2512.06694**, 2025.

[7] Chris Ding and Tao Li. Adaptive dimension reduction using discriminant analysis and k-means clustering. In **Proceedings of the 24th International Conference on Machine Learning**, pp. 521–528, New York, NY, USA, 2007. Association for Computing Machinery.

[8] Ken Lang. Newsweeder: Learning to filter netnews. In **Proceedings of the 12th International Conference on Machine Learning**, pp. 331–339. Morgan Kaufmann, San Francisco, CA, USA, 1995.

[9] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In **Advances in Neural Information Processing Systems 28**, pp. 649–657, Red Hook, NY, USA, 2015. Curran Associates, Inc.

[10] Stephan A. Curiskis, Barry Drake, Thomas R. Osborn, and Paul J. Kennedy. An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. **Information Processing & Management**, Vol. 57, No. 2, p. 102034, 2020.

[11] Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Vítor Silva, Leonardo Neves, and Francesco Barbieri. Twitter topic classification. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 3386–3400, Gyeongju, Republic of Korea, 2022. International Committee on Computational Linguistics.

[12] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In **Proceedings of the 22nd International Conference on World Wide Web**, pp. 1445–1456, New York, NY, USA, 2013. Association for Computing Machinery.

[13] Akash Srivastava and Charles A. Sutton. Autoencoding variational inference for topic models. In **Proceedings of the 5th International Conference on Learning Representations**, 2017.

[14] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. Topic modeling in embedding spaces. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 439–453, 2020.

[15] Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. Cross-lingual contextualized topic models with zero-shot learning. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 1676–1683. Association for Computational Linguistics, 2021.

[16] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. **arXiv: 2203.05794**, 2022.