Deep Learning-Based Estimation of SNS Users' Place of Residence Using Posts and Following Relationships

Hiroki Hiramatsu[1] and Kazuaki Ando[2,*]

[1]Graduate School of Science for Creative Emergence, Kagawa University, Kagawa, Japan
[2]Faculty of Engineering and Design, Kagawa University, Kagawa, Japan
Email: s23g361@kagawa-u.ac.jp (H.H.); ando.kazuaki@kagawa-u.ac.jp (K.A.)
*Corresponding author
Manuscript received November 21, 2024; revised January 24, 2025; accepted April 28, 2025; published August 6, 2025.

***Abstract*—In this paper, we propose a new method for estimating the place of residence of Social Networking Service (SNS) users living in Japan. In our previous paper, we proposed a model that estimates the users' place of residence at the prefecture level based on the content of their posts and the weather and earthquake information described in them. While the weather and earthquake information contributed to an improvement in precision, the information did not enhance recall, particularly in prefectures with a small number of users or those adjacent to large metropolitan prefectures. Therefore, we need to improve the recall for these prefectures in this task. To address this issue, we focus on that the users living in rural areas tend to have closer relationships than those in urban areas. We aim to improve recall, especially in rural prefectures, by utilizing the following relationships between SNS users. In addition, we introduce Bidirectional Encoder Representations from Transformers (BERT) model as feature extraction from tweet content to further improve precision. In our evaluation experiments using the same dataset as in our previous study, we achieved an F1-measure of 0.7747. This is the best result in our study so far. Furthermore, we observed an improvement in the recall of up to 5 points compared to the baseline model using only the tweet content.**

***Keywords*—social networking service, location estimation, following relationship**

I. Introduction

We are developing a system that collects and analyzes Japanese posts mentioning the onset of users' diseases and symptoms on social media for public health surveillance. This system visualizes the spread of infectious diseases and the occurrence of various symptoms by prefecture and over time [1–3]. By using this system, it is possible to capture trends related to various diseases and symptoms, including unknown infectious diseases, and to enable early intervention. The proposed system consists of a factuality analysis module that determines whether SNS posts containing diseases or symptoms are intended to indicate the onset of the poster's own symptoms, and a location estimation module that estimates the place of residence of users who have posted content related to diseases or symptoms. The factuality analysis module collects posts containing diseases or symptoms, and verifies whether the users who have posted are confirmed to be experiencing the diseases or symptoms. Then, for users who have been confirmed to have the diseases or symptoms, the system collects additional information and uses the location estimation module to estimate the place of residence by prefecture. The system visualizes the spread of infectious diseases and the occurrence of various symptoms by linking the estimated locations with the mentioned diseases and symptoms and overlaying them on a map of Japan. Users can detect outbreaks of various diseases and symptoms by tracking regions with a significant increase in posts mentioning health concerns over a specific time period.

In this paper, we focus on the location estimation module. To detect posts about diseases and symptoms by prefecture, it is necessary to determine the location information of the users who posted them. However, in X (formerly Twitter), which is used in this paper, explicit location information is known to be scarce. Hashimoto *et al.* [4] reported that the percentage of geotagged tweets in Japan is only about 0.18% of all tweets. Yamaguchi *et al.* [5] found that only about 25% of Japanese SNS users had entered their residence information in the "Location" field of their profiles. Therefore, it is difficult to estimate the place of residence of users who do not include location-related words in their tweets, as well as in their profile.

To address the issue, we proposed a method based on a deep neural network to estimate users' place of residence by using not only tweet content but also weather and earthquake information described in their tweets and external weather and earthquake data provided by official institutions in our previous studies [1, 2]. By using both weather and earthquake information in addition to tweet content, we achieved a best F1-measure of 0.7277. In particular, precision improved by about 6 points compared to using only tweet content, confirming that weather and earthquake information significantly contribute to improving the precision. On the other hand, weather and earthquake information did not enhance recall. Even the best recall did not exceed 70%, especially for users living in rural prefectures with fewer users.

In this paper, we propose a new method of the place of residence based on the friendship network between SNS users to improve recall. Users living in rural areas tend to have closer relationships than those living in urban areas. Therefore, if we can reflect this regionality through the following relationships between users, we expect to improve the estimation performance, particularly in rural prefectures. Furthermore, we introduce BERT as a feature extraction method from tweet content to enhance the estimation performance based on tweet content, and improve overall performance by using both the tweet content and the following relationships.

II. Related Work

Various methods have been proposed for estimating the place of residence of users, not only for public health surveillance but also for purposes such as public opinion polls

and local content recommendations. Particularly in the case of estimating the place of residence of X (formerly Twitter) users, three main types of information were used: tweet content, Twitter network, and tweet context. The Twitter network can be further divided into two types: a mention network, which is constructed based on the mention structure of tweets, and a friendship network, which is built based on following relationships between users. Tweet context indicates user profiles and posting times [6]. Recent efforts have discussed the use of external data, such as external corpora [7] or weather information [1, 2, 8]. Moreover, many existing methods adopt the multiview-based inference methods, which combine multiple features such as tweet content and a Twitter network rather than relying on a single feature [9]. The multiview-based methods are broadly categorized into the ensemble approach and the joint learning approach. In the ensemble approach, estimates are made separately based on each feature, and the final estimate is derived by integrating the individual estimates. The joint learning approach builds a single model that simultaneously learns from multiple features and makes estimates.

In our previous studies [1, 2], we proposed a location estimation method that jointly learns from tweet contents as internal data, and weather and earthquake information as external data. From the experimental results, we confirmed that using both weather and earthquake information in addition to tweet content resulted in the best values for a precision of 0.7835 and a F1-measure of 0.7277. In particular, the precision improved by 6 points compared to using tweet content only, confirming that the weather and earthquake information significantly contributed to the improvement in the precision. The best recall was obtained when only tweet content was used, and the weather and earthquake information was not useful for improving recall. Fig. 1 shows the recall by prefecture obtained using a method based only on tweet content. As shown in Fig. 1, recall varied by region and two main tendencies were observed. The first tendency is that the recall is low in rural prefectures where the number of users in the dataset is small. For users living in rural areas with a limited number of users, the model was undertrained, resulting in a tendency to prioritize estimates for more populous urban prefectures. The second tendency is that the recall is also low in prefectures adjacent to large metropolitan areas. In these regions, the social boundaries between the metropolis and its surrounding areas are blurred, making accurate location estimation difficult. Similar to the first tendency, there is the significant difference in the number of users between metropolitan prefectures such as Tokyo and Osaka and their neighboring prefectures, therefore the recall is lower. However, low recall in certain regions means that the spread of disease in those regions may not be accurately detected. Therefore, in this paper, we focus on following relationships between users as a new approach to improve the recall. Since users living in rural areas tend to have closer relationships than those living in urban areas, reflecting regional characteristics through the following relationships is expected to be particularly effective for the first tendency.

Several methods [10–14] have been proposed for location estimation using a friendship network. Backstrom *et al.* [10] proposed a method that calculates the probability of following relationships based on the geographical distance between users and selects the residence label with the highest likelihood. Davis Jr. *et al.* [11] proposed a method that selects the most frequently occurring place of residence among the locations of users with the direct following relationships. Jurgens *et al.* [13] evaluated location estimation methods that use a twitter network by comparing methods based on a mention network. Hironaka *et al.* [14] classified following relationships into four types: followee, follower, mutual, and linked relationships, and compared methods using them for Japanese users. They showed that follower relationships are the most effective for estimating the place of residence of Japanese users.

As a multiview-based method that utilizes different types of features, Miura *et al.* [15], Huang *et al.* [16] and Qiao *et al.* [17] proposed a location estimation method that learns from tweet content and the mention network. However, while ensemble approaches using tweet content and friendship networks [18, 19] have been proposed, joint learning approaches have not yet been proposed. The joint learning approaches often achieve higher performance than the ensemble approach by learning the dependencies between different features simultaneously and integrating the information more efficiently. Therefore, we propose a new method that jointly learns from tweet content and the friendship network.

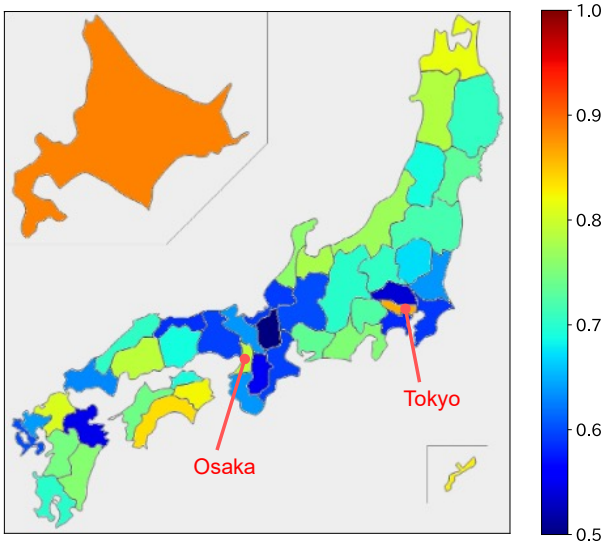

Fig. 1. Recall-based results by prefecture using the previous method based on only tweet content.

III. Estimation Method

Let the set of users be denoted as $U = \{u_1, u_2, \ldots, u_{|U|}\}$. Each user u_j has a set of tweets $T_j = \{t_{j1}, t_{j2}, \ldots, t_{j|T_j|}\}$, where each tweet $t_{ji} \in T_j$ represents the content posted by user u_j.

In many social networking services, users can follow other users to subscribe to their posts. In this paper, if user u_i follows user u_j, we refer to u_i as the "follower" and u_j as the "followee." We define two sets of users who have following relationships with user u_j: the set of users u_p that u_j follows, $Followee(u_j) = \{u_p \mid u_j \text{ follows } u_p\}$, and the set of users u_q who follow u_j, $Follower(u_j) = \{u_q \mid u_q \text{ follows } u_j\}$. Additionally, based on these two sets, we define the set of

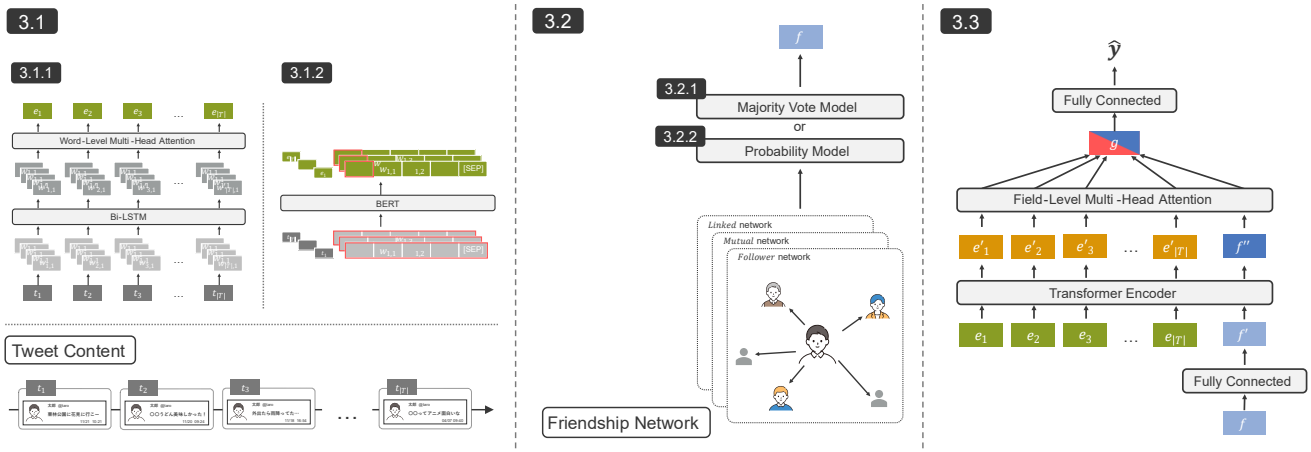

Fig. 2. Overview of the proposed method.

users u_r who mutually follow u_j, $Mutual(u_j) = \{u_r \mid u_j \text{ follows } u_r \wedge u_r \text{ follows } u_j\}$, and the set of users u_s who either follow or are followed by u_j, $Linked(u_j) = \{u_s \mid u_j \text{ follows } u_s \vee u_s \text{ follows } u_j\}$. These user sets represent the one-hop friendship network of user u_j.

Our task is to estimate the place of residence of user u_j at the level of 47 prefectures in Japan based on both the tweet content and the friendship network of u_j. Fig. 2 shows an overview of the proposed method. The details of the proposed method are explained in the following sections.

A. Models of Generating Tweet Content Embeddings

We will compare two models for generating a tweet embedding $\boldsymbol{e}_{ji}$ from the tweet content t_{ji} of user u_j. One is a conventional model based on the tweet content model [1], and the other is a BERT model. These models extract features from the tweet content of users and generate tweet embeddings that can be used as input to subsequent layers.

1) Conventional model (Word2Vec + Bi-LSTM + Word-Level Multi-Head Attention)

As the first model, we use a model proposed by Matsumoto *et al.* [1] to generate tweet embeddings from the tweet content. This model was constructed by simplifying the model proposed by Huang *et al.* [16] to reduce the time complexity. In this model, each tweet is divided into a sequence of words, and each word is vectorized using Word2Vec. Then, a Bi-LSTM layer is used to capture the contextual information of the word sequence, and the Word-Level Multi-Head Attention layer further aggregates the sequence of word vectors as tweet embeddings by focusing on words that are important for residence estimation.

2) BERT model

As the second model, we propose a BERT model for extracting context from the tweet content and generating tweet embeddings. Compared to the conventional model described in Section Ⅲ-A-1, BERT can capture more contextual information. Therefore, using the tweet embeddings obtained through BERT is expected to improve the estimation performance based only on tweet content. In this model, each tweet in the tweet sequence is individually input into BERT, and the last hidden state corresponding to the [CLS] token is used as the tweet embedding. By inputting tweets into BERT individually, we can maximize the use of the available information without being affected by the limitations of the input sequence length.

B. Generating Follow-Information Embeddings

We compare two models for encoding a user's friendship network of a user into a follow-information embedding $\boldsymbol{f}_j$. One is an extension model of the Majority Vote Model [11], and the other is an extension model of the Probability Model [10]. Both models originally aim to estimate the place of residence based only on the friendship network. However, we extend these models to generate follow-information embeddings in a format that can be integrated with tweet content. In this paper, we generate follow-information embeddings from three types of friendship networks based on $Follower(u_j)$, $Mutual(u_j)$, $Linked(u_j)$ using each model described below. Then, we investigate which following relationship is the most effective for residence estimation by comparing the performance of embeddings based on each network. Note that experiments using the $Followee(u_j)$ were not conducted because followee relationships often include one-way connections to celebrities or official accounts, which are not useful for residence estimation.

1) Majority vote model

The majority vote model [11] selects the most frequently occurring place of residence among the locations of the users who have a direct following relationship with the target user. This model is based on the assumption that most of the user's friends live in the same location. Hironaka *et al.* [14] reported that this model achieved the highest accuracy for residence estimation at the municipality level for Japanese users. Rather than selecting a single residence label, the majority vote model in this paper counts the number of occurrences of each residence label within the friendship network, and creates a 47-dimensional vector representing the distribution of residence labels by prefecture. This vector is normalized using the L2 norm and used as the follow-information embedding.

2) Probability model

The probability model [10] calculates the probability of the following relationship between users based on the geographical distance between them and selects a residence label with the highest likelihood. In this paper, we adopted the equation proposed by Hironaka *et al.* [14]. Eq. (1) shows the model representing the probability $p(d)$ of a following

relationship being formed when the distance between users is d. In the equation, a, b, and c are real-valued parameters. The user's place of residence is estimated using Eq. (2). Here, L is the set of all users, N_u is the set of users who have direct following relationships with user u, l_u is the true residence of user u, and $dist(a,b)$ is the distance between residence a and residence b.

$$p(d) = a(d+b)^c \tag{1}$$

$$\gamma_l(l) = \prod_{n \in L} \left[1 - p\left(dist(l, l_n)\right)\right]$$
$$\gamma(l,u) = \prod_{n \in N_u \cap L} \frac{p\left(dist(l, l_n)\right)}{1 - p\left(dist(l, l_n)\right)} \gamma_l(l) \tag{2}$$

The probability model used in this paper calculates the likelihood for each residence label within the friendship network and creates a 47-dimensional vector based on these values. The vector elements corresponding to a residence label with no neighboring nodes within the friendship network are set to 0. The vector obtained through this process is used as the follow-information embedding.

C. Estimation of Users' Place of Residence by Prefecture

The tweet embeddings $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, …, $\boldsymbol{e}_{|T_j|}$ and the follow-information embedding $\boldsymbol{f}_j$ are integrated to estimate the user's place of residence. In this paper, we utilize the Transformer Encoder layer and Field-Level Multi-Head Attention layer by referring the method proposed by Huang *et al.* [16] to capture the dependencies between tweet content and follow-information.

First, the follow-information embedding $\boldsymbol{f}_j$ based on each of the following relationship is linearly transformed into the same dimension as the tweet embeddings, and concatenated column-wise with $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, …, $\boldsymbol{e}_{|T_j|}$. Next, this concatenated matrix is input into the Transformer Encoder layer, and then the Field-Level Multi-Head Attention layer is applied to aggregate it into a single distributed representation $\boldsymbol{g}$. By applying a fully connected layer and a Softmax function to $\boldsymbol{g}$, we obtain the multiclass estimate probability vector $\hat{\boldsymbol{y}} = [\hat{y}_0, \hat{y}_1, \ldots, \hat{y}_{46}]$. Each element $\hat{y}_i$ represents the probability that the user lives in prefecture i.

$$\hat{\boldsymbol{y}} = softmax\left(\boldsymbol{g} \cdot \boldsymbol{W_g} + \boldsymbol{b_g}\right) \tag{3}$$

where $\boldsymbol{W_g}$ is a learnable parameter matrix, and $\boldsymbol{b_g}$ is a bias term. The loss function is defined to minimize cross-entropy as:

$$\mathcal{L}(\Theta) = -\frac{1}{N} \sum_{j=1}^{N} \sum_{i=0}^{46} y_{ji} \log\left(\hat{y}_{ji}\right) \tag{4}$$

where Θ represents all trainable parameters of the neural network.

Additionally, to ensure that the Transformer Encoder layer appropriately handles the different types of features, such as tweet content and follow-information, a Feature Type Embedding is concatenated row-wise. In the model proposed by Huang *et al.* [16], a feature type embedding of the same shape as the input was added to various features like tweet content, mention network, and profile descriptions. However, since we confirmed that concatenation contributed more to performance improvement than addition in our preliminary experiments, we adopted concatenation.

We compare the performance of our proposed method with a baseline model that estimates the place of residence based on tweet content without the follow-information. In the baseline model, the embedding is set to a zero vector before it is input into the model.

IV. Experiments

A. Dataset and Evaluation Settings

1) Dataset

We use the dataset constructed in our previous study [1]. This dataset consists of 204,965 Twitter users who posted tweets mentioning the diseases or symptoms and who explicitly stated their place of residence in their profiles. We further collected additional information on each user's followers and followees, and constructed a friendship network with users who explicitly stated their place of residence in their profiles as well. We were unable to obtain follow-information for some users because their accounts had been set to private or had been deleted. However, we successfully obtained follow-information for 191,764 users. Out of these 191,764 users, 171,764 are assigned to the training data, and 10,000 users are used for the validation and test data, respectively.

2) Model configuration

We compare two models for generating tweet embeddings: the conventional model and the BERT model. For the word embeddings in the conventional model, we use pre-trained word vectors provided by Hotlink Corporation [20]. For unknown words, we initialize them randomly using a uniform distribution $U(-0.25, 0.25)$. Adam is used as the optimizer, with an initial learning rate set to 10^{-5}. The conventional model is trained with a maximum of 15 epochs and a batch size of 16. For other hyperparameters, we follow the settings of the paper [1]. On the other hand, the BERT model uses the Japanese pre-trained model tohoku-nlp/bert-base-japanese-v3 [21]. The maximum input sequence length is set to 120, and sequences exceeding this length are truncated. AdamW is used as the optimizer, with the initial learning rate set to 10^{-5}. The BERT model is trained for a maximum of 5 epochs. Due to the structure of the model, batch inputs on a per-user basis are not possible. Therefore, gradient accumulation is used to effectively simulate a batch size of 16.

We also compare the performance of two follow-information embeddings based on the majority vote model and the probability model, respectively. In the calculation of $p(d)$ for the probability model, we adopt the values $a = 0.0019$, $b = 0.196$, and $c = -0.15$ as proposed in the paper [10], and the distance between two locations $dist(l, l_n)$ is calculated based on the shortest distance between prefectural government offices using the geodetic distance on the ellipsoid (GRS80).

Moreover, we compare three patterns for the friendship network $Follower$, $Mutual$, and $Linked$, as well as a pattern where only tweet content is used without the friendship network. When using the conventional model for generating tweet embeddings and not using friendship

networks, the structure is the same as the tweet content model [1] that does not use either weather or earthquake information. This model is used as the baseline model for the experiments.

The tweet content to be input into the model consists of the last 300 tweets from each user. The epoch with the smallest loss on the validation data is selected, and the estimation performance of the model is compared using the test data.

3) Evaluation metrics

The place of residence described in the location field of the user profile is used as the ground truth label. The estimated place of residence is compared at the prefecture level to determine exact matches. In this study, we treat the task as a 47-class classification problem (one class for each prefecture). For each prefecture i $(0 \leq i \leq 46)$, we define:

- TP_i (True Positive): The number of users whose true label is i, and whose predicted label is also i.
- FP_i (False Positive): The number of users whose true label is not i, but whose predicted label is i.
- FN_i (False Negative): The number of users whose true label is i, but whose predicted label is not i.
- TN_i (True Negative): The number of users whose true label is not i, and whose predicted label is also not i.

We computed $\mathrm{precision}_i$, recall_i, $\mathrm{F1\text{-}measure}_i$ for each prefecture i using the standard one-vs-rest approach:

$$\mathrm{precision}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i} \tag{5}$$

$$\mathrm{recall}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} \tag{6}$$

$$\mathrm{F1\text{-}measure}_i = \frac{2 \times \mathrm{precision}_i \times \mathrm{recall}_i}{\mathrm{precision}_i + \mathrm{recall}_i} \tag{7}$$

We then take the average of each metric over all 47 prefectures to obtain macro-average precision, recall, and F1-measure. Additionally, accuracy is defined as the ratio of correctly classified users (i.e., $\sum_i \mathrm{TP}_i$) to the total number of users.

Each model's performance is evaluated by macro-average precision, recall, F1-measure, and accuracy scores for the 47-prefecture estimates in a single run.

B. Experimental Results

Table 1 shows the macro-average precision, recall, F1-measure, and accuracy for the 47-prefecture estimates in a single run. Figs. 3 and 4 show the results visualized in precision and recall for the baseline model and the model that achieved the best macro-average values for each prefecture, respectively. The model ① in the table is the baseline model.

From Table 1, the model ④ achieved the best Recall of 0.7472, showing an improvement of about 5 points compared to the baseline model. Furthermore, the model ⑨ achieved the best Precision of 0.8350. This model also yielded the best values for both F1-measure and accuracy.

When comparing the methods for generating tweet embeddings in terms of precision, the BERT models ⑧ to ⑭ consistently showed higher performance than the conventional models ① to ⑦. In the conventional model, no improvement in the precision was observed in the Koshinetsu region, regardless of the friendship network used. However, even the BERT model ⑧ using only tweet content achieved high precision in the Koshinetsu region. This result suggests that the BERT model is better at capturing region-specific dialects, place names, and landmark names that the conventional model could not fully recognize. Moreover, the model ⑨ that achieved the best precision by utilizing the $Follower$ network demonstrated a particularly significant improvement among the models that combined the BERT model with the friendship network. This result is consistent with the findings in the paper [14]. As shown in Fig. 3, this model achieved particularly high precision in the Tohoku, Shikoku, and Kyushu regions.

In terms of recall, our previous methods [1, 2] have the issue of lower performance in rural areas and prefectures adjacent to major metropolitan areas. However, as shown in Fig. 4, the recall was significantly improved, especially in rural prefectures such as Fukui, Shimane, and Miyazaki. In addition, we defined the top 10 prefectures in the dataset by the number of users as "urban" and the bottom 10 prefectures as "rural," and compared the recall improvements between the baseline model and the model ④, which achieved the highest recall. The results showed that the average recall in urban prefectures increased by only about 0.5 points (indicating minimal change), whereas rural prefectures showed an average improvement of about 3.7 points. Despite the relatively small number of user samples in these rural prefectures, this significant improvement underscores the effectiveness of our approach. Additionally, some improvement in the recall was observed in prefectures adjacent to major metropolitan areas, indicating that the following relationships between users were effective. On the other hand, we confirmed that the factors of "which friendship network to use" and "which tweet embedding method to use" were not very important for improving recall. As shown in Table 1, while there is a significant performance difference between estimates using only tweet content and those combining it with a friendship network, no significant performance differences were observed based on the specific friendship network used.

When comparing the follow-information embedding methods, the majority vote model consistently demonstrated higher performance than the probability model. Although the probability model tended to perform better than estimates based on tweet content only, it did not exhibit as significant an improvement as the majority vote model. Since the majority vote model does not consider geographical biases or the relative locations of prefectures, we introduced the probability model to address this limitation. However, while there were cases where the probability model worked effectively, overall, the majority vote models were able to more accurately estimate the place of residence for a larger number of users.

V. Conclusion

In this paper, we proposed a new method for estimating the place of residence of Twitter users living in Japan. Our previous research proposed a hybrid approach that combined the tweet content with the weather and earthquake

Table 1. Macro-average precision, recall, F1-measure, and accuracy for 47-prefecture estimates[1]

Model	Tweet embedding method	Follow-information embedding method	Friendship network type	Precision	Recall	F1-measure	Accuracy
①	Conventional model	—	—	0.7654	0.6963	0.7246	0.7416
②		Majority vote model	Follower	0.7964	0.7460	0.7676	0.7685
③			Mutual	0.7911	0.7356	0.7597	0.7604
④			Linked	0.8061	**0.7472**	0.7729	0.7687
⑤		Probability model	Follower	0.7581	0.7141	0.7324	0.7402
⑥			Mutual	0.7676	0.7190	0.7395	0.7436
⑦			Linked	0.7648	0.7160	0.7359	0.7456
⑧	BERT model	—	—	0.8065	0.7017	0.7461	0.7478
⑨		Majority vote model	Follower	**0.8350**	0.7304	**0.7747**	**0.7748**
⑩			Mutual	0.8018	0.7461	0.7689	0.7711
⑪			Linked	0.7978	0.7430	0.7649	0.7700
⑫		Probability model	Follower	0.7786	0.7111	0.7381	0.7538
⑬			Mutual	0.8145	0.6902	0.7414	0.7534
⑭			Linked	0.7972	0.6981	0.7380	0.7503

[1] The experimental results were based on a single run due to constraints on computational resources. Based on the results of preliminary experiments in which the proposed models were trained with different dataset partitions, we confirmed that the order of performance among models was consistent, and no significant fluctuations in performance metrics occurred.

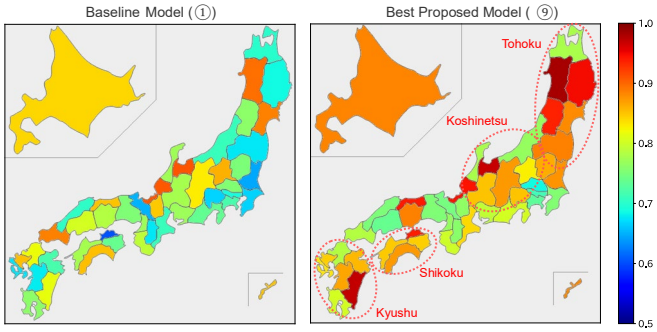

Fig. 3. Precision-based results by prefecture using the baseline and best proposed models.

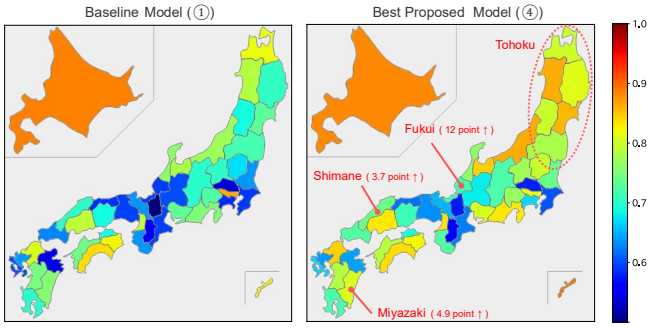

Fig. 4. Recall-based results by prefecture using the baseline and the proposed models.

information, however, it has the issue of low recall in rural areas and prefectures adjacent to large metropolitan areas. Therefore, in this paper, we focused on following relationships between users and proposed a new estimation method that jointly learns the tweet content and the friendship network. Additionally, we introduced BERT as the feature extraction and embedding generation from the tweet content to further improve precision.

The experimental results showed that the model with the friendship network improved recall, resulting in an improvement of up to 5 points compared to the baseline model using only tweet content. In particular, recall was significantly improved in rural prefectures, confirming that the method using both the tweet content and the friendship network proposed in this paper contributed greatly to the improvement in recall. Furthermore, the BERT model significantly improved Precision achieving an F1-measure of 0.7747, which is the best score recorded in our previous study.

As future work, identifying and removing noisy users from the dataset remains an important task. The proposed method estimates residence labels even for users whose tweets do not contain information such as place names or landmarks useful for estimation. Among these users are so-called noisy users, including bots or spam accounts aimed at criticizing politics or society, which are estimated to make up at least 3 to 5% of the dataset. These noisy users may negatively affect the learning of the model. Since the BERT model is susceptible to noise, it is important to remove such users from the dataset. We also focus on removing these noisy users and retraining the model to develop a more robust model.

Data Availability Statement

Our dataset is available at the following GitHub repository, in accordance with the API Terms of Service of X (formerly Twitter): https://github.com/Hirarince/dataset-for-estimation-of-sns-users-place-of-residence. If you use this dataset, please cite this paper and the previous work [1] appropriately.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Hiroki Hiramatsu: Methodology, data curation, software, formal analysis, investigation, and writing-original draft preparation. Kazuaki Ando: Conceptualization, writing-reviewing and editing, and supervision. All authors had approved the final version of the manuscript.

Funding

Part of this work was partially supported by JSPS KAKENHI Grant Number 22K12294.

References

[1] M. Matsumoto and K. Ando, "A deep learning model of estimating user's place of residence using tweets and weather information," in *Proc. IEEE Asia-Pacific Conf. Comput. Sci. Data Eng. (CSDE)*, 2022, pp. 1–6.

[2] M. Matsumoto and K. Ando, "Prediction of Twitter user's location using tweets, weather, and earthquake information," in *Proc. 85th Nat. Conv. IPSJ*, 2023, pp. 499–500. (in Japanese)
[3] T. Ando and K. Ando, "Factuality analysis of SNS posts containing diverse symptom expressions for public health surveillance," in *Proc. 9th Int. Conf. Syst. Informat. (ICSAI)*, 2023, pp. 1–5.
[4] Y. Hashimoto and M. Oka, "Statistics of geo-tagged tweets in urban areas," *J. Jpn. Soc. Artif. Intell.*, vol. 27, no. 4, pp. 424–431, 2012. (in Japanese)
[5] Y. Yamaguchi, Y. Ikawa, T. Amagusa, and H. Kitagawa, "User location inference using local events in social media," *J. Inf. Process. Soc. Jpn.*, vol. 6, no. 5, pp. 23–37, 2013. (in Japanese)
[6] X. Zheng, J. Han, and A. Sun, "A survey of location prediction on Twitter," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 9, pp. 1652–1671, Sept. 2018.
[7] P. Zola, C. Ragno, and P. Cortez, "A Google trends spatial clustering approach for a worldwide Twitter user geolocation," *Inf. Process. Manage.*, vol. 57, no. 6, 102312, 2020.
[8] Y. Kondo, M. Hangyo, M. Yoshida, and K. Umemura, "Home location estimation using weather observation data," in *Proc. 2017 Int. Conf. Adv. Informat., Concepts, Theory, Appl. (ICAICTA)*, 2017, pp. 1–6.
[9] X. Luo, Y. Qiao, C. Li, J. Ma, and Y. Liu, "An overview of microblog user geolocation methods," *Inf. Process. Manage.*, vol. 57, no. 6, 102375, 2020.
[10] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: improving geographical prediction with social and spatial proximity," in *Proc. 19th Int. Conf. World Wide Web (WWW '10)*, 2010, pp. 61–70.
[11] C. A. Davis Jr., G. L. Pappa, D. R. R. de Oliveira, and F. de L. Arcanjo, "Inferring the location of Twitter messages based on user relationships," *Trans. GIS*, vol. 15, pp. 735–751, 2011.
[12] A. Sadilek, H. Kautz, and J. P. Bigham, "Finding your friends and following them to where you are," in *Proc. 5th ACM Int. Conf. Web Search Data Mining (WSDM'12)*, 2012, pp. 723–732.
[13] D. Jurgens, T. Finethy, J. McCorriston, Y. Xu, and D. Ruths, "Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice," in *Proc. Int. AAAI Conf. Web Soc. Media*, 2021, vol. 9, no. 1, pp. 188–197.
[14] S. Hironaka, M. Yoshida, M. Okabe, and K. Umemura, "Analysis of social network generation methods for home location estimation in Japan," *Trans. Jpn. Soc. Artif. Intell.*, vol. 32, no. 1, pp. WII-M_1–1, 2017. (in Japanese)
[15] Y. Miura, M. Taniguchi, T. Taniguchi, and T. Ohkuma, "Unifying text, metadata, and user network representations with a neural network for geolocation prediction," in *Proc. 55th Ann. Meeting Assoc. Comput. Linguistics (ACL)*, 2017, vol. 1, pp. 1260–1272.
[16] B. Huang and K. Carley, "A hierarchical location prediction neural network for Twitter user geolocation," in *Proc. 2019 Conf. Empirical Methods Nat. Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 4732–4742.
[17] Y. Qiao, X. Luo, J. Ma, M. Zhang, and C. Li, "Twitter user geolocation based on heterogeneous relationship modeling and representation learning," *Inf. Sci.*, vol. 647, 119427, 2023.
[18] A. Rahimi, T. Cohn, and T. Baldwin, "Twitter user geolocation using a unified text and network prediction model," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics (ACL-IJCNLP)*, 2015, vol. 2, pp. 630–636.
[19] S. Ribeiro and G. L. Pappa, "Strategies for combining Twitter users' geo-location methods," *GeoInformatica*, vol. 22, pp. 563–587, 2018.
[20] S. Matsuno, S. Mizuki, and T. Sakaki, "Constructing of the word embedding model by Japanese large scale SNS + Web corpus," in *Proc. 33rd Annu. Conf. Jpn. Soc. Artif. Intell.*, 2019, pp. 1–3. (in Japanese)
[21] BERT base Japanese (unidic-lite with whole masking, CC-100 and jawaki-20230102). [Online]. Available: https://huggingface.co/tohoku-nlp/bert-base-japanese-v3