BERTopic 文書分類器の分類安定性に関する評価手法の提案

櫻井　勇気†　　小林　亜樹††

† 工学院大学大学院工学研究科 電気・電子工学専攻　〒 163–8677　東京都新宿区西新宿 1-24-2
†† 工学院大学情報学部情報通信工学科　〒 163–8677　東京都新宿区西新宿 1-24-2
E-mail: †cm23034@g.kogakuin.jp, ††aki@cc.kogakuin.ac.jp

あらまし　BERTopic は 2020 年に提案された，埋め込み表現と分類，次元圧縮，可視化手法の組み合わせを自由に設定するトピックモデリング手法であり，Transformer ベースの学習済み言語モデルと任意の分類アルゴリズムを用いた分類器を作成することが可能である．BERTopic によって作成される分類器は LDA など従来の分類手法よりも高い性能を持つとされている．しかし，予備実験の結果，分類処理毎に異なる分類クラスターを出力し，このときの分類結果の一致性は必ずしも高くない．そこで本稿では，分類器の分類結果出力の安定性を評価する枠組みを提案し，対象文書の種類などによる分類結果の安定性の違いを表現できることを目指す．提案手法は，異なる 2 文書の文書対を単位データと考え，同一の対象文書集合に対する独立した相異なる分類処理結果において，文書対の分類結果が同一となるか否かの比率として分類の安定性を表すものである．いくつかの評価対象文書集合において提案評価手法に関する評価を行い，指標としての有効性などを検証する．

キーワード　テキスト分類，LLM，言語モデル，文書要約，情報抽出

1　はじめに

大規模言語モデルで注目を浴びる Transformer を文書分類に応用した手法として，2020 年に BERTopic [1] が提案されている．BERTopic は自由に文書分類器を作成する手法であり，文書を分類するにあたって必要な埋め込み表現の獲得や，次元圧縮，分類，可視化に用いる手法を自由に組み替えることが可能である．Abeer らは Transformer をベースとした複数の自然言語処理モデルと密度ベースのクラスタリングアルゴリズムである HDBSCAN [2] を組み合わせた分類器による文書分類を行い，LDA や NMF といった他の文書分類手法による分類結果と比較して良好な結果が得られたと報告している [3]．HDBSCAN はクラスタの数や形状を柔軟に変更できるが，k-means のような従来のアルゴリズムは事前のクラスタ数の決定を必要とする。そのため、教師無し学習を用いて未知の文書を分類していく場合，HDBSCAN は適したアルゴリズムであると考えられる．しかし HDBSCAN を採用した分類器を用いる場合分類を行うたびにクラスターが変化するため，複数回の分類で内容の一貫した出力を得ることは難しい．本論文では BERTopic を用いて，Sentence-bert [4]，USE [5]，Flair [6]，SpaCy [7] [8] の 4 つの異なる埋め込み表現を持つ分類器を作成し，これらを用いて同一データセットに対し分類を行うことで BERTopic が文書集合に対してどれほど安定した結果をもたらすのか調査を行い，結果を考察する．

1.1　BERTopic

本章では BERTopic による分類器の作成について説明する。BERTopic によって作成される分類器の概要を図 1[1]に示す．BERTopic の具体的な仕組みとして，最初に文書を埋め込み表現に変換する．この時文書が多くの語彙を含むほど埋め込み表現に変換された場合には高次元の表現となる．そのため次に次元圧縮を行い，埋め込み表現の次元を減らすことでより低次元化した埋め込み表現を得る．こうして得た低次元の埋め込み表現を分類（クラスタリング）することで，似た特徴を持つ文書による文書集合（クラスタ）を複数得る．最後にクラスタを一つの文書と見做してトークン化したうえで，ラベルの付与などを経て人間が理解できる表現に可視化される．このように，BERTopic は埋め込み表現，次元圧縮，分類手法，可視化手法を組み合わせて文書分類器を作成する手法の一つであるが，その特徴は各手順で用いる具体的な手段を作成者が任意に変更できる点にある．各手順は独立しているため，文書の特徴や分類に必要なコストに応じて各手順で用いる手段を変更することで，作成者が望むような分類器を作成することが出来る．

1：https://maartengr.github.io/BERTopic/algorithm/algorithm.html

2　用語定義

本章では本稿で用いる独自の用語について説明する．

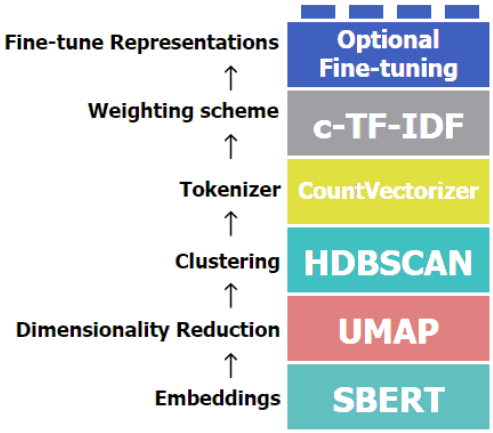
2.1　分類放棄文書

分類器は一般に複数のクラスタへ文書分類を行う．このとき，これらのクラスタのいずれにも属さないと判定された文書を分類放棄文書と呼ぶ．本稿で扱う分類器では，HDBSCAN の密度閾値未満の文書が該当する．

2.2　文書対

分類対象文書集合に含まれる相異なる 2 文書のことを文書対と呼ぶ．

. width.



width

図 1　BERTopic 分類器の概要

表 1　作成分類器

手法	種類
埋め込み表現	Sentence-bert，USE，Flair，Spacy
次元圧縮	UMAP
分類	HDBSCAN
トークン化	CountVectrizer
ラベル付与	c-tf-idf

2.3　分類一致文書対

一般に複数回の分類において，文書対に含まれる両文書の分類結果が一致する文書対のこと．各回の分類が完全に安定している場合，分類一致文書対によって全ての文書が被覆される．

3　予 備 実 験

本章では，評価手法の提案に際して行われた予備実験の結果から，BERTopic による分類器の特性を示す．

処理対象とする文書集合として，scikit-learn で配布されている 20 グループからなる英ニュース記事によるデータセット [9] と，Google Colaboratory で配布されているドナルド・トランプにまつわる英 tweet によるデータセット [10] の 2 種類を用いた．本論文では以降，英ニュース記事を分類対象とした場合は N，英 tweet を分類対象とした場合は T と文字を割り振り，埋め込み表現とともに表記している．

予備実験では表 1 に示したように，BERTopic を用いて Sentence-bert，USE，Flair，SpaCy の 4 つの異なる埋め込み表現を持つ分類器を作成し，それぞれの分類器を用いてデータセットに対して 6 回の分類を行った．1 回目の分類結果と 2-6 回目の分類結果を比較し，同一の文書が比較において同一のクラスタに分類された回数を測定し，すべての比較回数における割合を測った．

1 回目の分類結果と 2-6 回目の分類結果の比較は図 2 に示した．また 1 回目の分類結果と 2-6 回目の分類結果から得られる割合の平均を表 2 に示した．

図 2 および表 2 に示したように，BERTopic 分類器による文書分類結果は，用いられた埋め込み表現や分類の対象となった

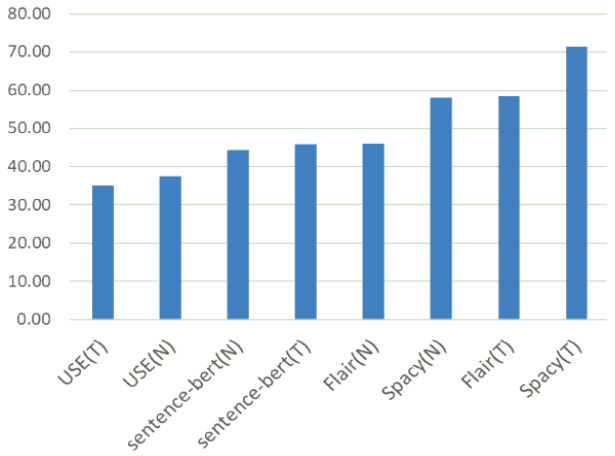


図 2　予備実験結果 (割合平均)

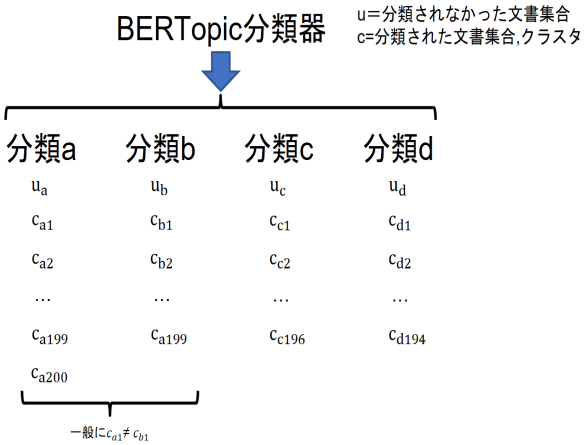


図 3　BERTopic 分類器の仕様

データセットにより変化し，その一致割合は必ずしも高くない．この仕様の原因となる BERTopic 分類器の挙動は図 2 に示したように，BERTopic 分類器は同一の文書集合を分類対象とする場合であっても，分類 a，b，c，d と分別されるような複数回の分類において分類処理毎に独立したベクトル表現や次元圧縮結果，分類クラスタを獲得するためである．よって分類ごとに得られるクラスタの総数や，各クラスタの形状は異なっており，複数回の分類において付与される番号やラベルが同一なクラスタが得られたとしても実際には異なるクラスタが生成されている．

本論ではこのように複数回の分類において異なる出力が得られる場合，それら出力の差はなるべく小さいほうが望ましいと考える．よって後述するように，文書対を用いた分類の安定性を評価する指標を提案する．

4　提 案 手 法

本稿で提案する文書分類器の安定性指標について説明する．BERTopic 文書分類器は内部で乱数要素を持つため分類結果は毎回異なる．分類器の性能評価には分類精度のほかこの分類結果の一貫性，安定性も重要であると考え，分類安定性指標を提案する．

複数の分類回の結果が完全に一致している場合，ある分類回で同一クラスタに分類された文書対は，他の分類回においても同一クラスタに分類されることとなる．そこで全文書対中，2回の分類それぞれにおいていずれも同一クラスタに分類された文書対の比率を用いて分類安定性指標とする手法を提案する．本指標は同一クラスタ数への分類において，値が大きいほど安定した（各回での結果の一致度合いが高い）分類を行っていることを示す．

まず，i 個の文書 d を含む文書集合 D を定義する．D は分類器によって処理される文書集合を想定している．

$$D = \{d_1, d_2, \ldots, d_i, \ldots d_j, \ldots\} \tag{1}$$

ある文書 d_i の独立した $r(r = \{0, 1\})$ 回目の分類クラスタ番号を

$$c(d_i, r) \tag{2}$$

で表す．

文書集合 D に属する相異なる 2 文書からなる文書対 p_k は，式 3 のように定義される．ここで，k は，文書対を識別する番号で，文書対集合 P の要素数 $|P|$ を用いて $1 \leqq k \leqq |P|$ である．$|P|$ は D を用いて式 4 で計算される．

$$P = \{p_k \mid p_k = (d_i, d_j) \in D^2, \quad i < j\} \tag{3}$$

$$|P| = \frac{1}{2}|D|(|D| - 1) \tag{4}$$

0 回目の分類において文書 d_i，d_j が同じクラスタに分類された場合，それらの文書からなる文書対 p_k は式 5 のようになり，同様に 1 回目の分類においても文書 d_i，d_j が同じクラスタに分類された場合，それらの文書からなる文書対 p_k は式 6 のようになる．

$$\{p_k \mid p_k \in P, c(d_i, 0) = c(d_j, 0), \quad i < j\} \tag{5}$$

$$\{p_k \mid p_k \in P, c(d_i, 1) = c(d_j, 1), \quad i < j\} \tag{6}$$

ここで，0 回目および 1 回目の両分類において文書 d_i，d_j が同じクラスタに分類された場合，文書 d_i，d_j からなる文書対 p_k を分類一致文書対と定義する．分類一致文書対の集合である分類一致文書対集合 T は式 7 のように示せる．

$$T = \{p_k \mid c(d_i, 0) = c(d_j, 0) \wedge c(d_i, 1) = c(d_j, 1), \quad i < j\} \tag{7}$$

以上から，文書対集合 P の要素数における分類一致文書対集合 T の要素数の割合を，分類器における安定性 s として式 8 のように定義し，分類器の分類結果出力の安定性を評価する手法として提案する．

$$s = \frac{|T|}{|P|} \tag{8}$$

5 分類器評価

分類器の評価実験にあたり，3 章と同様に，Sentence-bert，USE，Flair，SpaCy の 4 つの異なる埋め込み表現を持つ分類器を作成した．それぞれの分類器を用いて英ニュース記事および英 tweet の 2 つのデータセットに対して 10 回の分類を行い．そこから 5 つの文書対集合 P_1-P_5 を得た．さらに P_1-P_5 から 5 つの分類一致文書対集合 T_1-T_5 を得て，ここから求められる 5 通りの安定性 s の平均を最終的な評価値として，表 3 の最右列および図 4 に示している．また英ニュース記事を分類対象とした場合は N，英 tweet を分類対象とした場合は T と文字を割り振り，用いられた埋め込み表現とともに表記している．

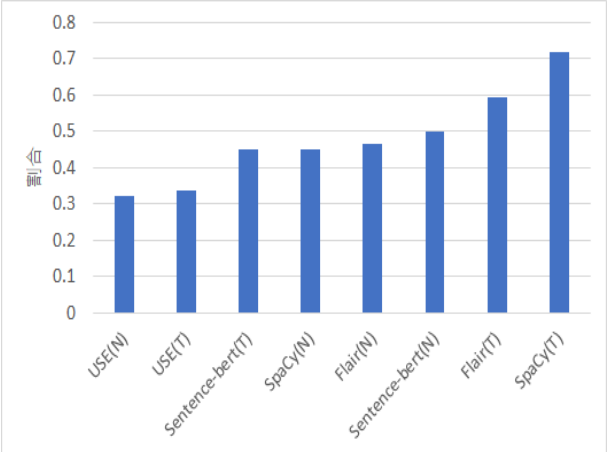


図 4　安定性評価

5.1 結　　果

評価手法を用いた評価結果を図 1 に示す。埋め込み表現に Sentence-BERT を用いた分類器は英ニュース記事を対象とした分類の安定性が 0.50、英 tweet を対象とした分類の安定性が 0.46 となり、分類対象文書による安定性の差は 0.06 となった。埋め込み表現に USE を用いた分類器は英ニュース記事を対象とした分類の安定性が 0.32、英 tweet を対象とした分類の安定性が 0.34 となり、分類対象文書による安定性の差は 0.02 となった。また埋め込み表現に SpaCy を用いた分類器は英ニュース記事を対象とした分類の安定性が 0.45、英 tweet を対象とした分類の安定性が 0.72 となり、分類対象文書による安定性の差は 0.27 となった。埋め込み表現に Flair を用いた分類器は英ニュース記事を対象とした分類の安定性が 0.46、英 tweet を対象とした分類の安定性が 0.60 となり、分類対象文書による安定性の差は 0.14 となった。

5.2 考　　察

Sentence-BERT 以外の埋め込み表現を用いた場合，英 tweet を対象とした分類のほうが英ニュース記事を対象とした分類よりも安定性が高くなった．これは Sentence-BERT の有する文章単位でトークンを獲得する特性から，含まれる語数の多い news 記事の分類においてより高い安定性を発揮したためと考えられる．Sentence-BERT と同様に文章単位でトークンを獲得する特性を持つ USE は英 tweet を対象とした分類の安定性が高く表れているが、分類対象文書による安定性の差は 0.02 と他分類器と比較して低いため、Sentence-BERT と同様

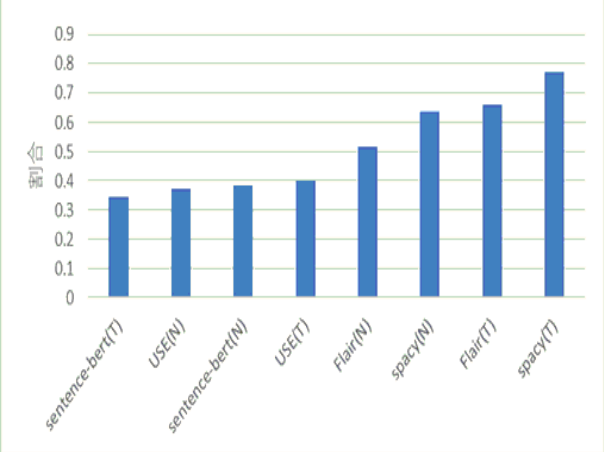


図 5　BERTopic 分類器における分類放棄文書割合

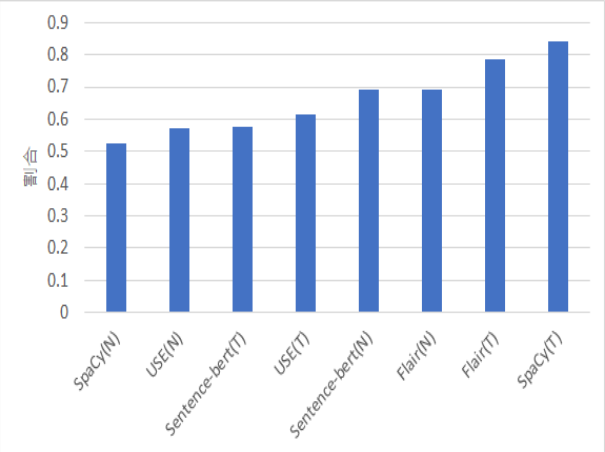


図 6　追加実験結果（分類放棄文書なし）

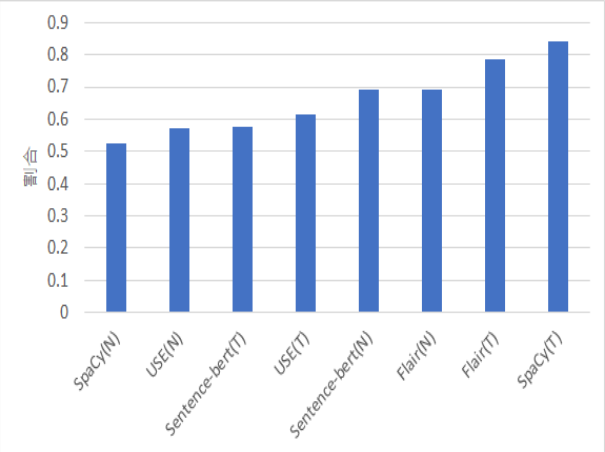


図 7　追加実験結果（分類放棄文書あり）

に含まれる語数の多い文書において高い安定性を発揮したと考えられる。一方で SpaCy，Flair を利用した分類器の場合，英 tweet を分類対象とした場合に安定性 s が 0.6-0.7 程度と比較的高くなったのは，これらの埋め込み表現が有する単語単位でトークンを獲得する特性のためであると考えられた．さらに，Sentence-BERT と USE を埋め込み表現として用いている分類器は他分類器と比較して分類対象文書による安定性の差が小さく、これらの埋め込み表現に共通する文章単位でトークンを獲得する特性は分類対象となる文書に限らず

このように，BERTopic 分類器の安定性には埋め込み表現の特性が関係しており，分類対象となる文書集合に対して適切な埋め込み表現を選択することで安定した出力が得られると考えられた．しかし，追加の調査から埋め込み表現による分類放棄文書の出現割合が安定性に寄与していることが考えられたため，さらに特殊な分類結果を想定した評価実験を行った．

6　追加評価実験手法

BERTopic 分類器による分類で得られる分類放棄文書の割合を表 4 および図 5 に示す．分類ごとに得られる分類放棄文書の数は変化するため，3 回の分類から得られる分類放棄文書の平均値を表 4 の最右列および図 5 に示している．図 5 では全文書における分類放棄文書の平均の割合を示しており，図 4 の結果と比較すると，分類放棄文書の割合が高い分類器は分類対象となるデータセットに限らず，比較的高い安定性を示しているのが分かる．ここから，分類放棄文書の出現割合が安定性に寄与していることが考えられる．よって，以下に挙げるような特殊な状況が分類において発生したことを想定した評価実験を行った．

- 分類放棄文書が出現しなかった場合

この状況を再現する際には，提案手法において分類放棄が含まれる文書対は数えず，文書対集合 P および分類一致文書対集合 T には分類放棄を含む文書対は存在しないものとした．

- 全ての文書が少なくとも一度は分類放棄文書とされた場合

分類放棄文書が出現しなかった場合とは逆に，この状況を再現する際には分類放棄が含まれない文書対は数えず，文書対集合 P および分類一致文書対集合 T には分類放棄を含まない文書対は存在しないものとした．

7　追加評価実験結果

7.1　分類放棄文書が出現しない分類における安定性評価

分類放棄文書が出現しなかった場合を想定した評価実験の結果を図 6 に示した．全ての埋め込み表現において英ニュース記事を対象とした分類のほうが英 tweet を対象とした分類よりも安定性が高くなった．また分類対象とするデータセットに限らず Sentence-BERT を用いた場合に安定性が 0.35 程度と最も高くなった．

7.2　全ての文書が少なくとも一度は分類放棄文書とされた分類における安定性評価

全ての文書が少なくとも一度は分類放棄文書とされた場合を想定した評価実験の結果を図 7 に示した．Sentence-BERT 以外の埋め込み表現を用いた場合は英 tweet を対象とした分類のほうが英ニュース記事を対象とした分類よりも安定性が高くなった．また英 tweet を分類対象とし，SpaCy，Flair を利用した場合に安定性が 0.8 程度と比較的高くなった．

8 考　　察

7.2 節より，少なくとも一度は分類放棄文書とされた場合の安定性は 0.5 から 0.8 程度と，他の実験結果と比較して高くなった．これは BERTopic のアルゴリズムの性質として，分類放棄文書と判定される文書は再度の分類においても分類放棄文書と判定されやすく，逆に分類放棄文書と判定されなかった文書についても再度の分類において分類放棄文書と判定されにくいためだと考えられる．よって BERTopic による文書分類は，分類が可能な文書か，あるいは不可能な文書かの 2 値に分類するような場合においては出力が安定しやすいために有用だと考えられる．しかし 7.1 節に示したように，分類が行われてクラスタに分類された文書は分類を行うごとに結果が変わりやすく，6 章で触れたように分類放棄文書の数は分類器によってさまざまであることから，5.1 節に示したような一般的な分類の安定性にも影響を与えると考えられる．これは決定されるクラスタの数や形状が分類を行うたびに変化する HDBSCAN のアルゴリズムの性質によるものと考えられ，安定化には工夫が必要である．例えば Sentence-BERT を用いた一部の分類結果を除いて英 tweet を分類対象とした場合に文書対の安定性が高くなっており，分類対象となるデータセットに対して適切な埋め込み表現を選択することで安定性を高められる可能性がある．また安定性は文書対の作成に用いる分類結果に依存することから，分類結果が大きく変動する分類器はその安定性も変動することが考えられ，分類器により安定性がどれほど変化するか調査を行う必要があると考える．

9 ま と め

本論文では，BERTopic を用いて Sentence-bert，USE，Flair，SpaCy の 4 つの異なる埋め込み表現を持つ分類器を作成し，2 種類のデータセットに対して複数回の分類を行った．そして出力された 2 つの適当な分類結果から文書ごとに文書対を作成し，分類結果が一致する割合を求めることで BERTopic 分類器の文書分類の安定性を評価した．結果として，分類放棄文書を含む文書対の安定性が高いことから，分類対象となる文書の分類が可能か否かの 2 値に分類する場合において安定した出力を行うと考えられた．しかし分類放棄文書と判定されなかった文書については，その文書を含む文書対の安定性は低くなったことから，決定されるクラスタの数や形状が分類を行うたびに変化する HDBSCAN のアルゴリズムの影響を受けるために分類結果は安定しないと考えられた．さらに分類器に用いる埋め込み表現によって分類結果が一致する割合が異なったことから，分類対象となる文書集合に対して適切な埋め込み表現を選択することで安定した出力が得られると考えられた．今後の展望として，本論で提案した安定性は文書対の作成に用いる分類結果に依存することから，分類器により安定性がどれほど変化するのか調査したい．

文　　献

[1] Grootendorst. Maarten.: *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv preprint arXiv:2203.05794（2022）.

[2] Ricardo J. G. B. Campello. Davoud Moulavi. Joerg Sander.: *Density-Based Clustering Based on Hierarchical Density Estimates*. Advances in Knowledge Discovery and Data Mining PAKDD 2013 Lecture Notes in Computer Science vol 7819. Pages 160–172（2013）.

[3] Abeer Abuzayed. Hend Al-Khalifa.: *BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique*. Procedia Computer Science Volume 189. Pages 191-194（2021）.

[4] Nils Reimers. Iryna Gurevych.: *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv preprint arXiv:1908.10084（2019）.

[5] Daniel Cer. Yinfei Yang. Sheng-yi Kong. Nan Hua. Nicole Limtiaco. Rhomni St. John. Noah Constant. Mario Guajardo-Cespedes. Steve Yuan. Chris Tar. Yun-Hsuan Sung. Brian Strope. Ray Kurzweil.: *Universal Sentence Encoder*. arXiv preprint 1803.11175（2018）.

[6] Akbik. Alan and Blythe. Duncan and Vollgraf. Roland.: *Contextual String Embeddings for Sequence Labeling*. International Conference on Computational Linguistics Volume 27 pages 1638–1649（2018）.

[7] *SpaCy itIndustrial-Strength Natural Language Processing*. https://SpaCy.io/

[8] Lester James Miranda. Ákos Kádár. Adriane Boyd. Sofie Van Landeghem. Anders Søgaard. Matthew Honnibal.: *Multi hash embeddings in SpaCy*. arXiv preprint arXiv:2212.09255v1（2022）

[9] scikit-learn：7.2.2. The 20 newsgroups text dataset. `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch\_20newsgroups.html`

[10] Google Colaboratory：Trump's Tweets. `https://colab.research.google.com/drive/1un8ooI-7ZNlRoK0maVkYhmNRl0XGK88f`

表 2　予備実験結果

分類器（データ）	比較 1	比較 2	比較 3	比較 4	比較 5	平均
sentence-bert(N)	53.12	40.77	39.63	50.06	38.16	44.35
sentence-bert(T)	52.75	41.83	43.18	46.27	44.71	45.75
Flair(N)	44.76	53.7	49.92	40.7	40.7	45.96
Flair(T)	57.05	59.32	58.46	58.97	57.95	58.35
USE(N)	32.97	50.85	28.43	37.01	49.7	37.42
USE(T)	33.21	35.8	35.06	32.83	37.55	34.89
Spacy(N)	71.79	68.3	7.13	71.03	71.58	57.97
Spacy(T)	72.73	71.33	70.29	71.55	70.54	71.29

表 3　安定性評価

分類器（データ）	T_1	T_2	T_3	T_4	T_5	s
sentence-bert(N)	10012	10569	8334	8817	8108	0.50068265
sentence-bert(T)	23926	20226	20220	15889	21943	0.450684599
USE(N)	6037	5585	6047	6101	5629	0.321107531
Flair(N)	8197	8248	8730	9642	7912	0.466703075
USE(T)	15061	15212	15753	14679	15794	0.337334362
spacy(N)	13147	1319	13022	12554	1393	0.452569494
Flair(T)	25874	27509	26946	27254	27396	0.595211112
spacy(T)	32988	32013	33300	32948	32031	0.720008819

表 4　放棄文書数

埋め込み表現 (データ)	放棄数 1	放棄数 2	放棄数 3	割合平均
Sentence-bert(T)	16826	14327	16402	0.34654
USE(T)	17882	18506	18340	0.398762
Flair(T)	29220	30261	30535	0.660655
SpaCy(T)	34329	34862	34843	0.769989
Sentence-bert(N)	7175	7201	7130	0.385178
USE(N)	6671	6904	6794	0.373956
Flair(N)	9105	9549	9326	0.514745
SpaCy(N)	13986	13723	13708	0.636475

表 5　追加実験結果（分類放棄文書なし）

埋め込み表現 (データ)	文書対 1	文書対 2	文書対 3	文書対 4	文書対 5	割合
Sentence-bert(N)	4643	5175	3118	3082	2614	0.356894
Sentence-bert(T)	10664	8897	8862	5520	9336	0.348625
USE(N)	1186	443	898	1405	781	0.097612
USE(T)	1165	1212	2285	438	2023	0.062728
Flair(N)	627	838	797	1042	324	0.103127
Flair(T)	15	904	462	727	724	0.048373
SpaCy(N)	780	349	869	399	329	0.152333
SpaCy(T)	1497	99	1116	870	407	0.107532

表 6　追加実験結果（分類放棄文書あり）

埋め込み表現 (データ)	文書対 1	文書対 2	文書対 3	文書対 4	文書対 5	割合
Sentence-bert(N)	5369	5394	5216	5735	5494	0.691453
Sentence-bert(T)	13262	11369	11358	10369	12607	0.574523
USE(N)	4851	5142	5149	4696	4848	0.570484
USE(T)	13896	14000	13468	14241	13771	0.612743
Flair(N)	7570	7410	7933	8600	7588	0.693588
Flair(T)	25859	26605	26484	26527	26672	0.785514
SpaCy(N)	12367	970	12153	12155	1064	0.525509
SpaCy(T)	31491	31914	32184	32078	31624	0.839792