BERT-based Classification of Four Major Dementias using Twitter Text Data

1st Kazuki Utsunomiya
Kogakuin University
Tokyo, JAPAN
em23007@ns.kogakuin.ac.jp

2nd Ryohei Banno
Kogakuin University
Tokyo, JAPAN
banno@computer.org

***Abstract*—In Japan, the declining birthrate and aging population have become a social problem, and it is predicted that the number of elderly people will reach about 40 million in 2040. As a result, the number of dementia patients is expected to increase. There are four major dementias, and appropriate care methods are different for each. In addition, early detection is important to suppress symptoms. Therefore, there is a need for a means to easily determine the type of dementia. In this study, we propose an automatic classification method for four major dementias. Using crowdsourcing, we extract text data of four major dementia symptoms from Twitter. By fine-tuning BERT with them, we obtain a classifier. Experimental result shows that the proposed method provides higher accuracy than random classification.**

***Index Terms*—Healthcare, Machine Learning, Text Classification, BERT, Dementia.**

I. Introduction

In Japan, a declining birthrate and an aging population have become a social problem.The number of elderly people is estimated to be about 40 million [1]. Regarding future estimates of the number and prevalence of dementia aged 65 and over [2], the number of elderly with dementia in 2012 was 4.62 million, about one in seven aged 65 and over. However, it is estimated that by 2025, it will be about one in five people. Therefore, the number of people with dementia is expected to continue to increase.

It is known that there are four major dementias, and appropriate care methods are different for each. In addition, early detection is important to suppress symptoms. Therefore, there is a need for a means of easily determining the type of dementia. Therefore, in recent years, dementia screening based on language ability has attracted attention, but manual screening may be difficult due to problems such as a decrease in medical personnel due to the declining birthrate and a shortage of hospitals. Existing studies include a method for extracting and classifying information from the field [3] and a method for constructing and evaluating machine learning models by performing explanatory work [4], which is expected to contribute to solving labor shortages. It has also been pointed out that data collection is difficult.

In this study, we focused on the possibility of using BERT to solve the shortage of workers and the possibility of solving the data shortage by collecting texts from Twitter. We propose a classification method by fine-tuning BERT with Twitter texts to reduce symptoms through early detection.

Identify applicable funding agency here. If none, delete this.

A. Four Major Dementias

Dementia is sometimes considered a single disease, but its causes, symptoms, and methods of coping with symptoms are diverse. Research until recent years has revealed that they have several tendencies. It is called four major dementias [5]. Here, we describe the characteristic symptoms of the four major dementias and the main care methods for each.

- Alzheimer's dementia
  It is caused by extensive brain atrophy centering on the hippocampus, which controls memory. Memory disorders and delusions mainly appear. After that, as the symptoms progress, hallucinations and disorientation become prominent. People with Alzheimer's dementia often talk about the same things over and over again and cannot recognize the names of things correctly. These could make family members who live together feel stressed, but it is essential not to deny such symptoms and create an environment that makes it easy for the person to spend time.
- Vascular dementia
  Vascular dementia involves both dementia and cerebrovascular disease, and also the temporal relationship between dementia and the onset of cerebrovascular disease. A characteristic symptom is executive dysfunction. Most of the patients are male, and the symptoms progress rapidly in stages. Physical function declines and patients may become bedridden, so physical care such as rehabilitation will lead to improvement of symptoms.
- Dementia with Lewy bodies
  It is caused by poor blood flow in a wide area from the hippocampus, which controls memory, to the occipital lobe, which controls vision, resulting in decreased function. Major initial symptoms are sleep disturbances and visual hallucinations. Autonomic symptoms such as constipation and abnormal sweating also appear. Visual hallucinations are what the patient sees, so denying them can confuse the mind and cause symptoms to worsen. It is important to coordinate conversations to prevent aggravation and alleviate symptoms.
- Frontotemporal dementia
  It is caused by atrophy of the frontal lobe, which controls personality, and the temporal lobe, which controls language. Symptoms include personality changes and stereotypies. As the disease progresses, selfish behavior increases and changes in diet become noticeable. Since

979-8-3503-0219-6/23/$31.00 ©2023 IEEE

the patients become sensitive to human voices and movements, it is important to prepare a quiet environment with little stimulation.

II. Existing Studies

Kosugi et al. [3] conducted an experiment using machine learning to automatically classify text information on behavioral and psychological symptoms of dementia posted on Chienowa net, a site for posting testimonials, into 10 categories. At the time of the experiment, approximately 3,300 pieces of text information related to the behavioral and psychological symptoms of dementia had been accumulated on the dementia Chienowa net, of which approximately 2,000 were targeted. They obtained an accuracy rate of about 50 % as a whole. Also, when deep learning was performed using the same data, it was found that the accuracy rate could be improved to about 73 %, but the collected testimonial data was biased, resulting in differences in classification accuracy for each category.

Shibata et al. [4] constructed an utterance data set from three tasks of image, episode, and animation explanation as a method of corpus construction and classification study of dementia potential, and used it to compare mild dementia patients and healthy subjects. As a result of the experiment, the classification performance was the highest when the average value of the feature values obtained in all tasks was used as the feature value, and it was possible to show the possibility of improving the classification performance when using multiple tasks. However, compared with previous research, the number of samples is 60, which is small, and there is a problem with reliability.

From the above, the number of data is a problem in these existing studies. In this research, we focused on the possibility of solving the problem of the amount of data from a huge number of posts by using Twitter as the information extraction destination.

III. Proposed Method

In this study, we propose a method to automatically classify each dementia by machine learning using BERT from tweets about four major dementias on Twitter. Fig.1 shows an overview of the proposed method. As a flow of the proposed method, we collect tweets about four major dementias, perform morphological analysis and encoding, and then create training data and test data from the tweet text. The created learning data is input to BERT, fine-tuning is performed, and accuracy is evaluated using the fine-tuned learning model and pre-sorted test data.

A. Obtaining Tweets and Preprocessing

We obtain tweets related to the four major dementias and concatenate them to create four texts for each of the four major dementias. As a pre-processing of the text, words are divided using a morphological analyzer, and the divided words are digitized by assigning IDs based on a morphological analysis dictionary called ipadic[6] as shown in Fig.2, and input to BERT.

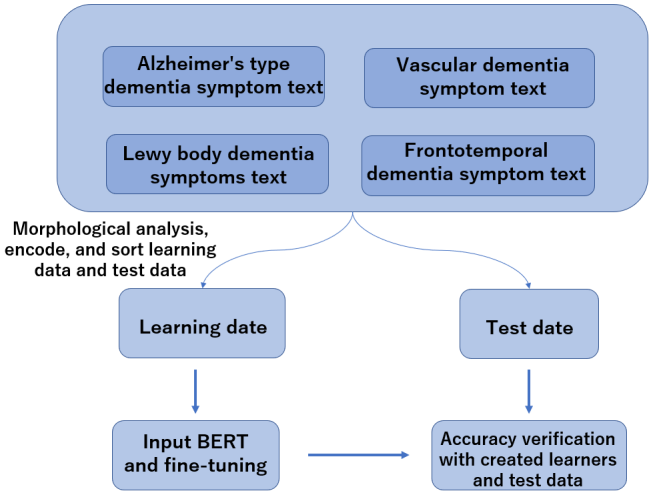

Fig. 1. Outline of proposed method

ID	Token
1	a
2	an
3	,
4	.
5	the
6	and
…	…

Fig. 2. Morphological analysis dictionary example

B. BERT

BERT is a natural language processing model published in a paper[7] by Jacob Devlin et al. of Google in 2018. It is a natural language processing model known for its high context understanding performance, with the highest scores recorded in various natural language processing fields such as translation, document classification, and question answering. As a mechanism of BERT, first, a sequence of word data is called a sentence or a sequence, and BERT predicts another sequence from the input sequence. BERT is a pre-trained model and does not have input labels. It learns by processing unnamed distributed representations. It can be used to overcome the lack of data because it can handle contextual understanding accuracy due to interactive learning, versatility due to transfer learning, and unlabeled data.

IV. Evaluation

To obtain tweets, we collected tweets describing specific symptoms and experiences about one of the four major dementias from Twitter using crowdsourcing. We excluded

tweets that mentioned multiple types of dementia. The acquired items are the user ID, the text of the tweet, and the accompanying tweet ID. The collection period was about 12 years from July 1, 2010 to December 10, 2022. A total of 753 tweets were collected, including 196 for Alzheimer's dementia, 197 for vascular dementia, 189 for Lewy body dementia, and 171 for frontotemporal dementia.

The extracted text was pre-processed on Google Colaboratory, and based on the grammar of the target language and part-of-speech information of words, morphological analysis was performed to break down the sentence into morphemes. As a morphological analysis tool, we used fugashi, which is a Japanese morphological analysis tool MeCab that can be handled from Python. After morphological analysis, the data were used for fine-tuning BERT. Then, the classification accuracy of the four major dementias was measured, and the recall and precision were calculated from the classification results using the macro-average method. The evaluation index F1 was also calculated.

A. Google Colaboratory

Google Colaboratory is a free service provided by Google for machine learning education and research that does not require installation and can immediately set up an environment for Python, machine learning, and deep learning. The Python execution environment required for machine learning can be used from the browser, and the necessary libraries are installed from the beginning, so there is no need to set up the environment, and the GPU required for learning BERT can be used free of charge[8].

B. Macro-averaging

The macro-averaging method is a method of calculating an evaluation index (precision rate, recall rate, etc.) for each class and then averaging them. By taking the average of the evaluation index for each class, the evaluation index can be calculated without being affected by the bias in the number of samples in each class. For example, suppose class 1 has 10000 samples, class 2 has 1000 samples, and class 3 has 1000 samples. In this case, if the macro-average method is adopted as the aggregation method for the precision, the precision for the entire multi-class classification problem is (0.5+0.8+0.8)/3 = 0.7. Thus, the macro-average weights all classes equally and is not overly sensitive to the most occurring class labels.

The macro-average method uses an evaluation index (precision rate, recall rate, etc.) for each classThis is a method of taking the average after computing . The evaluation index for each class By taking the average, it is affected by the bias in the number of samples in each class. The evaluation index can be calculated without For example, 10000 Samples for class 1, 1000 samples for class 2, 1000 samples for class 3, Class 1 has a precision of 0.5, class 2 has a precision of 0.8, and class 3 has a precision of 0.8. In this case the macro When the average method is adopted as the aggregation method of precision, the multi-class classification problem Overall precision is (0.5+0.8+0.8)/3 = 0.7. like this Since the macro-average weights all classes equally, It is not overly sensitive to class labels that appear.

Using Fig.3 as an example, the horizontal axis is the number of texts of text A, and the vertical axis is the number of texts that were actually classified as A. For example, there are 300 texts of text A at the top, 240 texts were actually classified as text A, 20 texts were misclassified as text B, and 40 texts were misclassified as text C. Recall and precision for class A can be expressed as: Then the F value is calculated from the calculated recall and precision. At this time, class A is positive, and the other classes are negative. The recall, precision and F-value for class B and class C were calculated as well.

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$F1 = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{3}$$

Based on the calculated recall, precision, and F value of classes A, B, and C, the macro recall, macro precision, and macro F value can be calculated from the following formulas. Here i represents a class.

$$MacroAverageRecall = \frac{\sum_{i=1}^{n}(\frac{TP}{TP+FN})_i}{i} \tag{4}$$

$$= \frac{\sum_{i=1}^{n} Recall_i}{i} \tag{5}$$

$$MacroAveragePrecision = \frac{\sum_{i=1}^{n}(\frac{TP}{TP+FP})_i}{i} \tag{6}$$

$$= \frac{\sum_{i=1}^{n} Precision_i}{i} \tag{7}$$

$$MacroAverageF1 = \frac{\sum_{i=1}^{n}(2 * \frac{Recall*Precision}{Recall+Precision})_i}{i} \tag{8}$$

$$= \frac{\sum_{i=1}^{n} F1_i}{i} \tag{9}$$

Initially, We thought that the micro-average method, which calculates the evaluation index considering the entire dataset, would be effective because there is not significant bias in the entire text.. However, the bias could be large when we divide data into the learning and test data.It might cause the prediction result to be good only for specific dementia and poorer results for other dementias.Therefore, we choose macro-averaging in this experiment.

C. Evaluation results

In this experiment, the batch size of training data is 32, the batch size of verification and test data is 256, and the number of epochs is 10. Verification was performed three times while changing the number of tokens, and the accuracy of the model was measured and evaluated. As a detailed breakdown, we divided the training and test data into a ratio of 8:2, fine-tuned the training data, and then measured the accuracy using the test data. Table 1 shows the results for each number of tokens. A certain accuracy was obtained, but the accuracy decreased as the number of tokens increased.

TABLE I
EVALUATION RESULTS

Token numbers	Precision	Macro recall	Macro precision	Macro F1
512	0.1667	0.3921	0.3480	0.3295
256	0.3333	0.2773	0.2152	0.2220
128	0.6667	0.4453	0.5517	0.4398

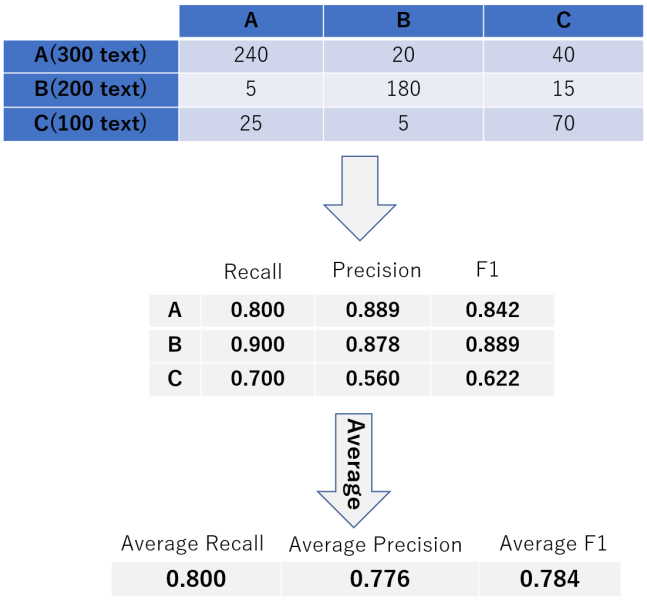

Fig. 3. Macro averaging example

D. Discussion

Table 1 of the experimental results shows that when the number of tokens is 128 and 256, the accuracy exceeds 25 %, which is the expected value for random classification, but when the number of tokens is 512, it falls short. There are several dementias with similar symptoms, and it is possible that learning similar symptoms led to misclassification.

Also, the accuracy decreases as the number of tokens increases. Accuracy improves as the number of tokens increases, but in this study, it was suggested that the accuracy of classification decreased as the number of tokens increased, because similar symptoms made the learning content more complicated. As the number of tokens increases, the recall rate becomes better than the precision rate. In general, recall and precision are in a trade-off relationship, and if one increases, the other decreases. The fact that the recall increases as the number of tokens increases means that the number of negative examples that can be easily judged as positive examples increases for the classifier, but negative examples can be judged correctly.

V. CONCLUSION

In this study, we proposed a machine learning method using BERT for the purpose of automatically classifying four main types of dementia from related tweets on Twitter. Crowdsourcing was used to collect tweets, fine-tuning was performed using training data for model verification, and then test data was used to measure the accuracy of the model and calculate the evaluation index. It can be said that the proposed method was able to obtain a certain degree of accuracy because it exceeded the expected value accuracy in the case of random classification.

As a result of the evaluation, we were able to conclude that as the number of tokens increased, similar symptom texts complicated the learning content and made it difficult to classify dementia. We would like to work on improving the accuracy using the multi-label classification method[9], which allows multiple selections so that it can be handled.

REFERENCES

[1] ISSN 1347-5428 Population Research Series, "Population Projections for Japan: 2016-2065", No.336,July 31, 2017 (in Japanese)

[2] Cabinet Office, White Paper on the Aging Society 2017, "Estimated number of elderly with dementia", 2017(in Japanese)

[3] National Research and Development Agency National Center for Geriatrics and Gerontology, Dementia initial intensive support team member training, "Comprehensive assessment of dementia", 2018.

[4] Naoko Kosugi, Shunsuke Sato, Kenji Yoshiyama, Hiromitsu Suchii, "Study on automatic classification of text information related to dementia care in dementia Chienowa net", DEIM, C21-5, 2021.(in Japanese)

[5] Daisaku Shibata, Kaoru Ito, Shoko Wakamiya, Eiji Aramaki, "Construction of a corpus of elderly people with a control group and development of screening technology for potential dementia patients", Transactions of the Japanese Society for Artificial Intelligence, Vol.34, No.4, pp.B–J11 1–9, 2019.(in Japanese)

[6] Masayuki Asahara , Yuji Matsumoto,"ipadic version 2.7.0 User's Manual",Nara Institute of Science and Technology,2003

[7] Jacob Devlin,"Pre-training of Deep Bidirectional Transformers for Language Understanding",Google AI Language,2018

[8] Takahiro Omi, Kentaro Kaneda, Makoto Morinaga, Ari Emami, Introduction to Natural Language Processing with BERT, Stockmark Co., Ltd., 2021, p44.

[9] Kosuke Yoshimura, Yukino Baba, Hisatsugu Kashima, "Multi-label classification by interdependence model", Japanese Society for Artificial Intelligence, 2K2-1in1, 2017.