Article

Using machine learning to information classification related on child-rearing of infants from Twitter

Mayinuer Zipaer [*] , Minoru Yoshida [*] , Kazuyuki Matumoto [*] , Kenji Kita [*]

Article

Using Machine Learning to Information Classification Related on Child-Rearing of Infants From Twitter

Mayinuer Zipaer [1], Yoshida Minoru [1] Matsumoto Kazuyuki [1] and Kita Kenji [1,*]

[1] Tokushima University,
*Correspondence: c501847005@tokushima-u.ac.jp

Abstract: It is difficult to obtain necessary information accurately from Social Networking Service (SNS) while raising children, and it is thought that there is a certain demand for the development of a system that presents appropriate information to users according to the child's developmental stage. There are still few examples of research on knowledge extraction that focuses on childcare. This research aims to develop a system that extracts and presents useful knowledge for people who are actually raising children, using texts about childcare posted on Twitter. In many systems numbers in text data are just strings like words and are normalized to zero or simply ignored. In this paper, we created a set of tweet texts and a set of profiles created according to the developmental stages of infants from "0-year-old child" to "6-year-old child". For each set, we used ML algorithms such as NB (Naive Bayes), LR (Logistic Regression), ANN (Approximate Nearest Neighbor algorithms search), XGboost, RF (random   forest), decision trees, and SVM (Support Vector Machine) to compare with BERT, a neural language model, to construct a classification model that predicts numbers from "0" to "6" from sentences. The accuracy rate predicted by the BERT (Bidirectional Encoder Representations from Transformers) classifier was slightly higher than that of the NB, LR, ANN, XGboost, RF , decision trees and SVM classifiers, indicating that the BERT classification method was better.

Keywords: Twitter; childcare information; machine learning; numerical classification

1. Introduction

This research aims to develop a system that extracts and presents useful knowledge to people who are actually raising children, using texts on childcare posted on SNS. The information is necessary for childcare changes from moment to moment according to the developmental stage of the child. Even in conventional media such as books and magazines, a lot of information about childcare is provided, but it is difficult to accurately obtain the necessary information at that time, and a system that presents appropriate information to the used is needed. For this problem, we aim to develop a system that collects information on childcare and appropriately presents it according to the user's situation, targeting SNS, especially Twitter. On Facebook, there are many long-form posts about business information and more detailed status reports. Compared to Facebook, many people use Twitter for the purpose of catching the latest information because the freshness of information is higher on Twitter. Users in their 20s to 40s, who are particularly sensitive to trends, use it because they can get real-time information. Since it is easy to post short 140-character posts that you can tweet as soon as you think of it, it seems that there is a lot of real-time information about infants posted by parents who are raising children in their 20s to 40s. Many live voices from parents who are actually raising children are posted on SNS, and it is thought that many posts are in the same situation as the user. It is thought that more useful knowledge can be presented to users by using the actual experiences of parents in similar circumstances.

There are still few examples of research on knowledge extraction focusing on child care, for example, there is research to predicting life events from SNS[1]. They contributed a codebook to identify life event disclosures and build regression models on the factors that explain life event disclosures for self-reported life events in a Facebook dataset of 14,000 posts [2]. Choudhur et al.

collected the users posted about their "engagement" on Twitter and analyzed the changes in words and posts used[3]. Burke et al. also analyzed users who have experienced "unemployed" by advertising or email on Facebook, and analyzed the activities on Facebook before changing stress and taking new jobs. but in this research, we collect texts specialized in childcare [4]. We aim to develop a more accurate method by conducting the analysis. As a method, we mainly use natural language processing technology using neural networks, which have been rapidly developing in recent years, especially.

Our contribution can be summarized as follows.

(1) We aim to develop a method that can perform more semantically accurate analysis by using techniques that can accurately handle numerical expressions related to childcare ("2-year-old child", "37 ° C", "100 ml", etc.). To the best of our knowledge, this is the first research on predicting the age of children appearing in text using the surrounding words.

(2) By using the profile information of the user who posted the text together with the text and grasping the attribute information of the poster, we aim to develop a method that emphasizes the text that is closer to the user's situation.

(3) We used BERT-based neural algorithm as well as several non-neural algorithms including SVM, NB, LR, ANN, XGboost, RF, decision trees, providing exhaustive evaluation on this task.

In these two points, we think that the research will be highly novel in terms of method.

2. Related Work

he relationship between numerals and words in text data has received less attention than other areas of natural language processing. Both words and numerals are tokens found in almost every document, but each has different characteristics. However, less attention is paid to numbers in texts. In many systems, numbers treated in an ad-hoc way, documents are just strings like words, normalized to zero, or simply ignored them.

Information Extraction (IE) is a question-answering task that asks the a priori question, "Extract all dates and event locations in a given document[5]." Since much of the information extracted is numeric, special treatment of numbers often improves performance on IE systems. For example, Anton Bakalov and Fuxman proposed a system to extract numerical attributes of objects given attribute names, seed entities, and related Web pages and properly distinguish the attributes having similar values[6].

those related papers described a study of the Nepali language, but this study focuses on the Japanese language. Chiranjibi et al. proposed a new hybrid feature extraction method that combines both syntax (word bags) and semantics (domain-specific and fastText-based) in the Nepali context[7]. for further improvement in natural language processing (NLP) research works in the Nepali language. Tej Bahadur et al. served different NLP research works with associated resources in Nepali language[8].

BERT is one of the pre-learning models in natural language processing for the large text corpus using the neural network called Pre-training of Bidirectional Trans-formers to fine-tune for each task[9]. The pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. The method of outputting a fixed-dimensional vector regardless of sentence length has the advantage that the more words input has, the more of each word will be in the output vector. In recent years, it has been attracting attention as a versatile language model because it has demonstrated the highest level of results in a wide range of natural language processing tasks. BERT is based on Transformer. The transformer is a neural machine learning model using the attention mechanism. Manually building large dataset with human attribute labels is ex-pensive, BERT does not need such a human label [10]. Zhang et al. identify contextual information in pre-training and numeracy as two key factors affecting their performance, the simple method of canonicalizing numbers can have a significant effect on the results[11].

3. Proposed Method

The method proposed in this research classifies tweets about child-rearing posted on Twitter into classes. The procedure is shown below.

- Acquire a large amount of Japanese text data from Twitter.
- Terms related to child care are selected, and two types of texts (tweets) containing those terms and profiles are collected. We created a set of texts divided into children's developmental stages.
- Using SVMs, etc (the NB,LR,ANN, XGboost, RF , decision trees and SVM), and the neural language model BERT, we build a classification model that predicts numbers from "0" to "6" from sentences.

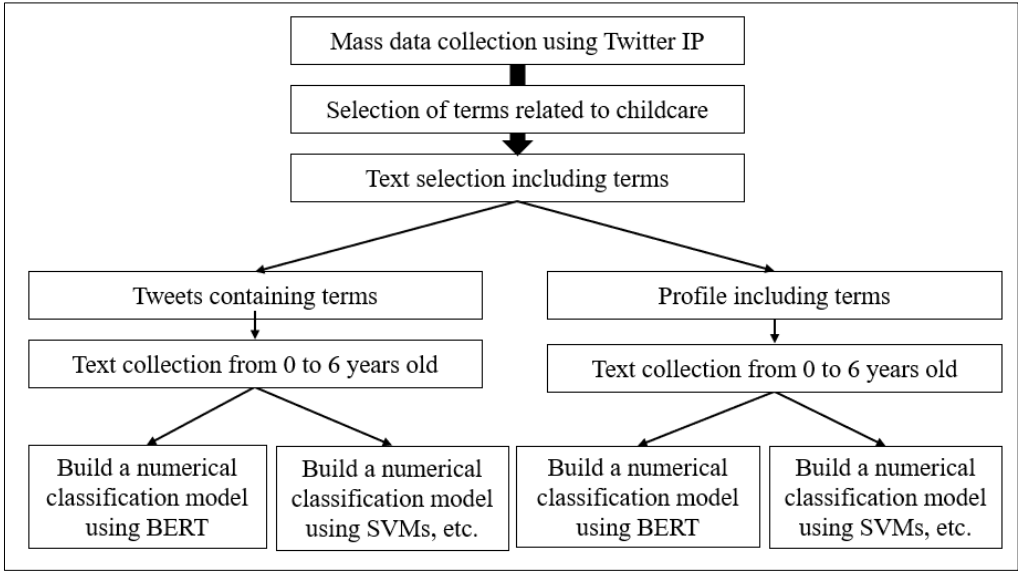Figure 1 shows a simple flow for classifying*+ tweets related to child-rearing.

Figure 1. Flow of proposed method.

3.1. Data collection

In this research, we use Twitter's API to obtain texts that target child-rearing information from Twitter. Next, we select terms related to parenting and screen texts (tweets) containing those terms. Parenting terms are shown in Table 1.

Table 1. The Parenting terms.

Search words used to collect text	Search terms used to collect profiles
child,0-year-old child One year old,1-year-old, Two years old,2-years old, Three years old,3-years old, Four years old,4-years old, Five years old, 5-years old, Six years old, 6 -years old, zero-year-old	raising children, Dad, Mother, Father, my dad, my mum, raising children, child

Two types of text collections are formed, one for containing terms in tweets and the other for containing terms in profiles.   Search for the character string "*-year-old child" and create a set of texts classified by the developmental stage of infants from "0-year-old child" to "6-year-old child". Table 2 and Table 3 show part of the tweet text set and profile set created according to the developmental stages of infants.

Table 2. part of the profile text set collection from 0 to 6 years old.

Old	Part of profile data collection
0	*I am a housewife raising a *-year-old child.*

6	*Full support for Hiroshima Cup in Tokyo. Takuya Kimura was my all-time favorite player. Mother of a * -year-old child. She is from Hiroshima Prefecture. She lives in Tokyo.*
2	*around four-working mother who is raising *-year-old child. I like music, Hanshin Tigers, Daichi Miura, and animals.*
4	*I live with Sunhora and Ayanogo.Adult.I'm quietly doing cosplay. I'm a mom of *-year-old child. RT is too much.\n Archive ID: *52993*
1	*An adventurer who explores cognitive science and sign language with a focus on linguistics.Specializes in cognitive linguistics.Postdoc in R.With *-year-old child. Working mama first grader. http:\/\/ask.fm\/rhetorico*
3	*With * year-old daughter,Shinma who is taking care of her family.\nTwitter is still not working.*
5	*Daily mumblings of an unfortunate rotten adult who likes anime, manga and sometimes games. Childcare (*Year-old child) Mutters here and there. Currently pregnant with second child. (Scheduled for the second half of October) Gestational diabetes, hospitalization for threatened premature labor, etc. I'm already full_ (:3" ∠) _*

Table 3. part of the tweet text set collection from 0 to 6 years old.

Old	*Part of original tweet texts*
4	*I'm coming Tokyo Disney, and when I saw the Coconut tree*year old boy said, "Mom! This is Hawaii!" and I laughed.*
5	**Year-old boy often shakes his head. Isn't the ex-nurse mother-in-law a tic ... https:\/\/t.co\/KstYS2lcJ1*
0	*My daughter's album has stopped until *-year-old 5 months ...that's bad... After all, I haven't had time since I returned to work.*
3	*@0246monpe in May become *-year-old! In the pitch-black room, I can hear the *year-olds humming (groaning?) (´− \`)\n I'm depressed that it will be crowded, but I'm looking forward to it and I have to take him (^◇^;)*
1	*@Alice_ssni*-year-old \n started crawling. he can get over my mother's body too. \n Even when my mother was lying on her back, I was able to come up to π and drink. \n thank you for the meal. Itadakimasu has a high probability of being done by yourself (when you wearing an apron then understand).*
2	*RT @FururiMama98: Parenting is hard and painful. Now, I have a *-year-old daughter, but it's really hard. It's really hard to kill myself and keep looking at others and hugging them. Even after chasing, cuteness: frustration = about 1:9. He is also an unbalanced eater, and is always overturned with toys around the house. How can I rest and comfort myself...*
6	*RT @unikunmama: 🎉happy birthday unnie 🍰\n I turned *-year-old today (o^^o) Let's have fun together from now on, https:\/\/t.co\/ybW9AcS4oP*

3.2. Preprocessing

For the data, the tweet text and profile are used as they are. Also, Therefore, information such as retweets were used as is. We have not done any processing such as removing hashtags.

Data pre-processing is performed in order to perform machine learning on the set of created tweet texts and profiles. In preprocessing, morphological analysis is performed with MeCab, a tool specialized for Japanese language analysis, and the document is vectorized using the Scikit-learn library TfidfVectorizer. we used Unigram features. Unigrams are standard features for text processing, especially for the case where the data size is small. We did not used bigram or trigram features as we did not have enough data for providing meaningful values for these advanced features.

Morphological analysis is the task of breaking down the words that make up a sentence into the smallest units and dividing and writing the sentences. Given a set of documents, TfidfVectorizer converts each document into a vector based on the TF-IDF values. TF (Term Frequency) represents the frequency of occurrence of words in a document. IDF (Inverse Document Frequency) is a score

that lowers the importance of a word that appears in a large number of documents. TF-IDF is a metric that is a product of TF and IDF [2].

3.3. Classification Method

ERT+fine tuning is used in this research, so it is neural supervised learning. We used ML algorithms such as NB, LR, ANN, XGboost, RF, decision trees, SVM model to compare with the pre-trained model with fine tuning BERT model. We used these algorithms because it is a standard non-neural algorithm.

Classifier A and classifier B are created using two classification methods, SVM and BERT.

3.3.1. SVM classification method

The classifier A is constructed using the data divided for each developmental stage of infants. For the text categorization task, we used the <u>NB, LR, ANN, XGboost, RF, decision trees,</u> SVM algorithm and the implementation used the Python machine learning library Scikit-learn to predict the numbers "0" to "6" from the sentences to build a classifier A. Classifier A using NB, LR, ANN, XGboost, RF, decision trees, SVM creates two types: Task Ts (tweet), which uses tweet sentences as data, and Task Ps (profile), which uses profiles.

3.3.2. Classification method by BERT

As a language model, the neural language model BERT has been actively studied in recent years. For the constructing classification model B, we used neural language model BERT to classify from "0" to "6". Two types of classifier B using BERT are created: task Tb (tweet), which uses tweet text as data, and task Pb (Profile), which uses profiles. The BERT model is required to create a classification model. However, it is difficult to prepare a sufficient amount of data set for pre-training to create a model specialized for numerical classification from "0" to "6", so it's a need to fine-tune the pre-trained model. A classifier is created by fine-tuning a pre-trained model.

In BERT, a model specialized for a specific task can be configured by fine-tuning using supervised data for each task, so performance improvement can be expected compared to applying a pre-trained model as it is. As a pretrained model uses a BERT-Base model with 110M parameters and a large model size. The data used for fine-tuning are a set of tweet texts and profile sets created by the developmental stages of infants from 0 to 6 years old. The Task Tb uses the tweet text set, and the Task Pb uses the profile set for fine-tuning.

4. Experimental setup and results

4.1. Classification method by BERT

4.1.1. Data used in experiments

Authors should discuss the results and how they can be interpreted from the perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

We Using Twitter 1.1 Streaming API, get real-time tweets randomly. There is data (21.8GB) of tweets acquired from July 5, 2017 to October 31, 2017. From here on, only tweets containing keywords are used. At that time, we collected tweet texts and profiles containing any of the words shown in Table 1.

Regarding the data size, the profile data is 1575 items and the size is 376 KB, The vocabulary size of profile is 78720 words. The data of the tweet text is 953 items, and the size is 238 KB, The vocabulary size of tweet is 55378 words. Items of each category of Tweet data profile data are shown in the Table 4.

Table 4. Items of each category of Tweet data and profile data.

data	age_0	age_1	age_2	age_3	age_4	age_5	age_6	total
Profile	206	572	255	170	80	85	207	1575
Tweet-	46	200	212	205	105	129	56	953

4.1.2. Experiment details

In the classification experiment, Task Tb and Pb described in Section 3.3.2 are used to classify tweet texts or profiles described in Section 4.1.1, and the performance is compared with classifiers other than BERT.

Classifier B, which uses BERT, classifies the data as described in the classifier. The Task Pb is fine-tuned on the profile set and classifies the profiles of users who are raising infants aged 0 to 6 years. One of the eight parts of the training data is used for verification and the rest for fine-tuning. We checked the classification results in each case. For each classifier output, the result output by the BERT classifier is the normalized probabilities ranging from 0 to 1 for each label, summing to 1. As for the classification results, the one with the highest probability of each label for the input is taken as the output.

In classifier A we also use ML algorithms such as NB, LR, ANN, XGboost, RF, decision tree, SVM for performance comparison. And Nonlinear SVM is to map nonlinear data to a space that becomes linearly separable and linearly separable on a hyperplane. In order to process the Japanese sentences of the tweets to be classified, the input text is converted to vectors. When we classify with SVM modules, we first normalized with StandardScaler and then used the default parameters. The nonlinear SVM classifies each label from "0" to "6" in the identification space according to which region it belongs.

We performed a 5-fold cross-validation on the train set using the train-test split on the profile data and the tweet body data. 80% of the tweets and profile used in the experiment were used as learning/verification data, and 20% as test data. Then, using stratified 5-fold cross-validation, the training data is divided into 5 so that the ratio of labels in each division is the same as the overall ratio, and 4 of the 5 are training data, 1 is used as validation data. to train and evaluate the model.

Once evaluation and training are completed, one of the four training data and validation data are replaced, and the model is trained and evaluated again. By doing this five times and obtaining the average classification accuracy of the five times, and using that value as the classification accuracy, it is possible to perform a robust evaluation that does not depend on the division of the data.

Regarding the adjustment of the parameters of each machine learning method, the above-mentioned cross-validation is performed for all combinations of parameters specified in advance using grid search, and the parameter model that shows the best classification accuracy is generated. Finally, we tested how well each generated model could classify the test data.

4.2. Experimental results

4.2.1. The classification results of the created classifier A.

We performed a 5-fold cross-validation on the train set using the train-test split on the profile data and the tweet body data. Finally, here are the test set results:
Profile data test set results:
Best score on validation set: 0.5834370306801115
Best parameters: {'gamma': 0.01, 'C': 100}
Test set score with best parameters: 0.5736040609137056
Results from a test set of tweet body data:
Best score on validation set: 0.27586920122131386
Best parameters: {'gamma': 0.01, 'C': 100}
Test set score with best parameters: 0.30962343096234307

The confusion matrix is shown in Table 5.

Table 5. The confusion matrix.

		True label	
		positive	negative
Prediction label by	positive	(A)True positive	(B)False positive
SVM	negative	(C)False negative	(D)True negative

True positive (A):   True label positive and prediction positive (correct answer).
False positive (B): True label negative and prediction positive (wrong answer).
False Negative (C): True label positive and prediction negative (wrong answer).
True negative (D): True label negative and prediction negative (correct answer).
Regarding the evaluation of test data, we used three values: accuracy, precision, recall and F1 score.

$$Accuracy=\frac{A+D}{A+B+C+D} \ , Recall=\frac{A}{A+C} \ , Precision=\frac{A}{A+B} \ , F1=2*\frac{precision*recall}{precision+recall}.$$

The classification results for each class from 0 to 6 years old by classifier A are shown in Tables 6 to 12 below.

Table 6. Classification results of 0 years old Class.

Age_0 / Classifier	Result of profile data			Result of tweet data		
	Precision	Recall	F1	Precision	Recall	F1
NB	0.62	0.40	0.48	0.00	0.00	0.00
LR	0.90	0.26	0.40	0.00	0.00	0.00
ANN	0.53	0.46	0.49	**0.11**	**0.05**	**0.07**
GXBoost	0.31	0.53	0.39	0.00	0.00	0.00
RF	0.35	**0.81**	0.49	0.00	0.00	0.00
Decision Tree	0.19	0.27	0.22	0.00	0.00	0.00
SVM	**0.94**	0.35	**0.52**	0.00	0.00	0.00

Table 7. Classification results of 1 years old Class.

Age_1 / Classifier	Result of profile data			Result of tweet data		
	Precision	Recall	F1	Precision	Recall	F1
NB	0.56	0.79	0.65	**0.54**	0.33	**0.41**
LR	0.45	**0.98**	0.62	0.50	0.05	0.10
ANN	0.62	0.62	0.62	0.32	0.45	0.38
GXBoost	0.78	0.57	**0.66**	0.32	0.31	0.32
RF	**0.96**	0.48	0.64	0.30	0.31	0.31
Decision Tree	0.78	0.49	0.60	0.16	**0.69**	0.26
SVM	0.48	0.94	0.64	0.47	0.16	0.24

Table 8. Classification results of 2 years old Class.

Age_2 / Classifier	Result of profile data			Result of tweet data		
	Precision	Recall	F1	Precision	Recall	F1
NB	0.53	0.41	0.46	0.15	0.20	0.17
LR	**0.93**	0.25	0.40	0.19	0.95	0.32
ANN	0.50	0.43	0.46	0.34	0.31	0.32
GXBoost	0.56	0.50	**0.53**	0.55	0.27	0.36
RF	0.30	**0.75**	0.43	0.64	0.26	0.37
Decision Tree	0.47	0.33	0.39	**0.96**	0.24	0.38

SVM	0.86	0.30	0.44	0.26	**0.77**	**0.39**

Table 9. Classification results of 3 years old Class.

Age _3	Result of profile data			Result of tweet data		
Classifier	Precision	Recall	F1	Precision	Recall	F1
NB	0.62	0.27	0.37	0.24	**0.37**	0.29
LR	**1.00**	0.20	0.33	**0.50**	0.02	0.04
ANN	0.45	0.51	**0.48**	0.21	0.24	0.23
GXBoost	0.33	0.41	0.36	0.14	0.21	0.17
RF	0.34	**0.74**	0.47	0.21	0.27	0.24
Decision Tree	0.05	0.29	0.08	0.05	0.33	0.09
SVM	0.68	0.37	**0.48**	0.38	0.36	**0.37**

Table 10. Classification results of 4 years old Class.

Age _4	Result of profile data			Result of tweet data		
Classifier	Precision	Recall	F1	Precision	Recall	F1
NB	0.33	0.21	0.26	0.05	0.04	0.04
LR	**0.67**	0.12	0.21	0.00	0.00	0.00
ANN	0.26	0.28	0.27	0.20	0.15	0.17
GXBoost	0.11	0.50	0.17	0.12	0.23	0.16
RF	0.25	0.62	**0.36**	0.12	**0.75**	**0.21**
Decision Tree	0.01	**1.00**	0.18	0.00	0.00	0.00
SVM	0.62	0.25	**0.36**	**0.50**	0.04	0.07

Table 11. Classification results of 5 years old Class.

Age _5	Result of profile data			Result of tweet data		
Classifier	Precision	Recall	F1	Precision	Recall	F1
NB	0.46	0.63	**0.53**	0.26	**0.19**	022
LR	**1.00**	0.14	0.25	**1.00**	0.04	0.07
ANN	0.56	0.40	0.47	0.13	0.09	0.11
GXBoost	0.33	0.46	0.39	0.09	0.11	0.10
RF	0.29	**1.00**	0.44	0.00	0.00	0.00
Decision Tree	0.14	0.43	0.21	0.00	0.00	0.00
SVM	**1.00**	0.29	0.44	0.43	0.13	0.20

Table 12. Classification results of 6 years old Class.

Age _6	Result of profile data			Result of tweet data		
Classifier	Precision	Recall	F1	Precision	Recall	F1
NB	0.82	0.55	0.65	**0.50**	0.11	0.18
LR	**0.91**	0.36	0.51	0.00	0.00	0.00
ANN	0.54	0.78	0.64	0.06	0.17	0.09
GXBoost	0.53	0.62	0.57	0.12	**1.00**	**0.21**
RF	0.46	0.76	0.58	0.00	0.00	0.00
Decision Tree	0.10	**1.00**	0.18	0.00	0.00	0.00
SVM	0.80	0.59	**0.68**	0.00	0.00	0.00

For analysis of classification results by age class

Age 1 was high in F1, rising to 60% overall, and F1 in Age 4 was low, below 36% overall. It is thought that characteristic texts appear more in children around the age of 1 than in children around the age of 4. The largest number of profile data for the 1-year-old class used for all classifiers is 572

items. Compared to that, the profile data of the 4-year-old class is less at 80 items. Consider that the classification result is affected by the amount of data.

4.2.2. Results of classifier B using BERT

Table 13 shows the classification results of the created classifier B. Classifier B training data, validation data, and evaluation data are randomly divided, so the results change each time the program is run. Table 13 shows the results obtained after many runs.

Table 13. Classification result of 6-year-old class by classifier B.

BERT Class	Result of profile data			Result of tweet data		
	Precision	Recall	F1	Precision	Recall	F1
Age_0	0.35	0.58	0.44	0.25	0.14	0.18
Age_1	**0.81**	0.55	0.66	0.25	**0.50**	0.33
Age_2	0.69	0.64	**0.67**	0.36	0.26	0.30
Age_3	0.65	0.69	**0.67**	0.24	0.28	0.26
Age_4	0.38	**1.00**	0.55	**0.64**	0.44	**0.52**
Age_5	0.12	0.50	0.20	0.31	0.44	0.36
Age_6	0.48	0.83	0.61	0.00	0.00	0.00

From these results of Tweet text data, it can be seen that the accuracy rate of classifier B is slightly higher than that of classifier A.

The results of BERT classification of profile text data are 67% for F1 and 68% for classifier A, so there is not much discrimination, but precision, recall, and f1 are overall more stable than classifier A.

4.2.3. Results by classifier

We have compared using more ML algorithms such as NB, LR, ANN, XGboost, RF, decision tree, SVM. The classification results for each classifier are shown in Table 14 below.

Table 14. The classification results for each classifier.

All class Classifier	Results of profile data				Result of the tweet data			
	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
NB	0.62	0.54	0.57	0.54	0.28	0.24	0.24	0.24
LR	0.74	0.53	0.47	0.53	0.40	0.20	0.10	0.20
ANN	0.54	0.54	0.54	0.54	0.24	0.24	0.24	0.24
XGBoost	0.62	0.54	0.57	0.54	0.36	0.26	0.28	0.26
RF	**0.79**	0.56	**0.60**	0.56	0.46	0.28	0.32	0.28
Decision Tree	0.61	0.43	0.49	0.43	**0.88**	0.26	**0.36**	0.26
SVM	0.70	**0.57**	0.55	**0.57**	0.35	**0.31**	0.26	**0.31**
BERT	0.70	**0.61**	**0.63**	**0.61**	0.34	**0.31**	0.31	**0.31**

About the result obtained from the classification result of Profile text data,

- The accuracy rate for Decision Tree was low at 43%, but there was not much difference for the others.
- RF and SVM showed stable and high accuracy.
- F1 had the highest RF at 60%.
  For analysis of Tweet text data classification results,
- Accuracy rate was generally lower than the accuracy rate of Profile. Here, F1 in Decision Tree was the highest at 36%.

5. Conclusions

In this study, we find that the results obtained by classifying the set of tweet texts with BERT are higher than the results obtained by classifying with SVM. the results obtained by classifying the set of profile with SVM are higher than the results obtained by classifying with BERT.

In order to improve the accuracy rate, we increase the amount of data and remove noise in the data preprocessing.

In recent years, there has been a lot of research into visualizing the learning results of neural network models (research on the explain ability of models) to solve the problem of how to present

language models that have actually been obtained to users. By incorporating these research results, we plan to present learning results to users.

The limitations, only one tweet is used, so it cannot be used when analysis of multiple tweets is required.

Future plans are as follows. Collect information about infants from Twitter to increase the size of the data. Data preprocessing is to denoise text data and to remove stopwords. We will try to apply more sophisticated models such as GPT for comparison in the future.

References

1. Maryam, Khodabakhsh.; Fattane, Zarrinkalam.; Hossein, Fani.; Ebrahim, Bagheri. Predicting Personal Life Events from Streaming Social Content, the 27th ACM International Conference on Information and Knowledge Management, Torino Italy October 22 - 26, 2018.
2. Koustuv, Saha.; Jordyn, Seybolt.; Stephen, M, Mattingly.; Talayeh, Aledavood.; Chaitanya, Konjeti.; Gonzalo, J, Martinez.; Ted, Grover.; Gloria, Mark.; Munmun De Choudhury.; What Life Events are Disclosed on Social Media, How, When, and By Whom? CHI '21: CHI Conference on Human Factors in Computing Systems, Yokohama Japan, May 8 - 13, 2021.
3. Choudhury, M.D.; and Massimi, M.; "She said yes!" Liminality and Engagement Announcements on Twitter, Proc. iConference 2015, Newport Beach, CA. USA, pp.1-13 (2015).
4. Burke, M.; and Kraut, R.; Using Facebook after Losing a Job: Differential Benefits of Strong and Weak Ties, Proc. 2013 Conf. Computer Supported Cooperative Work and Social Computing (CSCW 2013), San Antonio, TX, USA, pp.1419-1430 (2013).
5. Minoru, Yoshida.; Kenji, Kita.; Mining Numbers in Text: A Survey;2021；IntechOpen.
DOI: http://dx.doi.org/10.5772/intechopen.98540
6. Anton, Bakalov.; Ariel, Fuxman.; Partha, Pratim, Talukdar.; Soumen, Chakrabarti.; SCAD: collective discovery of attribute values. WWW 2011: 447-456.
7. Chiranjibi, Sitaula .; Tej, Bahadur Shahi.; Multi-channel CNN to classify nepali covid-19 related tweets using hybrid features;arXiv:2203.10286v1 [cs.CL] 19 Mar 2022.
8. Tej Bahadur , Shahi.; Chiranjibi ,Sitaula,; Natural language processing for Nepali text: a review;Artificial Intelligence Review (2022) 55:3401–3429; https://doi.org/10.1007/s10462-021-10093-1
9. Jacob, Devlin.; Ming-Wei, Chang.; Kenton, Lee.; and Kristina, Toutanova.; Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
10. Kohei, Yamamoto.; Kazutaka, Shimada.; Acquisition of Periodic Events with Person Attributes. IALP 2020, Kuala Lumpur, Dec 4-6, 2020.
11. Xikun, Zhang.; Deepak, Ramachandran.; Ian, Tenney.; Yan, Elazar.; Dan, Roth: Do Language Embeddings capture Scales? EMNLP(Findings) 2020: 4889-4896.