文書埋め込みとクラスタリングを組み合わせたトピック分析手法の提案

藤田葵 [1]　中山悠理 [1]　山本泰智 [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

トピック分析とは，文書群を内容によってグループに分類する技術であり、大規模テキストデータの分析に有用である．トピック分析の代表的手法として LDA (Latent Dirichlet Allocation) があり，政治学，文献計量学，ソーシャルメディア分析などの分野へ応用されてきた．一方，LDA を単語数が少ない文書に適用すると，人間が解釈しやすい分類結果を得ることが難しいという問題があった．本研究では，文書埋め込みとクラスタリングを組み合わせたトピック分析手法を提案する．4 つのデータセットを分析した結果，既存手法に比べ，提案手法はより人間に近いトピック分類を行うことが示された．

1　はじめに

インターネットをはじめとする情報技術の発展により，ニュース記事やソーシャルメディア投稿，学術論文など，多種多様なテキストデータが日々生成されている．これらのデータは，社会やビジネス，学術研究において重要な情報源となる一方，その膨大な量から人手による分析は現実的ではなく，自動化された分析技術の必要性が高まっている．

こうした背景のもとで，トピック分析はテキストデータの大規模な解析を可能にする手法として注目されている．トピック分析とは，文書群を内容によってグループに分類する技術であり、大規模テキストデータの分析に有用である．代表的な手法として LDA (Latent Dirichlet Allocation) [1] が挙げられる．LDA は単語の出現頻度分布に基づき，トピック分析を行う統計的手法であり，政治学，文献計量学など幅広い分野に応用されてきた．

特に，ソーシャルメディアから得られるテキストデータの分析は，東日本大震災 [2, 3] など災害時における情報拡散の動向や選挙 [4] や新型コロナワクチン [5] などの社会的課題に対する人々の認識を把握するための有力な手段となっている．しかし，ソーシャルメディアデータには投稿が短文であること，誤字やスラングなどの口語的表現が多いなどの難しさがある．実際に，LDA は，短い文書のデータセット [6] やソーシャルメディアデータ [7] のトピックを分類することが難しいことが指摘されている．

本研究では，文書の埋め込み表現とクラスタリングを組み合わせたトピック分析手法を提案する．そして，ニュース記事, 短文のニュースタイトル, ソーシャルメディアデータなど 4 つのデータセットを用い，提案手法の性能を評価し，既存手法と比較を行った．

2　提案手法

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

- 文書をベクトル空間に埋め込む．
- 得られたベクトルをクラスタリングする．

の 2 段階に基づく．以下では，各段階について説明する．

2.1　ベクトル空間への埋め込み

提案手法では，Sentence-BERT [8] を用いて文書をベクトルに変換する．個々の単語ではなく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．Sentence-BERT は，既存のラベル付きデータやラベル無しの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

で公開されている Sentence-BERT モデルの一種である all-MiniLM-L6-V2 [1] を用いて，それぞれの文書を 384 次元のベクトルに変換した．

2.2 クラスタリング

ここでは，文書から得られた埋め込みベクトルから，K 個のトピックに分類するのに用いたクラスタリング手法を説明する．本手法では，前処理を行うことで，予備的なトピック分類の結果を得る．その後，Adaptive Dimension Reduction [9] を適用することにより分類結果を洗練させていった．

まず，前処理について説明する．文書群から得られた埋め込みベクトル集合に主成分分析 (PCA) を適用することにより，384 次元から 64 次元に次元を削減した．次に，次元削減されたベクトル集合を混合ガウス分布でフィットする．つまり，K 個のガウス分布の平均 $\boldsymbol{\mu}_i$，共分散行列 Σ_i，混合比 π_i を推定した．そして，それぞれの文書に対応する次元削減されたベクトル $\boldsymbol{x}_{PCA}$ に対して事後確率 $p[z|\boldsymbol{x}_{\mathrm{PCA}}]$ (ただし，z はトピック番号: $1, 2, \cdots K$) を計算し，事後確率が最大になるトピックを予備的な分類結果とした．混合ガウス分布によるクラスタリングは，python パッケージ scikit-learn [10] を用いて行った．

次に，Adaptive Dimension Reduction (ADR) について説明する．ADR では，手順 1) 分類性能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ことにより，前処理で得られた結果をさらに洗練させる．手順 1) では，線形判別分析を適用することにより，分類に適した低次元空間を求める．線形判別分析 [11] を用いると，クラスラベルが既知のデータに対して，分類に適した次元削減を行うことができる．ここでは，クラスラベルとして現在のトピック分類結果を用いて，線形判別分析を適用することにより，文書ベクトルの次元を 64 次元から $K-1$ 次元に削減した． 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 [11]．手順 2) では，次元削減で得られた $K-1$ 次元の文書ベクトルに対して，混合ガウス分布に基づくクラスタリングを適用することにより，新しいトピック分類結果の結果を得る．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るか，この結果を最終的な分類結果とする．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．

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

表 1 本研究で用いたデータセット

データセット	平均単語数	文書数	トピック数
20News	135.2	18806	20
AgNewsTitle	5.2	119794	4
Reddit	34.6	37933	5
TweetTopic	12.3	6997	6

3 実験

3.1 データセット

本研究では 4 種類のデータセットを用いた:

1. 20News [12]: ニュース記事のデータセットであり，20 種類のトピック (医学，銃問題，中東政治など) に分類された結果も付与されている．
2. AgNewsTitle [13]: ニュース記事のタイトルのデータセットであり，4 種類のトピック (世界，科学/技術，スポーツ，ビジネス) に分類された結果も付与されている．
3. Reddit [7]: Reddit 上での投稿のデータセットであり，5 種類のトピック (パソコン，ニュース，映画，アメフト，人間関係) に分類された結果も付与されている．
4. TweetTopic [14]: Twitter 上での投稿のデータセットであり，6 種類のトピック (芸術と文化，ビジネス，ポップカルチャー，日常生活，スポーツとゲーム，科学技術) に分類された結果も付与されている．

これらのデータセットには，人間によって分類された結果も付与されているため，トピック分析手法による分類結果と比較し，類似度を評価することが可能となる．

20News はトピックモデルに関する先行研究で広く用いられてきたデータセットである．20News, AgNewsTitle はインターネット上のニュース記事のデータであるため文語的特徴を持ち，Reddit, TweetTopic はソーシャルメディア上で投稿されたデータであるためスラングや口語表現を多く含む．表 1 は，これらのデータセットの平均単語数，文書数，トピック数を示したものである．

3.2 評価指標

本研究では，テキストデータから抽出されたトピックの性能を評価するための指標として，AMI (Adjusted Mutual Information) [15] を用いた．AMI は 2 つの分類結果の類似度についての指標である．AMI を使えば，人間による文書の分類結果と人工的に分類された結果の類似度を測ることができる．AMI は 1 以下の値を取る指標であり，2 つの分類結果が完全に一致するときにのみ，最大値 1 をとる．また，類似度がランダムに分類した場合と同程度である場合には AMI は 0 をとる．AMI は Python ライブラリ scikit-learn [10] を用いて計算した．

3.3 比較に用いた手法

以下では，本研究で用いた 6 つの既存手法の概要を説明する．

1. **LDA (Latent Dirichlet Allocation)** [1]: 単語出現をいくつかのトピックの単語分布の混合によって表現するモデルを用いる手法．
2. **BTM (Biterm Topic Model)** [6]: LDA の派生で，トピックを単語ではなく単語ペアの共起パターンの分布であるとしたモデルを用いる手法．
3. **ProdLDA** [16]: LDA をやや複雑化したモデルを用いた手法．LDA 系のモデルに基づく推定に初めてニューラルネット (Variational AutoEncoder) を取り入れた手法として知られている．
4. **ETM (Embedded Topic Model)** [17]: 単語埋め込みに基づいてトピック内の単語分布が生成される確率モデルによる手法．
5. **CTM (Contextualized Topic Model)** [18]: SBERT の文書ベクトルを用いて ProdLDA の精度を向上させる手法．
6. **BERTopic** [19]: 提案手法と同様に SBERT を使って，ベクトル埋め込みを行い，得られたベクトルを HDBSCAN [20] を用いてクラスタリングすることでトピック分類を行う．

LDA, BTM, ProdLDA は単語分布に基づく手法であり，ETM は単語埋め込み，CTM, BERTopic は SBERT による文書埋め込みも用いてトピック分類を行う．また，BERTopic と提案手法を除く 5 つの手法においては，単語情報を基にトピックに分類するため，頻出単語（例: "the"，"is"，"and" など）や記号などを削除する前処理を行った．

表 2 人間のトピック分類結果との類似度 (AMI). **太字**は最も人間の分類結果と類似した手法である．

Method	20News	AgNewsTitle	Reddit	TweetTopic
LDA	0.378	0.038	0.191	0.038
biterm	0.397	0.350	0.225	0.169
ProdLDA	0.282	0.233	0.097	0.143
ETM	0.197	0.127	0.184	0.009
CTM	0.319	0.259	0.116	0.166
BERTopic	0.405	0.199	0.237	0.231
提案手法	**0.601**	**0.492**	**0.452**	**0.403**

3.4 人間によるトピック分類との類似性評価

表 2 は，4 つのデータセットについて，AMI を用いて，人間の分類結果とトピック分析手法が分類した結果の類似性を評価した結果である．提案手法は，4 種類のデータセット全てに対して，6 つの既存手法よりも高い AMI を示した．この結果は，提案手法のトピック分類結果が人間に最も近いことを示している．

LDA は，20News においては一定の性能（AMI=0.378）を示したが，AgNewsTitle, TweetTopic のように平均単語数が少ないデータセットでは AMI が 0 に近い値を取り，性能が著しく悪化した．この原因としては，LDA は単語分布を利用するため，短文データでは学習が難しいことが考えられる．BERTopic は，いずれのデータセットでも比較的高い性能を示した．AgNewsTitle, TweetTopic の AMI は LDA を大きく上回っている．しかし，提案手法の方が高い AMI の値になった．クラスタリング手法の違いが性能の差を生んでいると考えられる．

3.5 文書長がトピック分類性能に与える影響

前節から，AgNewsTitle や TweetTopic のような短文データセットでは，既存手法の LDA はトピック分類性能が著しく悪化すること，提案手法は性能の高いトピック分類を実現できることが示された (表 2)．しかし，4 つのデータセットは使っている単語や文脈も大きく異なるため，トピック分類の性能に影響を与えている可能性がある．そこで，本節では文書が比較的長い 20News データセットを用いて，文書長がトピック分類性能に与える影響を調べる．

1 つの方法として，ランダムに単語を削除することが考えられるが，文書が不自然になってしまうと

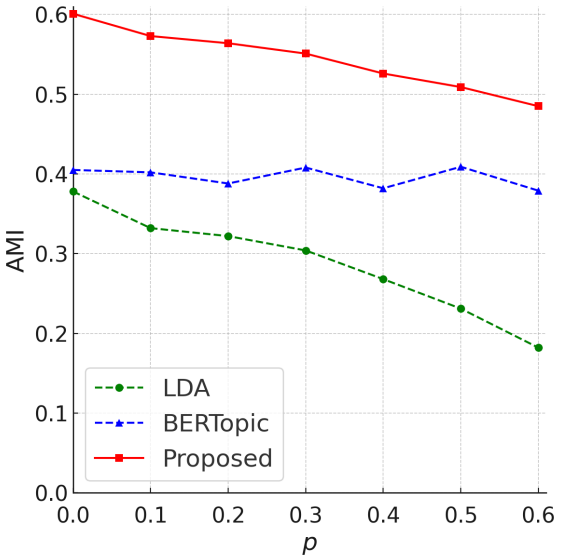

図 1 文書長削減が分類性能（AMI）に与える影響．20News データセットを用いた．

いう問題がある．そこで，本研究ではピリオド（.），感嘆符（!），疑問符（?）により文書を文に区切り，確率 p で文章を削除することによってデータセットの文書長を系統的に短くした．

図 1 は文章の削除確率 p を変えた時に，AMI を用いて人間のトピック分類結果との類似性を評価した結果である．LDA では，p が大きくなるとともに，AMI が減少していくことがわかる．一方，SBERT に基づく手法である提案手法や BERTopic では，p が大きくなっても高い性能を維持していることがわかる．この結果は，提案手法や BERTopic は短文データに対しても良い性能でトピック分類ができることを示唆している．

4 おわりに

本研究では，文書の埋め込み表現と反復的なクラスタリングを組み合わせたトピック分類手法を提案した．そして，人間によるトピックの分類結果が付与された 4 つのデータセットを用いて，性能評価を行なった．その結果，提案手法は，LDA や BERTTopic などの既存手法と比べて高い性能であること，つまり，より人間に近いトピック分類を行うことが示された．さらに，提案手法は短い文書に対しても，人間による分類と良く一致する結果を得られることを示唆する結果が確認できた．

謝辞

本研究は，JSPS 科研費 JP18K11560, JP21H04571, JP22H03695, JP23K21728, JP23K24950, AMED JP223fa627001，JST 創発 JPMJFR232O の支援を受けたものである．

References

1.BLEI, David M; NG, Andrew Y; JORDAN, Michael I. Latent dirichlet allocation. **Journal of machine Learning research**. 2003, vol. 3, pp. 993–1022.

2.TAKAYASU, Misako; SATO, Kazuya; SANO, Yukie; YAMADA, Kenta; MIURA, Wataru; TAKAYASU, Hideki. Rumor Diffusion and Convergence during the 3.11 Earthquake: A Twitter Case Study. **PLOS ONE**. 2015, vol. 10, no. 4, e0121443.

3.HASHIMOTO, Takako; SHEPARD, David Lawrence; KUBOYAMA, Tetsuji; SHIN, Kilho; KOBAYASHI, Ryota; UNO, Takeaki. Analyzing Temporal Patterns of Topic Diversity Using Graph Clustering. **The Journal of Supercomputing**. 2021, vol. 77, no. 5, pp. 4375–4388.

4.TUMASJAN, Andranik; SPRENGER, Timm; SANDNER, Philipp; WELPE, Isabell. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. **Proceedings of the International AAAI Conference on Web and Social Media**. 2010, vol. 4, no. 1, pp. 178–185.

5.KOBAYASHI, Ryota; TAKEDOMI, Yuka; NAKAYAMA, Yuri; SUDA, Towa; UNO, Takeaki; HASHIMOTO, Takako; TOYODA, Masashi; YOSHINAGA, Naoki; KITSUREGAWA, Masaru; ROCHA, Luis E. C. Evolution of Public Opinion on COVID-19 Vaccination in Japan: Large-Scale Twitter Data Analysis. **Journal of Medical Internet Research**. 2022, vol. 24, no. 12, e41928.

6.YAN, Xiaohui; GUO, Jiafeng; LAN, Yanyan; CHENG, Xueqi. A biterm topic model for short texts. In: **Proceedings of the 22nd international conference on World Wide Web**. 2013, pp. 1445–1456.

7.CURISKIS, Stephan; DRAKE, Barry; OSBORN, Thomas; KENNEDY, Paul. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. **Information Processing and Management**. 2019, vol. 57.

8. REIMERS, Nils; GUREVYCH, Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2019.
9. DING, Chris; LI, Tao. Adaptive dimension reduction using discriminant analysis and K-means clustering. **ACM International Conference Proceeding Series**. 2007, vol. 227.
10. PEDREGOSA, Fabian; VAROQUAUX, Gaël; GRAMFORT, Alexandre; MICHEL, Vincent; THIRION, Bertrand; GRISEL, Olivier; BLONDEL, Mathieu; PRETTENHOFER, Peter; WEISS, Ron; DUBOURG, Vincent; VANDERPLAS, Jake; PASSOS, Alexandre; COURNAPEAU, David; BRUCHER, Matthieu; PERROT, Matthieu; DUCHESNAY, Édouard. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**. 2011, vol. 12, no. 85, pp. 2825–2830.
11. FUKUNAGA, Keinosuke. **Introduction to statistical pattern recognition (2nd ed.)** USA: Academic Press Professional, Inc., 1990.
12. KO, Youngjoong. A Study of Term Weighting Schemes Using Class Information for Text Classification. In: **Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 2012, pp. 1029–1030.
13. ZHANG, Xiang; ZHAO, Junbo; LECUN, Yann. Character-Level Convolutional Networks for Text Classification. In: **Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1**. Montreal, Canada: MIT Press, 2015, pp. 649–657.
14. ANTYPAS, Dimosthenis; USHIO, Asahi; CAMACHO-COLLADOS, Jose; NEVES, Leonardo; SILVA, Vítor; BARBIERI, Francesco. **Twitter Topic Classification**. 2022. arXiv: 2209.09824 [cs.CL].
15. VINH, Nguyen Xuan; EPPS, Julien; BAILEY, James. Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary? In: **Proceedings of the 26th Annual International Conference on Machine Learning**. New York, NY, USA: Association for Computing Machinery, 2009, pp. 1073–1080.
16. SRIVASTAVA, Akash; SUTTON, Charles. **Autoencoding Variational Inference For Topic Models**. 2017. arXiv: 1703.01488 [stat.ML].
17. DIENG, Adji B.; RUIZ, Francisco J. R.; BLEI, David M. Topic Modeling in Embedding Spaces. **Transactions of the Association for Computational Linguistics**. 2020, vol. 8, pp. 439–453.
18. BIANCHI, Federico; TERRAGNI, Silvia; HOVY, Dirk; NOZZA, Debora; FERSINI, Elisabetta. Cross-lingual Contextualized Topic Models with Zero-shot Learning. In: MERLO, Paola; TIEDEMANN, Jorg; TSARFATY, Reut (eds.). **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**. Online: Association for Computational Linguistics, 2021, pp. 1676–1683.
19. GROOTENDORST, Maarten. **BERTopic: Neural topic modeling with a class-based TF-IDF procedure**. 2022. arXiv: 2203.05794 [cs.CL].
20. CAMPELLO, Ricardo JGB; MOULAVI, Davoud; ZIMEK, Arthur; SANDER, Jörg. Hierarchical density estimates for data clustering, visualization, and outlier detection. **ACM Transactions on Knowledge Discovery from Data (TKDD)**. 2015, vol. 10, no. 1, pp. 1–51.