Authorship Attribution by Attention Pattern of BERT with Topological Data Analysis and UMAP

1st Wataru Sakurai*
Artificial Intelligence Section
National Research Institute of Police Science
Kashiwa-shi, Japan
sakurai@nrips.go.jp

2nd Masato Asano
Artificial Intelligence Section
National Research Institute of Police Science
Kashiwa-shi, Japan
asano@nrips.go.jp

3rd Daisuke Imoto
Artificial Intelligence Section
National Research Institute of Police Science
Kashiwa-shi, Japan
imoto@nrips.go.jp

4th Masakatsu Honma
Artificial Intelligence Section
National Research Institute of Police Science
Kashiwa-shi, Japan
honma@nrips.go.jp

5th Kenji Kurosawa
Second Department of Forensic Science
National Research Institute of Police Science
Kashiwa-shi, Japan
kurosawa@nrips.go.jp

Abstract—Authorship attribution is the problem of identifying the author of a text based on its content, topic and stylistic features. BERT can visualize the weight of each word via self-attention and are effective in explaining the basis for the decision for authorship attribution. However, the basis is not easy to capture because of the large amount of self-attention in a model. Topological data analysis(TDA) is a method used to capture a set of points in space based on their topology. Using TDA for text data, we can capture how the words focus on each other, or the "pattern of recognition" by self-attention. In this study, we attempt to analyze the explanatory power of the model based on the attention patterns of BERT by extracting TDA-based features based on the zeroth-order persistent homology through a developed classification model and visualizations based on uniform manifold approximation and projection (UMAP). Based on experimental results, we can conclude that TDA-based features contain sufficient information to discriminate authors. The proposed method can capture the basis of BERT's authorship attribution more clearly, facilitating the explanation of the basis.

Index Terms—Deep learning,natural language processing, topological data analysis

I. Introduction

Authorship attribution is the problem of identifying the author of a text based on its content, topic, and stylistic features. A typical problem involves identifying the author of a text based on classifications from a list of candidates that has been narrowed down to a certain extent by related parties. Various texts written on digital media, such as social media, and those generated by artificial intelligence are widely encountered in our daily lives. However, they may be related to crime (e.g., spoofing and letter of responsibility for a crime); thus, an authorship attribution method must be established. Quantitative approaches based on natural language processing (NLP) techniques using handcrafted features [1] or deep learning (DL) techniques [2] have been proposed for authorship attribution. Some studies have shown the effectiveness of the former, such as n-grams, particularly for Japanese sentences [3]. In contrast, in the field of NLP, DL technology has made significant progress. DL is commonly being used for various language-related tasks and is now beginning to be used for authorship attribution with explanations provided by generative models [4]. For a convincing identification, it is necessary to provide the basis for the decisions of the model. BERT [5], a model for the downstream tasks of NLP, can visualize the weight of each word by self-attention and is effective in explaining the basis for its decisions. However, if a large amount of self-attention is included in a specific transformer model, then its analysis encounters challenges. The explanatory nature of bidirectional encoder representations from transformer (BERT) models has been analyzed from various perspectives. Among them, the knowledge of topological data analysis (TDA) is occasionally used. This method involves capturing a set of points in space based on topology. Using TDA in BERT, we can capture the connections between words and how the words focus on each other within the model. Various methods have been proposed to improve the accuracy of the model by adding new features [6] or to determine the various self-attentions that are effective for discrimination and use them for pruning [7]. Consider the meaning of using only these features; that is, the values obtained from only query and key among the various attentions. As mentioned above, the values obtained only from query and key are weights

for each word, which are based on the model's perception of the relationship between the words. This pattern is considered to be more closely related to sentence structure but not the meaning of each word.

II. Method

In this section, we explain the proposed method using TDA and Self-Attention. Fig.1 presents an overview of the proposed method.

A. Self-attention of BERT

In this section, feature calculation methods are explained, with a brief description of self-attention using TDA. The relationship between each word captured by BERT is quantified mathematically from the features for the 0 th-order persistent homology, which represents the connected components of the geometric model extracted from persistent diagrams, which is the most typical method in TDA. First, we explain the judgement basis of BERT. The i-th head of multi-head self-attention is defined as (1).

$$Attention_i(Q,K,V) = softmax(\frac{QW_i^Q(KW_i^K)^T}{\sqrt{d_k}})VW_i^V \tag{1}$$

Here, d is the embedding length of each token obtained from the sequence (e.g., sentences) input into BERT, and the sequence length is s; $Q, K, V \in \mathbb{R}^{s\times d}$ are matrices which is called query, key, and value matrices. In "multi-head" self-attention, these matrices are separated and calculated, respectively in each head via the matrices $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d\times d_k}(d_k < d)$. The basis of judgement, a, of the i-th head of self-attention, is defined as (2).

$$a_i = softmax(\frac{QW_i^Q(KW_i^K)^T}{\sqrt{d_k}}) \tag{2}$$

$a_i \in \mathbb{R}^{s\times s}$ is a matrix that represents the relationship between each token (e.g words) in a sequence. A BERT model have several transformer layers, which contain several multi-head self-attention, and it is difficult to analyze them.

B. TDA-based features

In this section, TDA is introduced into the self-attention value described above. Based on previous studies, we use an attention graph to represent each word in a sentence using self-attention as a geometric model [9]. In this method, words are represented as a group of points based on their self-attention values, the distances between them as an adjacency matrix, and an undirected graph, such that each node corresponds to each word. For a sentence comprising s words, let the self-attention from the i-th word to the j-th word be a_{ij} for w_i and w_j of the i-th and j-th words extracted from (1)$(i, j = 1 - s)$. The adjacency matrix $D = (D_{ij})$ is expressed as (3):

$$D_{ij} = \begin{cases} 1 - max(a_{ij}, a_{ji}) & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \tag{3}$$

The components matrix D, indicate the distances between words, which are defined from the self-attention values, an undirected graph of words is constructed. Let the set of words in a sentence be defined as S and the nodes are defined as x. Then, a simplical complex $C(r, S)$ can be defined from the words and their distances. By increasing the radius r, (defined in (4)) of each word, according to $(r_1, r_2, \cdots, r_n)(r_1 < r_2 < \cdots < r_n)$, simplical complex filtration, as that defined in (5), is obtained.

$$B(r) = \{x \in \mathbb{R}^2 | \quad |x - x_i| \leq r\} \tag{4}$$

$$C(r_1, S) \subset C(r_2, S) \subset \cdots \subset C(r_n, S) \tag{5}$$

Through filtration, a persistence diagram for the 0-th order persistent homology is calculated as shown in Fig.2. In the persistent diagram, the horizontal axis represents the radius at which the n-th order holes are created as the radius increases, whereas the vertical axis represents the radius at which they disappear. The vertical axis represents the radius of disappearance, whereas the 0-th order holes represent the connected components. Therefore, this time, distribution of the 0-th order persistent homology comprises the point clouds whose horizontal axis value ("birth" radius) is all 0 and the vertical axis ("death" radius) has different values for each sentence. A sequence of these vertical values is extracted and used to generate TDA-based features.

As mentioned earlier, self-attention is abundant in a BERT model because it has several transformer layers and its self-attentions are multi-head; in BERT, twelve heads exist for each of the twelve transformer layers, each of which has a self-attention. In proposed method, the self-attention of the layer closest to the output is considered, and the sequence of values based on the TDA calculated for each of these is concatenated and used as the feature value.

C. Authorship Attribution by deep learning using TDA-based features

A simple model was used to perform authorship attribution. A multi layer perception-based simple feed-forward NN was constructed, and a classification model was built to output the author classes from the TDA-based features. Although the number of dimension is over 1000 due to concatenation, the number of layers and units were taken to be small for simplicity.

D. Discrimination of authors using TDA-based features based on UMAP

An analysis of the relationship between features and authors based on TDA is performed. The relationship between authors and features was projected to lower dimensions using UMAP, a manifold learning method. Basic UMAP is an unsupervised method and comprises two stages, graph construction and graph layout. When data points are provided, the relevant weighted graph

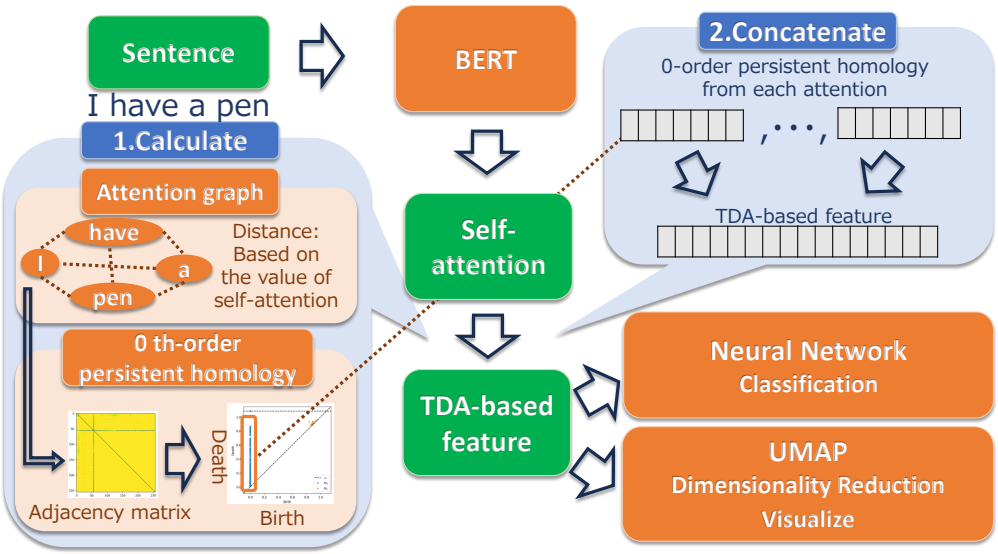

Fig. 1. Overview of the proposed method

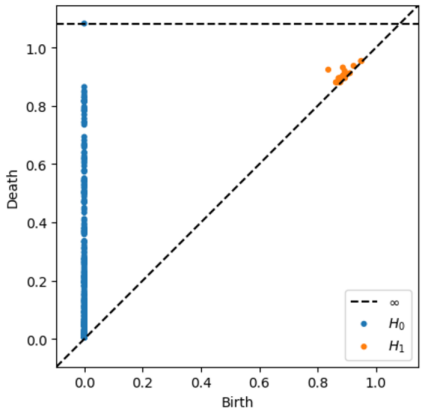

Fig. 2. Example of persistence diagram(H_0 means 0-th order,H_1 means 1st order),using Ripser [12]

(fuzzy simplicial complex) based on k-neighbors is constructed at the graph construction stage. In the graph layout, the dimension-reduced embeddings are calculated and obtained. In addition, we can use supervised UMAP and obtain low-dimensional representations of data points considering their labels. We obtained and visualized two- or three-dimensional (2D or 3D) representations of TDA-based features (high-dimensional data) using supervised and unsupervised UMAP methods by considering and not considering author labels, respectively. We can visually understand the relationship between the TDA-based features and authors and whether these features have sufficient information to distinguish authors without using raw sentences.

III. Experiment

We assumed a multi-class classification problem comprising eight authors, constructed from a balanced corpus of written Japanese (BCCWJ) [10]. The training data included 1485 paragraphs, while the test data comprised 185 paragraphs. The BERT model pretrained by Tohoku University [11] was fine-tuned for authorship attribution. The self-attention values after fine-tuning were extracted from the layer closest to the output layer according to the proposed method described in the previous section. Values were extracted from the layer closest to the output layer after fine-tuning, and the TDA-based features were computed. When calculating the TDA-based features, a persistent diagram was calculated using Ripser [12]. In addition, For the data input into the BERT model, each sample was divided into tokens and padding for 256 words was applied to align the number of tokens for each sample. (Special tokens, such as [PAD], were used to set the value of self-attention value from other words to zero.) The discrimination performance of the BERT test data showed 89% accuracy.

Next, we calculated the TDA-based features described above. The dimensions of the features extracted by one self-attention head were 255 (the final node was never connected or "die" and was thus excluded), and the dimension of the TDA features was $255 \times 12 = 3060$ (twelve features calculated from twelve self-attention heads were concatenated). The TDA features of one sample from the training data are displayed in Fig. 3.

Next, we constructed a simple feed-forward NN. An overview of the NN structure is provided in Fig.4.

Accuracy using TDA-based features and the NN was 71.8%, and the inference results of the two models were in 74.1% agreement.

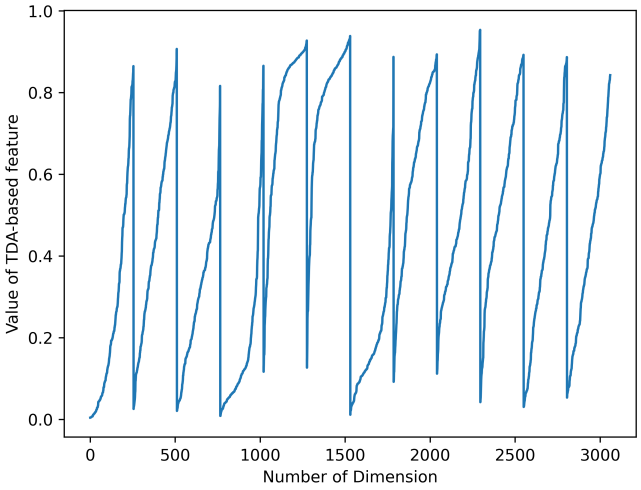

Fig. 3. Example of a TDA-based feature

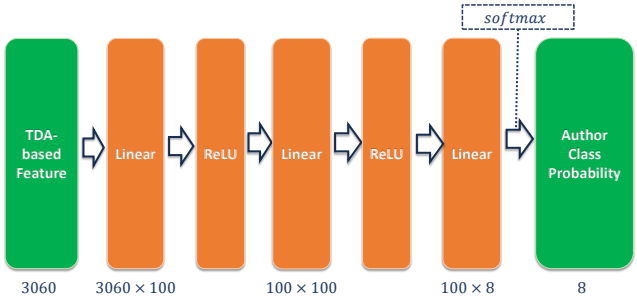

Fig. 4. Structure of the NN used

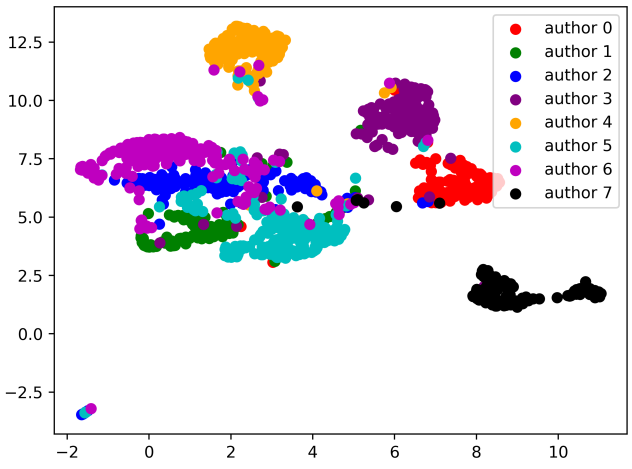

Fig. 5. 2D UMAP projection of the TDA-based features of the training data

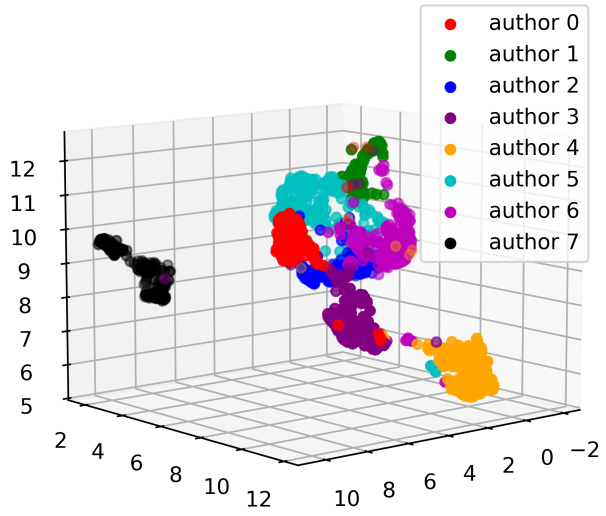

Fig. 6. 3D UMAP projection of the TDA-based features of the training data

The data were visualized via projection onto a 2D space using UMAP, as shown in Fig. 5. As a result, areas that were well separated and not so well separated became apparent. A 3D projection provides a clearer picture of the classification, as shown in Fig. 6.

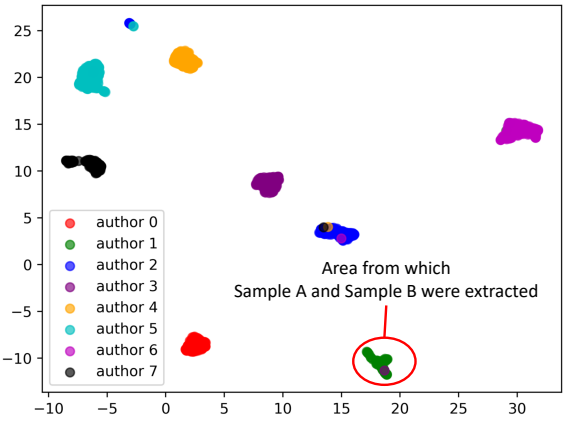

Fig. 7. 2D projection of TDA-based features using supervised UMAP (training data)

The results of manifold learning on the training data using a supervised UMAP that considers the author labels for each sample are presented in Fig. 7. The samples could be discriminated with an even higher accuracy than that obtained using the unsupervised UMAP described above. The model could discriminate some of the authors; however, the others could not be identified.

The performance of the supervised method could be further evaluated by applying it to the test data projected using the same UMAP model as used for the training data. As shown in Fig.8, the clusters are generally located in the same positions as for the training data; however,

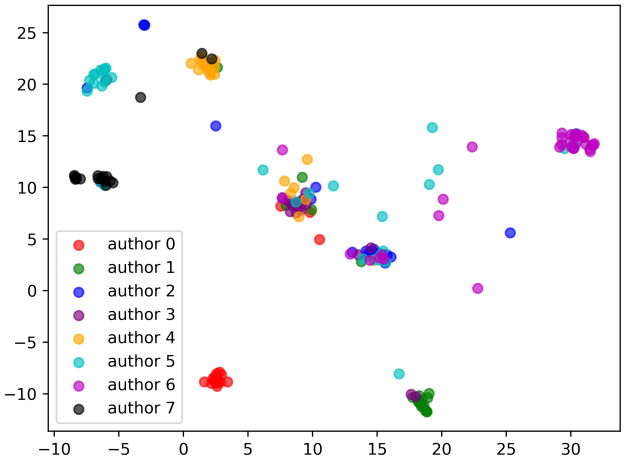

Fig. 8. 2D projection of TDA-based features using supervised UMAP (test data)

some sentences are located in positions where cluster formation is impossible. To examine the applicability limits of discrimination, we extracted and compared two overlapping samples of supervised UMAPs. Specifically, we selected a sample sentence written by an author (Sample A) that, when projected in two dimensions, was located inside a cluster formed by a different author. These sentences were compared with a sample extracted from the cluster (Sample B). The area from which Samples A and B were extracted is shown in Fig.7, and Samples A and B are shown in Figs.9 and 10, respectively. These sentences were taken from the BCCWJ [10], as described above.

'そうした年齢の女性を見ても、やはり女性には女性の美しさがあると感じられる男は幸福である。その気持は女性に対するやさしさとなって自然にあらわれてくるだろうからである。男が女性に対して示すやさしさには幾通りかの道があって、一つは男性的勇気から女性を騎士的にかばおうとする西洋流のやさしさがあったり、また、女性の美しさを両手でそっと包みこんであげようとするやさしさもあり、極端な場合には、女性を虫けらのように無視する蛮勇が、かえって女性からは男らしいやさしさと倒錯的に受とめられることすらある。その中でも、やはり女性を美しい存在として、その一挙手一投足、あるいはまばたき一つ、かすかな微笑の一つの中にも、女らしい美しさを感じとっていくやさしさを身につけていくことが、最も日常的であり、無理をしない美的感受性ではなかろうか。'

Fig. 9. Sentences of Sample A (taken from the BCCWJ [10])

'思春期には大なり小なり、何らかの「荒れ」が起こるのが普通なのだ。それが適切な「守り」のなかで生じてこそ、大人になる変革につながっていく。この守りは家庭、学校、社会などが受けもつのだが、以前に比べて家庭の重みが大きくなってきているのは、すでに述べたとおりである。'

Fig. 10. Sentences of Sample B (taken from the BCCWJ [10])

Sample A is a sentence about the behaviors of men and women, while Sample B is a sentence about adolescence. These sentences are different in topic and length, among other features, but similar in that the word endings and other stylistic features show typical characteristics used in Japanese critiques. This suggests that TDA features are more related to the stylometric features of sentences than the topic and length of sentences.

IV. Discussion

Authorship attribution was performed using a simple NN. The results suggest that the weighting pattern by self-attention is such that the data contain sufficient information for author identification. We believe that one of the reasons for the low accuracy of this model when compared to the accuracy of the BERT model is the effect of padding. Although the value of self-attention "toward" the special token used for padding was set to zero, that "from" the special token was not set to zero. Therefore, the distance between special tokens, which should not be considered, was calculated, and this may have had a negative impact on the TDA-based feature calculation. The projection using UMAP also visually confirmed that a correlation existed between authorship and TDA-based feature calculation. Furthermore, when we examined the samples of different authors in close proximity in the 2D space projected by the supervised UMAP, we confirmed that samples with similar sentence patterns existed, such as word endings, although the content and topic of the sentences differed, suggesting that this similarity in sentence patterns may be related to the application limit of this method. Thus, the similarity in sentence structure may be related to the limitation of the application of this method. This method contributes toward explaining the basis for judgments by BERT, a deep learning model, which is generally a black box in a way that its output is independent of the words used, such as the form of sentences. However, only the decision basis of the model was extracted, and the fact that not all the features were used is noteworthy. In addition, the accuracy of this model is lower, and the computational cost is higher than that of BERT. Therefore, it is necessary to discuss its practical applications based on the amount of data.

V. Conclusion

In this study, we proposed an authorship attribution method to analyze the self-attention of BERT by the "shape of the recognition pattern" using TDA and visualize its discrimination ability using manifold learning. In the future, we plan to improve the accuracy and interpretability of the proposed method.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP22K21320.

References

[1] M. Koppel and Y. Winter, "Determining if two documents are written by the same author," Journal of the Association for Information Science and Technology, vol. 65, no. 1, pp. 178–187, 2014.

[2] M. Fabien, E. Villatoro-Tello, P. Motlicek, and S. Parida, "BertAA : BERT fine-tuning for authorship attribution," in Proceedings of the 17th International Conference on Natural Language Processing (ICON), P. Bhattacharyya, D. M. Sharma, and R. Sangal, Eds. Indian Institute of Technology Patna, Patna, India: NLP Association of India (NLPAI), Dec. 2020, pp. 127–137. [Online]. Available: https://aclanthology.org/2020.icon-main.16

[3] K. Y. Matsuura Tsukasa, "Authorship detection of sentences by 8 japanese modern authors via n-gram distribution," IPSJ Transactions on Natural Language Processing, vol. 00-NL-137, pp. 1–8, 2000.

[4] B. Huang, C. Chen, and K. Shu, "Can large language models identify authorship?" arXiv preprint arXiv:2403.08213, 2024.

[5] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of naacL-HLT, vol. 1. Minneapolis, Minnesota, 2019, p. 2.

[6] A. Uchendu, T. Le, and D. Lee, "Topformer: Topology-aware authorship attribution of deepfake texts with diverse writing styles," 2024. [Online]. Available: https://arxiv.org/abs/2309.12934

[7] L. Balderas, M. Lastra, and J. M. Benítez, "Can persistent homology whiten transformer-based black-box models? a case study on bert compression," 2023. [Online]. Available: https://arxiv.org/abs/2312.10702

[8] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2020. [Online]. Available: https://arxiv.org/abs/1802.03426

[9] I. Perez and R. Reinauer, "The topological bert: Transforming attention into topology for natural language processing," 2022. [Online]. Available: https://arxiv.org/abs/2206.15195

[10] K. Maekawa, "Kotonoha and bccwj: development of a balanced corpus of contemporary written japanese," in Corpora and Language Research: Proceedings of the First International Conference on Korean Language, Literature, and Culture, 2007, pp. 158–177.

[11] Inui and S. L. at Tohoku University, "Pretrained japanese bert models," 2023, [Online; accessed 25-Nov-2024]. [Online]. Available: https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

[12] U. Bauer, "Ripser: efficient computation of Vietoris-Rips persistence barcodes," J. Appl. Comput. Topol., vol. 5, no. 3, pp. 391–423, 2021. [Online]. Available: https://doi.org/10.1007/s41468-021-00071-5