Social Network Analysis and Mining (2025) 15:32
https://doi.org/10.1007/s13278-025-01437-8

ORIGINAL ARTICLE

Leveraging social media for public health: NLP implementations for blood donation data analysis in Japan

Roberto Espinoza[1] · Kazumasa Kishimoto[2] · Chang Liu[2] · Luciano H.O. Santos[3] · Yukiko Mori[2] · Goshiro Yamamoto[2] · Tomohiro Kuroda[2]

Received: 30 August 2024 / Revised: 7 December 2024 / Accepted: 25 December 2024

Abstract
Blood donation is crucial for healthcare systems, yet maintaining an adequate supply is a persistent challenge. Traditional methods to understand public sentiment and donor behavior are often limited. Social media, particularly "X" (formerly Twitter), offers a promising alternative for real-time insights. This study explores the viability of using "X" data to analyze blood donation sentiment in Japan, considering the evolving perspectives of younger generations. We replicated previous study results using the Tohoku BERT model and tested a refined blood donation tweets for user classification (BDT-UC) dataset and another customized version of the model for better classification. We also compared various topic modeling methods, including latent Dirichlet allocation (LDA), non-negative matrix factorization (NMF), and BERT-based models, using two different preprocessing techniques. Finally, we integrated the classification into the Topic Modeling process, to explore the possible impact of the previous steps in such execution, for a final evaluation. Our findings indicate that although the refined dataset has an overall lower classification performance, it improved the implementation results, ensuring more balanced labeling across the data. Our refined model had a small reduction in overall precision (from 78.4% in the best evaluated model to 75.8% in the refined model). However, we improved the implementation results, ensuring more balanced labeling across the data. For topic modeling, BERT-based topic models, particularly those preprocessed with the MeCab library, achieved higher coherence and diversity scores than traditional methods. Additionally, there were significant differences when the dataset was processed following the categories of the BDT-UC study, which used specific categories related to the tweets role in blood donation. There was increased coherence and diversity for one of the categories but notably lower coherence values for the others. This study underscores the significance of initial classification and preprocessing for effective topic modeling approach when working with Japanese text, which impacts the viability of extracting insights from Japanese social media data. The developed methodologies could support more effective analysis of blood donation groups, and better targeted donation campaigns in Japan.

Keywords Blood donation · Japanese · Social networking sites (SNS) · Twitter data analysis · BERT · NLP · MeCab · Topic modeling · BERTopic

✉ Roberto Espinoza
espinozarob@kuhp.kyoto-u.ac.jp

Kazumasa Kishimoto
kishimoto@kuhp.kyoto-u.ac.jp

Chang Liu
liuchang@kuhp.kyoto-u.ac.jp

Luciano H.O. Santos
contact@luciano.ac

Yukiko Mori
yukimori@kuhp.kyoto-u.ac.jp

Goshiro Yamamoto
goshiro@kuhp.kyoto-u.ac.jp

Tomohiro Kuroda
tomo@kuhp.kyoto-u.ac.jp

1 Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku 606-8317, Kyoto, Japan

2 Kyoto University Hospital, Shogoin Kawaracho, Sakyo-ku 606-8507, Kyoto, Japan

3 Fitting Cloud Inc., 85-1 Mikura-cho, Nakagyo-ku 604-8166, Kyoto, Japan

Published online: 03 April 2025

Abbreviations

NLP	Natural language processing
ML	Machine learning
BERT	Bidirectional encoder representations from transformers
LDA	Latent Dirichlet allocation
NMF	Non-negative matrix factorization
BERTopic	BERT-based topic modeling
BERT-MiniLM	BERT model from sentence-paraphrase-multilingual-MiniLM-L12-v2
BERT-sbert	BERT model from colorfulscoop/sbert-base-ja
BERT-ja-mean	BERT model from sonoisa/sentence-bert-base-ja-mean-tokens-v2
BERT-mpnet	BERT model from sentence-transformers/paraphrase-multilingual-mpnet-base-v2
MeCab	A Japanese morphological analysis engine
OCTIS	Optimized clustering for topic inference and summarization

1 Introduction

Blood donation is a vital component of healthcare systems worldwide. Maintaining an adequate blood supply is a persistent challenge, with demand often outstripping supply due to various factors, including demographic changes, seasonal variations, and public health emergencies. Effective blood donation campaigns are essential to encourage voluntary donations and ensure a stable blood supply, especially in aging countries such as Japan, where the majority of regular donors belong to an aging demographic. However, newer generations have different and constantly changing perspectives, interests, and ideas regarding blood donation. Traditional methods for understanding donor behavior, such as surveys and focus groups, are limited in scope, reach, and timeliness. Understanding what younger citizens think about blood donation on a regular basis becomes a critical challenge.

Analyzing social media offers a compelling alternative, granting access to vast, real-time user-generated data. “X” (formerly Twitter), in particular, is a valuable platform for this purpose due to its public nature, high user engagement and frequent discussions on social issues, including blood donation. By analyzing “X” data, researchers can monitor public sentiment, track trends, and explore factors influencing blood donation behavior.

Given the vast volume of data available on this platform, topic modeling has been proposed as a method to automate the initial stages of analysis. Topic modeling employs unsupervised learning to group documents into predefined topics representing the data corpus (Wagner and Fernández 2015). However, analyzing social media data in non-Roman scripts, such as Japanese, presents unique challenges. The complexity of the language, including its use of multiple scripts (kanji, hiragana, and katakana) and its morphological richness, often requires more specialized preprocessing and analysis techniques, when compared to languages which use a phonetic alphabet only (Qin et al. 2016; Lind et al. 2022).

In this study, compared the effectiveness of various preprocessing strategies and topic modeling techniques, to identify optimal conditions that can provide value to future studies in the analysis of text in the Japanese language, particularly with large datasets obtained from SNS data.

In the first phase, we validated and refined an approach, proposed in a related study, for the classification of SNS data related to blood donation (Espinoza et al. 2023), using a manually labeled dataset (BDT-UC) for model training. However, the initial implementation of the proposed classifier had skewed results in our collected data, resulting in the absence of records for one expected category. To overcome this issue, we customized the Tohoku BERT model, a state-of-the-art language model for Japanese, incorporating mechanisms to address data imbalance and accommodate potential category expansions.

In the second phase, we compared the topic modeling methodologies of latent Dirichlet allocation (LDA) (Blei et al. 2003), non-negative matrix factorization (NMF) (Lee and Seung 1999), and BERTopic (Grootendorst 2022). Topic modeling is essential for identifying the main themes and topics discussed in large datasets of unstructured text, such as tweets. By evaluating these methods via the OCTIS framework (Terragni et al. 2021) (a framework for training, analyzing, and comparing topic models) and incorporating MeCab, a library for tokenizing Japanese text (MeCab 2024) in the preprocessing step, we aimed to identify the most effective approach for analyzing our collected Japanese “X” data.

Finally, we performed additional modeling tests, including the classification of the data using the BDT-UC categories as guidance. We explored whether an initial classification process and segmentation would reveal significant changes in regards to the topic modeling evaluation metrics.

2 Methods

The steps undertaken for the development of this project are outlined in Fig. 1. The following subsections provide a detailed description of the tasks associated with each step.

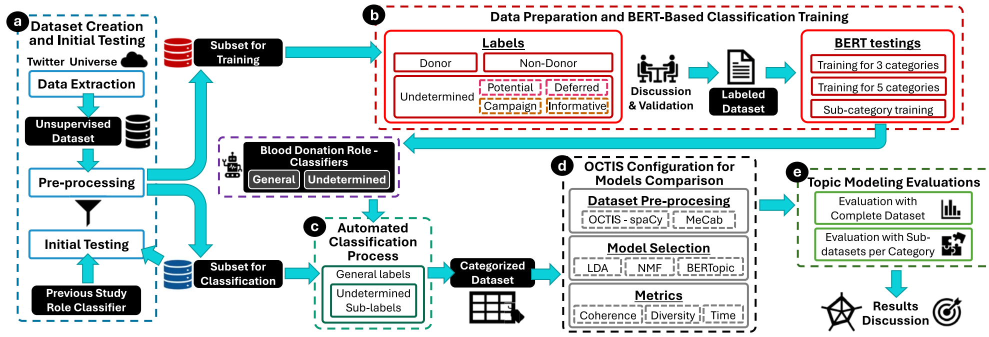

Fig. 1 Diagram illustrating the workflow of the study: **a** step focusing on the collection of data, initial dataset creation and initial classification testing; **b** step focusing on the labeled dataset preparation, and BERT-based models training in multiple tests to generate the final classifiers; **c** step focusing on the usage of the classifiers to classify the 412k dataset into the final 5 labels; **d** step focusing on the configuration of the OCTIS framework to test the topic modeling process with different models and preprocessing steps; **e** step focusing on the execution of the OCTIS Topic Modeling tests on the 412k dataset, and the subsets per labeled category.

2.1 Dataset creation and initial testing

As shown in the first step in Fig. 1, the initial task involved creating a dataset of Japanese text related to blood donation, collected from “X”. We used the search strings “(献血 OR kenketsu OR けんけつ OR ラブラッド OR LoveBlood OR #献血 OR #けんけつ OR #kenketsu OR #ラブラッド OR #LoveBlood) lang)” and focused on the time period from October 1st, 2022, to April 30th, 2023. We removed tweets recognized as retweets, resulting in an unsupervised dataset of approximately 412k tweets.

Following the acquisition of the initial dataset, the next step was to determine how to classify the data in such a way that it can be used for better understanding of blood donation-related topics. As we aimed to train a model for supervised learning, the data required manual labeling. We first defined potential labels that would add value to our dataset. One of the challenges addressed was managing the presence of noise, which required us to delineate what would be considered priority data versus non-priority data.

In that regard, we initially followed the approach of a previous project, that similarly worked with “X” data related to blood donation. In that study, data in Japanese was collected in the periods of January 1st, 2022 to June 30th, 2022, and from October 1st, 2022 to April 30th, 2023 (Espinoza et al. 2023). For their process, a randomized selection of tweets was compiled for manual labeling, considering a human-manageable but representative size of data, to create the BDT-UC dataset. The data was classified in the groups of Donor, Non-Donor and Undetermined, considering potential citizen roles in blood donation, and non-classifiable tweets. Implementations of BERT-Models were fine-tuned to work with the Japanese dataset, and their results compared, showing the most promising results with the implementations of the tohoku-based (BERT model implemented by the Tohoku university, pretrained on text in the Japanese language) and the twhin-based (multi-lingual BERT model trained on “X” data by the previous Twitter Research group) models.

For a first overview, we requested the files of the previously mentioned study’s best performing model (the tohoku-based one), deployed the provided model, and tested it on our collected unsupervised dataset to contrast the classification results. The initial results showed a classification distribution of 38.2% “Donor”, 0% “Non-Donor”, and 61.8% “Undetermined” for our dataset of 412k tweets, which underlined the model’s weakness regarding the “Non-Donor” category, related to the imbalanced nature of the dataset used for its training. This limitation led us to consider a multi-class classification approach and the exploration of additional labels to better balance our dataset.

2.2 Data preparation and BERT-based classification training

Aiming to address the biased categorization from the model of the previous study, we acknowledged the imbalanced distribution of labels in the original BDT-UC dataset. We refined the previous study’s approach, as to reduce the bias of the categorization results and possibly improve the classification accuracy.

To that end, as seen in the second step of Fig. 1, we generated our own version of the BDT-UC dataset, by further distributing the largest category (Undetermined) into other possible roles, which were defined as the categories “Potential,” and “Deferred.” We also added two additional subcategories for potential deeper analysis, “Campaign” and “Informative,” regarding topics not related to roles of citizens but to the activities of the blood donation institutions .

To standardize the labeling criteria for each record, the conditions for selecting each category were as follows:

- Donor: Tweets whose content expressed or implied past successful blood donation experience.
- Non-donor: Tweets whose content expressed or implied no previous blood donation experience or that expressed no interest in such activity.
- Undetermined: Tweets that initially were not included in the other categories. Optional subcategories were defined in this group for further validation. The additional options were:
  - Potential: Tweets whose content expressed the intention or possibility of donating blood but without confirmation of the action or past activity.
  - Deferred: Tweets whose content expressed disqualification from donating blood due to recent or past experiences or a recurrent medical or health-related disability.
  - Campaign: Tweets that explicitly described details of a future blood donation campaign or included requests for participation.
  - Informative: Tweets discussing results of a campaign, providing information about the blood donation process, or offering support related to blood donation.

For the labeling process, we separated a subset for our goal of training the classification model. For this subset, we considered a similar number of tweets as the previously mentioned study, representing approximately 1% of the complete dataset. From the unsupervised dataset, a randomized seeded distribution was performed to generate an initial sub-dataset that would still be representative of the collected data, which provided us with a sub-dataset of around 4k tweets. An additional column with English translation of the tweets text data was included for reference. The translations were generated through the use the translation tool DeepL, with the validation of the main researcher and one Japanese peer researcher.

The labeling process involved the main researcher selecting a category and subcategory (for the Undetermined category, when possible) for all the 4k tweets, while a group of peer researchers labeled smaller subsections. The choices were contrasted and discussed to reach a final decision for each tweet, defining the final version of the labeled dataset. Finally, during the final review of the sub-dataset, while in the Japanese language 1 character by itself can carry meaning, we considered that, from the validated data, tweets with less than four characters did not include enough meaning to form a coherent idea. We corroborated this occurrence with the data by also validating the same cases on tweets from the complete dataset. With that in consideration, we removed such tweets for the training purposes, as they could negatively affect the models. Samples of the final state of the tweets are available in Appendix 1.

For the automated classification, we followed a natural language processing (NLP) approach similar to the one suggested in the previously mentioned study (Espinoza et al. 2023). We focused on the implementation of the Tohoku BERT model, `bert-base-japanese-v3` (from here on referred to as tohoku3) from tohoku (2024), available on HuggingFace (Wolf et al. 2020).

Using the results from the previous study as a guide, we focused on implementing the task-specific layers (dropout and two linear layers) for transfer learning (Munikar et al. 2019) of the tohoku3 model to our version of the BDT-UC dataset. Regarding the hyperparameters, we used Batch Size: 12, Maximum Sequence Length: 150, and Learning Rate: 2×10^{-5}. We evaluated the results considering macro (since it is a multi-class scenario) and weighted (since the data was imbalanced) averages for the classification metrics, using the "classification_report" method from the sklearn library from Python.

Two stages were defined for the testing. The first stage involved attempting to train the customized model with our sub-dataset (4k tweets), for the initial three categories established in the previous study (Donor, Non-Donor, Undetermined). In this stage, we defined two tests: one with a version of the customized model without adding any mechanism to control the imbalanced nature of the data (from here on, identified as test "1A"), and another with a version of the customized model with a mechanism to handle the imbalanced data through class weights for each category (from here on, identified as test "1B").

For the second stage, we initially considered training the model with all the new sub-categories and distribution of the categories. However, there were issues to properly configure the model to handle seven labels, possibly related to the lower amount of data available for the new categories, and the apparent overlap in recurrent words used between the different categories, which negatively impacted the initial tests. With that in consideration, as since the redistribution of labels focused on the "Undetermined" category into more specific categories, we opted for a multi-step model approach.

We used a second customized model (with the same setup as the previous one) but fine-tuned it with only the "Undetermined" labeled records of our 4k sub-dataset. We performed two training tests in this stage. The first one focused on the redistribution of the "Undetermined" category into only the new citizen-related (Potential and Deferred) categories (from here on, referred as test "2A"). The second one focused on the redistribution of the "Undetermined" category into all the previously mentioned new sub-categories (from here on, referred as test "2B").

In the training tests of 1A and 1B, we segmented the 4k dataset into three sections, for training, validation and testing purposes, considering a ratio of 60% records for the training set, and 20% for the validation and test sets. Similarly, for the training tests of 2A and 2B, from the Undetermined segment of the data (2385 records), we also considered a ratio of 60%, 20% and 20% for the training, validation and test sets respectively. The distribution was performed with the "train_test_split" function of the sklearn library from Python, using a random seed for consistent selection in a single split, and stratifying the records considering the categories ratios.

2.3 Automated classification process

We continued with the stacked models approach, and considered the better results from the previous tests for the subsequent implementations, as shown in step 3 of Fig. 1. Initially, the first role classifier categorized the full 412k-tweet dataset into "Donor," "Non-donor," and "Undetermined," to validate the results in comparison to the classification values obtained with the classifier from the previous study.

Subsequently, we executed the second blood donation role classifier. Specifically, we used the model trained exclusively on the new citizen-related labels ("Potential" and "Deferred"), as the results for the activity-related labels ("Campaign" and "Informative") were inconclusive because of the small data sizes. Thus, the "Undetermined" category from the 412k dataset, was further divided into "Undetermined", "Potential", and "Deferred." Once the 412k dataset was updated to have the 5 categories, we proceeded with the setup of the OCTIS framework.

2.4 OCTIS configuration for models comparison

For the topic modeling comparison, we first prepared the environment required to execute multiple tests for different models, considering the different elements shown in step 4 of the Fig. 1.

For the models, we selected latent Dirichlet allocation (LDA), non-negative matrix factorization (NMF), and BERTopic. As described in Sect. 1, these are some of the most common methods used in the literature. BERTopic was selected for its ability to leverage pretrained language models for document and word representations, capturing complex relationships between words and context.

For comparison, we used the OCTIS framework (Terragni et al. 2021), which aims to allow researchers and practitioners to make fair comparisons between topic models of interest. Because BERTopic is not included in the framework as a default model, we transformed the resulting data to the input format required by OCTIS. Additionally, a tokenized file and a vocabulary file were also required for the comparison. Thus, we considered two test cases for the preprocessing: using the OCTIS built-in method (spaCy) to transform the corpus; and using a customized process with the MeCab library (MeCab 2024). The latter is a library widely used in the literature to tokenize Japanese texts, which warranted a comparison to validate possible differences in the topic modeling process.

To compare the models, we used topic coherence, topic diversity, and execution time as the main metrics. Coherence measures the interpretability, consistency, and meaningfulness of topic modeling outcomes, capturing the semantic interpretability of discovered topics on the basis of their corresponding description terms. Typically, a higher coherence value in topic modeling results indicates a more effective topic model (Röder et al. 2015). Additionally, coherence allows for the extraction of the optimal number of topics. We used the "c_v" coherence metric, which creates content vectors of words via their co-occurrences and calculates the score via normalized pointwise mutual information (NPMI) and cosine similarity. Diversity is the percentage of unique words in the top 25 words of all topics. A diversity close to 0 indicates redundant topics; a diversity close to 1 indicates more varied topics (Dieng et al. 2020).

We defined a pipeline to execute each of these models with a seed hyperparameter for reproducibility, and a topk = 10 (top number of words on which the topic diversity will be computed) for the metrics of comparison. The value of 20 topics, commonly used in previous studies, was chosen for the evaluation. However, fewer topics in large datasets, such as ours, can negatively affect the coherence. As a comparative measure, we also included the value of 50 topics for testing.

For further analysis, we compared multiple versions of the BERTopic implementations by using different sentence transformer models, which were accessed from the webpage HuggingFace: `sentence-paraphrase-multilingual-MiniLM-L12-v2` (referred to as BERT-MiniLM) (sentence-transformers 2024a), `colorfulscoop/sbert-base-ja` (referred to as BERT-sbert) (colorfulscoop 2024), `sonoisa/sentence-bert-base-ja-mean-tokens-v2` (referred to as BERT-ja-mean) (sonoisa 2024), and `sentence-transformers/paraphrase-multilingual-mpnet-base-v2` (referred to as BERT-mpnet) (sentence-transformers 2024b).

2.5 Topic modeling evaluations

After the preparation of both the OCTIS framework and the classified dataset, we prepared two instances of tests for the final validation, as seen in step 5 of Fig. 1. We first executed the topic modeling tests through the OCTIS framework, using the complete dataset (412k tweets), without considering the categories, to determine the model and setup that had more effective results with out data. For the second instance, we separated the complete dataset into five subdatasets based on their labels, and re-executed the Topic Modeling tests for each of them, using the best-performing configurations (model and setup) from the initial tests. For this instance with the sub-datasets, we focused only on the coherence results for more specific comparisons, using once again the values of 20 and 50 topics for the tests. From these results, we aimed to validate if there were significant differences between the execution of the topic modeling process on the complete dataset against its execution on the more specific subsets.

3 Results

3.1 Datasets distributions

The first result of this study was the creation of a training dataset for categorization, building on the recommendations of the previous study mentioned in Sect. 2.1. The initial distribution of our BDT-UC dataset comprised 2399 "Undetermined," 1407 "Donor," and 199 "Non-Donor" tweets. However, as mentioned in Sect. 2.2, we removed tweets that did not contain at least 4 unique characters to avoid excessive random noise, which affected only the "Undetermined" category. The updated distribution of the dataset with the initial 3 categories was of 2385 "Undetermined," 1407 "Donor," and 199 "Non-Donor" tweets.

Once the new labels were included, the dataset distribution changed as follows, solely affecting the "Undetermined" category:

- With only the new citizen-related categories, the distribution was of 1524 *Undetermined*, 1407 *Donor*, 720 *Potential*, 199 *Non-Donor* and 141 *Deferred* tweets.
- When using all the new sub-labels, the final distribution was of 1407 *Donor*, 1395 *Undetermined*, 720 *Potential*, 199 *Non-Donor*, 141 *Deferred*, 83 *Informative* and 46 *Campaign* tweets.

3.2 Classification performance

For the first stage of the classification tests, using the test subset (799 records) from our 4k dataset, we were able to achieve similar classification results (available in the Table 8 of Appendix 3) to the ones of the previous study (Espinoza et al. 2023) when executing the configuration of test 1A. When the trained classifier from test 1A was implemented, the distribution of the collected 412k tweets included 254,862 records labeled as "Undetermined," 157,784 records as "Donor," and 0 records as "Non-donor".

For the test 1B, in which we used the customized model with the mechanism to handle imbalanced data, the metrics per class are shown in Table 1, providing a general accuracy of 0.7584, demonstrating improved balance across categories, For the case of the classifier of test 1B, the distribution was of 186,608 records for the "Undetermined" category, 185,765 records for the "Donor" one, and 40,273 records for the "Non-donor" one.

Table 1 Test 1B metrics results (model with imbalanced data control)—distribution by user category

Categories	Precision	Recall	F1-score	Support
Undetermined	0.7505	0.8689	0.8054	412
Donor	0.8014	0.7107	0.7533	318
Non-donor	0.5500	0.3188	0.4037	69
Accuracy			0.7584	799
Macro avg	0.7006	0.6328	0.6541	799
Weighted avg	0.7535	0.7584	0.7500	799

Table 2 Test 2A metrics results (model focusing on the citizen-related subcategories)—distribution of undetermined 3 subcategories

Categories	Precision	Recall	F1-score	Support
Undetermined	0.9082	0.8738	0.8907	317
Potential	0.7292	0.7955	0.7609	132
Deferred	0.7143	0.7143	0.7143	28
Accuracy			0.8428	477
Macro avg	0.7839	0.7945	0.7886	477
Weighted avg	0.8473	0.8428	0.8444	477

As established, We initially tested the 1B setup with additional labels, but performance was significantly lower (see Tables 9 and 10 in Appendix 3, Sect. 3.1). Thus, to address the low performance in direct multi-label classification, we adopted the multi-step classification approach.

For the second stage of the classification tests, as mentioned in Sect. 2.2, we used the test subset (477 records) of the "Undetermined" data of the 4k dataset. We present the results of the implementation for test 2A in Table 2, with the distribution of results when only "Potential" and "Deferred" were used as the additional subcategories. The classifier had a general accuracy of 0.8428.

Table 3 shows the metric results of the implementation for test 2B using the same test subset of 477 records, with the sub-distribution of five categories, with an overall accuracy of 0.6855.

Among the results, the "Campaign" and "Informative" categories had the lowest scores. Furthermore, the false positives overlapped with each other and the main "Undetermined" category, issue which was also present when we trained the model with the 7 labels directly (both distributions are shown in Figs. 7 and 8 respectively in Appendix 3, Sect. 3.1). The occurrences implied possible issues in detecting noticeable differences between these specific groups, for which we decided to omit these specific labels from the next steps of the study.

Table 3 Test 2B metrics results (model focusing on all the subcategories)—distribution of undetermined 5 subcategories

Categories	Precision	Recall	F1-score	Support
Undetermined	0.6774	0.8363	0.7485	226
Potential	0.7292	0.7664	0.7473	137
Deferred	0.8571	0.4898	0.6234	49
Campaign	0.5556	0.1351	0.2174	37
Informative	0.2353	0.1429	0.1778	28
Accuracy			0.6855	477
Macro avg	0.6109	0.4741	0.5029	799
Weighted avg	0.6753	0.6855	0.6606	799

3.3 Topic modeling results

To compare the topic modeling methods, we used the basic implementation of the BERTopic models. We focused first on the results from processing the dataset as a whole. In this case scenario, we first collected the results when using the preprocessing step incorporated in the OCTIS framework. Table 4 shows the comparison results for the coherence (c_v) metric, for the topic diversity metric, and for the time required by each model to complete the modeling process.

Table 5 shows the results for the same structure of comparison when the MeCab library was used for the preprocessing step of the data. For the majority of the cases, the results with the Mecab preprocessing method follow a similar tendency as the OCTIS method but were comparatively slightly higher for the coherence metric, particularly for the BERT-MiniLM model.

From the metrics comparison, we selected the best performing BERTopic implementation (BERT-MiniLM) for further tests. We implemented additional tests using both preprocessing methods to explore if the number of topics initially selected (20 and 50), were the most optimal choices for the generation of topics. For these tests, we executed the topic modeling process with variations from 20 to 200 topics. In Fig. 2, we show the results of such tests with the standard (default) BERT-MiniLM implementation, comparing the metrics when implementing the OCTIS preprocessing method, and when implementing the MeCab preprocessing method.

3.4 Integrated validation results

The final distribution of the collected dataset, after application of the second classifier, was as follows: 185,765 tweets for the Donor category, 140,626 for the Undetermined one, 40,273 for the Non-donor one, 35,819 for the Potential one and 10,163 for the Deferred one.

Table 6 shows the coherence metrics when running the modeling process for each category, for 20 and 50 topics respectively, with the data preprocessed with the MeCab library. Table 7 follows the same structure for the results of the Diversity metrics.

Table 4 Comparison between models with the OCTIS preprocessing method

	Coherence (c_v)		Diversity		Execution time (s)	
Models	20 topics	50 topics	20 topics	50 topics	20 topics	50 topics
LDA	0.55	0.46	0.77	0.86	**144**	**198**
NMF	**0.61**	**0.59**	0.64	0.60	1008	3307
BERT-MiniLM	**0.62**	**0.63**	**0.84**	**0.87**	**628**	**674**
BERT-sbert	0.53	0.52	**0.86**	**0.87**	674	636
BERT-ja-mean	0.55	0.57	0.82	0.86	624	720
BERT-mpnet	0.54	0.59	0.83	0.84	746	667

Table 5 Comparison between models with the MeCab preprocessing method

	Coherence (c_v)		Diversity		Execution time (s)	
Models	20 topics	50 topics	20 topics	50 topics	20 topics	50 topics
LDA	0.53	0.53	0.80	0.80	**143**	**142**
NMF	0.61	0.63	0.55	0.55	1021	3439
BERT-MiniLM	**0.75**	0.68	**0.87**	**0.86**	745	742
BERT-sbert	0.56	0.55	0.82	0.85	910	655
BERT-ja-mean	0.69	**0.69**	0.80	0.80	755	848
BERT-mpnet	0.59	0.60	0.85	0.81	797	1090

Fig. 2 Coherence and diversity comparison between standard BERT-MiniLM and different preprocessing methods.

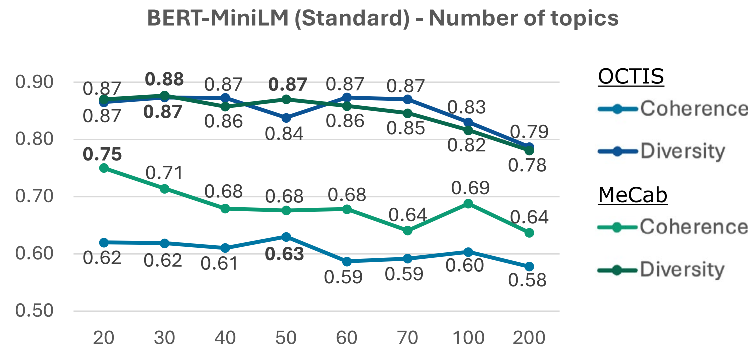

Table 6 Comparison of coherence results between categories and number of topics with MeCab preprocessing

Models	Donor		Non-donor		Undetermined		Potential		Deferred	
	*c_v*20	*c_v*50	*c_v*20	*c_v*50	*c_v*20	*c_v*50	*c_v*20	*c_v*50	*c_v*20	*c_v*50
LDA	0.51	0.51	0.42	0.42	0.58	0.58	0.34	0.34	0.36	0.36
NMF	0.56	**0.58**	**0.50**	**0.46**	0.70	0.68	**0.48**	**0.40**	**0.41**	0.38
BERT-MiniLM	**0.56**	0.51	0.34	0.39	**0.72**	**0.72**	0.40	0.39	**0.41**	0.38
BERT-sbert	0.43	0.46	0.38	0.37	0.52	0.52	0.32	0.35	0.39	0.35
BERT-ja-mean	0.55	0.51	0.42	0.40	0.68	0.69	0.37	0.36	0.38	0.38
BERT-mpnet	0.51	0.51	0.38	0.38	0.64	0.65	0.35	0.35	0.38	**0.40**

Table 7 Comparison of diversity results between categories and number of topics with MeCab preprocessing

Models	Donor		Non-donor		Undetermined		Potential		Deferred	
	div20	div50	div20	div50	div20	div50	div20	div50	div20	div50
LDA	0.81	0.81	0.78	0.78	0.77	0.77	**0.86**	**0.86**	0.76	0.76
NMF	0.56	0.50	0.53	0.45	0.60	0.56	0.58	0.46	0.50	0.39
BERT-MiniLM	**0.89**	**0.87**	0.77	0.82	**0.87**	**0.85**	0.75	0.85	**0.89**	**0.88**
BERT-sbert	0.83	**0.87**	0.78	0.82	0.82	**0.85**	0.74	**0.86**	0.76	0.81
BERT-ja-mean	0.83	0.79	0.79	0.81	0.83	0.79	0.81	**0.86**	0.78	0.84
BERT-mpnet	0.86	0.86	**0.83**	**0.85**	0.82	**0.85**	0.83	0.82	0.83	0.82

4 Discussion

In this study, we evaluated different strategies for analyzing SNS Japanese data with topic modeling, and assessed their impact on overall model precision, topic diversity, and topic coherence. The results provide several insights into the performance and applicability of advanced NLP techniques in this domain, and may serve as a foundation for future studies aiming to delve deeper into the results of a topic modeling project with Japanese data.

4.1 Classification performance

The first stage of our classification tests successfully generated similar results to the ones from the previous study, using their recommended setup, highlighting the robustness of the Tohoku BERT model for this task. However, our customized model revealed a slightly lower overall accuracy (0.7584) compared to the original study results (0.7840), underscoring the challenges of balancing and refining classification processes. This result is consistent with other studies that reported similar difficulties in achieving high classification accuracy in imbalanced datasets Johnson and Khoshgoftaar (2019). Nonetheless, while the accuracy of our proposed version was lower, it yielded more practical results by generating usable "Non-Donor" labeled records, in contrast to the implementation of the initial classifier.

Our customized model showed varied performance across different categories as shown in Table 1. The "Undetermined" and "Donor" categories achieved higher precision, recall, and F1-scores compared to the "Non-donor" category scores. This suggests that while our model largely maintained its effectiveness in identifying clear donor-related content, it still struggles with Non-donor tweets due to the low amount of training data. These findings indicate a need for further refinement, particularly in distinguishing non-donor tweets, which could require more sophisticated feature extraction, additional training data, or a more stringent fine-grained entity recognition approach (Zhang et al. 2020). Such potential for improvement was further highlighted on our validation tests when training the model to classify the data directly to either 5 or 7 categories, as the performance significantly declined, as shown in Tables 9 and 10 of Appendix 3, Sect. 3.1.

The second stage of the classification tests demonstrated significant improvements with the inclusion of subcategories in a separate model, following a multi-step approach, as shown in Tables 2 and 3. The 2A implementation, which added "Potential" and "Deferred" subcategories, achieved a higher overall accuracy of 0.8365. This underscores the value of incorporating more granular categories to capture nuances in the data. However, the 2B implementation with five subcategories highlighted the complexity of further subdivision, as evidenced by lower precision and recall for "Campaign" and "Informative" categories. This suggests that while more categories can provide detailed insights, they also introduce challenges in model performance due to overlapping features and increased ambiguity (Lorena et al. 2019).

Nonetheless, it is considered that the multi-step approach or an stacked models approach have inherent weaknesses regarding classification. While we trained the second model in data manually labeled (the "Undetermined" subsection of our 4k dataset) to avoid possible compound errors during the training phase, some misclassified records were detected during the final classification of the 412k dataset. Determining the final percentage of the compound errors to compare it with the possible misclassification rate from a regular approach could provide us better evidence about which approach might be more suitable for future projects.

4.2 Topic modeling comparison

Our comparison of topic modeling methods revealed that BERT-based models generally outperformed traditional methods such as LDA and NMF in terms of coherence and diversity metrics for our dataset, as shown in Tables 4 and 5. Specifically, the BERT-MiniLM model consistently achieved higher coherence scores, especially when preprocessed with the MeCab library. This finding aligns with previous research suggesting that transformer-based models are more adept at capturing semantic nuances in short text bodies compared to traditional topic modeling techniques (Grootendorst 2022; Gan et al. 2023).

Preprocessing methods significantly impacted the performance of topic modeling methods. The MeCab library, designed for Japanese text, yielded superior coherence scores across most models (with a more detailed example shown in Fig. 9). We consider that the specialization in the Japanese language from the MeCab library could have contributed to the better performance of the tests with that setup, emphasizing the importance of tailored preprocessing for datasets in languages that do not use a single phonetic alphabet. This is particularly relevant for researchers working with multilingual data, highlighting the need to consider language-specific preprocessing tools to enhance model performance (Martin et al. 2019; Lind et al. 2022). While BERTopic does not require a preprocessing step, it may prove beneficial to include such a step for comparison when working with other Japanese datasets to ensure coherence (Egger and Yu 2022).

The execution time varied across models, with NMF requiring significantly more computational resources compared to BERT-based models. While BERT models such as BERT-MiniLM and BERT-sbert showed moderate execution times, their superior performance in coherence and diversity metrics suggest that they are preferable choices for large-scale Japanese text analysis. This balance between computational efficiency and model accuracy is crucial for practical applications, especially in real-time data processing scenarios (Devlin et al. 2018).

However, regarding the setup of the BERT-MiniLM model itself, there is still room for improvement, specially for the selection of the number of topics. While we defined that such number should change according to the amount of records in the dataset, our current results were not conclusive in regards to the most optimal number for our dataset. We should also highlight that the coherence values we obtained were not possible to reproduce in the exact values in every test. There was a ± 0.04 margin of change in the results for each number of topics (possibly related to the stochastic nature of the topic modeling process of BERTopic), which indicates that further validation is still recommended.

In that regard, we also conducted the same tests with a customized version of the BERT-MiniLM model (which can be seen in Fig. 9) to explore if the margin of change can be controlled through more advanced setup of the model. The results were more consistent in the tests, but the coherence decreased considerably, possible because of a suboptimal configuration. A better understanding of all the available modules of the BERTopic model can potentially create more optimal results in future tests. Nonetheless, independently of these last points, we consider that the current metrics do suggest that the proposed implementation of BERT-MiniLM for this study can generate valuable results in comparison to the other analyzed models.

While the current scope of the study did not delve into the exploration and analysis of the topic modeling results, it allowed to identify a suitable setup for such study in the future for a dataset with characteristics such as ours. For an initial overview of such potential analysis, we generated a quick list of the detected topics for the BERT-MiniLM tests with 20 and 50 topics (shown in Figs. 10 and 11 respectively), currently only available in Japanese.

4.3 Integration of classes for topic modeling

The coherence metrics for each category shown in Table 6, which were calculated using different models an topic counts, revealed insights into the effectiveness of different NLP techniques in topic modeling performance for our data.

- **Donor category**:
  - The LDA model shows moderate coherence with a score of 0.51 for both 20 and 50 topics, similar to the general analysis.
  - NMF performs better than LDA with coherence scores of 0.56 and 0.58 for 20 and 50 topics, respectively. However, they are lower than those in the general analysis
  - BERT-MiniLM and BERT-ja-mean achieve higher coherence scores in the BERT group, but are slightly lower than the NMF results in this case.
- **Non-donor category**:
  - The NMF model has the highest coherence scores, suggesting that it effectively identifies distinct topics within this category.
  - LDA also outperforms the BERT models with coherence scores of 0.42 (for both 20 and 50 topics), indicating a weakness of BERT models when working with data of this category.
- **Undetermined category**:
  - The coherence scores for this category are generally higher across all the models when the results were compared with those of the general analysis.
  - Furthermore, the BERT-mini results remained the highest ones, suggesting its ability to capture some underlying themes in the undetermined category.
- **Potential and deferred categories**:
  - The coherence metrics for these categories are less conclusive, with scores varying across models and topic counts and being considerably lower than those in the general analysis.
  - The BERT models showed promise with higher coherence scores for the deferred category, indicating their potential in identifying themes related to deferral reasons.

Regarding diversity, all the models show some consistency in their results independently of the category and number of topics as shown in Table 7, which were also close to their results in the general analysis. BERT models remained the best performers regarding diversity, with BERT-Mini being the best model from the group.

The differences in the results when executing topic modeling on the full dataset versus by specific categories suggest that there are underlying challenges in the ability to capture nuanced themes. While BERTopic models, specifically the BERT-Mini implementation, appeared to be the best option to implement for our data analysis, it is not as suitable as initially implied if we aim to work with the more specific sections of the dataset. One of the reasons for these differences in coherence values could be related to the size of the specific data corpus per category, since the smallest categories have lower specialized coherence. As mentioned before, BERTopic is expected to perform better when analyzing large data corpus, which does not apply to imbalanced categories that have a low number of records.

However, with that in consideration, the donor category was expected to have higher coherence than the ones obtained from the implementation. A possible interpretation for this difference is that, as the undetermined category had further subcategorization, the remaining elements had more cohesion for the coherence analysis. In that context, performing a similar subcategorization process for the donor category could improve the coherence result. Another possible interpretation is that the donor category ideas were not adequately segmented with the selected number of topics for the tests. If the selected number of topics was too low, the model might have not identified accurate classifiers; if the number was too high, the model could have forced excessive segmentation that made interpretation less accurate (Zhao et al. 2015). Exploring the optimal number of topics for each category is considered for future work, as it can lead to a better and more coherent generation of topics per category, and thus, to easier identification of the key issues for each type of citizen.

Nonetheless, performing a specific analysis per category appears to be valuable in the context of the topic results, considering that the diversity values remained considerably high for each individual category. For example, the diversity scores in the Non-donor category, while lower than those in the other groups, still remained at approximately 0.80 when BERTopic models were used, suggesting that focusing on specific user groups with such models can lead to more insightful and actionable findings. When applying topic modeling to the entire dataset, the models often identify general topics that may overlook category-specific details. In contrast, categorizing the dataset and then applying topic modeling to each category separately can allow the models to uncover more refined and relevant themes within each group. However, given the differences in coherence metrics previously mentioned, additional hyperparameter optimization, as well as data preprocessing will be needed to achieve more consistent and coherent results.

Improving the metrics across categories (in addition to the undetermined category) could more reliably highlight diverse and more specialized concerns and motivations, which can be crucial for designing targeted public health interventions for specific categories of citizens regarding blood donation. Therefore, while analyzing the full dataset provides an overarching view, breaking down the data into specific categories can enhance the granularity and specificity of the insights, making them more practical for addressing particular public health challenges.

4.4 Implications and future work

The findings from this study have several implications for future work in regards to public health campaigns and social media analysis in the context of blood donation and Japan. The improved classification accuracy for blood donation-related tweets builds the foundation to enhance targeted outreach efforts by identifying and engaging potential donors more effectively. Additionally, identifying the model and setup that better performs with datasets of characteristics such as ours, reduces the effort required for preparation, facilitating the direct exploration of the thematic analysis itself. The integration of categorization and topic modeling analysis (with optimized setup) can lead to studies that generate more focused insights and can lead to the design of more compelling and relevant campaign messages, tailored to the themes and concerns prevalent in social media discourse (Harrell et al. 2022; Zhang and Liu 2024).

Future work could explore several avenues to build on these findings and address the current study limitations. First, integrating additional contextual features, such as user demographics and engagement metrics, could further refine the classification models. Second, applying transfer learning techniques to leverage pretrained models on related tasks could enhance performance, particularly for underrepresented categories (Ruder et al. 2019). Finally, extending the analysis to other social media platforms could provide a more robust and diverse source of information on public sentiment and behavior (Tuck and Thompson 2021), initially related to blood donation and subsequently to other welfare topics of interest for the Japanese community.

Another area for future research is the application of unsupervised and semi-supervised learning techniques to further improve the classification and topic modeling processes for Japanese text. These techniques can leverage large amounts of unlabeled data, which are abundant in social media contexts, to enhance model training and performance. While in this study we only focused on the coherence, diversity and time metrics, the topic modeling setup can still be further improved to make its results more consistent. Similarly, the topic modeling results can be further explored with our Japanese dataset to properly delve into the topics generated for the general dataset and the subdatasets per category, which we could not focus on in this study as it has lots of room for exploration.

Furthermore, our current work opens the path for future incorporation of sentiment and time analysis into the topic modeling framework, which could yield deeper insights into the emotional tone and public perception of blood donation campaigns (Yadollahi et al. 2017), ultimately benefiting resource management for blood donation centers in Japan. The study could be further extended with the usage of the English translations available in the dataset. While they were used only for reference in this study, it could be worth to explore possible differences in the topic modeling setup and analysis with different languages in the source dataset.

5 Conclusion

The integration of advanced NLP techniques with social media data offers significant potential for public health research and intervention. This study demonstrates the potential for such methods to provide meaningful insights into public attitudes and behaviors regarding blood donation, ultimately aiding in the design of more effective public health campaigns.

We demostrated the effectiveness of a customized Tohoku BERT model in classifying blood donation-related tweets and highlighted the superior performance (in coherence and diversity metrics) of BERT-based topic modeling methods over traditional approaches, such as LDA and NMF, especially when applied to categorized datasets. Our study also emphasizes the importance of tailored preprocessing methods in modeling implementations for datasets in languages that do not use a single phonetic alphabet, particularly in the context of Japanese social media analysis. Furthermore, the comparison between topic modeling on the full dataset and categorized subsets underscores the importance of granular analysis. Categorizing the data can affect the capture of specific themes relevant to each group, in which higher coherence can lead to more actionable and precise public health strategies.

Future research should further refine these proposed implementations, and explore their application across different health-related topics to optimize their impact on public health outcomes, such as blood donation.

Appendix 1: Labeled Tweets Samples

Sample tweets for each defined category are provided for reference in this Appendix. We included five tweets per each label category, including their their cleaned content and the translation of their content in the following list. We omitted the original content of the tweets from this submission, as there were conflicts in the Latex document when trying to include both the Japanese characters and the emoji at the same time.

- Campaign:

  1. Tweet: 1650053153386290000
     - Cleaned: carp公式サイトの新着news 助かるいのちがある「みんなの献血day」開催❗
     - Translation: New News from the official Carp website! There is a life to be saved "Blood Donation Day for Everyone"! #Carp. https://t.co/RD5RllYPXD
  2. Tweet: 1578250000000000000
     - Cleaned: 石川県学生献血推進委員会への匿名のメッセージを募集中❗ マシュマロを投げ合おう
     - Translation: We're looking for anonymous messages to the Ishikawa Student Blood Donation Promotion Committee. #Let's throw marshmallows at each other. https://t.co/s8WPOxfrKN
  3. Tweet: 1640319998651660000
     - Cleaned: 川口駅献血ルーム埼玉県川口市栄町3 1 240120 353 611
     - Translation: Kawaguchi Station Blood Donation Room 3-1-24 Sakae-cho, Kawaguchi-shi, Saitama 0120-353-611 https://t.co/6ADnvy18aG
  4. Tweet: 1591338448526660000
     - Cleaned: 早番さんからも献血 マルハンの他店舗からも、スタッフの家族や、お客さま、地域の方々 ご協力ありがとうございました マルハン甲府 献血
     - Translation: Blood donations from the early shift as well. We also received donations from other MARUHAN stores, staff families, customers, and community members. Thank you for your cooperation. #MARUHAN Kofu. #BloodDonation https://t.co/7YzrZ2TRC8
  5. Tweet: 1592851871009360000
     - Cleaned: ネイマール献血キャペーンでo型と判明 ネイマール代表49ゴールバルセロナ１シーズン３９点ロナウジーニョ代表33ゴールバルセロナ１シーズン２６点 血
     - Translation: Neymar blood donation campaign revealed as type O . https://t.co/bYN9caNnqs . Neymar 49 national team goals, 39 Barcelona 1 season, Ronaldinho 33 national team goals, 26 Barcelona 1 season . #blood type #A #B #O #AB. https://t.co/0SE7FwyaMb

- Deferred:

  1. Tweet: 1477944027106660000
     - Cleaned: 最近献血に失敗して数週間利き手が痛かった。
     - Translation: My dominant hand was sore for weeks after a recent failed blood donation
  2. Tweet: 1506985854602190000
     - Cleaned: 血管見えんてずーっと断られてるんや 人生で1度でいいから献血したい
     - Translation: I've been turned down for years because I can't see the blood vessels. I want to donate blood just once in my life.
  3. Tweet: 1492025681588670000
     - Cleaned: 仕事終えて買い物に寄ったスーパーに献血車があったから、『よし久々に』と張り切って検査したら、ヘモグロビンの数値が足りなくて出来ませんでしたお茶だけ頂いて帰宅しました
     - Translation: I stopped by the supermarket for shopping after work and saw a blood donor van, so I was all excited and said 'okay after a long time!!', but I couldn't do it because my hemoglobin level was insufficient, I just got some tea and went home. https://t.co/rhjeUiawmx.
  4. Tweet: 1495382814426280000
     - Cleaned: 最近全然できていなかったので開封の儀 まずはさんからの返礼品 お薬のせいで献血できないのが残念だけれども
     - Translation: I haven't been able to do this at all lately, so here's the opening ceremony. First of all, @BrightJapanFC returned my blood. Too

bad I can't donate blood because of my meds https://t.co/kQiseD1JEA

5. Tweet: 1633985027523610000

   - Cleaned: ウワァアア献血バスだ！！献血させろォ！！！！！(薬のせいで出来ない)(なんで)(どうして)(おれは社会貢献がしたいぞジョジョーッ！
   - Translation: Wow, a blood donation bus! Let me donate blood ！！！！！ (I can't because of my meds...(Why?)(Why?)(I want to contribute to society, JoJo!)

- Donor:

1. Tweet: 1519501574447730000

   - Cleaned: 暇やったから献血した。
   - Translation: I donated blood because I had time on my hands.

2. Tweet: 1505107642695390000

   - Cleaned: 置いてなかったので、拾った勇気で買いました。. 献血もしました。
   - Translation: They didn't have it, so I bought it with the courage I picked up. I also donated blood.

3. Tweet: 1509676887320140000

   - Cleaned: 自分の部屋掃除してたら数年前に献血した時貰った乃木坂のシール見つけた！数年前までアイドルなんて全然興味無かったのにな〜笑笑.
   - Translation: I was cleaning my room and found a sticker of Nogizaka that I got when I donated blood a few years ago! I never cared about idols until a few years ago. #Nogizaka https://t.co/8LQMpp9fCy

4. Tweet: 1485550838999250000

   - Cleaned: 月曜日だけど献血に行ったー！ なんと！前まで11 6とかだったギリギリの数値が12 5まで爆上がり！ 無事400mlの献血が出来ましたよっ おやすみの日にめっちゃ沢山食べたからですね！ 定期的に献血して、元気な血液を届けられるよう頑張ります！ そしてカルピスの原液をもらう。 飲めない！
   - Translation: It's Monday and I went to donate blood! What a surprise! My blood count, which used to be 11.6 but was just barely there, went up to 12.5! I was able to donate 400ml of blood safely. I ate a lot of food on my good night! I will do my best to donate blood regularly and deliver healthy blood! And I get undiluted Calpis. I can't drink it!

5. Tweet: 1516619934960270000

   - Cleaned: 献血しながら、gyao で配信している〜魔法先生ネギま！2〜見てる。第3話の構成が原作から大胆に入れ替わっていて、スラム編に2話かける。神鳴流宴会芸のタイミングが微妙に早い。世界観の説明に大事なシリーズだけど、後に続かない。
   - Translation: While donating blood, I'm watching #UQHOLDER! The structure of the third episode has been boldly switched from the original, spending two episodes on the slam version. The timing of the Kaminari-style banquet performance is subtle and early. The series is important for explaining the worldview, but it doesn't follow.)

- Informative:

1. Tweet: 1597553115007710000

   - Cleaned: みんなからの匿名質問を募集中！ こんな質問に答えてるよ ●好きな事 ●けんけつしてくだ ●死にたいと思った事ありますか？ ●「待ち人、超来たる」っていう漫... 質問箱匿名質問募集中
   - Translation: We're looking for anonymous questions from everyone! We answer questions like.... What do you like to do? I'd like to go to a blood donation center, but I'm afraid of getting a big needle in my arm. Have you ever wanted to die? I'm a comic called "Waiting is Super Coming". .... #Question Box #Anonymous Questions. https://t.co/D0vqiinbAj

2. Tweet: 1645249496434060000

   - Cleaned: 西暦2023年4月10日、中四国の献血状況ヲ....。 ☆400ml ａ型 ｏ型 ｂ型 ab型 ☆200ml ａ型 ｏ型 ｂ型 ab型 ☆成分献血 ａ型 ｏ型 ｂ型 ab型
   - Translation: On April 10, 2023 A.D., the blood donation situation in Chugoku and Shikoku wo....... Type A . Type O. Type B . Type AB . 200ml. type A . Type O . Type B . Type AB .

Component blood donation. Type A . Type O . Type B . Type AB .

3. Tweet: 1631462285259720000

   – Cleaned: モンタナ州で接種者の献血を犯罪へ』⇒アメブロより
   – Translation: Montana to criminalize blood donations from vaccinators:!?’. ⇒ https://t.co/UJGGIsG5Kh #Ambro from @ameba_official

4. Tweet: 1600084148659840000

   – Cleaned: 献血映画って…笑。確かにマックス血抜かれて爆走されてたね〜笑。あの映画モノクロ含め観すぎてちと食傷気味かも…苦笑。でも大好きね、あれ❗サンキュ❗勇気…僕の場合、今ならrrrかな〜さぬさん観てないならマジにダッシュで❗ヤバいからっ。今年最高…いやこの数十年…クラスだからあの映画❗
   – Translation: @sanucker A blood donation movie...lol. I’m sure Max was being blasted for being drained of blood...lol. I’ve seen so many of those movies, including the black and white ones, that I’m getting a little sick of them...lol. But I love that one! Thank you! Courage...in my case, I’d say RRRR now...if you haven’t seen it, Sanu-san, seriously, dash to it! It’s not good. Best of the year...or the last few decades...that movie is class!

5. Tweet: 1591583417527070000

   – Cleaned: ↓献血ルーム（献血センター）の求人の特徴↓献血ルーム（献血センター）の求人のメリットは、日勤のみで、ルーティンワークが多いため、パートやアルバイトでも働きやすいことです。しかし給料は低めな傾向があります。
   – Translation: ↓Characteristics of blood donation room (blood donation center) jobs The advantage of blood donation room (blood donation center) jobs is that they are day shift only, and since most of the work is routine, it is easy to work part time or even part time. However, salaries tend to be lower. https://t.co/xxksfxrdk9

- Non-Donor:

1. Tweet: 1535813790977200000

   – Cleaned: 献血はできないのよごめんねー
   – Translation: I can’t donate blood, sorry. https://t.co/9Kms55vDoU

2. Tweet: 1522799069093000000

   – Cleaned: 若気の至りで開けた舌ピが献血を阻む健康体なのにもうしわけない
   – Translation: The tongue piercing I opened in my youthful folly is preventing me from donating blood, and I’m sorry for that, even though I’m in good health.

3. Tweet: 1481512686387360000

   – Cleaned: 献血の呼びかけをしてるスタッフの方を見るたび、俺は輸血経験者だから力になれないんだ……すまん………って気持ちになる
   – Translation: Every time I see a staff member calling for blood donations, I feel like I can’t help because I’ve been transfused ...... sorry ....

4. Tweet: 1478297830435730000

   – Cleaned: 採血で失神するから献血なんて一生いけねえ
   – Translation: I’ll never be able to donate blood because I faint when they take blood.

5. Tweet: 1620121007104350000

   – Cleaned: 血液型/o型。献血に3回くらい行ったけど、断られました。ウルルンとか行ってたんで……(uniformers/2008)
   – Translation: Blood type/O. I went to donate blood about 3 times but was turned down. I was going to Ulleung and other places. ......(UNIFORMERS #4/2008)

- Potential:

1. Tweet: 1538684381245550000

   – Cleaned: 今日は献血行って映画2本社観る予定。あくまで予定。
   – Translation: I’m going to donate blood and watch 2 movies today. I’m not sure if I’m going to be able to do it.

2. Tweet: 1493408901307540000

   – Cleaned: 今すぐ献血行ってこい

  - Translation: @xWdyynhHl4hQv9T Go donate blood now.

3. Tweet: 1577080000000000000

  - Cleaned: 献血しに行ってこんなハイパーイケメン出てきたら血圧上がりまくってしまう最高です
  - Translation: When I go to donate blood and see such a hyper-handsome guy, my blood pressure goes up! It's the best!

4. Tweet: 1509452384636100000

  - Cleaned: 次回献血20回記念だし、 献血行ってチケット当てるぞー！
  - Translation: I'm going to donate blood next time and win a ticket! I'm going to donate blood and win tickets! https://t.co/zY3IVkVdya

5. Tweet: 1586560000000000000

  - Cleaned: バス到着に立ち会おうって思ってたのに 別のバスに乗っちゃってるよ 献血バスー
  - Translation: I was going to be there when the bus arrived. I'm on another bus. Blood donation bus.

- Unknown:

1. Tweet: 1581500000000000000

  - Cleaned: オタクの血なんか入れるな！って言ってる馬鹿は フェミニストの血なんか汚らわしいから献血行かないでくれる？って言われたら どうせ発狂するんやろ？ 人に言われて嫌な事は するな 言うな っていう最低限の教育すらスルーしてきた 腐った脳みそばっかだもんなお前等
  - Translation: Don't let any nerd's blood in! Don't let any nerd's blood in! Don't go donate blood because feminist blood is disgusting. If they were asked... You'd go crazy, wouldn't you? Don't do what you don't like people saying. Don't do it. Don't say it. That's the least we can do. You guys have rotten brains, don't you?

2. Tweet: 1477162806327780000

  - Cleaned: 「オタクの鑑」コミケ献血に長蛇の列所要時間150分も 2日で590人参加の大盛況 (j castニュース) ニュース
  - Translation: Long lines at Comiket blood donation event. 590 people participated in the two-day event (J-CAST News). #Yahoo News. https://t.co/3fBjUPQjrj

3. Tweet: 1639562971881010000

  - Cleaned: 何時に行けば献血してもらえる（何だよしてもらえるって）権利を得られるのこれ……
  - Translation: @magic_aikatsu What time do I have to go to get the right to donate blood (what do you mean, "donate"?) ......

4. Tweet: 1483220226955440000

  - Cleaned: 献血した人に肉親を殺されたので恨み骨髄なのかもしれんけど、なら輸血拒否の宗教にでも入られたらよろしいので、他人が献血したりすることにケチつけるな
  - Translation: I don't know if you have a bone of contrition because the person who donated blood killed your family member, but if so, join a religion that refuses blood transfusions and don't be stingy with other people donating blood.

5. Tweet: 1493179646191430000

  - Cleaned: qブランドはお前の祖父の血がなければパターンプリントが出来ない？ a正確には、抜き検査に完全部位指定があり、一部血を使う。何の動物かは絶対に明かすなだが、悪い場合は俺の祖父の献血の血全てだ。それを決めて、変えたり組み替えたり図式できるのは、決まった聖なる血を持つ者のみ。私の血統だ
  - Translation: Q Can't a brand make a pattern print without your grandfather's blood? A To be exact, there is a full part designation for the extraction test, and some blood is used. Never reveal what animal it is, but in the worst case scenario, it is all of my grandfather's donated blood. Only those who have a certain holy blood can determine it and change or recombine or figure it out. It's my bloodline. https://t.co/f2R3OQ2sBy

Appendix 2: Pipelines' Code Snippets

In this Appendix we share snippets of the pipelines' code sections for the initialization of variables and hyperparameters used for the different trainings and tests.

The code for each pipeline (data collection, BERT models training and deployment, OCTIS setup and testing) is currently stored in an internal repository. It will be made available on Github in the future, but will be provided by the authors on request until the public release.

Appendix 2.1: BERT models training and deployment

See Fig. 3.

Appendix 2.2: OCTIS setup

See Figs. 4, 5, and 6.

```
### Variable Settings ###
os.environ["TOKENIZERS_PARALLELISM"] = "false"
os.environ["CUDA_VISIBLE_DEVICES"] = "2"
task = 'tohoku3'
mode = 'general' ###Use 'general' to work with full labeled dataset, or 'category' when subdividing the Undetermined category
repeated = True
PRE_TRAINED_MODEL_NAME = BERT_TASKS[task]
batch_samples = 2
num_workers = 0
num_labels = 3 ### Also including additional sublabels for future work
batch_size = 15 if mode != 'general' else 12 ### multiple of the number of labels
MAX_LEN = 150 if task != "izumi" else 128
epochs = 40
patience = 5
learn_rate = 2e-5 ### Testing between 1e-5, 2e-5 & 5e-5 #### it was 1e-5 here for category
cleaned = True  ### If text gets processed for noise reduction (emoji, links, mentions, etc.)
balanced = False  ### If oversampling for balancing dataset
weight_test = True  ### If using weight value for loss calculation for balancing dataset
new_layer = True  ### When testing modbatmaxiels with the added new layers
layer_type = 2  ### Type of new layer, goes from 1 to 6
hidden = True  ### If including all the hidden layers
seed_val = 17
metric_method = 'macro' ## macro or weighted according to test (not needed after implementation of sklearn library)
```

Fig. 3 Code section for the initialization of the parameters and hyperparameters used for the BERT models training—current setup for test 1A

```
### Dataset setting ###
DIRECTORY = "/home/roberto/projects/dataFiles/"
set_type = "general" ### Use 'general' when testing the full dataset, use the label name when testing subdatasets
cleaned = True ### To use the cleaned text instead of the raw tweet values
stop_setting = True ### Variable to manage memory issues
models = [LDA, NMF] ### Select the models from the OCTIS framework
bert_models = [
                'sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2',
                'colorfulscoop/sbert-base-ja',
                'sonoisa/sentence-bert-base-ja-mean-tokens-v2',
                'sentence-transformers/paraphrase-multilingual-mpnet-base-v2',
                'sentence-transformers/stsb-xlm-r-multilingual'
            ]
cv = Coherence(texts=dataset1.get_corpus(),topk=10, measure='c_v') ### Setup of the coherence metric
topic_diversity = TopicDiversity(topk=10) ### Setup of the diversity metric
dicts = Dictionary(dataset1.get_corpus()) ### Setup of the Dictionary required for OCTIS validations
r = [20,50] ### number of topics to use for tests (options limited by memory and processing power)
hyper = [{'random_state': 42},{'random_state': 42}] ### seeds for the OCTIS available models
test_type = 'MeCab' ### Indicate if MeCab or OCTIS preprocessing method will be used for the test
```

Fig. 4 Code section for the initialization of the parameters and hyperparameters used for OCTIS comparison—setup for MeCab tests for the LDA, NMF and BERTopic models

```python
def bertopic_test(data, n_topics, model_name):
    # Create new `pandas` methods which use `tqdm` progress
    # (can use tqdm_gui, optional kwargs, etc.)
    tqdm.pandas()

    gc.collect()
    torch.cuda.empty_cache()

    japanese_bert_model = SentenceTransformer(model_name)
    embeddings = japanese_bert_model.encode(data, show_progress_bar=True)

    topic_model = BERTopic(
        language="japanese",
        embedding_model=japanese_bert_model,
        calculate_probabilities=False,
        verbose=True,
        nr_topics=n_topics,
        )

    start = time.time()
    print('Starting model: ' + model_name)
    bertopics, probs = topic_model.fit_transform(data, embeddings)
    end = time.time()
    print('Finished model: ' + model_name)
    computation_time = end - start

    return bertopics, topic_model, computation_time
```

Fig. 5 Code section for the general setup of the BERTopic model for the OCTIS evaluation

```python
results_bert_collection = []
results_model = []

for results in results_bert:
    for result in results:
        topics_group = {}
        topics_group["topic-word-matrix"] = result[1].get_topics()
        topics_test = [
        [word for word, _ in result[1].get_topic(topic) if word != ""] for topic in topics_group["topic-word-matrix"]
        ]
        topics_group["topics"] = topics_test
        topics_group["comp_time"] = result[2]
        results_model.append(topics_group)
    results_bert_collection.append(results_model)
    results_model = []
```

Fig. 6 Code section for the extraction and transformation of BERTopic results for their integration and comparison in the OCTIS evaluation

Appendix 3: Additional results: BERT tests

In this section we provide for reference the results of additional results reported through the multiple tests of the BERT model implementations.

Appendix 3.1: Classification performance

See Tables 8, 9, 10 and Figs. 7 and 8.

Table 8 Test 1A metrics results (model without imbalanced data control)—distribution by user category

Categories	Precision	Recall	F1-score	Support
Undetermined	0.8868	0.7951	0.8385	532
Donor	0.7128	0.7528	0.7322	267
Non-donor	0.0000	0.0000	0.0000	0
Accuracy			0.7810	799
Macro avg	0.5332	0.5160	0.5236	799
Weighted avg	0.8286	0.7810	0.8030	799

Table 9 Alternative test 1B metrics results (model with imbalanced data control) for 5 labels output—distribution by user category

Categories	Precision	Recall	F1-score	Support
Potential	0.6389	0.6917	0.6643	133
Deferred	0.4286	0.4000	0.4138	30
Unknown	0.5806	0.6835	0.6279	237
Donor	0.6773	0.7992	0.7332	239
Non-donor	0.5250	0.2333	0.3231	90
Accuracy			0.6095	799
Macro avg	0.5351	0.4422	0.4442	799
Weighted avg	0.6146	0.6095	0.5853	799

Table 10 Alternative test 1B metrics results (model with imbalanced data control) for 7 labels output—distribution by user category

Categories	Precision	Recall	F1-score	Support
Potential	0.6389	0.6917	0.6643	133
Deferred	0.4286	0.4000	0.4138	30
Unknown	0.5806	0.6835	0.6279	237
Donor	0.6773	0.7992	0.7332	239
Non-donor	0.5250	0.2333	0.3231	90
Informative	0.1176	0.1667	0.1379	12
Campaign	0.7778	0.1207	0.2090	58
Accuracy			0.6095	799
Macro avg	0.5351	0.4422	0.4442	799
Weighted avg	0.6146	0.6095	0.5853	799

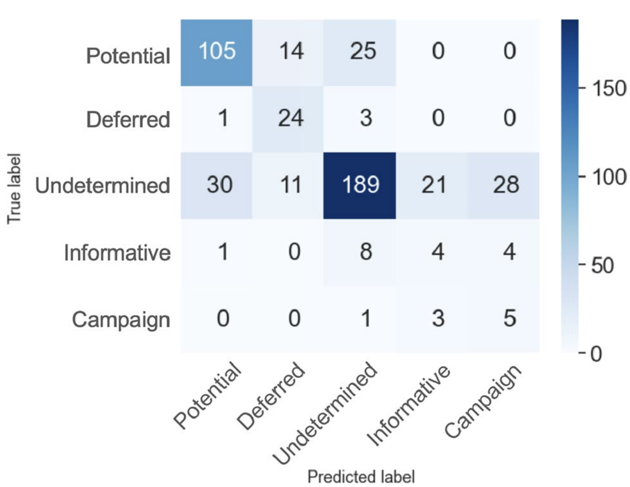

Fig. 7 Distribution of predicted labels—BERT model for subclassification of the undetermined category with 5 labels

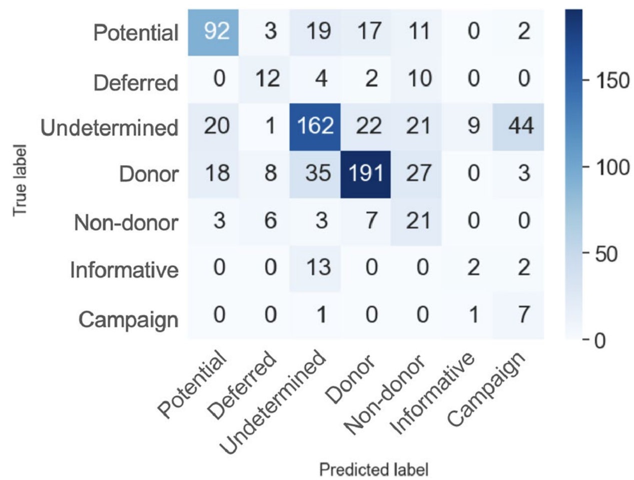

Fig. 8 Distribution of predicted labels—BERT model for general classification of data with 7 labels

Appendix 3.2: Topic modeling metrics

See Fig. 9.

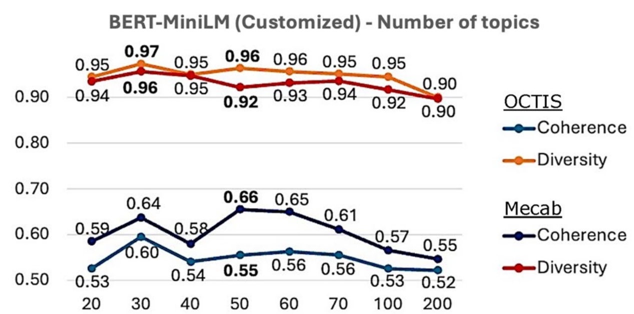

Fig. 9 Coherence and diversity comparison between customized BERT-MiniLM without CountVectorizer layer and different preprocessing methods

Appendix 3.3: Topic modeling examples

See Fig. 10, 11 and Table 11 and 12.

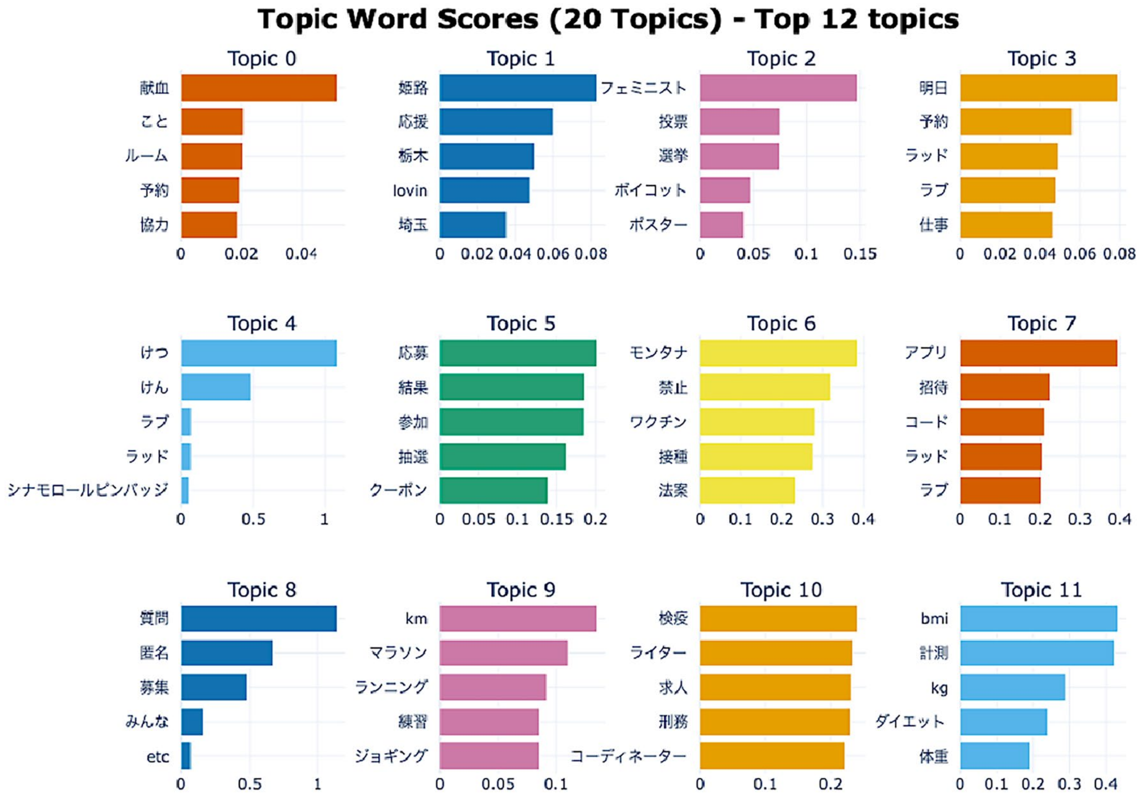

Fig. 10 Top words representation for top 12 topics—BERT-MiniLM with 20 topics and MeCab setup

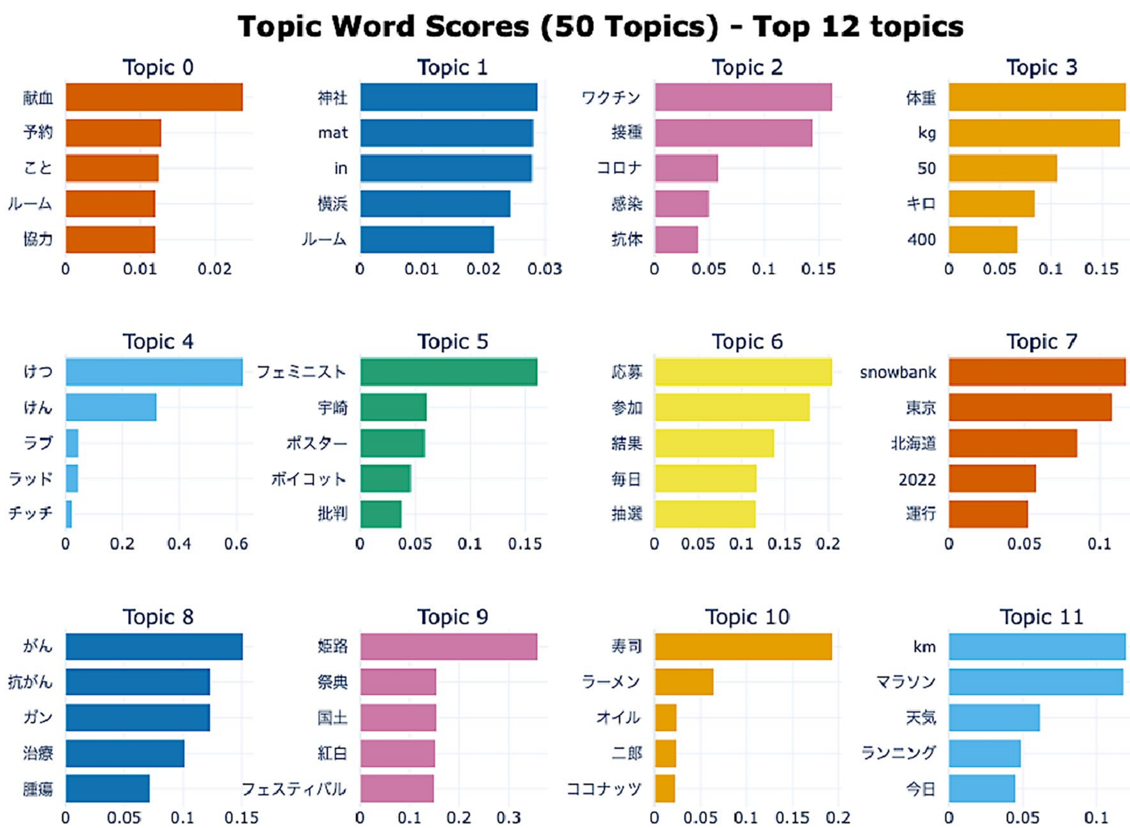

Fig. 11 Top words representation for top 12 topics—BERT-MiniLM with 50 topics and MeCab setup

Table 11 Translations of keywords (from top to bottom) for the top topics results from Fig. 10

Topic	Translations	Details of the keywords
0	Blood donation, situation, room, reservation, cooperation	Topic related to the blood donation activity in society
1	Himeji, support, Tochigi, lovin, Saitama	Topic related to supporting events from different Japan Prefectures
2	Feminist, vote, election, boycott, poster	Topic related to social issues regarding specific blood donation events
3	Tomorrow, reservation, "Blood", "Love", work	Topic related to the usage of the "LoveBlood" application for setting up blood donation schedule
4	Blood, donation, "Love", "Blood", "Cinnamonroll" pin badge	Topic related to the usage of the "LoveBlood" application for rewards exchange
5	Apply, result, participation, lottery, coupon	Topic related to raffle events in blood donation campaigns
6	Montana, prohibition, vaccine, inoculation, law	Topic related to possible restrictions in donations regarding the vaccine
7	App, invitation, code, "Blood", "Love"	Topic related to the usage of the "LoveBlood" application for referrals
8	Question, anonymous, recruitment, everyone, etc.	Topic related to consultations about blood donation activities
9	km, marathon, running, practice, jogging	Topic related to a marathon event for blood donation
10	Quarantine, writer, job offering, prison, coordinator	Topic related to job opportunities in medical fields, including blood donation centers
11	BMI, measurement, kg, dieting, body weight	Topic related to body health restrictions regarding blood donation

Table 12 Translations of keywords (from top to bottom) for the top topics results from Fig. 11

Topic	Translations	Details of the keywords
0	Blood donation, reservation, situation, room, cooperation	Topic related to the blood donation activity in society
1	Shrine, mat, in, Yokohama, room	Topic related to supporting events in a Yokohama temple
2	Vaccine, inoculation, Corona, infection, antibody	Topic related to possible restrictions for blood donation in regards to the Covid vaccine
3	Body weight, kg, 50, kilo, 40	Topic related to body weight restrictions regarding blood donation
4	Blood, donation, "Love", "Blood", "Chicchi"	Topic related to the usage of the "LoveBlood" application for rewards exchange
5	Feminist, "Uzaki", poster, boycott, criticism	Topic related to social issues regarding specific blood donation event about a series
6	Apply, participation, result, everyday, lottery	Topic related to raffle events in blood donation campaigns
7	Snowbank, Tokyo, Hokkaido, 2022, operation	Topic related to annual supporting events from different Japan Prefectures
8	Cancer, anticancer, "Cancer", treatment, tumor	Topic related to restrictions for blood donation in regards to cancer situation
9	Himeji, festival, national land, red & white, "Festival"	Topic related to annual end-of-the-year event in Himeji
10	Sushi, ramen, oil, "Jiro", coconut	Topic related to food self-rewards after blood donation participation
11	km, marathon, weather, running, today	Topic related to a marathon event for blood donation

Acknowledgements We express our thanks to the students who supported the main author during the labeling process of the dataset.

Author contributions R.E.C performed the literature research, design of the study, collection of the data, and its processing and analysis. K.K, C.L, L.S, Y.M, G.Y, and T.K supported and supervised the proposed content of the manuscript. All the authors derived points for discussion and implications regarding the results. All the authors drafted and revised the manuscript.

Funding The main author holds a Japanese government (Monbukagakusho: MEXT) scholarship for the doctoral program. This work was supported by JST (CREST Grant JPMJCR21M1), Japan.

Data availability The tweet datasets generated and analyzed in this study are available in their dehydrated form in the following repository site: https://zenodo.org/records/12541478. The pipelines' code generated for the study can be provided by the authors on request. Some snippets of the code can be found in the Appendix 2.

Declarations

Competing interests The authors declare that they have no competing interests.

Consent for publication Not applicable.

Ethics approval and consent to participate Not applicable.

References

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

colorfulscoop/sbert-base-ja. Hugging Face. Accessed 2 January 2024. https://huggingface.co/colorfulscoop/sbert-base-ja

Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Dieng AB, Ruiz FJR, Blei DM (2020) Topic modeling in embedding spaces. Trans Assoc Comput Linguist 8:439–453. https://doi.org/10.1162/tacl_a_00325

Egger R, Yu J (2022) A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts. Front Sociol 7:886498

Espinoza R, Liu C, Kishimoto K, Yamamoto G, Mori Y, Santos L, Kuroda T (2023) Adjusting Twitter data as a source for blood donation analysis: BDT-UC dataset and BERT implementations. In: 2023 IEEE EMBS special topic conference on data science and engineering in healthcare, medicine and biology. IEEE, pp 27–28

Gan L, Yang T, Huang Y, Yang B, Luo YY, Richard LWC, Guo D (2023) Experimental comparison of three topic modeling methods with LDA, Top2Vec and BERTopic. In: International symposium on artificial intelligence and robotics. Springer, pp 376–391

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

Harrell S, Simons AM, Clasen P (2022) Promoting blood donation through social media: evidence from Brazil, India and the USA. Soc Sci Med 315:115485

Johnson JM, Khoshgoftaar TM (2019) Survey on deep learning with class imbalance. J Big Data 6(1):1–54

Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401(6755):788–791

Lind F, Eberl J-M, Eisele O, Heidenreich T, Galyga S, Boomgaarden HG (2022) Building the bridge: topic modeling for comparative research. Commun Methods Meas 16(2):96–114. https://doi.org/10.1080/19312458.2021.1965973

Lorena AC, Garcia LP, Lehmann J, Souto MC, Ho TK (2019) How complex is your classification problem? A survey on measuring classification complexity. ACM Comput Surv (CSUR) 52(5):1–34

Martin L, Muller B, Suárez PJO, Dupont Y, Romary L, La Clergerie ÉV, Seddah D, Sagot B (2019) CamemBERT: a tasty French language model. arXiv preprint arXiv:1911.03894

MeCab (2024) Yet another part-of-speech and morphological analyzer. Accessed 2 January. https://taku910.github.io/mecab/

Munikar M, Shakya, S, Shrestha A (2019) Fine-grained sentiment classification using BERT. In: 2019 Artificial Intelligence for Transforming Business and Society (AITB), vol 1, pp 1–5. IEEE

Qin Z, Cong Y, Wan T (2016) Topic modeling of Chinese language beyond a bag-of-words. Comput Speech Lang 40:60–78. https://doi.org/10.1016/j.csl.2016.03.004

Röder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on web search and data mining, pp 399–408

Ruder S, Peters ME, Swayamdipta S, Wolf T (2019) Transfer learning in natural language processing. In: Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, pp 15–18

sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2. Hugging Face. Accessed 2 January 2024. https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

sentence-transformers/paraphrase-multilingual-mpnet-base-v2 (2024) Accessed 2 January. https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

sonoisa/sentence-bert-base-ja-mean-tokens-v2. Hugging Face. Accessed 2 January 2024. https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

Terragni S, Fersini E, Galuzzi BG, Tropeano P, Candelieri A (2021) OCTIS: comparing and optimizing topic models is simple! In: Proceedings of the 16th conference of the European chapter of the association for computational linguistics: system demonstrations, pp 263–270

tohoku-nlp/bert-base-japanese-v3 (2024) Hugging Face. Accessed 2 January. https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

Tuck AB, Thompson RJ (2021) Social networking site use during the Covid-19 pandemic and its associations with social and emotional well-being in college students: survey study. JMIR Formative Res 5(9):26513

Wagner S, Fernández DM (2015) Chapter 3—Analyzing text in software projects. In: Bird C, Menzies T, Zimmermann T (eds) The art and science of analyzing software data. Morgan Kaufmann, Boston, pp 39–72. https://doi.org/10.1016/B978-0-12-411519-4.00003-3

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M et al (2020)Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pp 38–45

Yadollahi A, Shahraki AG, Zaiane OR (2017) Current state of text sentiment analysis from opinion to emotion mining. ACM Comput Surv (CSUR) 50(2):1–33

Zhang Z, Liu Q (2024) Rational or altruistic: the impact of social media information exposure on Chinese youth's willingness to donate blood. Front Public Health 12:1359362

Zhang T, Wang Y, Wang X, Yang Y, Ye Y (2020) Constructing fine-grained entity recognition corpora based on clinical records of traditional Chinese medicine. BMC Med Inform Decis Making 20:1–17

Zhao W, Chen JJ, Perkins R, Liu Z, Ge W, Ding Y, Zou W (2015) A heuristic approach to determine an appropriate number of topics in topic modeling. In: BMC bioinformatics, vol 16. Springer, pp 1–10

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.